Rate and Distortion Modeling for Real-time MGS Coding and Adaptation

Abdul Haseeb*†, Maria G. Martini†, Sergio Cicalò* and Velio Tralli*
*University of Ferrara, Italy - †Kingston University, UK
Email: {abdul.haseeb, sergio.cicalo, velio.tralli}@unife.it - m.martini@kingston.ac.uk

***Abstract*—Scalable Video Coding (SVC) is the extension of the Advanced Video Coding standard (H.264/AVC) providing video compression with spatial, temporal and quality scalability. Scalability can be exploited in order to provide a better video quality for the end user in video transmission over wireless networks. In this paper we develop a parametric Rate Distortion (R-D) model for Medium Grain Scalability (MGS) SVC depending only on two indexes describing the spatial and temporal complexity of video sequences. The two indexes can be easily obtained from the original raw video, thus enabling real time video adaptation for transmission over channels with variable bandwidth such as wireless channels. The results from simulations show that the use of the proposed model for rate adaptation of multiple-videos sharing a common channel results in an end user video quality comparable to that obtained by using a more accurate non-real time rate distortion model.**

***Index Terms*—Rate-Distortion modeling, Scalable Video Coding, Rate adaptation, Video over wireless**

I. Introduction

Video streaming is one of the most popular applications of today's Internet. As the Internet is a best effort network, it poses several challenges especially for high quality video streams. The Advanced Video Coding (H.264/AVC) scalable extension, also called Scalable Video Coding (SVC), provides an attractive solution for the difficulties encountered when a video source is transmitted over wireless transmission systems. Such challenges include error prone channels, heterogeneous networks and capacity limitations and fluctuations [1]. Within SVC, each sequence is encoded with one base layer (BL) and several enhancement layers (ELs) which can be sequentially dropped by providing a graceful degradation. Three types of scalabilities, namely spatial, temporal and SNR scalability are supported by the standard, which allows to extract from the encoded video sub-streams of a suitable resolution, frame rate and quality matching various network conditions and terminal capabilities. SNR scalability is achieved by using Coarse Grained Scalability (CGS) or Medium Grained Scalability (MGS) [2]. In CGS a limited number of quality levels can be extracted, which is equal to the number of coded layers, while MGS provides a finer granularity of quality scalability by dividing each CGS layer into 16 MGS layers. In this paper we focus on SNR scalability with MGS layers.

Different video sequences have different complexities, hence the relationship between rate and quality differs from one video sequence to another. Assuming the same physical resources are shared among different video sequences, an equal rate allocation scheme would divide the available rate equally among the sequences, which could lead to a high or even unacceptable level of distortion for the most complex videos requiring higher rates. In order to optimize the transmission strategy based on the end user video quality, the rate should be allocated among the videos based on a fairness criterion [3][4]. Moreover, the trade-off between the goal of reducing the bit rate and the goal of keeping the distortion at acceptable levels can be afforded dynamically, in order to perform adaptation to different conditions.

Rate-distortion (R-D) models enable to predict the minimum bit rate required to achieve a target quality. The rate of a video sequence is expressed in bytes/s, while the distortion is defined in terms of Mean Square Error (MSE). The Peak Signal to Noise Ratio (PSNR) is more often used to express the quality of a video sequence. The time required to model the R-D curve for a given sequence may drive the decision on the methodology/algorithm to be adopted for the R-D modeling. On the other hand, the performance of the streaming system is directly affected by the accuracy of the R-D model [5]. R-D models are often categorized in analytical, semi analytical and empirical models. Analytical R-D models are used to predict rate and distortion of video sequences prior to the encoding process but they often incur in a loss of accuracy. Empirical models require the computation of all R-D points set resulting in a high complexity. Semi-analytical models aim at reducing such complexity by deriving parametrized functions that follow the shape of analytically derived functions, but are evaluated through curve fitting from a subset of the rate-distortion empirical data points.

For real time video streaming systems the computation of the model should be fast enough to deal with the timing constraints of the video stream. Hence, we investigate here techniques to further reduce the complexity of semi-analytical models. This is made possible by introducing new functions dependent only on the uncoded video streams. The coefficients of this new functions can be estimated off-line through a prior knowledge of the parameters of a set of sample video sequences, and then used for any future video sequence.

Many R-D models have been proposed in the literature for real time and non-real time video streaming, see for example, [5]–[9] and references therein. In [6] the authors present a detailed analysis of the R-D relationship in fine-grain scalable coders and provide an accurate square root R-D model, which requires at least two empirical points. Enhanced R-D models

for H.264/AVC were proposed for coded video sequences in [7]. However the parameter extraction is performed after transformation and quantization in the encoding process. The late extraction of the parameters can significantly affect real time applications such as video over wireless networks. An improved real time R-D model for medium grain scalable video coding was proposed in [8]. This model reduces significantly the dependency on the encoding process. In this model the delay is reduced by extracting the parameters before transformation.

In this paper we propose a new R-D model for real time MGS video streams. Our model only uses two parameters which are calculated taking into account the characteristics of the video sequences through a spatial and a temporal index extracted from the original raw video streams. Moreover, we also use these complexity indexes to calculate BL and EL rates of the given video stream.

The remainder of this paper has the following main contributions: Section II illustrates the proposed R-D model. A brief overview of the adaptation algorithm used for verification of our proposed model is illustrated in Section III. Section IV describes simulation and model verification, while conclusions are drawn in Section V.

II. A REAL-TIME R-D MODEL

In this section we propose a parametric R-D model for MGS SVC which is simple enough to be used by rate-adaptation techniques in real-time video streaming. The model depends on the Spatial Index (SI) and the Temporal Index (TI) of the original raw video sequence. After encoding, the GOP of the k-th generic video results in a finite discrete set of codes with rate r_k and distortion d_k. The R-D function which represents this set of points is often modeled as a continuous function $\mathcal{R}_k(D)$, because it can be more easily used to obtain simple rate adaptation algorithms. We consider as a reference R-D model the one introduced in [4] for MGS coded video, which is based on two parameters and has been proved as accurate as other state-of-the-art models:

$$\begin{cases} \mathcal{R}_k(D) = \frac{\alpha_k}{D} + \beta_k \\ \mathcal{R}_k(D) \geq \mathcal{R}_{k,BL} \\ \mathcal{R}_k(D) \leq \mathcal{R}_{k,EL} \end{cases} \tag{1}$$

where α_k and β_k are sequence dependent parameters of the k-th GOP while $\mathcal{R}_{k,BL}$ and $\mathcal{R}_{k,EL}$ are the BL and highest EL rates obtained from the encoded video. The drawback of this model is the fact that its parameters can only be evaluated by looking for the best fitting of at least 4 R-D points after the encoding process of the video, hence the model is not suited for real time applications.

The model proposed here replaces the parameters α_k and β_k with a function of the spatial index SI_k and the temporal index TI_k, as explained in the following:

$$\alpha_k = p_1 + p_2 SI_k + p_3 TI_k \tag{2}$$

$$\beta_k = q_1 + q_2 SI_k + q_3 TI_k \tag{3}$$

The same approach is used to express the BL and EL rates:

$$R_{k,BL} = r_1 + r_2 SI_k + r_3 TI_k \tag{4}$$

$$R_{k,EL} = s_1 + s_2 SI_k + s_3 TI_k \tag{5}$$

The values on the sets $\boldsymbol{p} = [p_1,\ p_2,\ p_3]$, $\boldsymbol{q} = [q_1,\ q_2,\ q_3]$, $\boldsymbol{r} = [r_1, r_2,\ r_3]$ and $\boldsymbol{s} = [s_1,\ s_2,\ s_3]$ are coefficients that can be calculated by using fitting methods in a sufficiently large set of GOPs from a set of video sequences (training set). As mentioned above, this process is executed off-line only once.

The SI and TI values are evaluated on the luminance component [10] of the video by means of Spatial Information and Temporal Information [11] of the k-th GOP as follows:

$$SI_k = \max_n std_\sigma\{Sobel[F_n(\sigma)]\} \tag{6}$$

$$TI_k = \max_n std_\sigma\{M_n(\sigma)\} \tag{7}$$

where $M_n(\sigma) = F_n(\sigma) - F_{n-1}(\sigma)$ is the motion difference, $F_n(\sigma)$ is the luminance component and n and σ are the temporal and spatial coordinates, respectively, of the frames used to encode GOP k.

To summarize, the R-D model is obtained by substituting in (1) the parameters α_n and β_n from (2) and (3), and $\mathcal{R}_{k,BL}$ and $\mathcal{R}_{k,EL}$ from (4) and (5), respectively:

$$\begin{cases} \mathcal{R}_k(D) = \frac{p_1 + p_2 SI_k + p_3 TI_k}{D} + q_1 + q_2 SI_k + q_3 TI_k \\ \mathcal{R}_k(D) \geq r_1 + r_2 SI_k + r_3 TI_k \\ \mathcal{R}_k(D) \leq s_1 + s_2 SI_k + s_3 TI_k \end{cases} \tag{8}$$

The proposed R-D model is verified by considering video sequences generated by the JSVM software [12]. We encoded six video sequences *i.e.* $Crew$, $Football$, $Coastguard$, $Soccer$, $City$, and $Mother\ and\ Daughter$ (MD) having different scene complexities, in CIF resolution with a frame rate of 30 fps. We denote this set as the training set. Two ELs are used to obtain SNR scalability where each layer is split into 5 MGS layers with vector distribution of [3 2 4 2 5]. All the videos are coded GOP by GOP with a GOP size of 8 to obtain sequences comprising 26 GOPs. The Quantization Parameter is set to 38, 32 and 26 to obtain the BL and two ELs.

Fig. 1 shows α_k, β_k, BL and highest EL models as in (2), (3), (4) and (5), respectively, using the spatial and temporal indexes. In the two upper figures the markers refer to the values of α_k and β_k derived according to model (1) and plotted for each GOP versus the corresponding value of SI_k and TI_k. In the two lower figures the markers refer to the BL and EL layer rates derived by encoding the sequences with JSVM [12]. It can be observed that the values of the parameters for all the models closely follow a linear behaviour. The metrics used to evaluate the goodness of the model in fitting the set of points are reported in the caption. The sets of coefficients, appearing in (2), (3), (4) and (5) of the proposed model, are calculated using the linear least square fitting method [13] with Least Absolute Residuals (LAR) [14] for robustness. The resulting values for the training set are the following:

$$\boldsymbol{p} = [-2.4 \times 10^4, 3975, 540.5] \qquad \boldsymbol{q} = [-246.1, 24.1, 3.3]$$

$$\boldsymbol{r} = [41.27, 17.09, 9.12] \qquad \boldsymbol{s} = [-237, 145.6, 34.02]$$

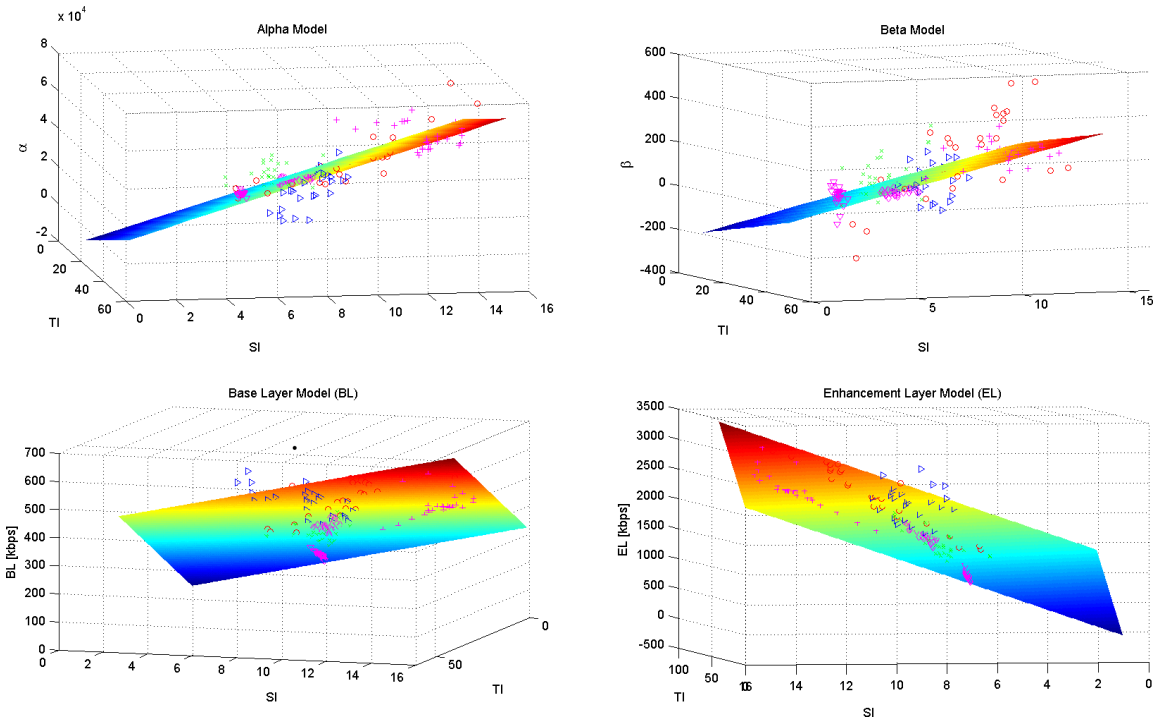

Fig. 1. Proposed Models for α, β, BL and EL rates. The parameters used for the goodness of the models are the coefficient of determination (R^2) and Root Mean Square Error ($RMSE$). (α) R^2= 0.987 $RMSE$ = 1598, (β) R^2 = 0.973 $RMSE$ = 21.2, (BL) R^2 = 0.979 $RMSE$ = 22.98, (EL) R^2 = 0.985 $RMSE$ = 79.36

In Fig. 2 the different R-D models are shown and compared for two sample GOPs of three video sequences. The accuracy changes GOP by GOP: the upper figure shows the result for a GOP with good matching between the proposed model and the model in (1), whereas the lower figure shows a result with poor matching. As shown in Section IV below, the GOPs with less accurate model do not have significant impact on the behaviour of rate adaptation strategies in real time multi-video transmission. To evaluate the goodness of BL and EL rate estimation, we compare in Fig. 3 the rates estimated with the model in (4) and (5) to the original rates obtained from the encoded sequences. We consider not only the video sequences in the training set but also the sequences outside the training set. More emphasis is given to BL rate as it is the minimum rate requirement of each video sequence when transmitted in bandwidth constrained channels. It can be observed from Fig. 3 that our model predicts the BL rate quite accurately for sequences outside the training set, as shown for $Mobile$ and $Foreman$. Moreover, it can be seen that the estimation is also good for EL rate.

III. Application to Real Time Rate Adaptation

The models discussed in Section III are useful to build up rate-adaptation algorithms that adaptively set encoding parameters or scale the video to suitably optimize the transmission in a bandwidth-constrained, time-variant channel shared by multiple video users. In this section we briefly illustrate the rate adaptation algorithms considered to validate the proposed model. We assume to have K different videos ($k = 1, \ldots, K$) characterized by different complexities and having the same GOP duration. The rate-adaptation algorithm selects GOP by GOP for each video the best R-D couple (r_k, d_k), given the estimated total bandwidth R_c available in the GOP interval. Two different adaptation strategies are considered. The first one is an Equal Rate (ER) scheme that tries to divide in equal parts the available bandwidth without taking care of the resulting distortion. It simply adjusts the rate to have:

$$\mathcal{R}_k(D_k) = R_c/K \qquad \text{if } R_c/K \in [\mathcal{R}_{k,BL}, \mathcal{R}_{k,EL}] \tag{9}$$

The second strategy is based on a fairness-oriented scheme (OPT), as the one proposed in [4] that tries to minimize the distortion of each video sequence while preserving fairness. The fairness metric is based on the GOP-by-GOP difference between the distortion of two videos, $i, j \in [1, \ldots, K]$, defined as:

$$\Delta(D_i, D_j) = \begin{cases} 0 & \text{if } (i,j) \in \mathbb{X}_D \vee \text{(j, i)} \in \mathbb{X}_{\mathcal{D}} \\ |D_i - D_j| & \text{otherwise} \end{cases} \tag{10}$$

where

$$\mathbb{X}_{\mathcal{D}} = \{(i,j) \in \mathbb{Z}^2 : (D_i = D_{max,i} \wedge D_j > D_i) \vee (D_i = D_{min,i} \wedge D_j < D_i)\}$$

and $D_{min,i} = \min\{d_{n,i}\} \approx \mathcal{R}^{-1}(\mathcal{R}_{i,EL}), D_{max,i} = \max\{d_{n,i}\} \approx \mathcal{R}^{-1}(\mathcal{R}_{i,BL})$. The operators $\wedge$ and $\vee$ are the logic "AND" and "OR", respectively. Assigning the distortion values to multiple video streams to have ideal fairness, *i.e.*,

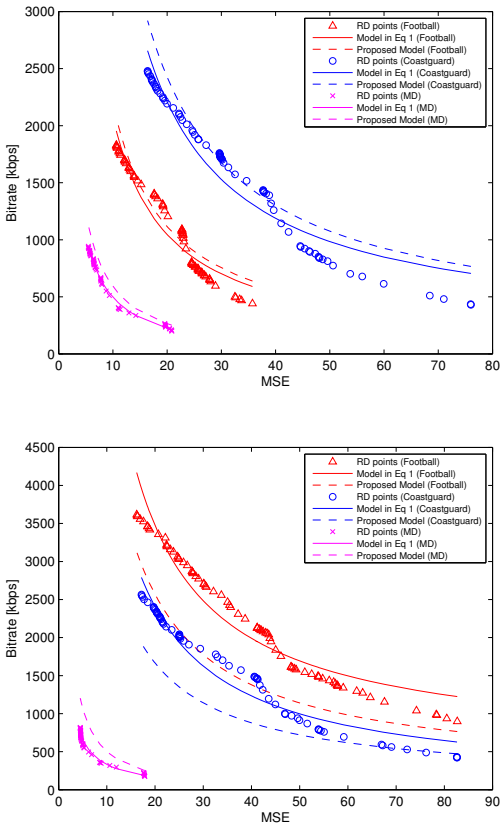

Fig. 2. R-D comparison among model in eq. (1) proposed model and actual values for two sample GOPs.

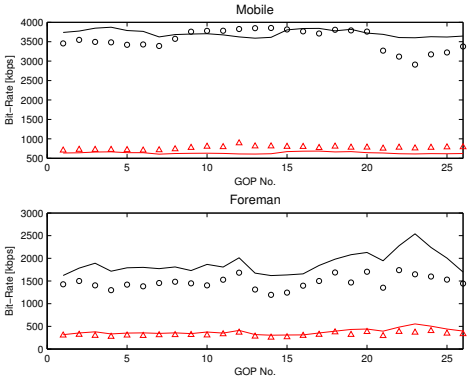

Fig. 3. BL and EL rates over 26 GOPs for two sequences in the training set (upper figures) and two sequences outside the training set (lower figures). The marker points refer to the original BL and EL rates, whereas the solid lines refer to rates estimated from (4) and (5), respectively.

$D_i = D_j, \quad \forall i \neq j$, is hard to achieve, due to (i) the discrete nature of R-D relationship and (ii) the presence of a minimum and a maximum distortion values for each stream directly related to the complexity of each video sequence, which can be very different from one video to another. The fairness metric takes the issue (ii) into account in the formulation of the distortion difference $\Delta(D_i, D_j)$ which suitably considers the minimum and the maximum constraints. The optimization problem of the OPT strategy is formulated as :

$$\min_{\boldsymbol{D}\in\mathbb{R}^{\boldsymbol{K}}} \sum_i \sum_{j<i} \Delta(D_i, D_j) \tag{11}$$

$$s.t. \quad \sum_{k=1}^{K} \mathcal{R}_k(D_k) = R_c \tag{12}$$

$$\mathcal{R}_{k,BL} \leq \mathcal{R}_k(D_k) \leq \mathcal{R}_{k,EL} \quad \forall k \tag{13}$$

The algorithms provide at least the minimum rate, which is the BL rate, to all video sequences only if:

$$\sum_{k=1}^{K} \mathcal{R}_{k,BL} \leq R_c \tag{14}$$

The BL rate $\mathcal{R}_{k,BL}$ and the EL rate $\mathcal{R}_{k,EL}$ are provided by (4) and (5), respectively, for our proposed model to allow the adaptation algorithms to check the minimum and maximum rate constraints.

IV. Simulation and Model Verification

In this section we verify the proposed R-D model in the transmission of multiple videos over a bandwidth constrained channel by using the rate adaptation algorithms outlined in Section III and described in detail in [4]. We propose results for both the videos in the training set and videos outside it. In the first case a bandwidth limited to $R_c = 3500$ kbps is considered. In the second case a set of 4 sequences, *i.e.*, $Foreman$, $Harbour$, $Container$ and $Mobile$, and a bandwidth limited to $R_c = 3000$ kbps are considered.

Tables I and II show the average MSE taken over the first 26 GOPs for the model (1) and our proposed model. It can be seen that the ER algorithm results in less distortion for the low complexity videos like MD, $City$ in the training set and $Foreman$ and $Container$ outside the training set, thus compromising the quality of more complex videos like $Football$, $Coastguard$ or $Harbour$ and $Mobile$. This behavior is mitigated by the OPT algorithm, as expected. Moreover, it can also be observed from both tables that the average MSE values for the proposed model closely follow the model (1) except for $Harbour$ and $Container$ in Table II with the OPT algorithm. For the ER algorithm in both table I and II, the results for model (1) and our proposed model show only slight differences mainly due to the fact that our estimated maximum and minimum rates which are the BL and EL rates are different from the original BL and EL rates.

A more detailed observation can be done through Fig. 4 which compares GOP-by-GOP the MSE obtained after rate adaptation for sequences of the training set, *i.e.*, $Football$, $City$ and outside the training set, *i.e.*, $Mobile$ and $Foreman$, with our proposed model and the model (1). It can be observed that, with the exception of some large deviations experienced in few GOPs of $City$ and $Mobile$, our model closely follow model (1). The exceptions suggest, in practical applications, that video servers determine off-line different models as in (6)

for a limited number of video classes having homogeneous characteristics.

V. Conclusion

In this work we proposed a new rate-distortion model using spatial and temporal indexes for MGS scalable video coded streams. The model has been developed aiming in particular at real time video streaming over wireless channels: it only needs information about the characteristics of the original uncoded video (as spatial and temporal indexes) to build the R-D relationship and to estimate BL and EL rates. The model has been compared to a state-of-the-art non real-time model in a scenario where multiple video sequences are transmitted over a bandwidth constrained channel with rate adaptation. Results show that both models lead to a similar end-user video quality.

TABLE I
RESULTING AVERAGE MSE OVER 26 GOPS OBTAINED WITH MODEL (1) AND PROPOSED MODEL. VIDEOS IN THE TRAINING SET.

Sequence	Model (1)		Proposed Model	
	ER	OPT	ER	OPT
Crew	36.99	38.00	36.24	44.09
Football	53.11	44.00	53.11	46.69
Coastguard	70.78	45.72	69.09	46.82
Soccer	39.93	42.15	38.21	34.72
City	37.61	51.05	35.41	49.10
MD	9.00	20.65	8.69	20.59

TABLE II
RESULTING AVERAGE MSE OVER 26 GOPS OBTAINED WITH MODEL (1) AND PROPOSED MODEL. VIDEOS NOT INCLUDED IN THE TRAINING SET.

Sequence	Model (1)		Proposed Model	
	ER	OPT	ER	OPT
Foreman	19.02	34.90	18.29	31.60
Harbour	79.78	57.86	79.18	81.11
Container	15.88	35.39	15.45	18.14
Mobile	103.84	65.44	103.84	72.76

Acknowledgments

The authors would like to acknowledge the support of the Royal Society, UK, in the framework of the International Joint Project "Cross-layer rate control and scheduling for video transmission over WiMAX" and the EU's 7^{th} Framework Programme under grant agreement no. 288502 (CONCERTO project).

References

[1] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 17, no. 9, pp. 1103–1120, 2007.

[2] M. Jacobs, S. Tondeur, T. Paridaens, J. B. R. V. de Walle, and P. Schelkens, "Statistical multiplexing using svc," *Proc. IEEE Int. Symp. Broad-band Multimedia Syst. Broadcast.*, p. 16, 2008.

[3] M. Martini and V. Tralli, "Video quality based adaptive wireless video streaming to multiple users," in *Proc. IEEE Int. Symp. Broad-band Multimedia Syst. Broadcast.*, Mar. 2008.

[4] S. Cicalò, A. Haseeb, and V. Tralli, "Fairness-oriented multi-stream rate adaptation using scalable video coding," *Elsevier Signal Processing: Image Communication*, 2012.

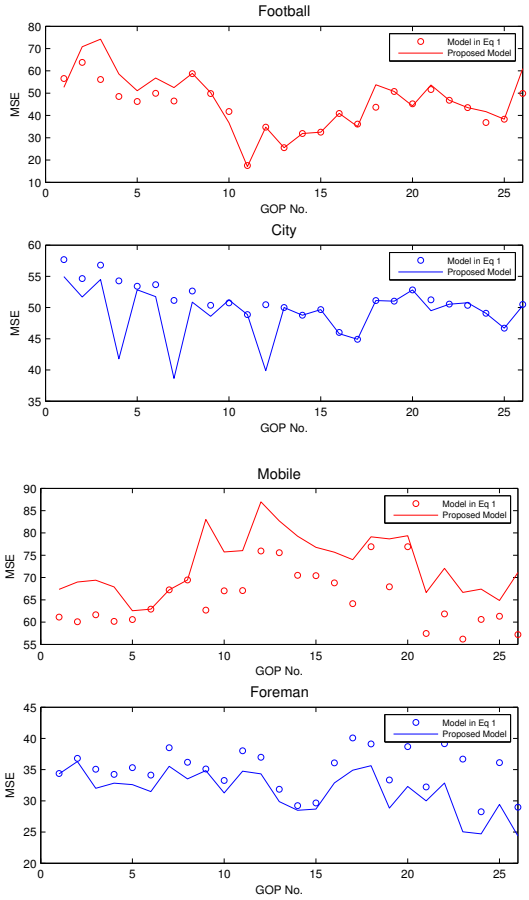

Fig. 4. Average MSE obtained GOP-by-GOP for sample videos transmitted over a bandwidth constrained channel with rate adaptation. The two upper figures refer to the transmission of the six videos of the training set (R_c = 3500 kbps), whereas the two lower figures refer to the transmission of four videos not included in the training set (R_c = 3000 kbps).

[5] H. Cheng-Hsin and M. Hefeeda, "On the accuracy and complexity of rate-distortion models for fine grained scalable video sequences," *ACM Trans. on Multimedia Computing, Communications and Applications*, 2006.

[6] M. Dai, D. Loguinov, and H. Radha, "Rate-distortion analysis and quality control in scalable internet streaming," *IEEE Trans. on Multimedia*, vol. 8 issue 6, pp. 1135–1146, 2006.

[7] K. Do-Kyoung, S. Mei-Yin, and K. C. C. Jay, "Rate control for H.264 video with enhanced rate and distortion models," *IEEE Trans. Circuits Syst. Video Tech.*, vol. 17, no.5, pp. 517–529, 2007.

[8] H. Mansour, V. Krishnamurthy, and P. Nasiopoulos, "Rate and distortion modeling of medium grain scalable video coding," *In Proc. of 2008 IEEE 15th Int. Conf. on Image Processing*, Oct. 12-15, 2008, San Diego.

[9] H. Seferoglu, O. Gurbuz, and Y. Altunbasak, "Rate-distortion based real-time wireless video streaming," *Elsevier Signal Processing: Image Communication*, vol. 22 Issue 6, pp. 529–542, 2007.

[10] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi, "A H.264/AVC video database for the evaluation of quality metrics," in *Proc. IEEE Int Acoustics Speech and Signal Processing (ICASSP) Conf*, Mar. 2010.

[11] ITU-T, "Subjective video quality assessment methods for multimedia application, recommendation," *ITU-T*, p. 910, Sept. 1999.

[12] *JSVM 9.19.11 Reference Software Febrary 2011*.

[13] P. Chaffe-Stengel and D. N. Stengel, *Working With Sample Data: Exploration and Inference*. Business Expert Press, Aug. 2011.

[14] Y. Dodge and J. Jureckov, *Adaptive Regression*, BPOD, Ed. Springer, 2000.