Additive Log-Logistic Model for Networked Video Quality Assessment

Fan Zhang, *Member, IEEE*, Weisi Lin, *Senior Member, IEEE*, Zhibo Chen, *Senior Member, IEEE*, and King Ngi Ngan, *Fellow, IEEE*

***Abstract*— Modeling subjective opinions on visual quality is a challenging problem, which closely relates to many factors of the human perception. In this paper, the additive log-logistic model (ALM) is proposed to formulate such a multidimensional nonlinear problem. The log-logistic model has flexible monotonic or nonmonotonic partial derivatives and thus is suitable to model various uni-type impairments. The proposed ALM metric adds the distortions due to each type of impairment in a log-logistic transformed space of subjective opinions. The features can be evaluated and selected by classic statistical inference, and the model parameters can be easily estimated. Cross validations on five Telecommunication Standardization Sector of International Telecommunication Union (ITU-T) subjectively-rated databases confirm that: 1) based on the same features, the ALM outperforms the support vector regression and the logistic model in quality prediction and, 2) the resultant no-reference quality metric based on impairment-relevant video parameters achieves high correlation with a total of 27 216 subjective opinions on 1134 video clips, even compared with existing full-reference quality metrics based on pixel differences. The ALM metric wins the model competition of the ITU-T Study Group 12 (where the validation databases are independent with the training databases) and thus is being put forth into ITU-T Recommendation P.1202.2 for the consent of ITU-T.**

***Index Terms*— Correlation and regression analysis, feature evaluation, image quality assessment, multivariate statistics.**

I. Introduction

MODELINGsubjective appraisal, such as opinion survey and credit scoring, is a common problem in the field of psychophysics, economics, education, cognitive science, and artificial intelligence. Considering a set of entities conforming to a predetermined population of interest, subjective appraisal about the focused entity's properties, often being rated from "best" to "worst" with limited scales, is expected to be modeled and thus predicted by a mathematical function with respect to the entities' features. In this study, we focus on the subjective quality of networked videos, which widely applies to video coding, video service diagnosing and optimizing, and multimedia service recommending [1]. Given the video clips compressed by a standard source codec and impaired by typical transmission errors, their perceived quality under regular viewing environments is to be modeled by a function about video features. A good model should match its predicted quality scores with the ground truth of subjective opinions.

Visual quality metrics are usually classified to the full-reference, reduced-reference, and no-reference metrics according to the availability of the reference (i.e., the original unimpaired video), and to the packet-layer, bitstream-layer, and pixel-layer metrics according to the accessible stage of video data. No-reference metrics have the widest scope of applications due to their possibilities to be used even when the reference is not available or too expensive to process [2].

Feature selection and model selection are the basic issues of modeling [3]. From the point of view of the features and models selected, we may have a different perspective about visual quality metrics. Modeling methodology for visual quality metric may be classified to (A) causal modeling, (B) regression, and (C) statistical learning. The features employed by the state-of-the-art metrics often include: (a) difference between the impaired video and the reference, (b) deviance from the impaired video to the priori, (c) impairment-relevant configurations in the processing chain of video data, and (d) other heuristic features of video.

To the researchers of (A) causal modeling, visual quality models must account for visual perception regularities. They attempted to model how the human visual system (HVS) responses to visual stimuli, especially from the early stage of visual perception. If the reference is available to the HVS, they exploited features (a), like pixel difference, statistic difference [4], etc. If the reference is absent, features (b) are natural candidates, such as the continuity among neighborhood pixels, the singularity across edges, the correlation between spatial-frequency sub-bands, etc. [1], [2]. Features (b) usually capture the artifacts like blockiness, blurriness, jerkiness, etc. [1], [2]. In the causal models, the select features were frequently inspired by visual-psychological experiments (like a recent example [5]) or derived by theoretical model (e.g., information-theoretical analysis [6]). The causal models generally concentrate on the ordinal match between objective predictions and subjective opinions. For numerical match, it requires a *monotonic map* from objective predictions to subjective opinions additionally. The inherent complexity of

Manuscript received March 11, 2012; revised November 16, 2012; accepted December 3, 2012. Date of publication December 11, 2012; date of current version February 6, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. E. Barth. This work was supported in part by the Singapore Ministry of Education Academic Research Fund (AcRF) Tier 2 under Grant T208B1218 and in part by the Research Grants Council of the Hong Kong SAR, China under Project CUHK 415712.

F. Zhang and Z. Chen are with the Department of Research and Innovation, Technicolor (China) Technology Co. Ltd., Beijing 100192, China (e-mail: fan.zhang@ieee.org; chenzb@ieee.org).

W. Lin is with the School of Computer Engineering, Nanyang Technological University, Singapore (e-mail: wslin@ntu.edu.sg).

K. N. Ngan is with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: knngan@ee.cuhk.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2012.2233486

the HVS, unfortunately, makes such models far from true causal models.

With access to an increasing number of subjective image/video quality databases, visual quality metrics can be learned from data, and thus the (B) regression and (C) statistical learning methods become more popular in this field. An additional *monotonic map* from objective predictions toward subjective opinions is not indispensable any longer, since it has been learned and contained in the statistical models. Moreover, the features can not only be designed empirically as in causal modeling, but also be evaluated and gauged by statistical inference. The difference between (B) and (C) is: in conventional regression the functional forms linking the quality to the features are predetermined before the fitting process, while in statistical learning the functional forms are generally arrived inductively from data [3].

Besides features (a) and (b), some configurations in the processing chain of video data [i.e., features (c)] were also exploited in (B) regression [8]–[10] and (C) statistical learning [11], [12] methods. For example, source configurations, like QP (quantization parameter), can forecast visual quality after compression, while channel configuration, e.g., packet loss rate, can forecast visual quality with packet loss. The source and channel configurations are the reasons of video impairments, and yet do not deterministically impair a video block or cause a distortion (for example, when the packet loss rate is known as 1% we cannot tell exactly which block will be lost and how bad the loss will be). Nevertheless, the configuration variables highly correlate with visual quality of video stream. They may be deemed as statistical descriptions of video. We call the feature (c)-based models descriptive models because visual quality is predicted according to statistical descriptions, following Zhu's conceptualization [13].

It is a step-wise roadmap of firstly modeling the quality attributes due to each type of impairments or artifacts and then combining them to a total quality (e.g., [10], [14]–[16]). However, for the video with hybrid impairments, which presents diversified artifacts, it is still an open problem how to combine multiple quality attributes to a composite prediction, since various impairments or artifacts may affect the perceived quality in different manners. Recent studies [10], [14] use a product function to combine the quality attributes in the subjective opinion space, this may work well in practice but is based on little ground. Another relevant method, multidimensional scaling (MDS), attempt to find an embedding from the subjective opinion space to a quality-attribute-relevant feature space such that distances were preserved [17]–[20]. MDS does not derive from raw data, but from the distance between every two data points. A simple distance metric (often Euclidean distance) in the attained feature space is guaranteed to be isotonic with the distance in the subjective opinion space. Thereby, how to combine quality attributes has been defined by the distance formulation. Yet, MDS is criticized as the attained feature spaces lack physical meanings [2].

The problem of quality-attributes combination is not explicitly addressed by (C) statistical learning either. Although machine learning tools including the support vector regression [15], [21] and the neural network [11], [22] present a general way to exploit high dimensional features from all of (a)∼(d) (e.g., metadata in bitstream [11], content measures [22], scene statistics [15], mel-cepstrum [21], etc.), they seldom provide explicit relationship about visual quality against features.

Some prior arts deemed quality assessment as a classification problem: whether visual artifacts [23] or target signals [24] can be accurately perceived or not. Statistical learning methods (C) are preferred for such problem formulations.

Our study differs from the prior modeling methods in two aspects. First, we aim at an explicit model (rather than a black box) with not only good prediction accuracy but also the plausibility in visual perception and the feasibility for quality optimization. Second, the features of model can be evaluated for selection and the modeling can thus be adaptive to the emerging applications and data. For this purpose, we keep the model simple (i.e., general linear) and complete (i.e., covering the monotonic map), so the parameters can be reliably estimated and the statistical significance of features can be inferred.

In this paper, first, we aim at formulating a new functional forms to better capture the relationship of visual quality against the features extracted from videos; second, we proposes a no-reference video quality metric considering both bitstream and pixel layer information of videos with hybrid impairments. We develop a step-wise regression framework with regularized procedures of feature selection and parameter estimation, for modeling subjective appraisal. That is, our methodology belongs to (B) regression, and our metric uses features (c) and (b). The proposed method is also expected to be applicable to full-reference measurement.

The rest of the paper is organized as following. The additive log-logistic model is derived in Section II, where the roadmap of step-wise regression is sketched in Section II-A, uni-type impairments are modeled in Section II-B, hybrid impairments are involved in Section II-C, and then the objective function of the modeling is formulated in Section II-D. Feature selection and parameter estimation for the framework are presented in Section III. The features in the proposed ALM metric are described in Section IV. The experimental results are reported in Section V, followed by the conclusion in Section VI.

II. Model Selection

A. Model Architecture

When a video suffers from hybrid types of impairments, subjects usually assess its quality according to all types of distortions in it. We introduce latent variables, $\{d\}$, to denote the distortion due to each type of impairment, and the final quality score comes from a combination of all the distortions.

The latent variables often are not measurable straightforwardly. However, given a set of videos which suffer from only one type of (i.e., uni-type) impairment, we can learn the quality attribute as only one latent variable is activated (i.e., taking a nonzero value). Thus, we can learn the quality model by two steps.

In the first step, we determine the quality attribute function for each uni-type impairment. In the second step, we optimize

the total function on video samples with hybrid impairments, under the constraints that the partial form of the total function should match the attribute functions determined in the first step. To some extent, the second step is similar with the problem of recovering the joint probability distribution from marginal distributions, or reconstructing a 3D structure from 2D projections. We select features in the first step while estimate parameters in the second step.

Without loss of generality, supposing there are totally three types of impairments, denoted by c, s, and f respectively. The above process can be formulated as: in the first step, we investigate the quality function $f_{\mathrm{c}}\left(d_{\mathrm{c}} \mid d_{\mathrm{s}}=d_{\mathrm{f}}=0\right)$ for c, $f_{\mathrm{s}}\left(d_{\mathrm{s}} \mid d_{\mathrm{c}}=d_{\mathrm{f}}=0\right)$ for s, and $f_{\mathrm{f}}\left(d_{\mathrm{f}} \mid d_{\mathrm{c}}=d_{\mathrm{s}}=0\right)$ for f; in the second step, we explore the total function $f\left(d_{\mathrm{c}}, d_{\mathrm{s}}, d_{\mathrm{f}}\right)$ under the following constraints:

$$\begin{cases} f\left(d_{\mathrm{c}}, d_{\mathrm{s}}, d_{\mathrm{f}} \mid d_{\mathrm{s}}=d_{\mathrm{f}}=0\right)=f_{\mathrm{c}} \\ f\left(d_{\mathrm{c}}, d_{\mathrm{s}}, d_{\mathrm{f}} \mid d_{\mathrm{c}}=d_{\mathrm{f}}=0\right)=f_{\mathrm{s}} \\ f\left(d_{\mathrm{c}}, d_{\mathrm{s}}, d_{\mathrm{f}} \mid d_{\mathrm{c}}=d_{\mathrm{s}}=0\right)=f_{\mathrm{f}}. \end{cases} \tag{1}$$

In order to satisfy the equations above, we introduce a transform g and assume:

$$g\left(f\left(d_{\mathrm{c}}, d_{\mathrm{s}}, d_{\mathrm{f}}\right)\right)=g\left(f_{\mathrm{c}}\left(d_{\mathrm{c}}\right)\right)+g\left(f_{\mathrm{s}}\left(d_{\mathrm{s}}\right)\right)+g\left(f_{\mathrm{f}}\left(d_{\mathrm{f}}\right)\right) \tag{2}$$

that is

$$f\left(d_{\mathrm{c}}, d_{\mathrm{s}}, d_{\mathrm{f}}\right)=g^{-1}\left(g\left(f_{\mathrm{c}}\left(d_{\mathrm{c}}\right)\right)+g\left(f_{\mathrm{s}}\left(d_{\mathrm{s}}\right)\right)+g\left(f_{\mathrm{f}}\left(d_{\mathrm{f}}\right)\right)\right). \tag{3}$$

Note that we have $f_{\mathrm{c}}(0)=f_{\mathrm{s}}(0)=f_{\mathrm{f}}(0)=q_{\mathrm{best}}$ since a video always achieves the best quality when no impairment happens. Then, if $g\left(q_{\mathrm{best}}\right)=0$, Constraints (1) will be easily satisfied. For example

$$\begin{aligned} & f\left(d_{\mathrm{c}}, d_{\mathrm{s}}, d_{\mathrm{f}} \mid d_{\mathrm{s}}=d_{\mathrm{f}}=0\right) \\ & \quad=g^{-1}\left(g\left(f_{\mathrm{c}}\left(d_{\mathrm{c}}\right)\right)+g\left(f_{\mathrm{s}}(0)\right)+g\left(f_{\mathrm{f}}(0)\right)\right) \\ & \quad=g^{-1}\left(g\left(f_{\mathrm{c}}\left(d_{\mathrm{c}}\right)\right)+g\left(q_{\mathrm{best}}\right)+g\left(q_{\mathrm{best}}\right)\right) \\ & \quad=g^{-1}\left(g\left(f_{\mathrm{c}}\left(d_{\mathrm{c}}\right)\right)\right) \\ & \quad=f_{\mathrm{c}}\left(d_{\mathrm{c}}\right). \end{aligned}$$

Assumption (2) is not strong, since we only require $g\left(q_{\mathrm{best}}\right)=0$ till now. There are many choices for g, e.g., the logarithm function. Yet, if we need to keep f flexible in fitting data and guarantee g easy to be solved, the design of g should be closely dependent with the forms of f_{c}, f_{s}, and f_{f}. In the next subsections, we firstly present the log-logistic regression [25] for f_{c}, f_{s}, and f_{f}, and then we design g.

B. Log-Logistic Regression for Uni-Type Impairment

This subsection addresses how to measure the uni-type impairment in the first step. We find that (univariate) log-logistic functions with respect to the configurations of the processing chain under consideration can measure uni-type impairments with a fairly good accuracy. And if taking into account the content features of videos, (multivariate) log-logistic functions can do better. Here, we focus on the general form of log-logistic models, and the select features will be described in detail in Section IV.

A uni-type impairment can be roughly measured by a univariate log-logistic model as

$$q=f(x ; a, b) \stackrel{\text{def}}{=} \frac{1}{1+a x^{b}}, \quad (x \geq 0, \quad a>0; \quad 0<q<1). \tag{4}$$

where a and b are the model parameters. The unique independent variable x may be the key-factor for impairment (such as *QP* for compression, impaired block rate for slicing, or freezing duration for freezing). It is easy to testify: a log-logistic curve is monotonic, and parameter b controls the existence and location (if existing) of the turning point of the log-logistic curve. A log-logistic model can be transformed to a logistic model as

$$f_{\text{logistic}}=\frac{1}{1+e^{a^{'}+b x^{'}}}, \quad (x^{'}=\log x, \ a^{'}=\log a). \tag{5}$$

It is true that logistic models are popular in statistical regression and machine learning. One of their popular but not necessary characteristics is to map a variable from $(-\infty, +\infty)$ to (0, 1) via a symmetric sigmoid curve. However, features are often bounded or at least semibounded in real applications. For example, *QP* normally ranges from 1 to 51 in the H.264 codec, and the impaired block rate is always nonnegative. Thus, x seldom takes a value on the whole range of $(-\infty, +\infty)$, but $\log x$ is more likely to suffice. And therefore, a log-logistic function is often more suitable than the naïve logistic function. With the extra logarithm transform of features, a log-logistic curve can fit with a more flexible shape than a logistic curve.

Considering that a uni-type impairment is not only controlled by the key-factor but also affected by the visual content, a log-logistic model can take into account multiple variables as

$$q \stackrel{\text{def}}{=} \frac{1}{1+a \cdot z^{b_1} \cdot x^{b_0}}, \quad (x, z \geq 0; \quad a>0). \tag{6}$$

where a, b_0, and b_1 are the model parameters, and z is thought of as the co-variate associated with the key-factor x. For example, z may be a feature of content complexity and play the role of masking effect; z helps to estimate the decay (growth) trend of q with respect to x. Indeed, multiplication is commonly used to simulate masking effect [7], probably because human's sensitivity to the distortion is thought of being attenuated by masking effect. Both x and z are the features, and we thus denote them by $\boldsymbol{x}=\{x, z_1, z_2 \cdots\}$, where more co-variates of course may be considered.

Like logistic models, log-logistic models belong to the generalized linear model (GLM) [26]. With a link function,

$$g(q)=\frac{1-q}{q}. \tag{7}$$

Model (6) can be rewritten as:

$$\log g(q)=b_0 x^{'}+b_1 z^{'}+\log a, \quad (x^{'}=\log x, \ z^{'}=\log z).$$

Every type of impairments in this study can be efficiently fitted by a log-logistic model. The model parameters are easily solved, and the features can be evaluated by statistical inference as introduced in Section III.

C. Additive Model for Multitype Impairment

Attribute functions f_c, f_s, and f_f share the same form as (6). It inspires us to introduce the transform g below for the additive model (2)

$$g(q) \stackrel{\text{def}}{=} \left(\frac{1-q}{q}\right)^{1/\beta} \tag{8}$$

where β is a positive real parameter. Then, the total function (3) becomes

$$f(d_c, d_s, d_f) = \frac{1}{1+(d_c+d_s+d_f)^{\beta}}$$
$$\text{where} \begin{cases} d_c = g(f_c) = \left(a_c z_c^{b_{c1}} x_c^{b_{c0}}\right)^{1/\beta} \\ d_s = g(f_s) = \left(a_s z_s^{b_{s1}} x_s^{b_{s0}}\right)^{1/\beta} \\ d_f = g(f_f) = \left(a_f z_f^{b_{f1}} x_f^{b_{f0}}\right)^{1/\beta}. \end{cases} \tag{9}$$

Framework (9) belongs to the generalized additive model (GAM) [27], with a linked form

$$g(f(d_c, d_s, d_f)) = d_c + d_s + d_f.$$

So we call it an *additive log-logistic model* (ALM).

The transform g with parameter β determines a metric space, where the distortions are addable. Such linear metric space is not that explicit and can scarcely be obtained from a fixed transform of the subjective opinion space. This is because quality appraisal has dependency with context. Subjects cannot assess an isolated entity without being demonstrated how the "best" and the "worst" ones look like. To be specific, a video tends to be assessed with a better MOS (mean opinion score) if the quality of other videos look worse; the MOS difference between a video pair tends to decrease if all the other videos look much worse than the video pair and also if all the other videos look much better. Therefore, equal distortions in perception may not correspond to equal MOSs in different context; equal distortion differences in perception may not correspond to equal MOS differences; and the transform between the subjective opinion space and the distortion space may be not fixed if the context is not consistent. Parameter β permits transform g to adapt to the data. If taking the total distortion $d_c + d_s + d_f$ as the independent variable, β controls the log-logistic curve shape like b in (4).

D. Objective Formulation

After selecting the GLM for uni-type impairments and the GAM for the multi-type impairments, we make an assumption on the priori distribution of opinion scores, and then can estimate parameters by the maximum likelihood method. In this paper, we assume binomial rather than Gaussian as the priori distribution of opinion scores, since:

1) The opinion scores are factually bounded.
2) The opinion scores on an individual sample show an approximately symmetrical distribution when they rank in the center (i.e., intermediate quality), but a skewed distribution with a heavier tail towards the center than towards the border when they rank in the border (i.e., the worst or best quality), as demonstrated in Fig. 1.

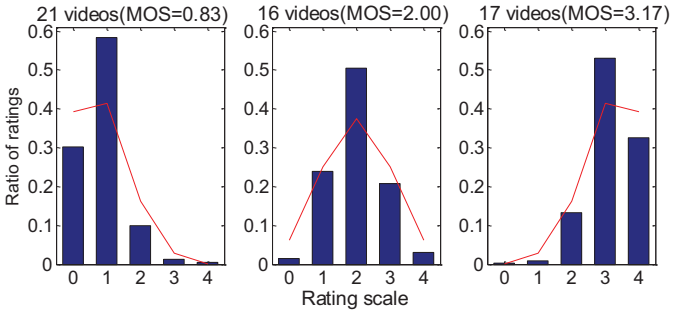

Fig. 1. Fitting histogram of opinions by binomial distribution. The video clips with equal MOS from the databases (see Section V-A) are regarded as an identical sample. Their opinion scores (on five-point scale: $0 \sim 4$) show an approximately symmetrical distribution for MOS $= 2$, but skewed distributions for MOS 0.83 and 3.17.

   The similar non-Gaussianities were also reported in psychological measurements [28], [29]. The skewed distribution occurs due to the flooring and the ceiling effects [30].
3) For the log-logistic regression, the assumption of binomial distribution may simplify the likelihood [as Eq. (13)] and thereby facilitate parameter estimation and feature evaluation, compared with Gaussian distribution.

Let us consider the binomial probability function for an opinion variable o, given by

$$\Pr(o = s) = \binom{S}{s} q^s (1-q)^{S-s}, \quad (s = 0, 1, \ldots, S) \tag{10}$$

where the opinion variable o takes a value on $(S+1)$-point scale ($S = 4$ in this study). Distribution (10) has a mean of qS, and q is thus the mean of the normalized scores. Then, the total opinion O on identical video clip rated by a total of M subjects conforms to:

$$\Pr(O = s) = \binom{MS}{s} q^s (1-q)^{S-s} \quad (s = 0, 1, \ldots, M \cdot S). \tag{11}$$

Distribution (11) has a mean of qMS. We often pay attention to the normalized mean opinion $s/(MS)$, i.e., the normalized MOS. $f(x_n)$ predicts not only the normalized MOS but also parameter q indeed, and hence we use the same symbol q to denote the parameter of binomial distribution and the visual quality.

For a total of M observed opinions on the n-th video clip, $\{o_{mn}\}$, their log likelihood as a function of q_n is thus derived as

$$L(q_n; o_{mn}) = \sum_{m=1}^{M} \left[o_{mn} \log\left(\frac{q_n}{1-q_n}\right) + S \log(1-q_n) \right] \tag{12}$$

where q_n will be predicted by $f(x_n)$, and the constant function of o_{mn} not involving q_n, namely $\sum \log\left(\frac{S}{o_{mn}}\right)$ has been omitted. In the rest of the paper, we always discuss the normalized MOSs, that is $\text{MOS}_n = \sum_m o_{mn} / (MS)$. The log likelihood normalized by MS for all MOSs is

$$L(q, \mathbf{MOS}) = \sum_{n=1}^{N} \left[\text{MOS}_n \log\left(\frac{q_n}{1-q_n}\right) - \log\left(1 + \frac{q_n}{1-q_n}\right) \right]. \tag{13}$$

TABLE I
KEY-FACTOR EVALUATION FOR UNI-TYPE IMPAIRMENT

	95% CI of b_{i0}
$f_c : (QP \mid \phi)$	-1.58 ± 0.39
$f_s : (ER \mid \phi)$	0.52 ± 0.16
$f_f : (FD \mid \phi)$	0.51 ± 0.23

TABLE II
CO-VARIATES EVALUATION FOR UNI-TYPE IMPAIRMENT

	$\Pr_{(\chi_1^2 > \Delta\text{Dev})}$	90% CI of b_{i1}
$f_c : (\log(CU+1) \mid QP)$	0.0986	-0.553 ± 0.551
$f_s : (\log(CU+1) \mid ER)$	0.0689	-0.326 ± 0.297
$f_f : (\log(MH+1) \mid FD)$	0.2562	0.211 ± 0.306

The likelihood (13) and the link function (8) share the same unit $q/(1-q)$. This is partly the reason why the logistic and the log-logistic functions are natural to model the binomial-distributed response variables. As discussed in Section III, likelihood (13) will be frequently used for parameter estimation and feature evaluation.

III. FEATURE EVALUATION AND PARAMETER ESTIMATION

A. Feature Evaluation

Before instantiation of the log-logistic models, the features need be evaluated and selected for each uni-type impairment respectively. Forward selection is adopted, that is, to initialize a null feature set and add the best unselected features at each stage until no further candidates satisfy the selection criteria. A new feature is accepted when:

Criterion 1: It brings a significant performance gain in goodness-of-fit.

Criterion 2: Its parameter b has a confidence interval not overlapped with 0, since $b = 0$ implies the feature is omitted.

For Criterion 1, we use the reduction in likelihood to judge the goodness-of-fit. Deviance is defined to be twice of the difference between the maximal log likelihood and the log likelihood attained under the fitted model [26]. The maximal log likelihood (13) is attained at the perfectly fitted points $\tilde{\boldsymbol{q}} = \mathbf{MOS}$. Under any given model, H_0, with fitted quality scores $\hat{q}_0$, the deviance function, denoted as Dev, is

$$\begin{aligned}\text{Dev}\left(\hat{\boldsymbol{q}}_0\right) &= 2l\left(\tilde{\boldsymbol{q}}; \mathbf{MOS}\right) - 2l\left(\hat{\boldsymbol{q}}_0; \mathbf{MOS}\right) \\ &= 2\sum_{n=1}^{N}\left[\text{MOS}_n \log\left(\frac{\text{MOS}_n}{\hat{\boldsymbol{q}}_{0n}}\right)\right. \\ &\quad \left. + (1-\text{MOS}_n)\log\left(\frac{1-\text{MOS}_n}{1-\hat{\boldsymbol{q}}_{0n}}\right)\right]. \quad (14)\end{aligned}$$

A lower deviance means a higher attained likelihood and thus a better fitted model.

Let the "null" hypothesis H_0 denote the model under test and the "alternative" hypothesis H_1 denote the extended model containing an additional feature. The corresponding fitted quality scores are denoted by $\hat{\boldsymbol{q}}_0$ and $\hat{\boldsymbol{q}}_1$ respectively. We judge the gain in goodness-of-fit, by the reduction in deviance

$$\begin{aligned}\Delta\text{Dev} &= \text{Dev}\left(\hat{\boldsymbol{q}}_0\right) - \text{Dev}\left(\hat{\boldsymbol{q}}_1\right) \\ &= 2\sum_{n=1}^{N}\left[\text{MOS}_n \log\left(\frac{\hat{\boldsymbol{q}}_{1n}}{\hat{\boldsymbol{q}}_{0n}}\right)\right. \\ &\quad \left. + (1-\text{MOS}_n)\log\left(\frac{1-\hat{\boldsymbol{q}}_{1n}}{1-\hat{\boldsymbol{q}}_{0n}}\right)\right]. \quad (15)\end{aligned}$$

The reduction in deviance, denoted by ΔDev, equals twice of the increase from the log likelihood under the test model to that under the fitted model with the addition of a new feature. It is approximately distributed as χ_1^2, i.e., chi-square distribution with the degree of freedom being equal to the number of the additional feature, i.e., 1. (p. 119 in [26]). The more reduction in deviance, the less probability of H_1 has a statistically indistinguishable performance as H_0, and thus the stronger evidence to support that there is a significant performance gain in H_1 over H_0. The significance of the reduction in deviance is often quantified in terms of the cumulative probability of that a χ_1^2-distributed variable is greater than the reduction in deviance, denoted by $\Pr_{(\chi_1^2 > \Delta\text{Dev})}$ and reported in Table II, Section IV-D.

For Criterion 2, we use the confidence interval (CI) of the corresponding parameter b_{ij}, as reported in Tables I and II of Section IV-D. An adaptation of the central limit theorem states that the sampling distribution of the estimated parameter $\hat{b}$ approximately conforms to a normal distribution [26]. The α confidence interval of $\hat{b}$ is calculated by

$$\hat{b} \pm t_{(1-\alpha/2)} \cdot I^{-1}(b) \quad (16)$$

where $I(b)$ is the second partial derivative of the log-likelihood with respect to b (p. 18 and p. 41 in [31]), $t_{(1-\alpha/2)}$ means the $(1-\alpha/2)$ percentile of student's t-distribution.

B. Parameter Estimation

After framework (9) is instantiated, the parameters are estimated. This is posed as a regression problem based on maximum likelihood:

$$\max_{\{a_i, b_{i,j}\}, \{\beta_u\}} \sum_{u=1}^{U}\sum_{n=1}^{N_u} L\left(q_{u,n}; \text{MOS}_{u,n}\right). \quad (17)$$

Referring to Eq. (13), the total likelihood is counted by the outer summation over totally U independent databases and the inner summation over totally N_u samples from the u-th database. Although parameters $\{a, b\}$ in framework (9) are fixed across databases, we suggest adaptive β_u for the u-th database and fitting a new β for a coming dataset. During training, adaptive β can balance the system deviation (e.g., the misaligned flooring and ceiling effects) due to the inconsistent configurations across databases (e.g., distortion range, impairment combination, and error pattern), and hence $\{a, b\}$ can focus on characterizing the intrinsic process of quality assessment (e.g., distortion pooling and combination).

The partial derivatives of the log-likelihood with respect to each element from $\{a, b\}$ and $\{\beta_u\}$ are available as

$$\frac{\partial \sum \log L}{\partial \log a_i} = \sum_{u=1}^{U}\sum_{n=1}^{N_u} \frac{\left(q_{u,n} - \text{MOS}_{u,n}\right)\beta_u d_i}{\sum_i d_i}$$

$$\frac{\partial \sum \log L}{\partial b_{i0}} = \sum_{u=1}^{U}\sum_{n=1}^{N_u} \frac{\left(q_{u,n} - \mathrm{MOS}_{u,n}\right)\beta_u d_i \log x_i}{\sum_i d_i}$$

$$\frac{\partial \sum \log L}{\partial b_{i1}} = \sum_{u=1}^{U}\sum_{n=1}^{N_u} \frac{\left(q_{u,n} - \mathrm{MOS}_{u,n}\right)\beta_u d_i \log z_i}{\sum_i d_i}$$

$$\frac{\partial \sum \log L}{\partial \beta_u} = \sum_{n=1}^{N_u} \left[\left(q_{u,n} - \mathrm{MOS}_{u,n}\right)\log\left(\Sigma_i d_i\right)\right] \quad (18)$$

where i indexes the type of impairments, e.g., $i \in \{\mathrm{c}, \mathrm{s}, \mathrm{f}\}$; d_i and $q_{u,n}$ can be computed by using the current estimations of $\{a, b\}$ and $\{\beta_u\}$ according to Eq. (9). Certainly, it is feasible to estimate a single β value for all databases, by

$$\frac{\partial \sum \log L}{\partial \beta} = \sum_{u=1}^{U}\sum_{n=1}^{N_u} \left[\left(q_{u,n} - MOS_{u,n}\right)\log\left(\Sigma_i d_i\right)\right]$$
$$\beta_u = \beta,\ (u = 1, 2, \ldots, U). \quad (19)$$

The Hessian matrix can be derived too. Note that we solve the logarithmic a_i, so as to guarantee them always positive. Consequently, the problem may be tackled by the conjugate-gradient method. The MATLAB code of the parameter estimation for additive log-logistic models is available at: http://ivp.ee.cuhk.edu.hk/projects/demo/piqm/index.html.

Framework (9) belongs to the GAM. As pointed out in [27], failure to converge is rarely a problem, unless the data are sparse and MOS = 0 or MOS = 1 for certain video clips, which can be avoided in the preprocessing.

Since β monotonically maps $(d_\mathrm{c} + d_\mathrm{s} + d_\mathrm{f})$ into [0, 1] as shown in Eq. (9), β does not affect the ordinal match between predicted scores and subjective opinions (as evaluated by the Spearman rank order correlation coefficient), and yet is still helpful for the numerical match (as evaluated by the Pearson linear correlation coefficient).

IV. Feature Extraction and Selection

Given the functional forms, modeling quality reduces to finding appropriate features. We firstly describe the select features and then briefly present the process of feature selection.

A networked video is encoded to a slice-by-slice bitstream (by the H.264 codec), packetized into transport packets, and transmitted usually through the UDP network. Visual quality is generally degraded due to the compression loss and the transmission errors.

A. Compression

Compared with transmission error, lossy compression often causes more uniform artifacts. Clip-wise features are efficient to measure uniform impairments. Clip-wise means that the feature is averaged over the video clip spatial-temporally. For example, clip-wise QP_T is the average of all the MBs' (macroblocks') QPs in the video clip.

QP may well describe the quality of the compressed videos which share the same content, but cannot balance the influence of the content complexity on visual quality. Meanwhile, visual distortions in a complex scene are more likely to be tolerated by human eyes than those in a simple scene, known as texture masking effect or contrast masking effect; hence we introduce CU (Content Unpredictability) to quantify the content complexity. For each MB (macroblock; with 16×16 pixels), its CU is the variance of the residuals in the luminance channel as

$$CU_r \overset{\mathrm{def}}{=} \frac{1}{K_{\mathrm{MB}}}\sum_{k=1}^{K_{\mathrm{MB}}}\left(I_{r,k} - \frac{1}{K_{\mathrm{MB}}}\sum_{k=1}^{K_{\mathrm{MB}}} I_{r,k}\right)^2 \quad (20)$$

where $I_{r,k}$ denotes the k-th pixel residual in the r-th MB. K_{MB} is the number of pixels of MB, i.e., 256. Generally, a greater CU after intra prediction implies a higher spatial complexity, while a greater CU after inter prediction suggests a higher temporal complexity. In summary, clip-wise CU_T may quantify the overall spatial-temporal complexity of the video clip. Visual distortions are more likely to be tolerated in complex videos with greater CU_T. We call it *global masking effect*.

B. Slicing

To alleviate packet loss, a decoder may utilize error concealment. The common techniques include a slicing mode and a freezing mode [32]. In the slicing mode, a decoder repairs the lost slice using the neighbor pixels which have been reconstructed, so not every impaired block yields a visible distortion. The visibility of local distortion is influenced by the error concealment efficiency and the local masking effect. A block can be well recovered when it is highly correlated with its temporal or spatial neighborhoods; otherwise, error concealment may yield visible borders around the recovered blocks, named *mosaic artifacts*. Besides, content itself may mask the artifacts to some extent. Therefore, it is necessary to identify the MB-wise visible artifacts and pool them into a clip-wise factor. We design the key-factor to slicing as

$$ER_T \overset{\mathrm{def}}{=} \sum\nolimits_t \left[\left(\sum\nolimits_r \frac{MA_{t,r}\, EP_{t,r}}{1 + TX_{t,r}^{c_2}}\right)^{c_1}\right] \Big/ (T \cdot R) \quad (21)$$

where ER is the short form for Error Rate; ER_T (visible Error Rate) is calculated by the inner summation over MBs (spatially); the outer summation is over pictures (temporally) and the normalization by (T•R), where r indexes the MB in a picture, t indexes the picture in a video clip, R is the total number of MBs in a picture and T is the total number of pictures in a video clip; c_1 relates to the temporal pooling strategy [33]; EP (Error Propagation flag) and MA (suspected Mosaic Artifact) are MB-wise Boolean features, respectively indicating whether the MB is propagated by packet loss or not and whether the MB presents suspected mosaic artifacts or not; TX (non-edge TeXture) quantifies the non-edge texture of an MB.

Generally, a textural region can hide more artifacts than a smooth region or edge one. We use bilateral filter to decompose each picture into a structure and a texture components. TX is quantified by the MB variance of the texture component and meanwhile is thresholded to 0 where the structure component presents any prominent edges.

Fig. 2 illustrates how to derive and merge TX, MA and EP into ER. First, each picture [Fig. 2(a)] is decomposed to a

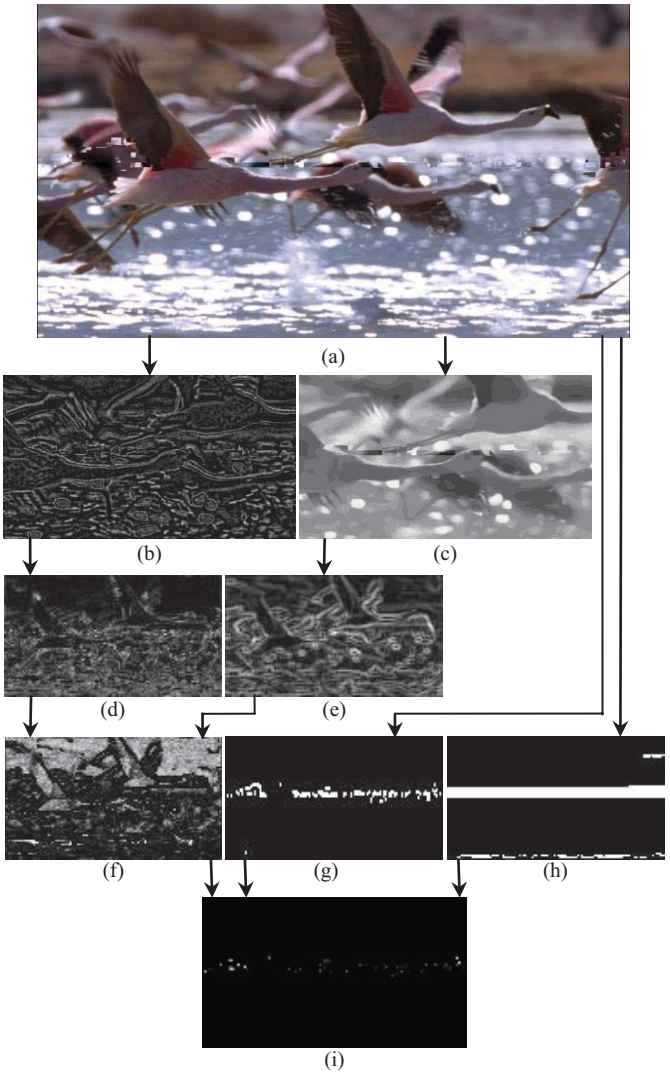

Fig. 2. Workflow of computing key-factor to slicing. (a) Impaired picture. (b) Texture component. (c) Structure component. (d) Texture strength. (e) Edge map. (f) Texture mask $1/(1 + TX)$. (g) Mosaic artifacts *MA*. (h) Error propagation *EP*. (i) Visible error rate map *ER*.

structure [Fig. 2(c)] and a texture components [Fig. 2(b)] by 7×7 bilateral filter [34]. The filtered picture is deemed as the structure while the difference between the original picture and its structure is thought of as the texture. Second, an edge map [Fig. 2(e)] is labeled where the Sobel filtering response on the structure component is above an empirical threshold of 150 for 8-bit depth. The vertical and the horizontal Sobel filtering are performed respectively and the absolute values of the two direction responses are summed. Before filtering, the structure component is downsampled at 1:16 so as to link an attained "pixel" to an MB (note that an MB is sized of 16×16 pixels). Third, texture strength [Fig. 2(d)] is calculated from the MBs' variance in the texture component [Fig. 2(b)], then is thresholded to 0 for the edge region in the edge map [Fig. 2(e)], and finally yields the texture mask [Fig. 2(f)] by a log-logistic mapping.

MA [Fig. 2(g)] is regarded to be true when unsmooth vertical gradients at the MB borders are detected as (22). The second-order gradient sums for each vertically adjacent MB pair are calculated. Thus, each MB corresponds to an upper and a lower gradient sums. The smaller one is compared with a threshold c_3 and determines whether the mosaic artifact occurs on the MB as (22). c_3 is set as 240 in this work

$$MA \stackrel{\text{def}}{=} \begin{cases} 1, & \text{if } \min\left\{\left|\nabla^2_{\text{upper}}\right|, \left|\nabla^2_{\text{lower}}\right|\right\} \geq c_3 \\ 0, & \text{else.} \end{cases} \tag{22}$$

EP [Fig. 2(h)] is true if the block is lost or the block (directly or indirectly) uses lost blocks for prediction. *EP* is parsed from the bitstream directly.

Lastly, $1/(1+TX)$ [Fig. 2(f)], *MA* [Fig. 2(g)], and *EP* [Fig. 2(h)] are MB-wisely merged into *ER* [Fig. 2(i)] as (21), where *TX* occurs in the denominator and is mapped to a masking multiplier to *MA* and *EP* by a log-logistic function; c_2 together with *TX* simulates the texture masking effect. c_2 is set as 1 after a grid search optimization, and the log-logistic mapping is decreasing, and thus texture masking is large in edge and smooth regions while small in textural ones.

C. Freezing

In the freezing mode, the decoder replaces the pictures which have been impaired or propagated by packet losses with their previous intact picture, until a decoded picture without errors has been received. Therefore, each pause is determined by the packet loss pattern and the picture types, and can be parsed from the bitstream directly. There may be several occurrences of pauses in a video clip. The total freezing duration (FD) is then calculated as

$$FD_T \stackrel{\text{def}}{=} \sum\nolimits_{\tau} FD_{\tau}^{c_4} \Big/ T \tag{23}$$

where FD_τ is the total number of pictures covered by the τ-th pause and FD_T is the total durations over all pauses normalized by T (the number of pictures). In (23), c_4 affects the temporal pooling strategy [33] and is set as 0.9 after a grid search optimization.

Each pause causes a content stagnation and then a content skip (it differs from buffering without skip). If a scene itself is still or nearly still, a factual freezing is often unnoticeable; however, if a scene has homogenous (even very slow) motions, it is easy for human eyes to identify. We introduce clip-wise MH_T to quantify the Motion Homogeneity:

$$MH_P \stackrel{\text{def}}{=} \sum\nolimits_{\tau=1}^{P} \max\{PH_\tau, ZH_\tau\} \Big/ P \tag{24}$$

where *PH* quantifies the Panning Homogeneity while *ZH* specifies the Zooming Homogeneity. MH_P is the mean of the maximums between *PH* and *ZH*, which is averaged over the total of P pauses. For each pause, *PH* is defined as the magnitude of the vector mean of motion field, while *ZH* is defined as the mean of motion's radial projections, as

$$\begin{aligned} PH_\tau &\stackrel{\text{def}}{=} \frac{1}{R}\sqrt{\left(\sum_{r\prec\tau} MV_{\text{h},r}\right)^2 + \left(\sum_{r\prec\tau} MV_{\text{v},r}\right)^2} \\ ZH_\tau &\stackrel{\text{def}}{=} \frac{1}{R}\left|\sum_{r\prec\tau} \langle \overrightarrow{MV}_r, \frac{\vec{r}}{|\vec{r}|}\rangle\right| \end{aligned} \tag{25}$$

$$\approx \frac{1}{R}\sqrt{\left|\sum_{r\prec\tau_{\mathrm{L}}} MV_{\mathrm{h},r} - \sum_{r\prec\tau_{\mathrm{R}}} MV_{\mathrm{h},r}\right|^2 + \left|\sum_{r\prec\tau_{\mathrm{T}}} MV_{\mathrm{h},r} - \sum_{r\prec\tau_{\mathrm{B}}} MV_{\mathrm{h},r}\right|^2}$$

where the motion field is approximated by motion vectors $\overrightarrow{MV}$; r is the spatial index of $\overrightarrow{MV}$, in the picture which just precedes the τ-th pause (denoted by $r \prec \tau$); $\overrightarrow{MV} = (MV_h,\ MV_v)$ and $\vec{r} = (r_{\mathrm{h}},\ r_{\mathrm{v}})$, both consisting of a horizontal and a vertical components; $(r_{\mathrm{h}} r_{\mathrm{v}})$ is the coordinate of r when the origin point is the center of picture; $<\bullet,\bullet>$ denotes the vector projection (i.e., dot product); R is the total number of r in the picture. *ZH* is a little complicated to compute, and hence we use a simpler method for approximation. Firstly, we minus the sum of horizontal motion vectors in the left half picture (denoted by $r \prec \tau_{\mathrm{L}}$) by those in the right half picture ($r \prec \tau_{\mathrm{R}}$), and minus the sum of vertical motion vectors in the top half picture ($r \prec \tau_{\mathrm{T}}$) by those in the bottom half picture ($r \prec \tau_{\mathrm{B}}$). Then, the magnitude of the attained vector is calculated and normalized by R.

PH is large when either a big object passes by the camera or the camera is panning, tilting, booming or tracking some objects. *ZH* is large when the camera is zooming or dollying.

D. Feature Selection

In this study, feature selection has involved a number of attempts. Apart from the aforementioned *QP*, *CU*, *ER*, *FD* and *MH*, the feature candidates included bit-rate, packet-loss-rate, packet-loss-frequency, the interval from the impaired block to the co-located one for each error concealment, the ratio of intra- prediction among MBs, the average number of the MB partitions for motion estimation, the mean of MVs, etc. For each feature candidate x, we also try its variation $h(x)$, where transform h can adjust the range and distribution of x for fitting and may take the forms such as:

1) Translation: $x - x_{\mathrm{l}}$ or $x_{\mathrm{h}} - x$, where x_{l} and x_{h} are the lower and the higher bounds of x respectively. For example *QP* in H.264 video is no greater than 51 and we find that $51-QP$ is more powerful than *QP*.
2) Logarithm: $\log(x - x_{\mathrm{l}}+1)$ or $\log(x_{\mathrm{h}} - x+1)$. The logarithm function may compact the maximum relative to the minimum, and attenuate the fluctuation of x. For example, we find that $\log(CU+1)$ is better than *CU*.
3) Exponential: e^x. The model then turns to a log-logistic model of f with respect to x.

The features for compression were selected by testing on the compressed videos without packet loss; the features for slicing corresponded to the videos with slight compression as well as the slicing mode of error concealment; and so forth for freezing.

We started from a null initial feature set. In the first round of feature selection, *QP*, *ER*, and *FD*, were found the most significant among the feature candidates for the uni-type of impairment, and thus accepted as the key-factors; in the second round, $\log(CU+1)$ and $\log(MH+1)$ were found more significant than other candidates and thus accepted as the co-variates; in the third round, the remaining features were found not significant and thus rejected.

Tables I and II list the statistical inference results about the select key-factors and co-variates, respectively. $f_i : (z \mid x)$ means to add z to the attribute function f_i, which has already contained x_i. The criteria which have been introduced in Section III-A are reported, including the cumulative probability of that a χ_1^2-distributed variable is greater than the reduction in deviance, and 90% confidence interval (CI) of corresponding b_{i1}. The statistical significance of a co-variate is indicated by a lower $\Pr_{(\chi_1^2 > \Delta\mathrm{Dev})}$ or a narrower CI.

The influence of *QP*, *ER*, and *FD* is clear, since the 95% CIs of b_{i0} are not overlapped with 0 in Table I. From Table II, it mainly concludes that the influence of co-variates is limited but in line with the well-known facts about subjective visual quality assessment. Both *CUs* (of f_{s} and f_{f}) satisfy that the 90% CIs are not overlapped with 0; while *MH* fails to satisfy. Actually, 70% CI of *MH* is not overlapped with 0. Therefore, we can only say that all the co-variates are weak features from the view point of conventional regression. However, *CU* corresponds to negative parameter b in both f_{c} and f_{s}. It confirms that complex contents (with greater *CU*) exhibit stronger masking effect and thus lead to higher distortion-resilient characteristic. *MV* having a negative b_{f1} confirms that contents with higher motion are more likely to expose the freezing impairments. Hence, we still use these co-variates and yet add no more co-variate. They do not to increase the risk of overfitting, as testified by the cross validations in Section V-B.

Finally, the attribute functions are devised as

$$\begin{aligned} f_{\mathrm{c}} &= \frac{1}{1 + a_{\mathrm{c}}\left[\log\left(CU_T + 1\right)\right]^{b_{\mathrm{c1}}} \left(51 - QP_T\right)^{b_{\mathrm{c0}}}} \\ f_{\mathrm{s}} &= \frac{1}{1 + a_{\mathrm{s}}\left[\log\left(CU_T + 1\right)\right]^{b_{\mathrm{s1}}} ER_T^{b_{\mathrm{s0}}}} \\ f_{\mathrm{f}} &= \frac{1}{1 + a_{\mathrm{f}}\left[\log\left(MH_T + 1\right)\right]^{b_{\mathrm{f1}}} FD_T^{b_{\mathrm{f0}}}} \end{aligned} \tag{26}$$

and framework (9) is instantiated as

$$f = \frac{1}{1 + \left(a'_{\mathrm{c}} z_{\mathrm{c}}^{b'_{\mathrm{c1}}} x_{\mathrm{c}}^{b'_{\mathrm{c0}}} + a'_{\mathrm{s}} z_{\mathrm{s}}^{b'_{\mathrm{s1}}} x_{\mathrm{c}}^{b'_{\mathrm{s0}}} + a'_{\mathrm{f}} z_{\mathrm{f}}^{b'_{\mathrm{f1}}} x_{\mathrm{f}}^{b'_{\mathrm{f0}}}\right)^{\beta}} \tag{27}$$

where $a'_i = a_i^{1/\beta}$, $b'_{ij} = b_{ij}/\beta$ ($i \in \{\mathrm{c}, \mathrm{s}, \mathrm{f}\}$; $j = 0, 1$). In (27), key-factors x_i are $(51-QP_T)$, ER_T, and FD_T while co-variates z_i are $\log(CU_T + 1)$, $\log(CU_T + 1)$, and $\log(MV_T + 1)$, for compression ($i = \mathrm{c}$), slicing ($i = \mathrm{s}$), and freezing ($i = \mathrm{f}$), respectively.

V. Experimental Results

A. Subjective Database

ITU-T SG 12 (Study Group 12 of Telecommunication Standardization Sector of International Telecommunication Union) recently built a set of subjectively-rated databases for standardization of multimedia quality models. We choose the databases which are specially designed for pure videos (other than audio-videos) in IPTV (other than mobile TV) scenario. Five databases meet our requirement, which contain a total of 1134 impaired video clips. The five databases were with separate source video selection, impairment generation,

TABLE III
NUMBER OF VIDEOS IN ITU-T SG12 DATABASES

DB #	09	10	08	11	15
Resolution	PAL	720 p	1080 p	1080 i	1080 p
Frame rate	25	50	30	30	25
Compression(c)	48	96	72	80	56
Slicing(s)+c	140	96	88	93	98
Freezing(f)+c	30	47	72	63	55
All	218	239	232	236	209

subjective quality assessment, and yet concerted requirements for the configurations. These requirements were:

1) Each database used eight source videos (including four common ones and four particular ones) as reference videos. They were of high quality, with resolution of 576, 720, or 1080 line, at frame rate of 25, 30, or 50 fps. The video contents were diversified along the spatial and temporal complexity.
2) The impairments included H.264 compression, H.264 compression plus slicing (EC mode), and H.264 compression plus freezing (EC mode). A predetermined range of bitrates (BR) and packet loss rates (PLR) gauged the level of compression loss and transmission error respectively. For SD (standard definition, e.g., PAL) videos, BR may vary within $0.5 \sim 9$ Mbps and PLR within $0.25 \sim 2\%$. For HD (high definition) videos, BR may vary within 1~30 Mbps and PLR within $0.02 \sim 1.5\%$. About 30 impaired videos were generated from each reference video.
3) Environments of subjective quality test accorded with [35]. Subjective opinions were reported on five-point discrete scale in a manner of ACR (absolute category rating). Excluding outlier subjects, exactly 24 subjects were required. For each video clip, the average score, termed as MOS (mean opinion score), was taken as the ground truth of the visual quality.

Referring to de facto configurations in real applications, the requirements also defined the group-of-pictures size, presence of hierarchical B mode, slice size, packetlization protocol, simulation model of transmission error, implementation of error concealment, etc. Such requirements restricted the system deviance across databases. The configurations of databases are briefly outlined in Table III. Note that the freezing impairment includes only the pausing with skipping but no buffering, i.e. pausing without skipping.

B. Model Comparison

We compare the proposed ALM (27) with the logistic model and the support vector regression, based on the same features. It aims at justifying the functional form of the ALM.

The logistic model was suggested by [8], [9]. We find that the inputs of $\{\log(51 - QP_T)$, ER_T, FD_T, $\log(CU_T+1)$, $\log(MH_T+1)\}$ are better than the raw features $\{QP_T, ER_T, FD_T, CU_T, MH_T\}$.

Although other choices of machine learning techniques are possible, in this paper, we compared with the SVR because it is well established technique and recommended by previous relevant studies [12], [15], [21]. We tried two SVR models. Model SVR I has the input of all the raw features in the ALM (27), including QP_T, ER_T, FD_T, CU_T, and MV_T. Each kind of features is linearly normalized into [0, 1]. Model SVR II is based on the input of f_c, f_s, and f_f, which are calculated by (26), and share exactly the same parameters with the ALM (27). No feature normalization is needed, since f_c, f_s, and f_f are already unbiasedly distributed within [0, 1]. We use the radial basis function (RBF) as the kernel of SVR, which is of the form $K(\boldsymbol{x}_v, \boldsymbol{x}) = \exp\left\{-\gamma \|\boldsymbol{x}_v - \boldsymbol{x}\|^2\right\}$ where $\boldsymbol{x}_v$ is the feature of a support vector, $\boldsymbol{x}$ is the feature of a test sample and the positive parameter γ controls the radius. The code is developed based on the ε-SVR mode of the MATLAB library for support vector machine [36], where the positive parameter ε controls the tube width to isolate support vectors and the positive multiplier c controls the regularization term of smoothing the SVR function. We find that SVR I performs best at $\gamma = 9$, $\varepsilon = 0.05$, and $c = 5$ while SVR II does at $\gamma = 72$, $\varepsilon = 0.05$, and $c = 8$.

Since the models need to be trained, we evaluate their performance by the k-fold cross validation. The data was split into k chunks, one chunk for test and the remaining $(k-1)$ chunks for training. The experiment was tested with each of the k chunks. The performance is evaluated by the mean accuracy of the tests over all the chunks. The data was carefully split in the following two ways. Since the data came from five databases, in the first way four databases were used for training and the remaining one database was for test. Since there were eight source videos in each database, in the second way data corresponding to every six source videos was used for training and the remaining data from the other two source videos was for test. The first splitting way led to $\binom{5}{1} = 5$ cases and the second way yielded $\binom{8}{2} = 28$ cases, so the accuracy was averaged on a total of 33 cases. Such meaningful splitting ways were suggested by [21]. The first way investigate the influence of the difference between databases, while the second way guarantees that the contents in test were absent in training.

The accuracy is reported in terms of three popular criteria namely: mean squared error (MSE), Pearson linear correlation coefficient (PLCC) and Spearman's rank order correlation coefficient (SROCC), between MOSs and objective predictions. For a perfect match between objective and subjective ratings, MSE = 0 and PLCC = SROCC = 1. Better accuracy is indicated by a lower MSE and a higher PLCC and SROCC. Before computing on a test chunk, linear mapping between objective outputs and MOSs was done for each database, to balance the slightly different system deviation across the five databases. For a fair comparison, we trained only one set of parameters $\{a, b, \beta\}$ for the ALM and applied them to all tests.

The accuracy is listed in Table IV. First, SVR II outperforms SVR I. It confirms that the log-logistic models about f_c, f_s, and f_f are effective forms to organize the raw features. It is also justified to introduce the latent variables d_c, d_s, and d_f. Second, the proposed model significantly outperforms SVR II, as the statistical significance is evaluated by the F-statistics about the ratio of $\text{MSE}_{\text{SVRII}}$ to MSE_{ALM} with the degree of freedom $1134-11$ (the cross validations used 1134 independent samples and the ALM metric has 11 parameters).

TABLE IV
AVERAGE ACCURACY OVER ALL DATABASES IN CROSS VALIDATIONS

	Logistic	SVR I	SVR II	**ALM**
MSE	0.0231	0.0198	0.0112	**0.0083**
PLCC	0.7979	0.8301	0.9071	**0.9328**
SROCC	0.8163	0.8375	0.8849	**0.9217**

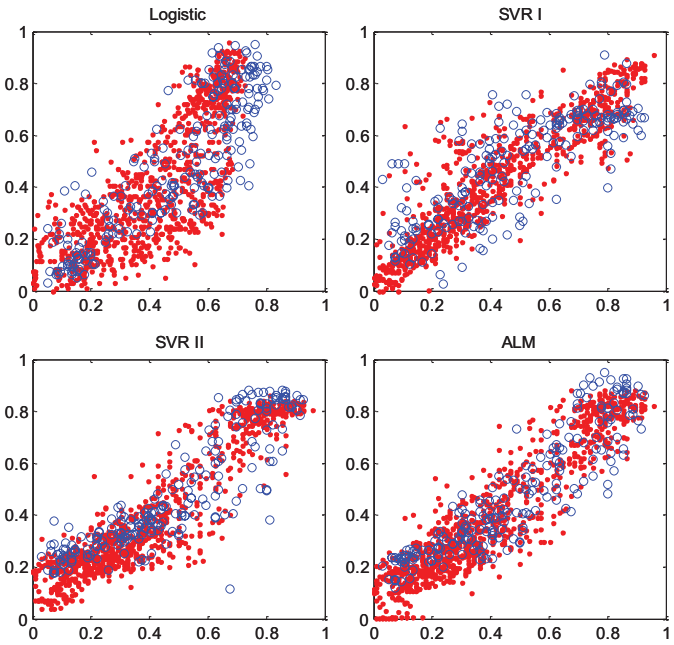

Fig. 3. Scatter plots of quality prediction in a cross validation (the train samples marked by (red) points and the test samples marked by (blue) circles).

It confirms that the additive model (27) is more efficient to combine multi-type impairments. Third, the logistic model is worse than the ALM. It confirms that the advantage of the ALM is partly attributed to the additive combination.

The scatter plots of objective predictions versus MOSs at one case of the cross validations are shown in Fig. 3. In this case, four databases (TR09, TR10, TR08, and TR11) were used for training, and the remaining database (TR15) was for testing. The test samples [marked by (blue) circles] obtained the similar prediction accuracy as the training samples [marked by (red) points] did. So, it seems that the resultant 4 models trained over 4 databases do not overfit on the 5th database.

C. Metric Comparison

We compared the no-reference ALM metric (27) with the full-reference metrics PSNR (peak signal noise ratio) and VQM [37], so that it could be justified whether video features were fully exploited in the ALM metric. PSNR is the most popular metric due to its simplicity, while VQM was adopted by the ANSI as a standard (ANSI T1.801.03-2003). Both metrics employ the pixel-layer information of videos. PSNR has no parameters to be tuned; VQM has its parameters trained during the previous study and we used its publicly-available source code. For a fair comparison, the ALM metric was trained on the other four databases and then tested on the target database, similar with the cross validations in Section V-B.

In order to compare the metrics for different impairments, the sub-datasets of compression, slicing plus compression, freezing plus compression, and the full dataset were tested respectively. We used SROCC to evaluate the metric performance. As shown in Fig. 4, the proposed metric outperforms the other two metrics for nearly all the data sets, except that it is inferior to VQM for the compression impairment in Database TR15. When comparing with PSNR and VQM, the ALM metric is more advantageous for the slicing and freezing than for the compression impairment. This implies that the proposed impaired block rate can better describe the visible artifacts and the freezing duration can better measure the annoyance due to visual pauses, compared with the video difference measurement in PSNR and VQM.

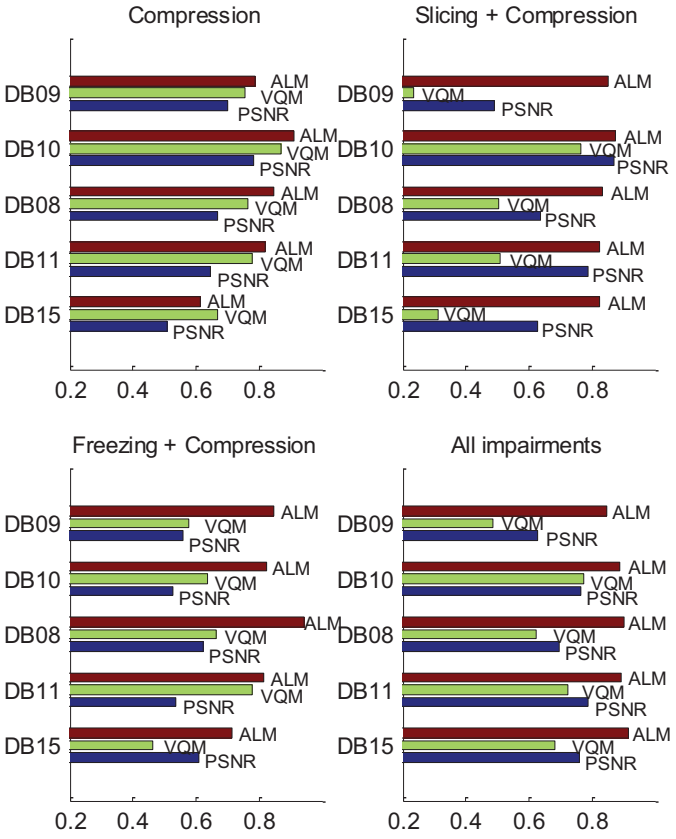

Fig. 4. SROCC of quality metrics on ITU databases.

TABLE V
PARAMETERS TRAINED IN EACH ITU DATABASE

	β_u	b_{c0}	b_{s0}	b_{f0}
TR09	0.440	−1.814	0.568	0.710
TR10	1.637	−1.868	0.764	1.297
TR08	0.357	−1.807	0.600	0.769
TR11	0.936	−1.609	0.641	0.915
TR15	0.887	−1.561	0.609	0.954
Variance	0.260	0.019	0.006	0.052

D. System Deviance Across Databases

The valuation stage in the subjective quality rating is usually dependent with the context of video clips. This can be testified by the fluctuation of the parameter which was trained in each database respectively, as adaptive $\{\beta_u\}$ is enabled in parameter estimation. As shown in Table V, although the five databases have shared concerted configurations, the tuned β_u

still varies a little bit more than b_{c0}, b_{s0}, and b_{f0} do, where the fluctuation of the parameter is quantified by the variance of the parameters tuned across databases. The different ranges of video impairments may cause the misaligned ceiling and flooring effects, and therefore β_u varies to fit with the valuation stage in each database. It implies that the additive space for d_c, d_s, and d_f should be adaptive to the data and may not be captured by a fixed transform of the subjective opinion space.

VI. Conclusion

The main results of this paper include the log-logistic models (26) to predict each quality attribute of a video with hybrid impairments and the proposed ALM (27) to predict the composite quality. The key points of the proposed framework include:

1) Since the binomial distribution is more appropriate for describing subjective opinions than the Gaussian distribution, the maximum likelihood based on binomial distribution is preferred as the goodness-of-fit criterion. This makes Point 2 below natural.
2) A log-logistic model can capture the visual quality attribute against the impairment-relevant feature. The multivariate log-logistic model with additional consideration of content features can improve the prediction accuracy further, and features can be evaluated and selected by means of statistical inference.
3) The ALM appropriately adds various quality attributes in a space, which is log-logistically transformed from the subjective opinion space. The devised quality metric is capable of accurate prediction.

The ALM metric has won the model competition in ITU-T Study Group 12 [38], where the validation databases were unknown when the metrics were designed. Up to now, we are drafting the standard recommendation P.1202.2 (containing the pseudo codes of the ALM metric) for the consent of ITU in near future.

How to combine multitype impairments (or artifacts) and composite quality attributes into a compositive prediction involves the middle or late stage of visual perception and recognition. Solving the problem by a pure causal model is still difficult, due to current lack of necessary knowledge about the related process in the HVS. Resorting to statistical learning methods is sometimes too general to exploit the special relationship among data. The proposed ALM provides an accurate functional description of the subjective appraisal, in the study of networked-video quality assessment. Based on the configurations in the processing chain and the content features, it outperforms the logistic model and the support vector regression model, and also enables a no-reference quality measurement to achieve comparable accuracy as full-reference metrics do. Like most regression tools, the ALM does not reveal the causal mechanisms underlying the coactions of video features.

Nevertheless, the ALM offers a concise, computationally tractable description of subjective visual quality. It relies on a parsimonious set of features and parameters, and keeps the parameters easy to be estimated. It allows us to measure the distortions due to every uni-type impairment as well as multi-type impairments, and clarify the relative contributions of impairments in the addable metric space. More generally, the model can be used to evaluate which feature is the most significant to predict quality. We expect this framework to extend to other subjective appraisal, and to have an important role in gaining more insights for the composition and valuation stage of perception and recognition.

Acknowledgment

The authors would like to thank Ericsson, Deutsche Telekom, Huawei, NetScout, NTT, Technicolor, Telchemy, and Yonsei University, Seoul, Korea, for building P.NBAMS databases and permitting publishing the relevant results. They would also like to thank Dr. N. Liao, Dr. X. Gu, Dr. D. Liu, and Dr. M. Narwaria for fruitful discussion, and K. Xie for technical support in video codec.

References

[1] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *J. Visual Commun. Image Represent.*, vol. 22, no. 4, pp. 297–312, 2011.
[2] S. Hemami and A. R. Reibman, "No-reference image and video quality estimation: Applications and human-motivated design," *Signal Process., Image Commun.*, vol. 25, no. 7, pp. 469–481, 2010.
[3] R. A. Berk, *Statistical Learning from a Regression Perspective*. New York: Springer-Verlag, 2008.
[4] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, pp. 011006-1–011006-21, 2010.
[5] P. J. Bex, "Sensitivity to spatial distortion in natural scenes," *J. Vis.*, vol. 10, no. 2, pp. 2301–2315, 2010.
[6] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.
[7] O. Schwartz and E. P. Simocelli, "Natural signal statistics and sensory gain control," *Nature Neurosci.*, vol. 4, no. 8, pp. 819–825, 2001.
[8] S. Kanumuri, S. G. Subramanian, P. C. Cosman, and A. R. Reibman, "Predicting H.264 packet loss visibility using a generalized linear model," in *Proc. Int. Conf. Image Process.*, 2006, pp. 2245–2248.
[9] S. Kanumuri, P. C. Cosman, A. R. Reibman, and V. A. Vaishampayan, "Modeling packet-loss visibility in MPEG-2 video," *IEEE Trans. Multimedia*, vol. 8, no. 2, pp. 341–355, Apr. 2006.
[10] *Opinion Model for Video-Telephony Applications*, ITU-R Standard G.1070, 2007.
[11] P. Gastaldo, S. Rovetta, and R. Zunino, "Objective quality assessment of MPEG-2 video streams by using CBP neural networks," *IEEE Trans. Neural Netw.*, vol. 13, no. 4, pp. 939–947, Jul. 2002.
[12] S. Argyropoulos, A. Raake, M. N. Garcia, and P. List, "No-reference video quality assessment for SD and HD H.264/AVC sequences based on continuous estimates of packet loss visibility," in *Proc. Int. Workshop Qual. Multimedia Exper.*, 2011, pp. 31–36.
[13] S.-C. Zhu, "Statistical modeling and conceptualization of visual patterns," *IEEE Trans. Pattern Anal. Mach. Iintell.*, vol. 25, no. 6, pp. 691–712, Jun. 2003.
[14] Y.-F. Ou, Z. Ma, T. Liu, and Y. Wang, "Perceptual quality assessment of video considering both frame rate and quantization artifacts," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 3, pp. 286–298, Mar. 2011.
[15] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.
[16] S. Li, L. Ma, and K. N. Ngan, "Full-reference video quality assessment by decoupling detail losses and additive impairments," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 7, pp. 1100–1112, Jul. 2012.
[17] J. S. Goodman and D. E. Pearson, "Multidimensional scaling of multiply-impaired television pictures," *IEEE Trans. Syst., Man Cybern.*, vol. 9, no. 6, pp. 353–356, Jun. 1979.
[18] J.-B. Martens, "Multidimensional modeling of image quality," *Proc. IEEE*, vol. 90, no. 1, pp. 133–153, Jan. 2002.
[19] B. Escalante-Ramirez, J.-B. Martens, and H. De Ridder, "Multidimensional characterization of the perceptual quality of noise-reduced computed tomography images," *J. Visual Commun. Image Represent.*, vol. 6, no. 4, pp. 317–334, 1995.

[20] L. T. Maloney and J. N. Yang, "Maximum likelihood difference scaling," *J. Vis.*, vol. 3, no. 8, pp. 573–585, 2003.
[21] M. Narwaria, W. Lin, and A. E. Cetin, "Scalable image quality assessment with 2D mel-cepstrum and machine learning approach," *Pattern Recognit.*, vol. 45, no. 1, pp. 299–313, 2012.
[22] P. Le Callet, C. Viard-Gaudin, and D. Barbara, "A convolutional neural network approach for objective video quality assessment," *IEEE Trans. Neural Netw.*, vol. 17, no. 5, pp. 1316–1327, Sep. 2006.
[23] H. H. Barrett, J. Yao, J. P. Polland, and K. J. Myers, "Model observers for assessment of image quality," *Proc. Nat. Acad. Sci.*, vol. 90, no. 21, pp. 9758–9765, 1993.
[24] P. Grother and E. Tabassi, "Performance of biometric quality measures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 531–543, Apr. 2007.
[25] S. Bennett, "Log-logistic regression models for survival data," *Appl. Stat.*, vol. 32, no. 2, pp. 165–171, 1983.
[26] P. McCullagh, *Generalized Linear Models*, 2nd ed. Boca Raton, FL: Chapman & Hall, 1989.
[27] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*. New York: Taylor & Francis, 1990.
[28] T. Micceri, "The unicorn, the normal curve, and other improbable creatures," *Psychol. Bullet.*, vol. 105, no. 1, pp. 156–166, 1989.
[29] S. Winkler, "On the properties of subjective rating in video quality experiments," in *Proc. Int. Workshop Qual. Multimedia Exper.*, San Diego, CA, Jul. 2009, pp. 1–6.
[30] A. Aron, E. N. Eron, and E. Coups, *Statistics for the Psychology*, 4th ed. Upper Saddle River, NJ: Pearson Education, 2006.
[31] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. New York: Wiley, 2000.
[32] *Requirement Specification for P.NAMS*, ITU-T Standard TD-467, 2011.
[33] H. De Ridder, "Minkowski metrics as a combination rule for digital image coding impairments," *Proc. SPIE 1666, Human Vis., Visual Process. Digit. Display III*, vol. 16, pp. 16–26, Aug. 1992.
[34] S. Paris and F. Durand, "A fast approximation of the bilateral filter using a signal processing approach," *Int. J. Comput. Vis.*, vol. 81, no. 1, pp. 24–52, 2009.
[35] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, ITU-R Standard BT.500-11, 2002.
[36] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.
[37] *Objective Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Full Reference*, ITU-T Standard J.144, 2004.
[38] *P.NAMS & P.NBAMS Model Performance Results*, ITU-T Standard TD 841 (GEN/12), May–Jun. 2012.

Fan Zhang (M'11) received the B.E. and Ph.D. degrees from the Huazhong University of Science and Technology, Wuhan, China, in 2002 and 2008, respectively, both in electronic and information engineering. He was a Visiting Student with Nanyang Technological University, Singapore, in 2008.

He was a Post-Doctoral Researcher with the Chinese University of Hong Kong, Hong Kong, from 2009 to 2010. He has been a Research Engineer with Technicolor Technology Co. Ltd., Beijing, China, since 2010. His current research interests include quality of experience and perceptual watermarking.

Weisi Lin (M'92–SM'98) received the B.Sc. and M.Sc. degrees from Zhongshan University, Guangzhou, China, and the Ph.D. degree from King's College, London University, London, U.K.

He was the Laboratory Head, Visual Processing, and the Acting Department Manager, Media Processing, Institute for Infocomm Research, Singapore. He is currently an Associate Professor with the School of Computer Engineering, Nanyang Technological University, Singapore. He has authored or co-authored more than 200 refereed papers in international journals and conferences. His current research interests include image processing, perceptual modeling, video compression, multimedia communication, and computer vision.

Dr. Lin is on the Editorial Board of the IEEE TRANSACTIONS ON MULTIMEDIA, the *IEEE Signal Processing Letters,* and the *Journal of Visual Communication and Image Representation.* He was the Lead Guest Editor of a special issue on perceptual signal processing, the *IEEE Journal of Selected Topics in Signal Processing* in 2012. He is the Chair of the IEEE MMTC Special Interest Group on Quality of Experience. He has been elected as a Distinguished Lecturer of APSIPA for 2012–2013. He is the Lead Technical Program Chair of Pacific-Rim Conference on Multimedia 2012, and the Technical Program Chair of the IEEE International Conference on Multimedia and Expo 2013. He is a Chartered Engineer in U.K., a fellow of the Institution of Engineering Technology, and an Honorary Fellow of the Singapore Institute of Engineering Technologists.

Zhibo Chen (M'01–SM'11) received the B.Sc., and Ph.D. degrees from the Department of Electronic Engineering, Tsinghua University, Beijing, China.

He has been with Technicolor Technology Co. Ltd., Beijing, since 2004, where he is currently a Principal Scientist with the Technicolor Research and Innovation Department, a Distinguished Fellow of the Technicolor Fellowship Program, the Manager of Media Processing Laboratory, and was the Media QoE Lead of Technicolor Research and Innovation. He was with Sony Research. He has authored or co-authored more than 60 papers in journals and conferences, and standard proposals. He holds more than 100 granted and filed EU and U.S. patent applications. His research on UMH Fast ME algorithm has been adopted by H.264 standard, widely used in standard reference software, and largely cited. His current research interests include media processing and coding, media quality of experience analysis and management for content delivery, and perceptual-based rendering.

He is a member of the IEEE Visual Signal Processing and Communications Committee and the IEEE Multimedia Communication Committee. He was an RC Member of ISCAS meetings from 2007 to 2013, a Key Member of the IEEE MMTC Special Interest Group on QoE. He was a TPC Member of the PCS and the VCIP, and the Chair of the ICME 2011 Multimedia track. He was a Co-Editor of the *IEEE Journal on Selected Areas in Communications QoE-Aware Wireless Multimedia Systems* in 2011, and a member on the Best Paper Selection Committee of the IEEE VCIP 2012.

King N. Ngan (F'00) received the Ph.D. degree in electrical engineering from the Loughborough University, Leicestershire, U.K.

He is currently a Chair Professor with the Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong. He was a Full Professor with the Nanyang Technological University, Singapore, and the University of Western Australia, Crawley, Australia. He holds honorary and visiting professorships with numerous universities in China, Australia, and South East Asia. He has authored or co-authored over 300 refereed technical papers in journals and conferences. He has authored three books, edited six volumes, and edited nine special issues in journals. He holds ten patents on image and video coding and communications.

Prof. Ngan was an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the *Journal on Visual Communications and Image Representation,* the *EURASIP Journal of Signal Processing: Image Communication,* and the *Journal of Applied Signal Processing.* He was the Chair of a number of prestigious international conferences on video signal processing and communications, and was on the advisory and technical committees of numerous professional organizations. He was the Co-Chair of the IEEE International Conference on Image Processing at Hong Kong in 2010. He is a Fellow of the IET (U.K.), and the IEAust (Australia), and was an IEEE Distinguished Lecturer from 2006 to 2007.