Accepted in the IEEE International Symposium on Multimedia (ISM 2013), 2013

Visual Quality and File Size Prediction of H.264 Videos and its Application to Video Transcoding for the Multimedia Messaging Service and Video on Demand

Didier Joset and Stéphane Coulombe
Department of Software and IT Engineering
École de technologie supérieure, Université du Québec
Montréal, Canada
e-mail: didier.joset.1@ens.etsmtl.ca ; stephane.coulombe@etsmtl.ca

***Abstract*—In this paper, we address the problem of adapting video files to meet terminal file size and resolution constraints while maximizing visual quality. First, two new quality estimation models are proposed, which predict quality as function of resolution, quantization step size, and frame rate parameters. The first model is generic and the second takes video motion into account. Then, we propose a video file size estimation model. Simulation results show a Pearson correlation coefficient (PCC) of 0.956 between the mean opinion score and our generic quality model (0.959 for the motion-conscious model). We obtain a PCC of 0.98 between actual and estimated file sizes. Using these models, we estimate the combination of parameters that yields the best video quality while meeting the target terminal's constraints. We obtain an average quality difference of 4.39% (generic model) and of 3.22% (motion-conscious model) when compared with the best theoretical transcoding possible. The proposed models can be applied to video transcoding for the Multimedia Messaging Service and for video on demand services such as YouTube and Netflix.**

Visual quality assessment; predictive models; H.264; video transcoding; Multimedia Messaging Service; video on demand

I. INTRODUCTION

The Multimedia Messaging Service (MMS) allows users with heterogeneous terminals to exchange structured messages composed of text, audio, images, and video [1], and represents a great source of revenue for mobile operators. According to Portio Research, 207 billion MMS messages were sent in 2011, and this number is expected to rise to 276.8 billion by 2016. Informa sees even higher MMS volumes by 2016, predicting that 387.5 billion MMS messages will be sent, representing 10.6% of global messaging revenues, valued at US $20.7 billion [2]. But the rapid development of mobile technologies is contributing to the rapid proliferation of mobile terminal types [3]. This situation creates interoperability problems, and the need to adapt images and videos to meet the target terminal's maximum resolution and file size constraints while maximizing visual quality [4-8]. Similar problems arise in video on demand services such as YouTube and Netflix where numerous versions of the same content need to be generated to satisfy the various devices' resolutions and their network bit rates.

Multimedia content adaptation (or transcoding) to meet terminal constraints is quite common in many applications. For visual content, previous works have considered professional applications using PowerPoint [5], as well as images for which a quality estimator [6], a file size estimator [7], and a complete algorithm using both estimators [8] have been proposed. For video, several studies have been conducted and numerous quality models proposed. In [9], a quality model function of the average PSNR and the frame rate is proposed for low bit rate QCIF videos. In [10], the model is extended to take motion information into account. In [11][12], the impact of frame rate on quality is studied, but the model parameters need to be adapted for each video sequence. In ITU-T G.1070 [13], a model is proposed which takes several parameters into account, including frame rate and quantization, not, however, changes in resolution. In [14], the authors propose an improved model, based on their study of the impact of resolution and display size in the context of G.1070. In [15], the authors claim that G.1070 cannot model perceptual video quality properly, especially at low bit rates, and therefore propose improvements to the model to take into account video motion. One drawback of these methods is that numerous model parameters need to be determined for every situation (e.g. a set of parameters for every resolution).

In [16][17], the authors propose the following complete quality model, which takes into account quantization, resolution, and frame rate:

$$QSTAR(s,t,f) = \left[\frac{1-e^{-c_q\times\frac{q_{min}}{q}}}{1-e^{-c_q}}\right]\left[\frac{1-e^{-c_s\times\left(\frac{s}{s_{max}}\right)^{0.74}}}{1-e^{-c_s}}\right]\left[\frac{1-e^{-c_f\times\left(\frac{f}{f_{max}}\right)^{0.63}}}{1-e^{-c_f}}\right]$$

where q, s, f, q_{min}, s_{max}, and f_{max} represent the quantization step size, spatial resolution, frame rate, minimum quantization step size, maximum resolution, and maximum frame rate respectively. c_q, c_s, and c_f are model parameters that need to be computed for every video sequence. Although the model is very accurate, tuning the parameters for every video is not practical for most applications.

With respect to video file size estimation, numerous papers have addressed rate models and rate control. They typically use intrinsic video characteristics, such as the distribution of transformed coefficients [18][19] or the percentage of zero quantized transformed coefficients (ρ-domain) [20]. Wang et al. [21] recently proposed a simple model based on frame rate

and quantization. However, we are not aware of a simple yet accurate model that is a function of the three parameters of interest (q, s, f).

In this paper, we address the problem of adapting video content under file size and resolution constraints, while conserving optimum visual quality. The challenge is to find the best compromise between resolution, quantization, and frame rate under a file size constraint. Although we could transcode a video using all possible combinations of parameter values (q, s, f) to find the best solution, this would require far too much computation to be useful in practice. To solve the problem, we have developed models that can predict the visual quality and file size of videos for various values of these parameters. Based on these predictions, we select the set of parameters that provides the best visual quality while meeting the constraints. We show that such an approach yields near optimal quality in a single transcoding operation.

Unlike previous work, this paper addresses the whole problem of video file adaptation, and proposes novel quality and file size models that are simple, yet accurate. Our paper is also unlike previous work in that the model's parameters do not need to be adapted for each video and take into account resolution, quantization, and frame rate (or a subset of these parameters). Another attractive feature of the proposed approach is that it is computationally light, which makes it appealing for large scale video adaptation systems.

This paper is organized as follows. Section II describes the proposed quality estimation models. Section III presents the proposed file size model. In Section IV, we apply the models to the problem of MMS video adaptation. In Section V, we describe how to apply the models to the problem of video on demand. In Section VI, we discuss the applicability of the proposed models in the context of other compression standards and applications. Section VII concludes the work.

II. Proposed Quality Estimation Models

A. Methodology and Assumptions

We first propose a model to estimate perceived visual quality as a function of resolution, quantization, and frame rate. To develop this model, we need to select an appropriate video dataset. Several video databases exist, such as the LIVE Video Database [22]. This database contains the mean opinion scores (MOS) of several videos compressed with the H.264 compression standard at different bit rates, however it doesn't contain MOS for various resolutions and frame rates. Therefore, we have selected the video database developed by Ou et al. at New York University's Polytechnic Institute [16][17]. The database comprises 10 videos encoded using H.264 with a resolution varying from QCIF to 4CIF, a quantization parameter (QP) ranging from 28 to 44, and a frame rate (FR) from 3.75 to 30 frames per second (fps). Specifically, the resolution values used are QCIF, CIF, and 4CIF; the QP values used are 28, 36, 40, and 44; and the frame rate values used are 3.75, 7.5, 15, and 30 fps. The database contains the MOS for each combination of these parameter values. For our model, we concentrate on seven of these videos as the *training set*, and use the remaining videos as the *test set* to validate the proposed models and the transcoding method. The training set comprises a representative mixture of videos: Akiyo, City, Crew, Football, Foreman, Ice, and Soccer. The test set contains slow to fast motion sequences: Harbor, Waterfall, and Flower Garden. The videos are available from [23].

First, because we want simple models, we assume that resolution, quantization, and frame rate affect quality independently and are non compensatory parameters. The first assumption means that the impact of modifying a parameter has the same effect on quality, regardless of the values of the other parameters. That is, the following three equations hold:

$$\frac{VQ(R,F,Q)}{VQ(R_{max},F,Q)} = f_R\left(\frac{R}{R_{max}}\right)$$

$$\frac{VQ(R,F,Q)}{VQ(R,F,Q_{min})} = f_Q\left(\frac{Q_{min}}{Q}\right)$$

$$\frac{VQ(R,F,Q)}{VQ(R,F_{max},Q)} = f_F\left(\frac{F}{F_{max}}\right)$$

where VQ represents the visual quality (MOS), R the resolution in pixels, Q the quantization step size (we convert the QP to quantization step size), F the frame rate, R_{max} the maximum resolution (4CIF in this database), Q_{min} the smallest quantization step size (28 in this database), F_{max} the maximum frame rate (30 fps in this database), and f_R, f_Q, f_F are generic functions to be determined by regression. The first equation shows that if we change the resolution R, the ratio of the resulting visual quality and the best visual quality obtained with R_{max} becomes a function of R/R_{max}. We make similar assumptions for the other parameters (with F/F_{max} for frame rate and Q_{min}/Q for quantization).

In the marketing field [24], compensatory models are considered when consumers evaluate their perception of the value of a product by performing a weighted sum of several attributes. In performing a cost/benefit analysis, attributes like good quality or great appearance can compensate for a poor attribute (e.g. high price). However, non compensatory models are considered when consumers immediately eliminate products they regard as inadequate (e.g. too expensive or ugly). In fact, with non compensatory models, consumers often use an elimination process to discard products that do not meet their minimal expectations for any attribute. In our case, we are clearly faced with non compensatory attributes, since a poor parameter can't be redeemed by selecting favorable attributes among the other parameters. For instance, high resolution or high frame rate can't compensate for coarse quantization, as the resulting quality will remain poor. Therefore, we propose a quality model comprising the product of the functions of each of the three parameters ($R, Q,$ and F) normalized with the extreme values that lead to the best quality ($R_{max}, Q_{min},$ and F_{max} respectively). The visual quality model we propose is defined as follows:

$$VQ(R,Q,F) = V_R\left(\frac{R}{R_{max}}\right)\times V_Q\left(\frac{Q_{min}}{Q}\right)\times V_F\left(\frac{F}{F_{max}}\right) \quad (1)$$

where $0 \le VQ(R,Q,F) \le 1$ is the normalized visual quality (we normalize the MOS values of the database to between 0 and 1). V_R, V_Q, and V_F are the normalized visual quality functions for resolution, quantization and frame rate respectively. Each function takes a value between 0 and 1 and represents the impact of each of the three parameters on global quality.

B. Visual Quality Estimation Function for Resolution

We first develop the normalized visual quality function for resolution, denoted V_R, to measure the impact of resolution on quality. In order to create a simple model, we choose to focus our research on functions with two parameters. After experimenting with several types of functions on a subset of the training set, we obtained the best fit with nonlinear regression using a logistic function:

$$V_R\left(\frac{R}{R_{max}}\right) = \frac{1}{1+e^{\alpha_R-\beta_R\left(\frac{R}{R_{max}}\right)}} \quad (2)$$

The best results were obtained with $\alpha_R = 0.89$ and $\beta_R = 8.5956$ (using regression). Unlike previously proposed models, where parameters such as α_R and β_R must be adapted for each video [16][17], the proposed model can operate with the same values for any video with good prediction accuracy.

Table I shows the Pearson Correlation Coefficient (PCC), and the mean absolute error (MAE) between the actual MOS and the predicted value using V_R for the selected subset of the training set when varying the resolution. The PCC measures the linear relationship between two variables (the closer to 1, the stronger the linear relationship). We can observe that the proposed model is quite accurate.

C. Visual Quality Estimation Function for the Quantization Step Size

We now develop the normalized visual quality function for the quantization step size, denoted V_Q. After experimenting with several types of functions on another subset of the training set, we obtained the best fit with nonlinear regression, again using a logistic function:

$$V_Q\left(\frac{Q_{min}}{Q}\right) = \frac{1}{1+e^{\alpha_Q-\beta_Q\left(\frac{Q_{min}}{Q}\right)}} \quad (3)$$

The best results were obtained with $\alpha_Q = 1.0293$ and $\beta_Q = 7.2729$ (using regression), which are used for all the videos. Table II shows PCC and MAE for the proposed function. Again, we can observe that the proposed model is quite accurate. Note that the training subsets in Tables I and II are not exactly the same. The reason is that we wanted to build the visual quality estimation models for resolution and quantization step size using independent subsets (or as independent as possible) and later see if, once combined, the models were still accurate on the complete training set (as will be shown). The goal is not to evaluate the performance of each individual model but the performance of the global model including resolution, quantization step size and frame rate.

TABLE I. ESTIMATION ACCURACY BETWEEN THE NORMALIZED MOS AND THE VISUAL QUALITY ESTIMATION FUNCTION FOR RESOLUTION V_R

	City	Crew	Fore-man	Ice	Soccer
PCC	0.970	0.977	0.957	0.963	0.948
MAE	0.050	0.039	0.057	0.052	0.056

TABLE II. ESTIMATION ACCURACY BETWEEN THE NORMALIZED MOS AND THE VISUAL QUALITY ESTIMATION FUNCTION FOR QUANTIZATION STEP SIZE V_Q

	Akyio	Crew	Football	City
PCC	0.992	0.988	0.930	0.974
MAE	0.015	0.084	0.066	0.036

TABLE III. ESTIMATION ACCURACY BETWEEN THE NORMALIZED MOS AND THE VISUAL QUALITY ESTIMATION FUNCTION FOR FRAME RATE V_F

	Akiyo	City	Crew	Foot-ball	Soccer	Fore-man	Ice
PCC	0.895	0.948	0.952	0.970	0.916	0.915	0.867
MAE	0.072	0.070	0.048	0.037	0.046	0.037	0.040

D. Visual Quality Estimation Function for Frame Rate

Finally, we develop the normalized visual quality function for the frame rate, denoted V_F. After experimenting with several types of functions on the training set, we obtained the best fit using a natural logarithmic function:

$$V_F\left(\frac{F}{F_{max}}\right) = \beta_F \times ln\left(\frac{F}{F_{max}}\right) + 1 \quad (4)$$

The best results were obtained with $\beta_F = 0.18368$ (using regression), which is used for all the videos. Table III shows the PCC and MAE for the proposed function. Again, we can observe that the proposed model is quite accurate. Like in the case of the resolution, we tested other functions, keeping in mind the necessity of having a simple model with at most two parameters (e.g. parameters a and b and models of the form: ax^b, ae^{bx}, $\frac{a}{1-e^b}$, $\frac{a\,x}{b+x}$ and $a \ln x + b$).

E. Proposed Visual Quality Estimation Model combining Resolution, Quantization and Frame Rate

The proposed normalized visual quality estimation model taking into account resolution, quantization and frame rate is obtained by replacing the terms in (1) with their equivalents from (2), (3), and (4):

$$VQ(R,Q,F) = \left(\frac{1}{1+e^{\alpha_R-\beta_R\left(\frac{R}{R_{max}}\right)}}\right)\times \left(\frac{1}{1+e^{\alpha_Q-\beta_Q\left(\frac{Q_{min}}{Q}\right)}}\right)\times\left(\beta_F\times ln\left(\frac{F}{F_{max}}\right)+1\right) \quad (5)$$

with the parameter values as presented previously. Table IV shows PCC and the Spearman Rank-Order Correlation Coefficient (SRCC) between the actual MOS and the predicted values using VQ for all the videos in the training set when varying all the parameters. The SRCC measures the correlation between the rankings of two datasets (the closer to 1, the closer the rankings). We can observe that the proposed model is very accurate for sequences with low levels of movement, and less so for sequences with significant motion, such as Football and Foreman. This motivates us to improve the model further, to take into account motion information.

F. Improving the Visual Quality Estimation Model to account for motion

The generic model performs well, but can be improved. In fact, the frame rate will not have the same impact on quality if we have a sequence with slow movements, such as Akiyo, or a sequence with fast and complex movements, such as Football, which would require a higher frame rate to maintain visual quality. Consequently, we wish to refine the proposed model to account for motion.

We modify the parameters of our generic model, in order to obtain different variations, each taking into account different motion amplitudes. As we want to keep the model as simple as possible, we decided to create three classes of motion, the first for a low level of motion, the second for a moderate level of motion, and the third for a high level of motion. We calculate three motion characteristics for every video to cluster (group) the videos into the three classes. The motion vectors are determined using the three-step search [25] with a block size of 8×8. First, we determine the mean motion vector for the entire video. Then, we calculate the standard deviation. Finally, we determine the mean value of the largest 25% of the motion vectors. Using these three features and the features in Table IV (PCC, SRCC), we group the sequences into three clusters using the k-means clustering algorithm [26], where each feature is normalized. We then optimize the model parameters for the three classes of videos on the training set. In the low motion class are Akiyo, City, and Ice. In the moderate motion class are Crew and Foreman. In the high motion class are Football and Soccer.

The optimized parameters, using regression, for both the generic and the motion-conscious models are presented in Table V. Bold values indicate that, for the motion-conscious model, the same values were optimal as for the generic model. Table VI shows PCC and SRCC for the motion-conscious model.

TABLE IV. ESTIMATION ACCURACY BETWEEN THE NORMALIZED MOS AND THE VISUAL QUALITY ESTIMATION FUNCTION $\boldsymbol{VQ(R,Q,F)}$

	PCC	SRCC
Akiyo	0.9823	0.9853
City	0.9755	0.9814
Ice	0.9836	0.9780
Crew	0.9715	0.9682
Foreman	0.9505	0.9408
Football	0.9351	0.8971
Soccer	0.9711	0.9695
Average	0.9671	0.9600

TABLE V. OPTIMIZED MODEL PARAMETERS FOR THE PROPOSED QUALITY ESTIMATION MODELS

		Generic	Motion-conscious		
			Low	Med.	High
Resolution	α_R	0.89	0.89	0.7744	0.872
	β_R	8.5956	8.5956	9.4514	8.5024
QP	α_Q	1.0293	1.0293	1.0293	1.0293
	β_Q	7.2729	7.2729	7.2729	7.2729
Frame rate	β_F	0.18368	0,1544	0.1548	0.2375

TABLE VI. ESTIMATION ACCURACY BETWEEN THE NORMALIZED MOS AND THE MOTION-CONSCIOUS VISUAL QUALITY ESTIMATION FUNCTION $\boldsymbol{VQ(R,Q,F)}$

	PCC	SRCC
Akiyo	0.9832	0.9882
City	0.9824	0.9878
Ice	0.9838	0.9823
Crew	0.9851	0.9829
Foreman	0.9634	0.9597
Football	0.9707	0.9618
Soccer	0.9779	0.9805
Average	0.9781	0.9776

TABLE VII. PERFORMANCE RESULTS FOR THE PROPOSED VISUAL QUALITY ESTIMATION MODELS ON THE VIDEO TEST SET

	Generic		Motion-conscious	
	PCC	SRCC	PCC	SRCC
Harbor	0.9505	0.9505	0.9510	0.9652
Waterfall	0.9801	0.9930	0.9865	0.9965
Flower Garden	0.9372	0.9475	0.9404	0.9420
Average	0.9559	0.9637	0.9593	0.9679

G. Visual Quality Estimation Model performance on the video test set

We tested the performance of the proposed quality models on the test set (generic and motion-conscious), and present the results in Table VII. Overall, we can observe that both models are very accurate, although the motion-conscious model performs noticeably better than the generic one. Note that Harbor is a sequence with slow motion, Waterfall a sequence with moderate motion, and Flower Garden a sequence with high/complex motion.

III. Proposed File Size Estimation Model

We now develop a file size model that predicts the file size of a video clip as a function of resolution, quantization step size, and frame rate, and the file size of the original video clip. To develop the file size estimation model, we select 5 videos from the training set and encode them with a fixed QP using JM 18.4 Baseline [13] (an I frame followed by P frames). The changes in parameters are the same as those used for the quality estimation model, and so each video is encoded 48 times (3 resolutions × 4 QPs × 4 frame rates). Using a similar approach and similar assumptions as for the visual quality model, we have, for the generic file size estimation model:

$$S(R,Q,F) = 0.999\times Size(R_{max},Q_{min},F_{max})\times \left(\frac{1}{1+e^{\mu_R-\theta_R\left(\frac{R}{R_{max}}\right)}}\right)\times\left(\mu_Q\left(\frac{Q}{Q_{min}}\right)^{\theta_Q}\right)\times\left(\mu_F\left(\frac{F}{F_{max}}\right)^{\theta_F}\right) +0.001\times Size(R_{max},Q_{min},F_{max}) \quad (6)$$

where R_{max}, Q_{min}, and F_{max} have been defined before, and $Size(R,Q,F)$ is the function that returns the actual file size of the video as a function of R,Q,F. The largest file size is obtained for $S(R_{max},Q_{min},F_{max})$. As before, the values for parameters μ_R, θ_R, μ_Q, θ_Q μ_F, and θ_F were obtained through regression (i.e. comparing the size obtained using the model with the sizes obtained from actual transcodings using various parameter values of R,Q,F). They are presented in Table VIII. The term $0.001\times Size(R_{max},Q_{min},F_{max})$ represents video headers which size remains constant regardless of the parameters. The performance results for the proposed file size estimation model on the test set are presented in Table IX. Contrary to what we did in the quality model, we only compute the PCC here, because we only wish to estimate the file size as closely as possible (we are not interested in the ranking, i.e. the Rank Order Correlation Coefficient (ROCC)).

For file size estimation, we do not take motion into account. However, we studied the effect of motion on our model and concluded that it did not improve its precision significantly. This is not to say that motion has no impact on file size, because obviously it does. But this impact is already embedded in the term $Size(R_{max},Q_{min},F_{max})$, which multiplies the other terms in (6) (i.e. higher motion will lead to larger values of $Size(R_{max},Q_{min},F_{max})$ which in turn will lead to larger estimated size values). Therefore, although not explicitly considered, motion is implicitly considered. Our findings for the images in [4] were similar.

TABLE VIII. PARAMETER VALUES FOR THE PROPOSED FILE SIZE ESTIMATION MODEL

Resolution		Quantization		Frame rate	
μ_R	θ_R	μ_Q	θ_Q	μ_F	μ_F
-3.856	10.383	1.0044	-1.0996	0.9942	0.9942

TABLE IX: ESTIMATION ACCURACY BETWEEN THE ACTUAL VIDEO FILE SIZE AND THE NORMALIZED FILE SIZE ESTIMATION FUNCTION $S(R,Q,F)$

	PCC
Harbor	0.9874
Waterfall	0.9818
Flower Garden	0.9838
Average	0.9844

IV. Application to MMS Video Adaptation

A. The importance of video adaptation for MMS

To accommodate the various types of terminals and support their technological evolution, the Open Mobile Alliance has defined several MMS content classes [1]. For instance, the *Video Basic* class can support videos with a maximum file size of 100 kB, while this limit is 300 kB for the *Video Rich* class and 600 kB for the *Content Rich* class. Therefore, to ensure interoperability and continued MMS profitability, video adaptation for MMS is required [4]. Since in MMS, the maximum file size supported by the destination terminal is known, the challenge is to find the combination of resolution, quantization step size, and frame rate leading to the best visual quality within exceeding the maximum file size constraint. The proposed visual quality estimation and file size estimation models are key to solve this problem. Indeed, although it is clear that lower resolution, larger quantization step size and lower frame rate all contribute to reducing the file size, it is unclear in what amount they need to be modified to lead to the best visual quality. The optimal combination will depend on the amount of motion in the video as well as how aggressively the file size needs to be reduced.

B. The video transcoding system

To solve the video adaptation problem, we use the architecture presented in Fig. 1. We first extract the features of the original video, denoted *i,* that we want to transcode (R,Q,F, and optionally motion information if we use the motion-conscious model). Using the proposed models to estimate visual quality and file size, we determine the combination of parameters that lead to the best estimated quality under the constraints of file size and resolution. This estimated optimal combination of parameters R_o,Q_o,F_o is then used to transcode the final video, denoted *f*.

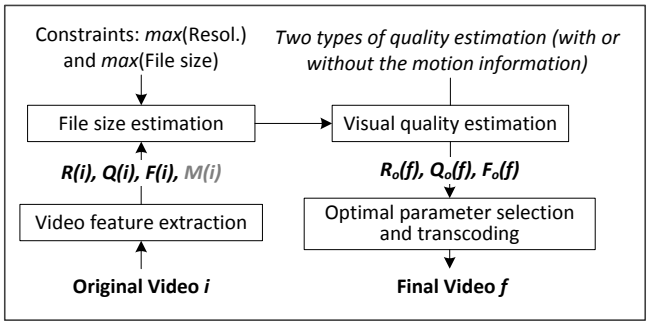

Fig. 1: Transcoding architecture of the proposed solution.

C. Experimental results

We now validate the performance of the proposed video adaptation approach based on the novel video quality and file size prediction models. The proposed validation architecture is presented in Fig. 2. We first extract the original test video's features and apply the file size estimation model (see the left part of the system). We keep only the candidate parameter combinations that meet the maximum resolution and file size constraints. Then, we apply the visual quality estimation model on the candidate combinations retained, and select the combination associated with the best quality (estimated optimal combination of parameters R_o, Q_o, F_o). We finally transcode the original video using this estimated optimal combination. To validate the performance, we undertake similar steps (see the right part of the system), but using oracular information (ground truth). This means that, instead of estimating file size and visual quality, we perform an extensive number of transcodings with all the possible combinations of parameters to compute the actual file size and visual quality for each combination. We keep the candidate parameters that meet the resolution and file size constraints, and from them we select the combination yielding the best visual quality (R_E, Q_E, F_E). Finally, we compute the difference in quality of the video clips obtained using the oracle and the proposed method. We do not compute their relative file size differences, since we are interested in the resulting video quality, not the actual file size. However, we did validate the results, in terms of meeting the file size constraints.

In the experiments, we varied the values of the maximum resolution between QCIF, CIF, and 4CIF, and the maximum file size over an extensive number of values from 10 kB to the maximum video size (several Mbytes). The range of variation of these parameters was chosen to correspond to realistic MMS problems. As mentioned, the candidate resolution values include: QCIF, CIF, and 4CIF. The candidate QP values include: 28, 36, 40, and 44. The candidate frame rate values are: 3.75, 7.5, 15, and 30 fps.

The mean quality errors, defined as the average differences between the optimal quality and the one obtained with the proposed system, for the test videos are presented in Table X for the generic and motion-conscious models (see the *Full* model, with and without motion). We observe a relative quality error (with respect to the optimal value) of 4.4% for the generic model and of 3.2% for the motion-conscious model. We can conclude that the system provides quality close to that obtained by the oracle with a single transcoding operation, while the oracle obtained the optimal solution at the cost of performing 48 transcoding operations (all the combinations of parameters). The proposed adaptation solution requires few computational resources to achieve good quality, which makes it appealing for MMS video adaptation. In cases where a strict maximum video file size must be met, such as in MMS, the system can dynamically adjust the quantization or frame rate while performing transcoding to satisfy such size limit (since the resolution can't change). The models help deciding which adjustments yield the best quality.

TABLE X. MEAN ABSOLUTE QUALITY ERROR BETWEEN THE ORACLE AND THE PROPOSED METHOD FOR MMS ADAPTATION

Model	Motion	Harbor	Waterfall	Flower Garden	Avg.
Full	No	5.206	1.738	6.235	4.393
	Yes	4.89	1.738	3.038	3.222
Known size	No	3.008	0.134	3.387	2.176
	Yes	2.864	0.130	0.051	1.015

In Table X, we can also observe the impact of our file size model on estimation. We performed another series of tests where we assumed that the file size was known (we still estimate the quality, but use the file size from the oracle). This is indicated as the *Known size* model in Table X. The results show a significant reduction in error. We conclude that our file size estimation model is responsible for more than half the errors. In a real system, this error could be reduced by adjusting the QP from frame to frame. Nevertheless, it is crucial to achieve the best compromise between resolution, quantization, and frame rate initially, since resolution and frame rate are typically not changed once they are set.

V. Application to Video on Demand

The proposed models and methods can be applied to transcoding for video on demand services such as YouTube and Netflix. For instance, Netflix is preparing 120 different versions of each movie to deal with the various devices' requirements [27]. Selecting the version with the best visual quality satisfying a device's constraints is very challenging. The difference between this problem and the one of MMS is that instead of dealing with video clips, we deal with video segments with specific duration (e.g. several groups of pictures). Since most decoders will not support changing the resolution dynamically, frame rate and quantization step size can be modified for each segment to maximize the visual quality while satisfying a desired bit rate. We first convert the file size estimation model (see Eq. (6)) to a constant bit rate estimation model as follows:

$$B(R,Q,F) = 0.999 \times BR(R_{max}, Q_{min}, F_{max}) \times \left(\frac{1}{1+e^{\mu_R - \theta_R\left(\frac{R}{R_{max}}\right)}}\right) \times \left(\mu_Q \left(\frac{Q}{Q_{min}}\right)^{\theta_Q}\right) \times \left(\mu_F \left(\frac{F}{F_{max}}\right)^{\theta_F}\right) + 0.001 \times BR(R_{max}, Q_{min}, F_{max}) \quad (7)$$

where $B(R,Q,F)$ is the estimated video bit rate and $BR(R,Q,F)$ is the function that returns the bitrate of the video segment as a function of R, Q, F. Eq. (7) results of the fact that for constant bit rate, $Size(R,Q,F) = BR(R,Q,F) \times time$. In a video transcoding context, only the original video properties are known and therefore only $BR(R_{max}, Q_{min}, F_{max})$ is known, which represents the largest bit rate. The bitrate estimation model parameters are the same as in Eq. (6).

To achieve our goal, we first need to find the best initial resolution, frame rate and quantization step size parameters

using the proposed approach, then keep the resolution constant and use the visual quality and bit rate estimation models to find the best adjustments of frame rate and quantization for each video segment to meet the bitrate constraint while maximizing the visual quality. The parameters may change for each video segment as in the case of real-time video streams transmitted over the Internet or over mobile network where the bitrate constraints might change with time. When some parameters are modified dynamically and a motion-conscious model is considered, the motion features (mean motion vector, standard deviation, mean value of the largest 25% of the motion vectors, etc.) can be determined over a window of several frames.

VI. Applicability of the Models

Although the proposed models have been developed for H.264 compression, they can be applied, at least in part, in the context of other compression standards such as MPEG-2, H.263 and HEVC (after possibly modifying some parameters of Table V). For instance, the visual quality models do not depend on the compression standard for resolution and frame rate (but they may to some extent for quantization step size). The models will also need to be readjusted and validated if larger resolutions than 4CIF are considered. The file size model needs to be validated and adapted as required for other compression standards. Finally, the proposed models may be applied to other applications where video clips or video streams are transcoded.

It is important not to confuse the proposed visual quality estimation models with the visual quality assessment (VQA) methods (that are abundant in the literature). The notion of visual quality estimation is used in the literature in the context of *no reference* VQA where no information related to the original content is known and therefore the visual fidelity needs to be *estimated*. But our goal is not to assess the quality of videos after they have been transcoded using various resolution, frame rate and quantization step size parameters. We rather want to *estimate* the visual quality without having to perform any transcoding operation. This allows us to determine, with very light computations, the best combination of parameters to use for the actual transcoding operation that we will then perform. Therefore our estimation is different from the one used in the context of no reference VQA. Actually, we can't compare our models with the abundant work on VQA presented in the literature. We could have only compared our models with the ones presented in the introduction. But since they are incomplete or have to be tuned for every video, we do not believe that the comparison would have been fair.

Finally, further validation of the models is required in future work. Indeed, the New York University's Polytechnic Institute video database [16][17] that contained MOS scores for the various video modifications we needed (resolution, quantization step size, and frame rate) only contained 10 videos clips that had to be used for both training and validation. A richer database will need to be developed.

VII. Conclusion

In this paper, we have proposed two new visual quality models for video that take into account resolution, quantization, and frame rate. The first is generic and the second takes into account video motion information. Not only are the proposed models accurate, but the parameters are fixed and do not need to be adapted to the characteristics of the videos. We have also proposed a new video file size prediction model. With these models, we present a complete video adaptation system that estimates the optimal combination of the three parameters, and we have shown that they yield quality close to that of the oracle in a single transcoding operation. The proposed method, with its near optimality and very light computations, is promising for use for MMS and video on demand transcoding.

Acknowledgments

This work was funded by Vantrix Corporation and by the Natural Sciences and Engineering Research Council of Canada under the Collaborative Research and Development Program (NSERC-CRD 428942-11).

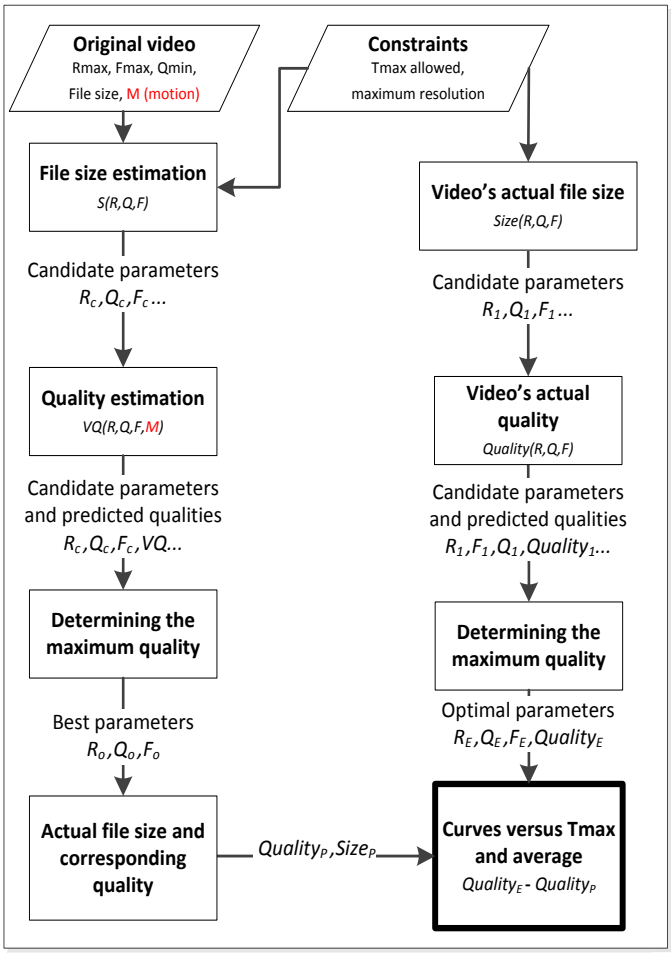

Fig. 2: Performance validation architecture.

REFERENCES

[1] Open Mobile Alliance, "Multimedia Messaging Service Conformance Document," OMA-TS-MMS_CONF-V1_3-20110913-A, 2011.

[2] Informa telecoms and media, "SMS will remain more popular than mobile messaging apps over next five years," Online http://blogs.informatandm.com/4971/press-release-sms-will-remain-more-popular-than-mobile-messaging-apps-over-next-five-years/ [Last accessed: 5 July 2013], 2012.

[3] A. Smith, "Smartphone adoption and usage," Online http://www.pewinternet.org/Reports/2011/Smartphones.aspx [Last accessed: 5 July 2013], 2011.

[4] S. Coulombe and G. Grassel. "Multimedia Adaptation for the Multimedia Messaging Service," *IEEE Communications Magazine* 42 (7): 120–126, July 2004.

[5] H. Louafi, S. Coulombe, and U. Chandra, "Quality Prediction-Based Dynamic Content Adaptation Framework Applied to Collaborative Mobile Presentations," *IEEE Transactions on Mobile Computing,* vol.12, no.10, pp.2024-2036, Oct. 2013.

[6] S. Coulombe and S. Pigeon, "Quality-Aware Selection of Quality Factor and Scaling Parameters in JPEG Image Transcoding," *IEEE 2009 Computational Intelligence for Multimedia, Signal, and Video Processing*, pp. 68-74, 2009.

[7] S. Pigeon and S. Coulombe, "Computationally Efficient Algorithms for Predicting the File Size of JPEG Images Subject to Changes of Quality Factor and Scaling," *24th Queen's University Biennial Symposium on Communications,* pp. 378-382, 2008.

[8] S. Coulombe and S. Pigeon, "Low-Complexity Transcoding of JPEG Images with Near-Optimal Quality Using a Predictive Quality Factor and Scaling Parameters," *IEEE Trans. Image Processing,* vol. 19, no. 3, pp. 712-721, March 2010.

[9] G. Hauske, T. Stockhammer, and R. Hofmaier. "Subjective image quality of low-rate and low-resolution video sequences," In *Proceedings of the 8th International Workshop on Mobile Multimedia Communications*. pp. 5-8, 2003.

[10] R. Feghali, F. Speranza, D. Wang, and A. Vincent, "Video quality metric for bit rate control via joint adjustment of quantization and frame rate," *IEEE Trans. Broadcasting,* vol. 33, no. 1, pp. 441-446, 2007.

[11] Q. Huynh-Thu and M. Ghanbari, "Temporal aspect of perceived quality of mobile video broadcasting," *IEEE Trans. Broadcast.*, vol. 54, no. 3, pp. 641-651, 2008.

[12] Y.-F. Ou, T. Liu, Z. Zhao, Z. Ma, and Y. Wang, "Modeling the impact of frame rate on perceptual quality of video," *IEEE 15th Int. Conference on Image Processing,* pp. 689-692, 2008.

[13] International telecommunication Union (ITU-T), "Recommendation G.1070: Opinion model for videotelephony applications," Geneva, 2007.

[14] B. Belmudez and S. Moller, "An Approach for Modeling the Effects of Video Resolution and Size on the Perceived Visual Quality, Multimedia," *2011 IEEE International Symposium on Multimedia (ISM),* pp. 464-469, 2011.

[15] J. Joskowicz and J. Ardao, "Enhancements to the opinion model for video-telephony applications," in Proceedings of the 5th International Latin American Networking Conference, pp. 87-94, 2009.

[16] Y.-F. Ou, Y. Xue, and Y. Wang, "Q-STAR: A Perceptual Video Quality Model for Mobile Platforms Considering Impact of Spatial, Temporal, and Amplitude Resolutions," Polytechnic Institute of NYU, Tech. Rep, 2012.

[17] Y.-F. Ou, Y. Xue, Z. Ma, and Y. Wang, "A perceptual video quality model for mobile platform considering impact of spatial, temporal, and amplitude resolutions," *IEEE 10th IVMSP Workshop*, pp. 117, 122, June 16-17, 2011.

[18] J. Sun, Y. Duan, J. Li, J. Liu, and Z. Guo, "Rate-distortion analysis of dead-zone plus uniform threshold scalar quantization and its application-part I: fundamental theory," *IEEE Trans. Image Processing*, vol. 22, no. 1, pp. 202-14, Jan. 2013.

[19] J. Sun, Y. Duan, J. Li, J. Liu, and Z. Guo, "Rate-Distortion Analysis of Dead-Zone Plus Uniform Threshold Scalar Quantization and Its Application-Part II: Two-Pass VBR Coding for H.264/AVC.," *IEEE Trans. Image Processing*, vol. 22, no. 1, pp. 215-28, Jan. 2013.

[20] Z. He and S.K. Mitra, "Optimum bit allocation and accurate rate control for video coding via ρ-domain source modeling," *IEEE Trans. Circuits and Systems for Video Technology,* vol.12, no.10, pp. 840, 849, Oct. 2002.

[21] Y. Wang, Z. Ma, and Y.-F. Ou, "Modeling rate and perceptual quality of scalable video as functions of quantization and frame rate and its application in scalable video adaptation," *17th International Packet Video Workshop,* pp. 1-9, 2009.

[22] LIVE Video Quality Database, Online http://live.ece.utexas.edu/research/quality/live_video.html [Last accessed: 5 July 2013], 2009.

[23] Xiph.org Foundation, "Xiph.org Video Test Media [derf's collection]," Online http://media.xiph.org/video/derf/ [Last accessed: 10 March 2013].

[24] W. D. Hoyer and D. J. Macinnis, "Consumer Behavior," 5th ed., South Western Educational Publishing, 2008.

[25] T. Koga, K. Iinuma, A. Hirano, Y. Iijima, and T. Ishiguro, "Motion-compensated interframe coding for video conferencing," Proc. NTC 81, pp. C9.6.1-9.6.5, New Orleans, LA, Nov./Dec. 1981.

[26] J. B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281-297, 1967.

[27] How a movie on NETFLIX comes to life? http://vimeo.com/52637219 [Last accessed: 5 July 2013].