Review of ITU-T Parametric Models for Compressed Video Quality Estimation

Yankai Liu*, Li Song†, Xiaokang Yang†, Rong Xie†, Wenjun Zhang†
*Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China
† Future Medianet Innovation Center, Shanghai, China
E-mail: liuyankai, song_li, xkyang, rongxie, zhangwenjun@sjtu.edu.cn

Abstract— **This paper presents a review of parametric models for video quality estimation standardized in the past few years. The focus of this paper is the estimation of quality degradation caused by the compression artifacts. The models introduced in the paper contain the work of Telecommunication Standardization Section of the International Telecommunication Union (ITU-T) Study Group 9 and Study Group 12, standardized as ITU-T Rec. G.1070, ITU-T Rec. P.1201 series, ITU-T Rec. P.1202 series and ITU-T Rec. J.343.1. The core module estimating coding quality of each model is described with key algorithms and parametric formulas. The review of each model is presented and the strengths and weakness of each model are remarked. Finally, future work towards the development of an updated parametric QoE model for latest HEVC/H.265 is discussed.**

I. INTRODUCTION

With the proliferation of video services among networks and mobile devices, applications such as Internet Protocol television, video conferencing system, video-on-demand, home and commercial surveillance are widely used. The service providers of such applications are dedicated to deliver the video that satisfies or even impresses the end users. This promotes the development of video quality assessment (VQA) aiming at providing tools for video real-time monitoring as well as adjustment of transmission and encoding setups to provide an overall satisfied quality of experience (QoE) from the users' perspective.

In recent years, different modeling and standardized efforts have been devoted and are consistently on going in order to assess or predict the perceived video quality depending on the applications and specific scenarios. Thus various evaluation models are necessary.

Based on the presence of reference in the quality evaluation, the VQA models can be categorized into full-reference (FR), reduced-reference (RR), no-reference (NR) models. In the case of FR models, the original and the degraded video are directly compared, because of the access of the original video, FR is also called intrusive method. In the NR scenario, the evaluation of the degraded video is estimated without the original video, thus this category is called non-intrusive method. The RR scenario is somewhere between above mentioned two, i.e., only partial information about the original video is used in the evaluation process. NR VQA can be further divided into packet-layer model (PLM), bitstream-layer model (BLM) and hybrid model (HM). In PLM methods, the analysis is based on the IP (Internet Protocol) and RTP (Real Time Protocol) packet header information without the access to the decoded bitstream, which makes this model suitable for in-service quality management. BLM method estimates the subjective quality using partially or fully parsed bitstream information (i.e. payload information), which takes the content dependency of video quality into account that is not available in the PLM method. The HM method combines the previous two models and thus exploits the information both from the packet header and bitstream. For more detailed introduction to VQA, see [1].

NR VQA methods have a wide application scope since no additional information of original signal is required by the algorithm, thus the evaluation process can be carried out solely on the receiving end without causing intrusive effects to the encoding and delivering channel. Video coding is the preliminary step before transmission to limit the amount of data transmitted; NR VQA method is proved to be effective estimating the coding artifacts in this scenario. Other types of video artifacts including those caused by transmission channel or packet-loss are beyond the scope of this work, interesting readers can refer to [2], [3].

Parametric models predict the perceived video quality in the form of mathematical formulas based on the extracted set of parameters, relevant to the encoding process and video content or network conditions. Parametric models are effective and easily implemented since there is no need of full access to the original video source and the estimation is obtained as the direct result of the mathematical formulas.

Based on the work of Video Quality Expert Group (VQEG) and other contributors, ITU-T has standardized some NR VQA parametric models. Among them, the ITU-T Rec. G.1070 standardized in 2007 provides a PLM method based on the measurable parameters of the encoding process [4]. ITU-T Rec. P.1201 series, titled *Parametric non-intrusive assessment of audiovisual media streaming quality* (*P.NAMS*) and standardized in 2012 [5], provides a PLM method based on the payload information as it is available from packet headers and additional side-information. ITU-T Rec. P.1202 series, titled *Parametric non-intrusive bit-stream assessment of video media streaming quality* (*P.NABMS*) and standardized in 2012 [6], proposes a BLM method using parsed bit-stream but without fully parsed pixel information. ITU-T Rec. J.343.1 standardized in 2014 [7], provides a HM method analyzing packet header information and video image data captured at the video player.

Joskowicz et al. provide a review of 10 VQA NR parametric modes in 2012 [8], however only Rec. G.1070 model is included as a standardized effort. Yang et al. provide a general review of Rec. G.1070, P.NAMS and P.NBAMS model in 2012 [9], but the analysis was conducted at the draft stage of Rec. P.1201 and Rec. P.1202 without the detailed algorithms and parametric formulas. Yamagishi and Gao provide a review of Rec. P.1201.1 [10], Garcia et al. provide a review of Rec. P.1201.2 [11], and Chen et al. proposed a review of Rec. P.1202.2 [12]. Up to the time of this paper, there is no work analyzing Rec. J.343.1 and there lacks a comprehensive review and comparison of ITU-T consistent efforts concerning the parametric VQA of compression artifacts.

In this paper, we present a review of the above mentioned parametric NR models standardized by ITU-T recently. These models predict video quality in terms of mean opinion score (MOS) on a five-point absolute category rating (ACR) scale according to ITU-T P.910 [13]. The models' parameters and comparisons are presented and the strengths and weakness of each model are remarked and summarized towards the development of a parametric model for HEVC/H.265 codec quality estimation.

This paper is organized as follows: Section II describes the details of different parametric models. In Section III the performance of each model is presented. Section IV concludes the characteristics of each model and discusses activities towards an updated parametric model for the latest video coding standard - HEVC/H.265. Section V summarizes the results and the main contribution of this paper.

II. Parametric Models

In this section, different parametric models are presented. These models have been officially standardized by ITU-T. Each model is briefly described and the key parametric formulas and algorithms are detailed.

A. ITU-T Rec. G.1070

This recommendation describes a computational model for point-to-point interactive videophone applications over IP networks [4].The module estimates the video quality ***QV*** with compression artifacts as:

$$QV = 1 + I_{coding} \tag{1}$$

Where ***Icoding*** represents the basic video quality affected by the coding distortion under a combination of video bitrate ***brv*** (kbit/s) and video framerate ***frv*** (fps) according to (2).

$$I_{coding} = I_{ofr}\ \exp\left\{ -\frac{\left(\ln(fr_v) - \ln(O_{fr})\right)^2}{2D_{fr_v}{}^2} \right\} \tag{2}$$

The ***Ofr*** is an optimal frame rate that maximizes the video quality at each video bitrate, ***Iofr*** represents the maximum video quality at each video bitrate and ***Dfrv*** represents the degree of video quality robustness due to frame rate.

$$I_{ofr} = c_1 - \frac{c_2}{1+\left(\frac{br_v}{c_3}\right)^{c4}}, \quad 0 \le I_{ofr} \le 4 \tag{3}$$

$$O_{fr} = c_5 + c_6 \cdot br_v, \quad 1 \le O_{fr} \le 30 \tag{4}$$

$$D_{fr_v} = c_7 + c_8 \cdot br_v \tag{5}$$

In this model, the video content is not taken into account. As stated in the recommendation, the model handles video size between VGA (Video Graphic Array, 640x480 pixels) and QQVGA (Quarter Quarter VQA, 160x120 pixels).

B. ITU-T Rec. P.1201.1

Recommendation ITU-T P.1201 [5] provides the framework for actually two algorithmic models described separately in ITU-T Rec. P.1201.1 [14] and ITU-T Rec. P.1201.2 [15]. These algorithms are aimed at monitoring the audio, video and audiovisual quality of IP-based video services with packet-header information, the former standard focuses on lower resolution application areas such as mobile TV and the latter focuses on the higher resolution application areas such as IPTV. The two standards provide different VQA methods which are introduced in detail.

The video quality estimation model in Rec. P.1201.1 estimates video quality ***QV*** as follows: If video frame rate is larger than 24 fps, then compute ***QV*** with (6), else with (7).

$$QV = (5 - Qcod) \tag{6}$$

$$QV = (5 - Qcod) \cdot \left(1 + c_1 \cdot cpx_{video} - c_2 \cdot cpx_{video} \cdot \log\left(\frac{1000}{fr_v}\right)\right) \tag{7}$$

Here ***Qcod*** denotes the video distortion due to compression and is calculated in equation (8) to (10), and ***cpxvideo*** denotes the video content complexity factor.

$$Qcod = \frac{4}{1+\left(\frac{normbr_v}{c_3 \cdot cpx_{video} + c_4}\right)^{(c_3 \cdot cpx_{video} + c_4)}} \tag{8}$$

$$normbr_v = \frac{br_v \cdot 8 \cdot 30}{1000 \cdot \min(30, fr_v)} \tag{9}$$

$$cpx_{video} = \min\left(\sqrt{\frac{br_v}{AvgByte_{I\text{-}frame}}}, 1.0\right) \tag{10}$$

Video content complexity factor $\boldsymbol{cpx_{video}}$ is the factor describing the content spatiotemporal complexity. The maximum value is 1.0, the initial value is 0.5, and the calculation is shown in (10). $\boldsymbol{AvgByte_{I\text{-}frame}}$ is the average number of bytes per I-frame. Since the model analyzes only packet header information, the computational power of the model is very light, and the model can be applied to encrypted packets. Interested readers can refer to [10] for further information.

C. ITU-T Rec. P.1201.2

The video quality model in Rec. P.1201.2 [15] is decomposed as follows:

$$QV = 100 - Qcod - Qtra \qquad (11)$$

Where $\boldsymbol{QV}$ is the predicted video-quality. $\boldsymbol{Qcod}$ is the quality distortion due to video compression, and $\boldsymbol{Qtra}$ is the quality distortion due to video packet loss, the latter is set to 0 since we are only interested in the quality evaluation with compression artifacts.

$$Qcod = c_1 \cdot \exp(c_2 \cdot bitPerPixel) + c_3 \cdot cpx_{video} + c_4 \qquad (12)$$

$$bitPerPixel = \frac{br_v \cdot 10^6}{numPixels_{frame} \cdot fr_v} \qquad (13)$$

The quality distortion of video compression is given by equation (12) where $\boldsymbol{bitPerPixel}$ denotes the averaged bits of one pixel in the video sequence. $\boldsymbol{cpx_{video}}$ estimates both the temporal and spatial complexity of the error-free encoded content:

$$cpx_{video} = \frac{\sum_{sc=1}^{Z} w_{sc} \cdot N_{sc}}{\sum_{sc=1}^{Z} S_{sc}^{I} \cdot w_{sc} \cdot N_{sc}} \cdot \frac{numPixels_{frame} \cdot fr_v}{1000} \qquad (14)$$

$\boldsymbol{S_{sc}^{I}}$ is the averaged I-frame size for scene $\boldsymbol{sc}$ (the first I-frame of the first scene is ignored). $\boldsymbol{Z}$ is the number of scenes in the video sequence, and $\boldsymbol{N_{sc}}$ is the number of Groups Of Pictures (GOPs) in scene $\boldsymbol{sc}$. For the scene having the lowest $\boldsymbol{S_{sc}^{I}}$ value, $\boldsymbol{w_{sc}} = 16$, otherwise $\boldsymbol{w_{sc}} = 1$. Since the value $\boldsymbol{S_{sc}^{I}}$ differs more between contents in the case of low-to-medium bitrates, the $\boldsymbol{cpx_{video}}$ parameter distinguishes the influence of the content complexity for low-to-medium bitrates only. Moreover, the use of a higher $\boldsymbol{w_{sc}}$ value for the lowest $\boldsymbol{S_{sc}^{I}}$ value highlights the impact of poor quality frames in the sequences, for the lowest $\boldsymbol{S_{sc}^{I}}$ yields the lowest quality and therefore has the highest weight on the overall quality [11].

The highlight of Rec. P.1201.2 is that it introduces the technique of scene detection to improve the overall model performance [16]. The proposed method for scene change detection in a video sequence is based on the assumption that modern encoders can detect abrupt scene changes and reset the GOP structure upon such detection, i.e. the scene cut picture is always encoded as an I-frame. Therefore, after the estimation of frame size and frame type of video sequence, the algorithm compares the size of each I-frame with that of previous I-frames. The proposed method is presented in detail below.

Let $\boldsymbol{F_k^T}$ denotes the size of the kth picture of type $\boldsymbol{T}$, where $\boldsymbol{T}$ denotes I, P or B-frame. The scaling term $\boldsymbol{I_s^k}$ of the kth I-frame under consideration is computed by:

$$I_k^s = \frac{\text{median}(F_{l-n_P}^P, \ldots, F_l^P)}{\text{mean}(F_{l-L_k+1}^P, \ldots, F_l^P)}, \quad k > 2 \qquad (15)$$

Where $\boldsymbol{l}$ is the index of the P-frame prior to the I-frame under examination, $\boldsymbol{L_k}$ is the number of P-frames in the previous GOP, and $\boldsymbol{n_P}$ is the number of P-frames to be considered for the computation of the median. Subsequently, the following ratios of frame sizes per type are computed, where $\boldsymbol{m}$ is the index of the B-frame prior to the I-frame under examination, and $\boldsymbol{n_B}$ is the number of B-frames to be considered for the mean.

$$r_k^I = \frac{F_k^I}{F_{k-1}^I \cdot I_k^s} \qquad (16)$$

$$r_k^P = \frac{\text{mean}(F_{l-n_P}^P, \ldots, F_l^P)}{\text{mean}(F_{l+1}^P, \ldots, F_{l+1+n_P}^P)} \qquad (17)$$

$$r_k^B = \frac{\text{mean}(F_{m-n_B}^B, \ldots, F_m^B)}{\text{mean}(F_{m+1}^B, \ldots, F_{m+1+n_B}^B)} \qquad (18)$$

Algorithm 1 Algorithm for scene change detection in P.1201.2

```
set $k$ to the third I-frame index
while ($k > 2$ && not last I-frame) do
    compute $I_k^s$ based on (15)
    compute $r_k^I$ , $r_k^P$, and $r_k^B$ based on (16) to (18)
    if ($r_k^I > I_1$ || $r_k^I < I_2$) then
        if ($P_1 < r_k^P < P_2$) &&($b_1 < r_k^B < b_2$) then
            continue;
        else
            denote I-frame as scene change;
        end if
    else if ($r_k^I > I_3$ || $r_k^I < I_4$) then
        if ($P_3 < r_k^P < P_4$) &&( $b_3 < r_k^B < b_4$) then
            continue;
        else
            denote I-frame as scene change;
        end if
    end if
    k++; %(move to the next GOP)
end while
```

If the ratio of the I-frame sizes $\boldsymbol{r_k^I}$ is within a specific range (cf. parameters I_1 and I_2 in the algorithm), then the criteria for the similarity between P/B-frames are more relaxed. Otherwise, a stricter set of thresholds is used [16].

D. ITU-T Rec. P.1202.1

The calculation of the compression quality score $\boldsymbol{QV}$ in Rec. P.1202.1 [17] is made by combining a set of compression quality related parameters. Thus the objective quality measure is based on extracted video coding dependent parameters as denoted in (20).

$$QV = \min\left(\max\left(Qcod, 1\right), 5\right) \tag{19}$$

$$Qcod = kfr_{impact} \cdot \exp\left(c_1 \cdot \left(QP\_fr_{impact} + cpx_{encoder} + motion_{impact}\right)\right) \tag{20}$$

The quantization parameter (QP) is a good first indication of the compression quality as it determines the amount of quantization to be applied on the transform coefficients of the video bitstream. The frame level quantization parameter $\boldsymbol{QP_{pic}}$ is computed by averaging the macroblock level QP. $\boldsymbol{QP_{pic}}$ together with ***intraflicker*** and frame rate ***fr*** forms the comprehensive parameter $\boldsymbol{QP\_frame_{impact}}$ as in (22). The intra-picture flicker algorithm determines a measure of QP for each picture in the received video bitstream and identifies a quality defect when an abrupt change in the quantization parameter occurs. Estimated key frame rate ***kfr*** is calculated as the frame rate of the sequence divided by the number of pictures between two intra pictures; its impact on compression quality is formulized in (21).

Additional parameter used is coding complexity indicator $\boldsymbol{cpx_{endoder}}$ and sequence motion indicator $\boldsymbol{motion_{impact}}$. The former is leveled indicator of the number of used reference pictures and the existence of various macroblock partition sizes in H.264 codec, more precisely the existence of inter 8x4 partitions, inter 4x8 partitions and inter 4x4 partitions. The latter is computed with the $\boldsymbol{avgMV_{pic}}$ which is the product of absolute vertical and horizontal motion vectors per macroblock averaged over all frames in the sequence as stated in (23).

$$kfr_{impact} = c_2 \cdot kfr + c_3 \tag{21}$$

$$\begin{aligned} Qp\_fr_{impact} = {} & c_4 + c_5 \cdot \left(\frac{avgQP_{pic}}{51}\right)^{c_6} + c_7 \cdot \frac{1}{fr} + c_8 \cdot intraflicker \\ & + c_9 \cdot (maxQP_{pic} - minQP_{pic}) \end{aligned} \tag{22}$$

$$motion_{impact} = c_{10} \cdot avgMV_{pic} \cdot \left(1 - \frac{fr}{30}\right) \tag{23}$$

$$cpx_{encoder} = c_{11} \cdot encoderCmplesityLevel \tag{24}$$

E. ITU-T Rec. P.1202.2 Model 1

Rec. P.1202.1 consists of one model and Rec. P.1202.2 consists of two models: model 1 and model 2, which both are no-reference models [18]. Mode 1 refers to a parsing mode; the model operates by analyzing information in the video bitstream without fully decoding the bitstream (i.e., no pixel information is used) for MOS estimation. Mode 2 refers to a full decoding mode, in addition to the bitstream information which mode 1 uses, it can also decode parts or all of the video bitstream (i.e., pixel information is used) for MOS estimation.

The formula to calculate compression quality is presents as (25), it combines the influence of video level $\boldsymbol{QP_{video}}$ and video content complexity. $\boldsymbol{QP_{video}}$ is the average value of all slice level QP of the sequence.

$$QV = c_1 + \frac{c_2}{c_3 + \left(\dfrac{QP_{video}}{c_4 - c_5 \cdot cpx_{video}}\right)^{c_6}} \tag{25}$$

The slice content complexity $\boldsymbol{sliceCpx_k}$ is calculated according to its correctly decoded quantization parameter $\boldsymbol{sliceQP_k}$ and bytes per pixel $\boldsymbol{sliceBytepp_k}$ as below:

$$sliceCpx_k = c_7 \cdot sliceQP_k \cdot sliceBytepp_k + c_8 \cdot sliceQP_k \tag{26}$$

$$cpx_{frame} = \frac{\sum_{k=1}^{numSlice} sliceCpx_k}{numSlice} \tag{27}$$

$$cpx_{video} = \frac{\sum_{j=1}^{\substack{numErrorFree \\ IntraFrame}} cpx_{frame}}{numErrorFreeIntraFrame} \tag{28}$$

The slice level content complexity is averaged first over all the slices and then over all the error free intra frames to aggregate video level content complexity.

F. ITU-T Rec. P.1202.2 Model 2

In Rec. P.1202.2 model 2 [18], the estimated levels of distortion for different artifact types are aligned to the same scales as the subjective MOS in (29) and the impact of compression artifacts on video quality is present as (30). Video level features are effective in estimating the uniform impairments caused by lossy compression. The contributors use content unpredictability (CU) to quantify the content complexity. A MB's CU is defined as the variance of the prediction residuals in the luminance channel. Clip-wise CU, $\boldsymbol{CU_{video}}$, is the weighted average of MB-wise CUs over all correctly parsed MBs of the video clip in H.264 codec as shown in (31) [12]. Here, I, P, B represents I-picture, P-picture, B-picture, respectively, r denotes the *r*th MB in certain picture. Clip-wise parameter $\boldsymbol{QP_{video}}$ is acquired as the same manner of $\boldsymbol{CU_{video}}$ in (32).

$$QV = c_1 \cdot Qcod + c_2 \tag{29}$$

$$Qcod = \frac{1}{1 + c_3 \cdot \left(\log(CU_T + 1)\right)^{c_4} (51 - QP_{video})^{c_5}} \tag{30}$$

$$CU_{video} = \frac{1}{T}\left(\sum_{t\in\{I\}} \sum_{r\in t} \frac{CU_{MB}}{20.6 \times |r|_t} \right) + \sum_{t\in\{P\}} \sum_{r\in t} \frac{CU_{MB}}{3.52 \times |r|_t} + \sum_{t\in\{B\}} \sum_{r\in t} \frac{CU_{MB}}{|r|_t} \tag{31}$$

$$QP_{video} = \frac{1}{T} \sum_{t} \sum_{r\in t} \frac{QP_{MB}}{|r|_t} \tag{32}$$

Generally, Huma Visual System (HVS) are more likely to tolerate visual distortions in complex scenes which is commonly known as the texture masking effect. Thus the proposed method could evaluate the compression quality with such consideration.

G. ITU-T Rec. J.343.1

The model in this recommendation measures the visual effect of spatial and temporal degradations as the result of video coding. The estimation of coding quality is based mainly on complexity and motion statistics – derived from the video frames – and on the total frame-size per scene – derived partly from the packet headers of the bitstream [7]. Equation (33) presents the estimate of visual quality on the 1 to 5 MOS value. Quality distortion of compression is first estimated over scenes of the video as ***Qcod_s*** and then average over the sequence to obtain ***Qcod$_{video}$***.

$$QV = 4 \cdot Qcod_{video} + 1 \tag{33}$$

$$Qcod\_s = \frac{bit\_s \cdot duration\_s}{cpx\_s + numFrame\_s \cdot fracMotion\_s \cdot v\_s} \tag{34}$$

bit_s denotes the total bits of certain scene and the ***duration_s*** is the lasting time in seconds of current scene. The complexity statistic of frames in current scene ***cmp_s*** is computed as local inter-frame and intra-frame dissimilarity of frame i and i-1 using Algorithm 1 as shown below.

$$v\_s = \left(\left(1 - c_1\right) + c_2 \cdot \left(1 - p\_s\right) \cdot cpx\_s\right) \tag{35}$$

$$p\_s = c_3 \cdot fracMotion\_s + c_4 \cdot confMotion\_s \tag{36}$$

The parameter ***fracMotion_s*** represents the moving fraction of local regions in one frame and ***confMotion_s*** is the confidence in motion estimation. The main idea to compute the complexity and motion features is to observe 3x3 blocks of the video frame, calculate the similarity of spatially and temporally adjacent blocks as well as the predictability of a block, by blocks of the previous frame at the same spatially adjacent locations. Algorithm 1 and 2 provide details of the motion estimation and complexity computing as shown blow.

Algorithm 1 Algorithm for computing local intra-frame complexity

```
set $Y_0$,$Y_1$ to consecutive frames, choose 40x40 equally spaced points in the frame (ignoring a border of 4 to avoid border problems) and compute at each position (i, j):
$\mathrm{S[i, j]} = \exp(\frac{-\mathrm{RMSE}\left(\boldsymbol{a}(i,j), \boldsymbol{b}(i,j)\right)}{c_5})$ ,
where a denotes the 3x3 block around (i, j) in $Y_0$
where b denotes the 3x3 block around (i, j) in $Y_1$
if (S[i, j] >=1)
    count++
end if
compute $prob\_eq = \frac{count}{\mathrm{S}.height \cdot \mathrm{S}.width}$
compute $cmp\_s = 1 - \mathrm{mean(S)}$
```

Algorithm 2 Algorithm for motion estimation per frame

```
set $Y_0$,$Y_1$ to consecutive frames
set dxdy the array containing displacement vectors (dx, dy) which indicate four quadrants in counter clockwise direction
denote $\langle \mathrm{a}, \mathrm{b} \rangle$ as the inner product of a and b
for ( i=2; i<=$Y_0$.width - 2; i++)
  for ( j=2; j<=$Y_0$.height - 2; j++)
      denote a as the 3x3 block around (i, j) in $Y_0$
      denote b as the 3x3 block around (i, j) in $Y_1$
      compute d_ab =b - a
      compute msqd as mean of d_ab* d_ab
      if ( msqd > 0.1)
      set mark(i, j) =1, indicate (i, j) as motion computed
          for (m=0; m < length of dxdy; m++)
              denote a_s as the shifted vision of a among($dx_m$, $dy_m$)
              denote b_s as the shifted vision of b among($dx_{m+1}$, $dy_{m+1}$)
              compute d_s= a_s - a , d_t= b_s - a
              r = $\left|\langle \mathbf{d\_ab}, \mathbf{d\_s} \rangle\right|$ |
              compute s = $\left|\langle \mathbf{d\_ab}, \mathbf{d\_t} \rangle\right|$
              compute err = r*d_s + s*d_t - d_ab
              $p\_dist_{\mathrm{m}}(i,j) = \exp\left(\frac{-\mathrm{mean}\left(\mathbf{err} * \mathbf{err}\right)}{10}\right)$
          end for
      else set mark(i, j) =0
      end if
```

```
            $fracMotion = \frac{\sum_i \sum_j mark(i,j)}{Y_0.width \cdot Y_1.height}$
            while ( $mark(i,j) > 0.5$ ) do
                compute $p_m(i,j)$ based on (38), find the $\max(p_m(i,j))$ over all quadrants
            end while
            $confMotion$ is computed as the mean of all $\max(p_m(i,j))$
        end for
    end for
```

The estimated motion vector ***Dx, Dy*** is given by (37). Furthermore, the confidence value ***p*** is given by (38).

$$Dx_{i,j}, Dy_{i,j} = \sum_m \left(p\_dist_m(i,j) \cdot \left(r_m \cdot dxdy_m + s_m \cdot dxdy_{m+1} \right) \right) \quad (37)$$

$$p_m(i,j) = \left\langle p\_dist_m(i,j), log\left(p\_dist_m(i,j) \right) \right\rangle \quad (38)$$

Scene detection technique has been utilized in Rec. J.343.1 to better capture the content complexity as well as the motion feature of the sequence. The method decomposing the video clip into scenes is illustrated in Algorithm 3. The scene detection is conducted on the deepest 2x2 down-sampled version of the current frame with the down-sample depth controlled by a pre-defined parameter. Then k x k tiles are extracted out of the frame sequence to compute histograms over each tile. The detail of mapping the frame sequence to the timeline can be found in the standard [7].

Algorithm 3 Algorithm for scene change detection

```
for (i = 1; i < total number of frames; i++)
    for (j = 1; j < min(i,window); j++)
        cut k x k tiles out of the frame i and frame i - j
        compute matrix $H_0$ for frame i, $H_1$ for frame i - j containing histograms of all the tiles in both frames
        for (m = 1; m < k; m++)
            for (n = 1; n < k; n++)
                tile_dissimilarity= $\sqrt{\text{mean}\left(\left(H_0(m,n) - H_1(m,n)\right)^2\right)}$
            end for
        end for
        scene_statistic(i) = median(tile_dissimilarity)
    end for
end for
find the nearest frame q prior to current frame p within the pre-defined minimum scene time duration.
compute s = median(scene_statistic(q : p))
for (i = q ; i < p; i++)
    if (scene_statistic(i) > s · $T_1$)
            && (scene_statistic(i) > $T_2$)
        scene start detected
    end if
end for
```

III. Comparison And Discussion

The models introduced in section II have all been formally standardized and the models' performance have been validated separately as shown in Table I with the PC (Pearson Correlation), the RMSE (Root Mean Square Error) indicators.

Table I
Models performance of coding distortion

	PC	RMSE	samples
G.1070	0.955	—	—
P.1201.1	0.830	0.535	1430
P.1201.2	0.902	0.461	6138
P.1202.1	0.918	0.391(0.284)	982
P.1202.2 Model 1	0.938	0.357(0.325)	3069
P.1202.2 Model 2	0.940	0.353(0.337)	3069
J.343.1	0.795	0.595	—

Rec. G.1070 is verified on 4 databases for MPEG-4 and H.264 codec. Rec. J343.1 is verified on 10 databases of various resolutions with H.264 codec. The validation set of J.343.1 contains the most affluent category of artifacts among all mentioned standards including compression and transmission ones, thus the overall PC value is relatively low compared to others because of the difficulty in measuring such a rich set of artifacts. Only Rec. P.1202 series has provided the statistics considering only compression conditions (marked with parentheses in Table I) besides the overall model performance. Note that the evaluation of different models is not directly comparable because of non-uniform testing databases. Table II provides a comprehensive summary of the models.

As shown from model details in section II and Table II, for PLM methods, the complexity is relatively low and the input parameters are easily available from the packet layer in the network even with encrypted payloads, but at the cost of lacking accuracy to specific scenarios. The BLM methods are generally moderate in complexity and their accuracy significantly depends on the level of access to the bitstream. Moreover the estimation of content complexity and the decomposition of sequence into scenes have positive impacts on the overall model performance. The demerit mainly is the unsuitability to services with encrypted payloads. HM method is a combination of the previous two methods, thus has a better capability tackling deferent service scenarios. The computing complexity of HM method is comparatively high.

IV. Future Work

Based on the previous analysis, bitrate, frame rate, content complexity estimation and motion estimation are all relevant to provide a satisfied coding quality evaluation for a given video sequence. The scene detection technique has been proved to improve the overall model performance and strengthen the model robustness to various service scenarios.

However, the validation of standardized model concerning coding distortion is designed and tested with no more advanced codec than H.264/AVC. ITU-T Study Group 9 has initialized drafting the recommendation *Objective perceptual video quality measurement methods for H.265*, while ITU-T Study Group 12 has begun the program G.OM_HEVC, namely *Opinion model for network planning of HEVC media streaming quality*, both are still under prudent study. Thus it is a necessity to first adjust and evaluate all the previously introduced models with uniformed HEVC/H.265 database containing video clips transmitted over real-application networks of various resolutions and contents, and then provide a comprehensive HEVC/H.265 parametric model to predict coding QoE from the end users' perspective.

According to the results, a more general model evaluating HEVC/H.265 coding distortions may derive from Rec. P.1202.2 and Rec. J343.1 especially for HD and upper resolutions. The proliferation of UHD video content and virtual reality applications generates an acute demand for objective quality assessment of networked videos for system design, QoE planning, and quality benchmarking and monitoring.

V. CONCLUSIONS

We present in this work a review and analysis of up-to-now ITU-T no reference video quality evaluation models concerning coding distortions. Detailed analysis of core algorithms and calculation steps of different models are illustrated. From the results obtained, trading-off between the model performance and computational complexity is crucial to the applicability of the proposed models. The combination of packet level method and bitstream method, namely model 2 in Rec. P.1202.2 and J.343.1, represents the direction of quality evaluation of more advanced video codecs and more challenging service scenarios.

TABLE II
MODELS COMPARISON

	Equations	Bit Rate	Frame Rate	QP	Content Complexity	Motion Estimation	Scene Detection	Encrypted Bitstream	Level	Tested Conditions
ITU-T G.1070	1-5	Yes	Yes	No	No	No	No	Yes	PLM	VGAQ,VGA,QQVGA; MPEG4 ,H.264
ITU-T P.1201.1	6-10	Yes	Yes	No	Yes	No	No	Yes	PLM	HVGA, QVGA, QCIF; MPEG4, H.264
ITU-T P.1201.2	11-18	Yes	Yes	No	Yes	No	Yes	Yes	PLM	SD, HD; H.264
ITU-T P.1202.1	19-24	Yes	Yes	Yes	Yes	Yes	No	No	BLM	QICF, QVGA, HVGA; H.264
ITU-T P.1201.2 Model1	25-28	Yes	Yes	Yes	Yes	No	No	No	BLM	SD, HD; H.264
ITU-T P.1201.2 Model2	29-32	Yes	Yes	Yes	Yes	No	No	No	BLM	SD, HD; H.264
ITU-T J.343.1	33-38	Yes	Yes	No	Yes	Yes	Yes	Yes	HM	VGA, WVGA, HD; H.264

REFERENCES

[1] J. G. Apostolopoulos and A. R. Reibman, "The challenge of estimating video quality in video communication applications [In the Spotlight], " IEEE Signal Process. Mag., vol. 29, no. 2, pp. 158–160, Mar. 2012.

[2] A. R. Reibman, V. A. Vaishampayan and Y. Sermadevi, "Quality monitoring of video over a packet network," in *IEEE Transactions on Multimedia*, vol. 6, no. 2, pp. 327-334, April 2004.

[3] G. Valenzise , S. Magni , M. Tagliasacchi and S. Tubaro, "Estimating channel-induced distortion in H.264/AVC video without bitstream information", *Proc. 2nd Int. Workshop Quality Multimedia Experience*, pp. 100-105, 2010.

[4] ITU-T Recommendation G.1070: *Opinion model for video-telephony applications*, 2012.

[5] ITU-T Recommendation P.1201: *Parametric non-intrusive assessment of audiovisual media streaming quality, 2012.*

[6] ITU-T Recommendation P.1202: *Parametric non-intrusive bitstream assessment of video media streaming quality*, 2012 *assessment of video media streaming quality*, 2012.

[7] ITU-T Recommendation J.343.1: *Hybrid-NRe objective perceptual video quality measurement for HDTV and multimedia IP-based video services in the presence of encrypted bitstream data*, 2014.

[8] J. Joskowicz, R. Sotelo and J. C. Lopez Arado, "Comparison of parametric models for video quality estimation: Towards a general model," *IEEE international Symposium on Broadband Multimedia Systems and Broadcasting*, Seoul, 2012, pp. 1-7.

[9] F. Yang and S. Wan, "Bitstream-based quality assessment for networked video: a review," in IEEE Communications Magazine, vol. 50, no. 11, pp. 203-209, November 2012.

[10] K. Yamagishi and S. Gao, "Light-weight audiovisual quality assessment of mobile video: ITU-T Rec. P.1201.1,"Multimedia

Signal Processing (MMSP), 2013 IEEE 15th International Workshop on, Pula, 2013, pp. 464-469.

[11] M. N. Garcia et al., "Parametric model for audiovisual quality assessment in IPTV: ITU-T Rec. P.1201.2," Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on, Pula, 2013, pp. 482-487.

[12] Z. Chen; N. Liao; X. Gu; F. Wu; G. Shi, "Hybrid Distortion Ranking Tuned Bitstream-layer Video Quality Assessment," in *IEEE Transactions on Circuits and Systems for Video Technology* , vol.PP, no.99, pp.1-1

[13] ITU-T Recommendation P.910: *Subjective video quality assessment methods for multimedia applications*, 2008.

[14] ITU-T Recommendation P.1201.1: *Parametric non-intrusive assessment of audiovisual media streaming quality – lower resolution application area*, 2012.

[15] ITU-T Recommendation P.1201.2: *Parametric non-intrusive assessment of audiovisual media streaming quality – higher resolution application area*, 2012.

[16] S. Argyropoulos, P. List, M. N. Garcia, B. Feiten, M. Pettersson and A. Raake, "Scene change detection in encrypted video bit streams," *2013 IEEE International Conference on Image Processing*, Melbourne, VIC, 2013, pp. 2529-2533.

[17] ITU-T Recommendation P.1202.1: *Parametric non-intrusive bitstream assessment of video media streaming quality – Lower resolution application area*, 2012.

[18] ITU-T Recommendation P.1202.2: *Parametric non-intrusive bitstream assessment of video media streaming quality – Higher resolution application area*, 2012.