From QoS to QoE: A Tutorial on Video Quality Assessment

Yanjiao Chen, *Student Member, IEEE*, Kaishun Wu, *Member, IEEE*, and Qian Zhang, *Fellow, IEEE*

***Abstract*—Quality of experience (QoE) is the perceptual quality of service (QoS) from the users' perspective. For video service, the relationship between QoE and QoS (such as coding parameters and network statistics) is complicated because users' perceptual video quality is subjective and diversified in different environments. Traditionally, QoE is obtained from subjective test, where human viewers evaluate the quality of tested videos under a laboratory environment. To avoid high cost and offline nature of such tests, objective quality models are developed to predict QoE based on objective QoS parameters, but it is still an indirect way to estimate QoE. With the rising popularity of video streaming over the Internet, data-driven QoE analysis models have newly emerged due to availability of large-scale data. In this paper, we give a comprehensive survey of the evolution of video quality assessment methods, analyzing their characteristics, advantages, and drawbacks. We also introduce QoE-based video applications and, finally, identify the future research directions of QoE.**

***Index Terms*—Quality of experience, subjective test, objective quality model, data-driven analysis.**

I. Introduction

WITH the exponential growth of the video-based services, it becomes ever more important for the video service providers to cater to the quality expectation of the end users. It is estimated that the sum of all forms of videos (TV, video-on-Demand (VoD), Internet, and P2P) will be around 80%∼90% of global consumer traffic by 2017 [1]. Video streaming over the Internet, especially through mobile network, is becoming more and more popular. Throughout the world, Internet video traffic will be 69% of all consumer Internet traffic by 2017 [1], and mobile video traffic will be over one third of mobile data traffic by the end of 2018 [2].

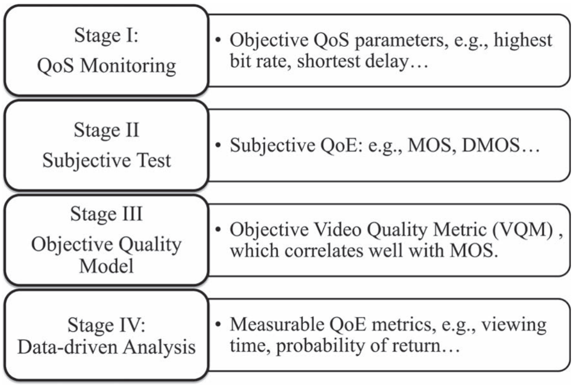

Stage I:
QoS Monitoring
• Objective QoS parameters, e.g., highest bit rate, shortest delay…
Stage II
Subjective Test
• Subjective QoE: e.g., MOS, DMOS…
Stage III
Objective Quality Model
• Objective Video Quality Metric (VQM) , which correlates well with MOS.
Stage IV:
Data-driven Analysis
• Measurable QoE metrics, e.g., viewing time, probability of return…

Fig. 1. Video quality assessment evolution.

In early works, researchers were trying to increase user perceptual video quality by appropriately selecting QoS parameters (such as video compression optimization [3]–[5] and network bandwidth allocation [6]–[8]). In [5], the authors study the relationship between the peak signal-to-noise ratio and quantization parameter, and propose a linear rate-quantization model to optimize quantization parameter calculation. In [8], the authors present a dynamic network resource allocation scheme for high-quality variable bitrate video transmission, based on the prediction of future traffic patterns. While monitoring and controlling QoS parameters of the video transmission system is important for achieving high video quality, it is more crucial to evaluate video quality from the users' perspective, which is known as Quality of Experience (QoE), or user-lever QoS. QoE-based video quality assessment is difficult because user experience is subjective, hard to quantify and measure. Moreover, the advent of new video compression standards, the development of video transmission systems, and the advancement of consumer video technologies, all call for a new and better understanding of user QoE. Video quality assessment has gone through four stages, as shown in Fig. 1. Table I gives a comparison of these video quality assessment methods.

QoS monitoring for the video traffic includes two parts: QoS provisioning from the network and QoS provisioning from the video application. QoS support from the network, especially wireless or mobile network, is essential for video delivery over the Internet. Three major approaches are congestion control, error control and power control. The challenges facing network QoS support include unreliable channels, bandwidth constraints, heterogeneous access technologies. QoS support from the video application includes advanced video encoding scheme, error concealment and adaptive video streaming protocol. A survey of video QoS provisioning in mobile network

Manuscript received February 25, 2014; revised June 5, 2014 and August 5, 2014; accepted September 21, 2014. Date of publication October 22, 2014; date of current version May 19, 2015. This work was supported in part by the 973 Project under Grant 2013CB329006; by the China NSFC under Grant 61173156; by RGC under Contracts CERG 622613, 16212714, HKUST6/CRF/12R, and M-HKUST609/13; by the Huawei-HKUST Joint Laboratory, Program for New Century Excellent Talents in University under Grant NCET-13-0908; by the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant S20120011468; by the Shenzhen Science and Technology Foundation under Grant JCYJ20140509172719309; and by the New Star of Pearl River on Science and Technology of Guangzhou under Grant 2012J2200081.

Y. Chen and Q. Zhang are with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Kowloon, Hong Kong.

K. Wu is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China, and also with Fok Ying Tung Research Institute, The Hong Kong University of Science and Technology, Kowloon, Hong Kong.

Digital Object Identifier 10.1109/COMST.2014.2363139

TABLE I
COMPARISON OF VIDEO QUALITY ASSESSMENT METHODS

	Direct measure of QoE	Objective or subjective	Real-time	Wide application	Cost
QoS monitoring	No	Objective	Yes	Wide	Not sure
Subjective test	Yes	Subjective	No	Limited	High
Objective quality model	No	Objective	Yes/No	Limited	Low
Data-driven analysis	Yes	Objective	Yes	Wide	Not sure

is given in [9], mostly from the network point of view. Error-concealment schemes are investigated in [10]. [11] and [12] consider both network and application QoS support. In this tutorial, we mainly focus on Stages II $\sim$ IV of video quality assessment. In the main text, we will not discuss Stage I, and interested readers can refer to the above surveys for more information.

Subjective test directly measures user QoE by soliciting users' evaluation scores under the laboratory environment. Users are given a series of tested video sequences, original ones and processed ones included, and then required to give scores on the video quality. Detailed plans for conducting subjective tests have been made by the Video Quality Expert Group (VQEG) [13]. Though being viewed as a relative accurate way of measuring user QoE, subjective test suffers from three major drawbacks. First, subjective test has high cost in terms of time, money, and manual effort. Second, subjective test is conducted in the laboratory environment, with limited test video types, test conditions, and viewer demography. Therefore, the results may not be applicable to video quality assessment in the wild. Third, the subjective test cannot be used for real-time QoE evaluation.

In order to avoid high cost of subjective test, objective quality models are developed. The major purpose is to identify the objective QoS parameters that contribute to user perceptual quality, and map these parameters to user QoE. Subjective test results are often used as ground truth to validate the performance of the objective quality models. Most of the objective quality models are based on how the Human Visual System (HVS) receives and processes the information of the video signals. One of the commonly used methods is to quantify the difference between the original video and the distorted video, then weigh the errors according to spatial and temporal features of the video. However, the need to access original video hinders online QoE monitoring. In order to develop QoE prediction models that do not depend on original videos, network statistics (such as packet loss) and spatiotemporal features extracted or estimated from the distorted video, are leveraged. Though some of the objective quality models can realize real-time QoE prediction(e.g., [14]–[27]), it is still an indirect way for QoE prediction. Most of the objective quality models rely on subjective test results to train model parameters. Therefore, these models cannot be widely applied due to limitations of the subjective test.

Data-driven video quality analysis emerges as a promising way of solving the problems faced by the previous methods. Video streaming over the Internet has made large-scale of data available for analyzing user QoE. How to effectively leverage these valuable data is both challenging and promising. There are two ongoing trends for data-driven video quality assessment. The first trend is from user quality of "experience" to user quality of "engagement". In stead of user opinion score, which can only be obtained from subjective test, QoE metrics that can be easily quantified and measured without much human interference are being explored, for example, the viewing time, the number of watched videos and the probability of return. The second trend is from small-scale lab experiments (e.g., VQEG FRTV-I subjective test involved 287 viewers [28], LIVE database involved 31 viewers [29]) to large-scale data mining (e.g., [30] contains 40 million video viewing sessions). Sophisticated models with high computational complexity may work well on small-scale data, but are very likely to be outperformed by simple models on large-scale online QoE evaluation. Developing light-weight, efficient and reliable QoE prediction models based on big data is the future direction.

There have been several surveys on video quality assessment [31]–[34], mostly focusing on objective quality models. This survey paper differs from all the previous survey papers as it provides a comprehensive overview of the evolution of QoE-based video quality assessment methods. As far as we know, we are the first to include the data-driven QoE analysis models, which have newly emerged and raised research interest.

The rest of the paper is organized as follows. In Section II, we provide the background of video quality assessment, and identify factors that may influence user QoE. In Section III, we give a detailed description of subjective test. In Section IV, we classify existing objective quality models and introduce representative ones in each class. In Section V, we present the new research progress on the data-driven QoE analysis models. In Section VI, applications of video QoE models are reviewed. Future research directions on QoE are discussed in Section VII. We finally summarize our work in Section VIII.

II. Background

In this section, we give a brief introduction of the video transmission system, focusing on the factors that may have an influence on user experience by causing video distortions or affecting viewing environment. In the subjective test, these factors are often considered as test conditions; in the objective quality models, these factors are often used as input for computing the final objective metrics; in the data-driven analysis, these factors are often collected in the data set for QoE prediction.

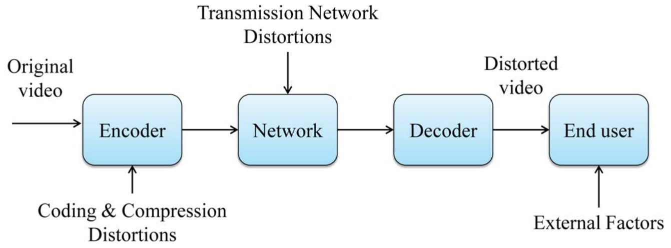


Fig. 2. Video transmission path.

The video transmission path from the server side to the client side includes: encoder, transmission network, decoder and display, as shown in Fig. 2. Each of these four places may introduce distortions or impairment that will affect the viewers' perception of the video quality. The resulting distorted videos usually exhibit the following typical visual distortions [35]:

- *Blocking effect*. Blocking effect refers to the discontinuity at the boundaries of two adjacent blocks. The reason for blocking effect is that the video coding is block-based, that is, individual blocks are coded separately, resulting in different types and levels of coding errors.
- *Blurring*. Blurring refers to the loss of spatial information or edge sharpness, especially for roughly textured areas or around scene object edges.
- *Edginess*. Edginess refers to the distortions happened at the edges of an image. The differences between the edge characteristics of the original video and those of the distorted video are often given special attention.
- *Motion jerkiness*. Motion jerkiness refers to the time-discrete intermission of the original continuous, smooth scene. This often happens due to delay variance (also known as "jitter"), which will be explained in Section II-B.

The visual impact of the above distortions does not only depend on the absolute quantization error, but also on the spatiotemporal features of the video sequence, both in the local level and in the global level. The threshold, above which the distortion is perceivable, is often referred to as Just Noticeable Difference (JND) [36], [37]. In the JND model, the following characteristics of the Human Visual System (HVS) are most commonly considered [38], [39]:

- Low-level characteristics:
  - *Frequency-dependent sensitivity*. The HVS has different sensitivity to motion, shape, depth, color, contrast, and lumination. Therefore, different errors will receive different sensitivity from the HVS. The HVS sensitivity decreases as the spatial or temporal frequency increases [40]. Many models use low-pass filter or band-pass filter to simulate such a feature [36], [41]–[44].
  - *Masking effect*. Under masking conditions, the perception of the visual target will be weakened by the masking stimulus in temporal or spatial proximity. A review of the research on visual masking can be found in [45].
- Mid- to higher-level characteristics include attention, eye movement and different unpleasantness towards different distortions. For example, looking at an image, the HVS first perceives the global structure, and then observes the detailed specifics. This coarse-to-fine-grained process is known as Global precedence, one important feature of the HVS [46].

TABLE II
COMPARISON OF VIDEO COMPRESSION FORMATS

Standard	Lossy/ Lossless	Major Applications
MPEG-2	Lossy	DVD, HDTV, Blu-ray Disc
MPEG-4 Part 2	Lossy	DVD, HDTV, electronic surveillance systems
H.264/MPEG-4 AVC	Lossy	Low/high resolution video, broadcast, DVD, RTP/IP packet networks
MPEG-C Part 3	N/A	Auxiliary video data format, e.g., stereoscopic video
MPEG-H Part 2 High Efficiency Video Coding (HEVC)	N/A	Ultra HD video
MPEG-DASH	N/A	Multimedia over the Internet
Multiview Video Coding (MVC)	N/A	Stereoscopic video, free viewpoint television, multiview 3D television
Scalable Video Coding (SVC)	N/A	Video storage, streaming, broadcast, conferencing, surveillance
VP8	Lossy	Real-time applications like videoconferencing
Dirac	Lossy/ Lossless	High-quality video compression for Ultra HDTV and beyond

Interested readers can refer to [47] for a detailed description of the artifacts of video compression and the mechanism of HVS [47].

A. Coding and Compression

In order to transmit rich video content through a capacity-limited network, the original video information needs to be reduced by compression. The compression methods may be lossy or lossless: lossless compression method can restore the original video while the lossy method may lead to video quality degradation. Video compression formats define the way to represent the video and audio as a file or a stream. Video codec, a device or software, encodes (compresses) or decodes (decompresses) a digital video based on the video compression format. The encoded video is often combined with an audio stream (encoded based on the audio compression format) to fit in a multimedia container format[1] such as FLV, 3GP, MP4, and WebM. Table II gives a comparison of commonly-used video compression formats.

The video compression formats, such as MPEG or H26x, significantly influence the video quality, because they decide how a video is coded. The following coding-related factors are often taken into consideration for QoE evaluation.

- *Bitrate*. Bitrate is the rate at which the codec outputs data. Constant bitrate (CBR) and variable bitrate (VBR) may be

[1] A container format can contain different types of video and audio compression. The container format may also include subtitles, chapter-information, and meta-data.

used. CBR is simple to implement, but it may not allocate enough data for more complex part of the video. VBR fixes the problem by flexibly assigning different bitrates according to the complexity of the video segments, but it takes more time to encode. Moreover, the instant bitrate of the VBR may exceed the network capacity. Efficient compression formats can use lower bitrates to encode video at a similar quality. Moreover, it is shown that high bitrate does not always lead to high QoE (e.g., frequent bitrate switching annoys viewers [30], [48]). Therefore, bitrate alone is not reliable to measure the video quality.

- *Frame rate*. Frame rate is the number of frames per second. The human visual system (HVS), can analyze 10 to 12 images per second [49]. The frame rate threshold, beyond which the HVS perceives no interruption, depends on both the content (e.g., motion) and the display (e.g., lighting). Given a fixed encoding bitrate subject to bandwidth limitation, higher frame rate means lower number of bits for each frame, therefore higher coding and compression distortions. It is shown that the frame rate affects QoE depending on the temporal and spatial characteristics of the video content [50].
- *Temporal and spatial features* of the video. Videos with different temporal and spatial features will have different degree of perceptual quality. For example, videos with low temporal complexity, where the frames are very similar to each other, may suffer less from jitter or packet loss as the viewers may not notice the delayed or missing frames. However, videos with high temporal complexity, where frames are quite different from each other, may be sensitive to jitter or packet loss because much information will get lost. A classification of video content, based on their temporal (e.g., movement) and spatial (e.g., edges, blurriness, brightness) features, is given in [51].

B. Transmission Network

Common transmission networks that are considered in the video QoE research include television broadcasting network and the Internet. For television broadcasting network, video quality assessment is usually conducted for different display resolutions, such as standard-definition television (SDTV), enhanced-definition television (EDTV), high-definition television (HDTV) and ultra-high-definition television (UHDTV). For video over the internet, special attention has been paid to IP network and wireless network, the latter including cellular network (or mobile network), wireless local area network (WLAN), sensor network and vehicular network. The video may be delivered by client-server video distribution or P2P video sharing.

Transmission network condition will greatly affect the video quality. Fig. 3 gives a brief illustration of the end-to-end video transmission between the server and the client. There are three major factors that will lead to video quality degradation.

- *Packet loss*, which is due to unreliable transmission.
- *Delay*, which depends on the network capacity.
- *Jitter*, also called delay variance, refers to irregular delays.

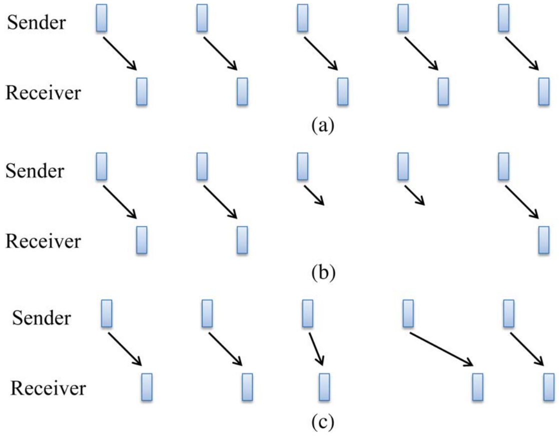


Fig. 3. Video transmission. (a) Delay. (b) Delay + Packet loss. (c) Delay + Jitter.

If there is only transmission delay (no packet loss or jitter), the video can be played smoothly with the help of a buffer. With packet loss, the most recent frame may freeze, then jump to the next inconsecutive frame that arrives. Packet loss can be compensated by retransmission at the cost of increased delay and jitter. Retransmission is a tradeoff between decreased packet loss and increased delay and jitter. With jitter, the most recent frame may freeze, until the belated frame arrives. Jitter can be mitigated through buffering, where the receiver plays the frames in the buffer with more steadiness. Choosing the optimal buffer size is a tradeoff between decreased jitter and increased delay. Some research found that jitter has nearly the same effect on the QoE as packet loss [52].

C. External Factors

Apart from distortions, there are other factors that will affect QoE. These external factors, some of which may not have direct impact on the video quality, influence users' experience by affecting viewing environment. The following are some typical external factors:

- *Video service type*, whether the video is live streaming video or Video-on-Demand (VoD). In [30], [53], it is assumed that viewers may have different quality expectations of VoD and live streaming video. By separating the two types of videos, the QoE prediction can be improved.
- *Viewer demography*. The characteristics of the viewers such as age, gender, occupation, nationality or even education background and economic factors will all have some impact on their perceived quality.
- *Viewer geography*. Studies show that people from different countries have different patience when faced delay of a service [54].
- *Video length*. It has been verified that viewer behaviors are different towards long videos (e.g., more than 10 minutes) and short videos (e.g., less than 10 minutes). For example, viewers are likely to be more tolerant of distortions when watching long videos than short videos.
- *Video popularity*. Viewers tend to be more tolerant of bad QoS for popular videos. However, there is also an interesting finding that more popular video has short viewing

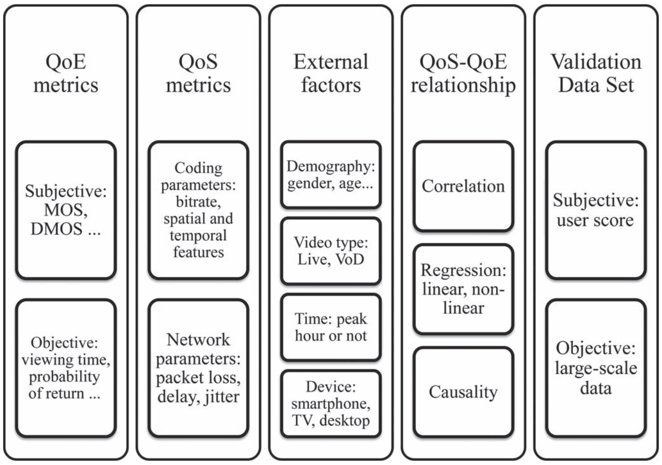

QoE metrics
Subjective: MOS, DMOS ...
Objective: viewing time, probability of return ...
QoS metrics
Coding parameters: bitrate, spatial and temporal features
Network parameters: packet loss, delay, jitter
External factors
Demography: gender, age...
Video type: Live, VoD
Time: peak hour or not
Device: smartphone, TV, desktop
QoS-QoE relationship
Correlation
Regression: linear, non-linear
Causality
Validation Data Set
Subjective: user score
Objective: large-scale data

Fig. 4. A summary of existing perceptual video quality assessment works.

session [55]. Possible explanation is that popular videos may be viewed from other sources, and viewers quit not wanting to watch repeated sessions.

- *Device*. The devices on which viewers can watch the video include TV, desktop computer, laptop, tablet, smartphone, etc. Specifically, the fast-growing popularity of smartphone and tablet draws attention to study on viewer experience on these devices. Viewers may have different expectations when they watch video on different devices. Device also determines the screen size. Typical screen sizes include QCIF, CIF, VGA, SDTV or HDTV [56].
- *Time of the day & day of the week*. User experience may be different when they watch video in peak hours or idle hours. It is estimated that viewers may have better viewing experience in the evening and on weekends, when they are more relaxed and are expected to watch the videos for a longer time.
- *Connectivity*. The major concern is usually the last-mile connection, for example, fiber, cable, DSL, 3G/4G, etc.

Before we discuss each stage of video quality assessment, we first give a brief summary of the related works in Fig. 4.

III. Subjective Test

Subjective test directly measures QoE by asking human assessors to give their scores for the quality of the video sequences under test. Subjective test results are often used as the ground truth for validating the performance of the objective quality model in Section IV. In this section, we first describe the conventional procedures of conducting subjective test in the laboratory context. Then, we give special instructions to the requirement of subjective test for 3D videos. Finally, we introduce subjective test crowdsourcing through Internet crowdsourcing platforms.

The flow of the subjective test is shown in Fig. 5.

A. Test Preparation

Test preparation includes checking the test environment, setting up equipment, selecting source videos, processing source videos, and recruiting assessors [57].

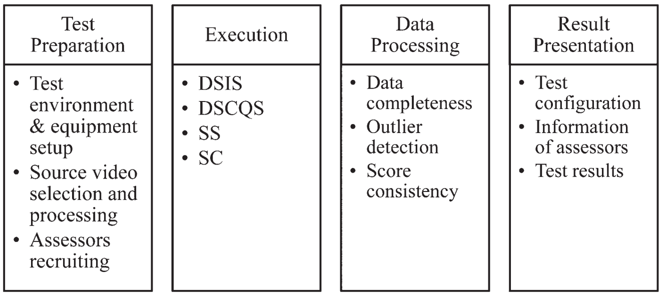


Fig. 5. Flow of the subjective test.

TABLE III
Test Environment Requirement [57]

Requirement	Laboratory	Home
Ratio of inactive screen luminance to peak luminance	≤ 0.02	≤ 0.02
Ratio of background luminance to picture's peak luminance	≈ 0.15	N/A
Peak luminance	N/A	$200cd/m^2$
Ratio of screen only black level luminance to peak white luminance	≈ 0.01	N/A
Display brightness and contrast	PLUGE [58], [59]	PLUGE
Background chromaticity	D_{65}	N/A
Environmental illuminance on the screen	N/A	$200lux$
Other room illumination	Low	N/A
Maximum observation angle relative to the normal	30°	30°
Screen size	N/A	Meet rules of PVD
Monitor	N/A	No digital processing; Meet resolution requirement

1) Test Environment: The subjective test can be conducted in two kinds of environment: laboratory environment and home environment, yet nearly all the subjective tests are conducted in the laboratory environment. Table III shows the requirement of both environment specified by the International Telecommunication Union (ITU) Recommendation ITU-R BT. 500-11 [57].

While the laboratory environment is easier to control, the home environment is more close to the users' real viewing experience. The screen size affects the preferred viewing distance (PVD), at which the viewers have the optimal viewing experience. Therefore, in the test, the viewing distance should be adjusted to satisfy the PVD determined by the screen size. It is suggested that the maximum and minimum resolutions of the monitor be reported, especially the consumer TV sets used in the home environment.

2) Source Video Selection: As we discussed before, the video content will influence user viewing experience. When selecting the source materials, the following factors have to be taken into consideration.

- Color.
- Luminance.

  - High luminance
  - Low luminance.
- Motion and spatial features.
  - Still images or video sequences.
  - Moving directions of the objects.
- Source origin, e.g., film, news, sports.
- Other factors, e.g., avoiding cultural or gender offensive materials.

3) Source Video Processing: The experimenters have to choose the Hypothetical Reference Circuits (HRC), such as the encoding bitrate and packet loss rate, to process the source videos. Firstly, the encoder encodes the video with a certain video compression format, during which the encoder's distortions are applied. Secondly, the video goes through the (often simulated) transmission network, during which the network's distortions are applied. Finally, the processed video can be obtained after decoding. If more than one distortion factors are considered (let $F_1, F_2, \ldots F_k$ denote the various factors, and F_i has n_i levels $f_{i,1}, f_{i,2}, \ldots, f_{i,n_i}$), "reasonable" range for each distortion factor (i.e., $f_{i,1}, f_{i,2}, \ldots, f_{i,n_i}$) should be determined, and the maximum and minimum values be specified. There are two ways to process the videos:

- Each processed video represents a level of one factor, while other factors are fixed at a chosen level. For instance, for factor F_i, we have processed videos $\{(f_{1,0}, \ldots, f_{i,j}, \ldots, f_{k,0})\}_{j=1,\ldots,n_i}$, in which $f_{1,0}, \ldots, f_{k,0}$ are reference levels.
- All combinations of the factor levels are considered, that is, we have processed videos $\{(f_{1,j_1}, \ldots, f_{k,j_k})\}_{j_i=1,\ldots,n_i}$.

After the video processing, the processed videos need to be normalized to eliminate "deterministic" differences from the source videos. The normalization includes temporal frame shift, horizontal and vertical spatial image shift, and chroma and luma scaling and alignment. The amount of normalization is estimated from the source and processed videos, and will be applied uniformly to all the video sequences. The accuracy of the alignment can be verified by MSE.

4) Assessor Recruitment: It is required that at least 15 non-expert assessors should be recruited for the tests. The assessors should be tested on visual acuity, color vision and familiarity of the language used in the test. Since the demography of the assessors may have influence on the final evaluation results, their personal information should be collected as broadly as possible such as age, gender, occupation, education, etc. Before the test sessions start, the assessors should be given instructions on:

- The flow of the test, e.g., training subsessions and test subsessions;
- The presentation of each trial, e.g., double stimulus or single stimulus;
- The possible quality impairment, e.g., color, brightness, depth, motion and "snow";
- The evaluation scale, e.g., continuous or categorical.

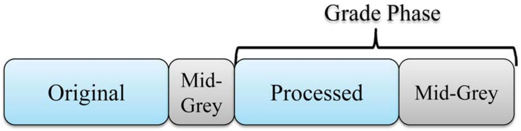


Fig. 6. DSIS video/image presentation sequence option I.

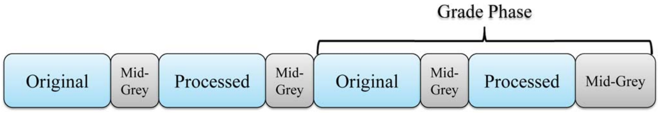


Fig. 7. DSIS video/image presentation sequence option II.

TABLE IV
DSIS SCALE

1	2	3	4	5
Imperceptible	Perceptible, but not annoying	Slightly annoying	Annoying	Very annoying

B. Test Execution

Test execution includes conducting the subjective tests and collecting the test results (e.g., user scores) [57]. Each test session should last fewer than 30 minutes, consisting of three subsessions:

- *Training subsession* is used to give instructions to the assessors about the sequence and timing of the test.
- *Stabilizing subsession* is used as a "warm-up" for the assessors to stabilize the following assessment. The assessment in this subsession will not be included as the results for further analysis.
- *Main test subsession* is the formal test phase, the results of which will be used for further analysis.

The order of the video presentation should be randomized, covering all the possible impairment conditions that are under study. In the main test subsession, there are several test methods that can be applied:

1) Double-Stimulus Impairment Scale (DSIS) Method (The EBU Method): For the DSIS, the assessors are first presented the source video, then presented the processed video. The assessors only grade the processed video, based on his knowledge or impression of the source video. For the assessment of a certain video, the presentation sequence has two options as shown in Figs. 6 and 7. In Fig. 6, the source video and the processed video are presented to the assessor only once, and the assessor can grade the video at the start when he sees the processed video. In Fig. 7, the source video and the processed video are presented to the assessors twice, and the assessor can grade at the start when he sees the source video for the second time. The scale for DSIS is discrete grades from 1 to 5 as shown in Table IV, indicating how the assessors evaluate the impairment of the processed video. It is found that the DSIS results are more stable for small impairment than for large impairment.

2) Double-Stimulus Continuous Quality-Scale (DSCQS) Method: In DSCQS, the assessors are also presented both the source video and the processed video. Let "PS" and "SP" denote the order of "first processed video, then source video" and"first source video, then processed video" respectively. It

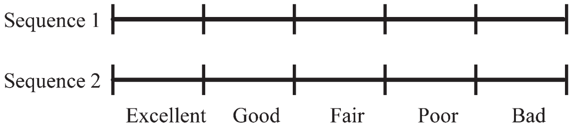


Fig. 8. DSCQS scale.

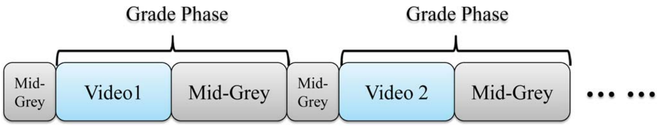


Fig. 9. SS video/image presentation sequence option I.

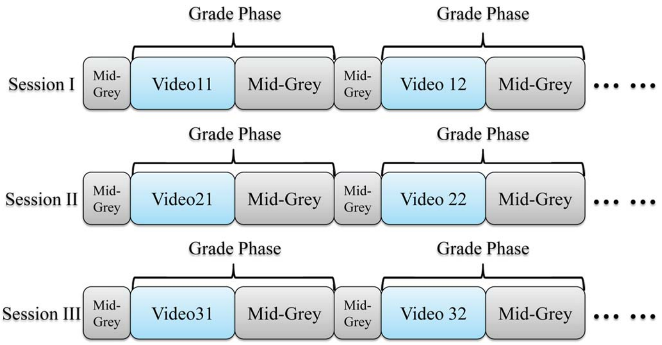


Fig. 10. SS video/image presentation sequence option II.

should be followed that the same video with different test conditions are not presented consecutively. The number of consecutive "PS" presentation order should be no more than a threshold, the same is for the "SP" presentation order. In addition, the number of events that two video sequences are presented consecutively should be no more than a threshold. Compared with DSIS, DSCQS is different in the following aspects:

- For the same video, both the source version and the processed version are presented to the assessors, but the assessors do not know which one is the source version.
- The assessors are asked to grade both versions of the same video. The scale for DSCQS grading is different (as shown in Fig. 8) in two aspects:
  - It has continuous grade bars;
  - It has two bars for the same video.
- For DSCQS grades, it is not the absolute value, but the difference between the two values for the same video, that matters.

3) Single-Stimulus (SS) Method: In SS, only the processed videos are presented to the assessors. The presentation can have two forms:

- Each processed video is shown once to the assessors (as shown in Fig. 9). The order to present the processed videos is random.
- Each processed video is shown three times in three sessions to the assessors (as shown in Fig. 10). The order to present the processed videos in each session should be different. Only the results in the last two sessions are counted for final results. The first session is to stabilize assessors' grading.

The grading scheme for SS can have three different forms:

- Categorical grading. The assessors categorize the videos into pre-defined categories. The category can be given numerically (e.g., category "1", "2"...,"10") or verbally (e.g., category "Excellent", "Good", "Fair", "Poor", "Bad").
- Numerical grading. The assessors give marks, for example, 1 $\sim$ 100.
- Performance-based grading. While the above two methods solicit assessors' grading directly, the video quality can be indirectly inferred by asking assessors to give video-related information.

Compared with the Double-Stimulus (DS) method, the Single-Stimulus method has the following advantages:

- For DS, if the source and processed videos are presented simultaneously on split screens, the assessors attention may be distracted [56].
- For DS, if the source and processed videos are presented consecutively, more time is required for one pair of video sequence. Since it is required that one session should not exceed 30 minutes, the possible pairs of video sequences tested in one session have to be reduced. Therefore, multiple sessions may be conducted, leading to the problem of how to best combine the results from different sessions.

4) Stimulus-Comparison (SC) Method: In the SC, two (processed) videos are presented to the assessors, and the assessors grade the *relationship* of the two videos. The grading scheme for SC also has three different forms:

- Categorical grading. The assessors categorize the relationship between the two videos into pre-defined categories. The category can be given numerically (e.g., category (the second video is) "−3, much worse", "−2, slightly worse",...,"3, much better") or verbally (e.g., category "Same","Different").
- Numerical grading. The assessors give (continuous) grades, for example, 1 $\sim$ 100 to the difference degree of the two videos.
- Performance-based grading. Assessors are asked to identify whether one video has more or less of a certain feature than the other video.

C. Data Processing

Data processing includes checking data completeness, screening the outliers and inconsistent assessors. To start with, the assessors' grading can be processed into two user score metrics:

- *Mean Opinion Score (MOS)* is for the single-stimulus tests. It is calculated as the average of the grades for a processed video. MOS is often used to validate the performance of the no reference objective quality models, which will be introduced in Section IV.

- *Difference Mean Opinion Score (DMOS)* is for the double stimulus tests. It is calculated as the average of the arithmetic difference between the grades given to the processed video and the grades given to the source video. DMOS is often used to validate full reference objective quality models and reduced reference objective quality models, which will be introduced in Section IV.

Then, the results should be screened as follows.

- Check the completeness of the data: whether an assessor gives score to every video; whether an assessor grades both source and processed video in the double stimulus score.
- Remove assessors with extreme scores (outliers).
- Remove assessors with unstable scores.

Check the data completeness is easy to do. Now we introduce how to screen the outliers and inconsistent assessors in more details. The basic assumption is that the data collected from the subjective test follow a certain distribution within the scoring range (e.g., 1 $\sim$ 5, or 1 $\sim$ 100), with variations due to differences in assessors, video contents, and so on. Let OS be the individual opinion score, i be the assessor index (a total of I assessors), j be the test condition index (a total of J test conditions), k be the video sequence index (a total of K video sequences). First, let's define some key parameters:

- *Mean Score*. The mean score for the jth test condition and kth video sequence is

$$MOS_{jk} = \frac{1}{I} \sum_{i} OS_{ijk} \tag{1}$$

- *Standard Deviation*. The standard deviation of MOS_{jk} is

$$S_{jk} = \sqrt{\sum_{i} \frac{(MOS_{jk} - OS_{ijk})^2}{I-1}} \tag{2}$$

- *95% Confidence Interval*. The 95% Confidence Interval of MOS_{jk} is

$$[MOS_{jk} - \delta_{jk}, MOS_{jk} + \delta_{jk}] \tag{3}$$

  in which $\delta_{jk} = 1.96 S_{jk}/\sqrt{I}$.
- *Kurtosis Coefficient*. The Kurtosis Coefficient, $\beta_{2,jk}$, used to verify whether the data distribution of the jth test condition and kth video sequence is normal, can be calculated as

$$\beta_{2,jk} = \frac{I \sum_i (MOS_{jk} - OS_{ijk})^4}{\left[\sum_i (MOS_{jk} - OS_{ijk})^2\right]^2} \tag{4}$$

1) Data Screening for DS: The data screening for DS is mainly to screen outliers, using algorithm 1. The detailed explanation is as follows:

- Step 2: Verify whether the data distribution of the jth test condition and kth video sequence is normal. If $\beta_{2,jk} \in [2, 4]$, the data distribution is regarded to be normal, otherwise, it is not.
- Step 3 $\sim$ 16: Compare the individual user score OS_{ijk} with two reference value $MOS_{jk} + 2S_{jk}$ and $MOS_{jk} - 2S_{jk}$ for normal distribution, or $MOS_{jk} + \sqrt{20}S_{jk}$ and $MOS_{jk} - \sqrt{20}S_{jk}$ for non-normal distribution. Individual user scores that are outside the range $[MOS_{jk} + 2S_{jk}, MOS_{jk} - 2S_{jk}]$ or $[MOS_{jk} + \sqrt{20}S_{jk}, MOS_{jk} - \sqrt{20}S_{jk}]$ will be recorded in $High_i$ and Low_i.
- Step 18$\sim$21: Decide whether to remove assessor i or not based on $High_i$ and Low_i.

Algorithm 1 Data Screening for DS

```
1: for all i, j, k do
2:   if $\beta_{2,jk} \in [2, 4]$ then
3:     if $OS_{ijk} \geq MOS_{jk} + 2S_{jk}$ then
4:       $High_i + +;$
5:     end if
6:     if $OS_{ijk} \leq MOS_{jk} - 2S_{jk}$ then
7:       $Low_i + +;$
8:     end if
9:   else
10:     if $OS_{ijk} \geq MOS_{jk} + \sqrt{20}S_{jk}$ then
11:       $High_i + +;$
12:     end if
13:     if $OS_{ijk} \leq MOS_{jk} - \sqrt{20}S_{jk}$ then
14:       $Low_i + +;$
15:     end if
16:   end if
17: end for
18: $Ratio_1 = \frac{High_i + Low_i}{JK};$
19: $Ratio_2 = \left|\frac{High_i - Low_i}{High_i + Low_i}\right|$
20: if $Ratio_1 > 0.05 \&\& Ratio_2 < 0.3$ then
21:   Remove assessor $i$;
22: end if
```

2) Data Screening for SS: The data screening for SS is two-folds: to screen the outliers who deviate from the average behavior, and to screen the assessors whose behavior is inconsistent. The difference between the screening process for DS and for SS is: for DS, we test each (condition, sequence) configuration; for SS, we test each (condition, sequence, time window) configuration. Let m be the index of time window (A total of M time windows).

- *Screen outliers*: also use Algorithm 1, but replace the OS_{ijk} with OS_{ijkm}, and modify the Kurtosis Coefficient $\beta_{2,jkm}$ and standard deviation S_{jkm} correspondingly. Further make the changes $Ratio_1 = \frac{High_i}{JKM}$ in Step 18, $Ratio_2 = \frac{Low_i}{JKM}$ in Step 19, and the condition for removing assessor i is $Ratio_1 > 0.2$ or $Ratio_2 > 0.2$ in Step 20.
- *Screen inconsistent assessors*

  The variable under test is

$$\widetilde{OS}_{ijkm} = OS_{ijkm} - MOS_{ijk} + MOS_{jk} \tag{5}$$

in which

$$MOS_{jk} = \frac{\sum_i \sum_m OS_{ijkm}}{I \times M}$$

$$MOS_{ijk} = \frac{\sum_m OS_{ijkm}}{M} \tag{6}$$

The corresponding Kurtosis Coefficient is

$$\widetilde{\beta}_{2,jkm} = \frac{I \sum_i (\widetilde{OS}_{ijkm})^4}{\left(\sum_i \widetilde{OS}^2_{ijkm}\right)^2} \tag{7}$$

The screening process is: use Algorithm 1, but replace OS_{ijk} with $\widetilde{OS}_{ijkm}$, $\beta_{2,jk}$ with $\widetilde{\beta}_{2,jkm}$, and modify the standard deviation $\widetilde{S}_{jkm}$ correspondingly. Further make the changes $Ratio_1 = \frac{High_i+Low_i}{JKM}$ in Step 18, $Ratio_2 = \frac{|High_i-Low_i|}{|High_i+Low_i|}$ in Step 19, and the condition for removing assessor i is $Ratio_1 > 0.1$ or $Ratio_2 < 0.3$ in Step 20.

D. Results Presentation

The final results should include the following:

- Test configuration;
- Test video sequences information;
- Types of video source;
- Types of display monitors;
- Number and demographic information of assessors;
- Reference systems used;
- The grand mean score for the experiment;
- The mean and 95% confidence interval of the statistical distribution of the assessment grades.

A common data format is desirable for inter-lab data exchange, because usually large-scale subjective tests will be carried out in different laboratories in different countries, maybe with assessors speaking different languages.

E. Subjective Test for 3D Videos

In [60], the ITU gives the guidance for subjective test for stereoscopic television pictures. Apart from the assessment factors for conventional monoscopic television pictures, there are additional factors to be considered for stereoscopic television pictures.

- *Depth resolution and depth motion*. Depth resolution is the spatial resolution in the depth direction; and depth motion is the movement along the depth direction.
- *Puppet theater effect* refers to the distortion in the reproduced 3D image, that the objects appear unnaturally large or small.
- *Cardboard effect* refers to the distortion in the reproduced 3D image, that the objects appear unnaturally thin.

In [61], the authors argue that the subjective test specified by the ITU may not simulate the home environment where the actual viewing is happening. In the standard ITU subjective test, short video sequences are often used, whose contents may not be interested to the viewers. Therefore, in [61], the authors propose to use long video sequences, with the test methods shown in Fig. 11. The same long video is continuously played with alternating processed and original segments, and assessors grade the video quality during the period when the original(unprocessed) segments are being played.

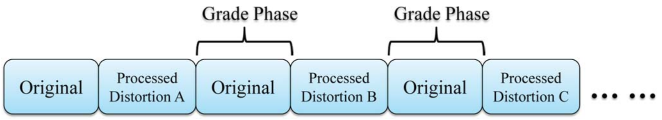


Fig. 11. Proposed 3D video evaluation method in [61].

F. Subjective Test Crowdsourcing

Conventionally, the subjective test is conducted in a lab or several cooperating labs, which is labor-intensive, time-consuming and expensive. A more cost-effective alternative is to conduct subjective test through Internet crowdsourcing platforms, such as Amazon Mechanical Turk (MTurk) [62].

One problem with the crowdsourcing subjective test is to detect the outliers, because the online assessors are performing the evaluation tasks without supervision. For example, if the test lasts a long time and the assessors get impatient, they may input random evaluations. In [63], the authors propose to verify the consistency of the ratings based on the *transitivity property*, that is, if the assessor prefers A to B, B to C, then he should prefer A to C. But the method cannot work when the data is incomplete. To solve this problem, the authors in [64] propose an outlier detection algorithm based on Hodge Decomposition theory, which is first proposed in [65] to check data consistency from incomplete and imbalanced data. In [66], paired comparison is proposed as a simpler rating scheme to replace MOS. In paired comparison, the assessors are given a pair of images or videos, and they only have to decide which one has better quality. A cheat detection mechanism based on the transitivity property is given to check and screen inconsistent assessment.

G. Discussion

Although the subjective test directly measures QoE by asking assessors for their evaluations, it suffers from some significant drawbacks:

- *High cost*. The subjective test is time-consuming, money-consuming and manpower-consuming.
- *Limited assessors*. Usually, no more than 100 assessors are involved in the subjective test due to its high cost. These assessors can only represent the demographic features of a very small fraction of the entire viewer population.
- *Controlled environment*. The subjective test is often conducted in the laboratory environment, which is not the usual place where the common viewers watch video. The results may not be an accurate reflection of viewers' true viewing experience in the wild, where other factors, such as delay, may also have an influence on QoE.
- *Limited distortion types*. The lab-processed distortion types are representative but cannot account for all parameters that have an impact on the QoE. Some of the conditions are hard to test in the laboratory environment, such as transmission network induced delay and jitter, or external factors such as different locations where viewers watch the video.

- *Distortion factor correlation*. One problem about video processing is that many of the distortion factors are correlated in reality. Some combinations of factors would not happen in the real environment. For example, if bitrate and frame rate are chosen as distortion factors, it is unlikely that the processing of (high bitrate, low frame rate) will happen in real environment.
- *Hard to account for frames of different importance in a video*. A video can be regarded as an aggregation of images (or frames), whose quality can be assessed by both double stimulus and single stimulus subjective tests. However, the quality of the video does not simply equal the sum of the quality of all its images. For example, some frames in a video is less visually important than others. Moreover, in video compression, certain frames (e.g., I-frame) contain more information than others (e.g., P-frame and B-frame).
- *Not applicable for online QoE estimation*. The subjective test cannot be used for real-time QoE monitor or prediction. Thus, it cannot provide instrumental guidance for real-time system adaptation.

IV. Objective Quality Model

To give relatively reliable QoE prediction but avoid the necessity of doing subjective test, researchers develop objective quality models. Objective quality models compute a metric as a function of QoS parameters and external factors. The output metric should correlate well with the subjective test results, which serve as the ground truth QoE. In this section, we first introduce representative objective quality models. Then, we describe the process of validating the performance of objective quality models. Finally, we introduce projects and international standards for objective quality models.

In previous survey papers on objective quality models, there are three major classification methods:

- The “psychophysical approach” and the “engineering approach” [47]. The two approaches are also termed as vision-based model and signal-driven model in some articles. The psychophysical approach is mainly based on characterizing the mechanisms of the HVS, such as masking effect, contrast sensitivity, and adaptation to color and illumination. The engineering approach is based on extracting and analyzing certain distortion patterns or features of the video, such as statistical features, structural similarity (SSIM) and compression artifacts (e.g., blockiness, edginess).
- Reference-based classification method [47]. Based on whether the reference to the original video is needed, the objective quality models are classified as Full Reference (FR) model, Reduced Reference (RR) mode and No Reference (NR) model.
  - *Full Reference (FR) Model*. Full access to the source video is required.
  - *Reduced Reference (RR) Model*. Partial information of the source video is required.
  - *No Reference (NR) Model*. No reference model does not need the access to the source video.

  The full reference and reduced reference models need to refer to the original video for quality comparison and assessment, making them less suitable for online QoE estimation. They are “intrusive” models in the sense that they insert additional load to the network or service [67]. No reference model is non-intrusive, adding no load to the network or service, thus more suitable for online QoE evaluation and system adaptation. When choosing a no reference model or metric for online QoE evaluation, real time performance and speed are also the deciding factors.
- Input data-based classification method [68]. Based on the type of the input data, there are five categories of models:
  - Media-layer models, whose input is the media signal.
  - Parametric packet-layer model, whose input is the packet header information.
  - Parametric planning model, whose input is quality design parameters.
  - Bitstream layer model, whose input is packet header and payload information.
  - Hybrid model, the combination of any of the other models.

The first two classification methods are most commonly adopted, and often used to complement each other. In general, psychophysical approach usually belongs to the FR, while RR and NR are mostly based on the engineering approach. Many survey papers mention both classification methods, but usually follow one of them. For example, [32], [69] mainly follow the psychophysical/engineering approach classification method; [31], [70]–[72] mainly adopt the reference-based classification method, and [47] adopts a combination of the two. The third classification methods is proposed in [68] and referenced in [31]. In [73], the objective models are classified as pixel-based model (e.g., PSNR and MSE), vision-based single-channel model, vision-based multi-channel model and specialized model, yet this is not a commonly adopted classification method.

The main purpose of this tutorial paper is to introduce the evolution of the video quality assessment methods on the whole, and in particular, to point out potential future directions. We will just adopt the existing classification methods for the objective quality model. Fig. 12 gives a summary of the objective quality models that we mainly focus on. We use FR/RR/NR as the first-tier classification, psychophysical/engineering approach as the second-tier classification, and other more specific criterion as the third-tier classification. It should be noted that some classification is non-exclusive. For example, similarity structural (SSIM) is an engineering approach, but many variations of SSIM also incorporate psychophysical features in the design. In this case, we still classify these variations as engineering approach as their major basis is SSIM. We believe that as the research on objective quality models advances, there will be a need for an evolution of the classification methods, but this is not the focus of this tutorial paper.

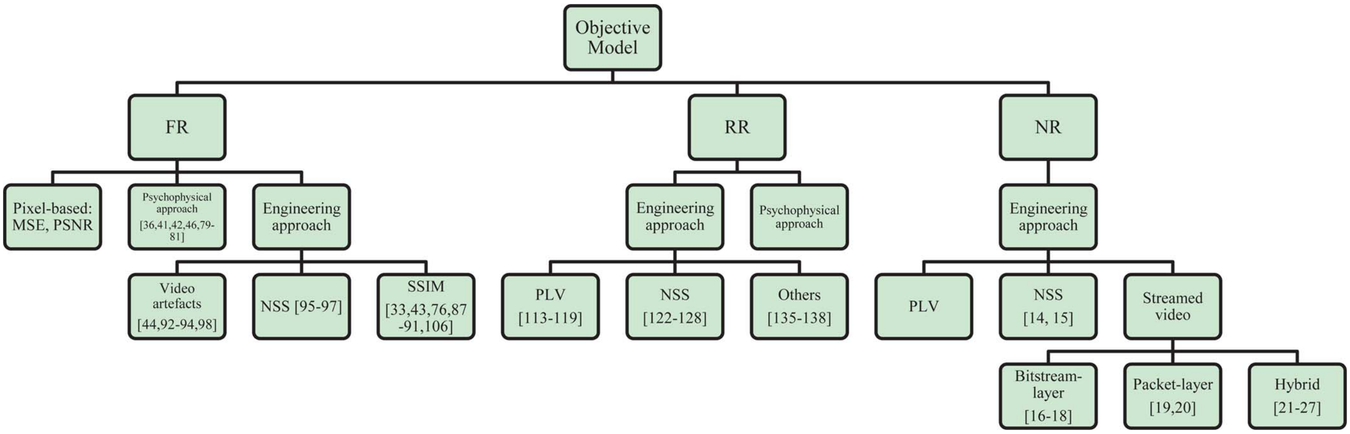


Fig. 12. An overview of objective quality models.

TABLE V
PSYCHOPHYSICAL APPROACH MODELS

CC: Correlation Coefficient; RMS: Root Mean Squared Error; OR: Outlier Ratio; SRCC: Spearman Rank-order Correlation Coefficient

	Year	Model Basis	Validation Database	Performance factors			
				CC	RMS	OR	SRCC
MPQM [36]	1996	Human spatio-temporal vision, contrast sensitivity and masking	3 video sequences with MPEG-2 and H.263	N/A	N/A	N/A	N/A
DVQ [41], [42]	2000	Human spatial-temporal contrast sensitivity	Data set in [78]	N/A	14.61	N/A	N/A
PVQM [79]	2002	Luminance "edginess", color error and temporal decorrelation	Medium to high quality database with digital codec distortions and analog PAL, VHS and Betacam distortions	0.934	N/A	N/A	N/A
VSNR [46]	2007	Contrast sensitivity, visual masking and global precedence	LIVE database	0.889	7.390	N/A	0.889
MOSp [80]	2009	Edginess and masking effect	Video sequences with H.264 at different bitrate	0.947	N/A	0.402	N/A
AFViQ [81]	2013	Contrast sensitivity, foveated vision, visual attention	LIVE database, VQEG HDTV Phase I database [82]	0.83	8.74	N/A	N/A

A. Full Reference Model

In this section, we mainly introduce three kinds of full reference models: simple pixel-based models, psychophysical approach and engineering approach. In the engineering approach, we further introduce models based on video artefacts, natural scene statistics (NSS) and structural similarity (SSIM).

1) Pixel-based Models: Two most basic objective quality models are Mean Squared Error (MSE) and Peak-Signal-to-Noise Ratio (PSNR), which are simple to compute and therefore usually serve as the benchmark for evaluating more advanced models.

- *MSE.* MSE can be calculated as

$$MSE = \frac{1}{N}\sum_i (y_i - x_i)^2 \tag{8}$$

in which x_i is the ith sample of the original signal, and y_i is the ith sample of the distorted signal.

- *PSNR.* PSNR is defined as

$$PSNR = 10 \log_{10} \frac{MAX}{MSE} \tag{9}$$

in which MAX is the maximum signal energy [74].

The advantage of pixel based model is simplicity. However, neither model consider the features of the HVS and viewing conditions, and are poorly correlated to subjective results [75]–[77].

2) Psychophysical Approach: Objective quality models of the psychophysical approach are based on the features of the HVS, which is related to visual perception, for instance, contrast sensitivity, frequency selectivity, spatial and temporal features, masking effects, and color perception [72]. Table V gives a summary of the psychophysical approach models. Note that the performance factors given in the table are highly dependent on the database used for evaluation and different model parameters; the values provided in the table only serve as a reference.

- Moving Picture Quality Metric (MPQM)
  MPQM is based on two features of human perception: contrast sensitivity and masking effect [36]. Contrast sensitivity means that a signal is visible only if its contrast is higher than a *detection threshold*, which is a function of spatial frequency. The inverse of the detection threshold is defined as the contrast sensitivity, which is usually denoted by the Contrast Sensitivity Function (CSF). The contrast

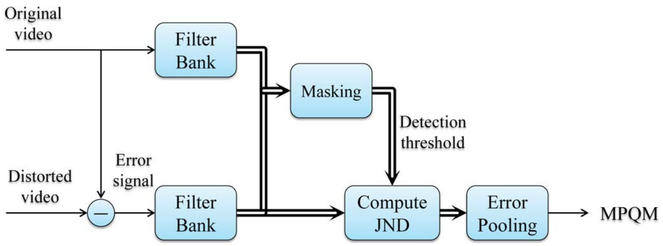


Fig. 13. Flow of MPQM.

sensitivity function proposed by Manos and Sakrison is [83]

$$A(f)=2.6(0.0192+0.114f)\exp\left[-(0.114f)^{1.1}\right]=\frac{1}{D_0} \tag{10}$$

in which f is the spatial frequency, D_0 is the detection threshold of the distortion without masking effect. One of the characteristics of the HVS is contrast masking: the visibility of a signal is highly affected by its background. The detection threshold of the foreground signal is a function of the contrast of the background. The distortion can be viewed as the foreground signal on the background original image. The foreground distortion may be highly visible, or partly/completely masked by the background original image. Let D denote the detection threshold of the distortion with the masking effect, C_b denote the contrast of the background. The masking effect model gives the following function of the D depending on D_0 and C_b:

$$D=\begin{cases} D_0, & C_b<D_0\\ D_0\left(\frac{C_b}{D_0}\right)^{\eta}, & C_b\geq D_0 \end{cases} \tag{11}$$

in which η is a constant parameter. Fig. 13 shows the flow of calculating the MPQM metric. The thick lines represent multi-channel output or input. Firstly, the original video and the error signal (the difference between the original and distorted videos) go through the filter bank, which decomposes them into multiple channels according to the orientation, spatial frequency and temporal frequency. Secondly, the detection threshold under the masking effect is calculated according to (10) and (11), for each channel. Thirdly, the error signal is divided by the detection threshold to get the Just Noticeable Difference (JND), which will be pooled over all channels by Minkowski summation (with exponent β) to get the final metric as follows:

$$MPQM=\left(\frac{1}{N}\sum_{f=1}^{N}\left(\frac{1}{N_xN_yN_t}\sum_{x,y,t}|e(x,y,t,f)|\right)^{\beta}\right)^{\frac{1}{\beta}} \tag{12}$$

in which $e(x,y,t,f)$ is the computed JND at position (x,y), time t and channel f.

- Digital Video Quality (DVQ)
DVQ calculates the visual difference between the original and distorted video sequences using Discrete Cosine Transform (DCT). It incorporates spatial and temporal filtering, spatial frequency channels and contrast masking [41], [42]. The flow of calculating DVQ is illustrated in Fig. 14. Pre-processing includes sampling, cropping, and color transformations to restrict the later processing into the Region of Interest (RoI). Then, blocked DCT is performed on the processed video sequence. Local contrast is obtained by dividing the DCT coefficient with the DC coefficients. Temporal filtering and JND conversion implement the temporal and spatial feature of the CSF respectively. After the contrast masking process, the results are pooled by Minkowski summation as in (12).

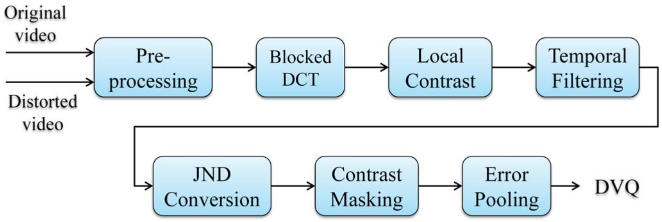


Fig. 14. Flow of DVQ.

- Perceptual Video Quality Measure (PVQM)
PVQM calculates the following three indicators:

  - *Edginess indicator E*. HVS is sensitive to the edge and local luminance change. The local edginess can be approximated by the local gradient of the luminance signal. The difference between the edginess of the distorted video and the original video can be viewed as sharpness loss (if the edginess of the distorted video is smaller) or distortion (if the edginess of the distorted video is higher). The introduced edginess difference is more obvious in areas with less edginess than in areas with much edginess. The edginess indicator is the local edginess of the distorted video minus the local edginess of the original video, then divided by the local edginess of the original video.
  - *Temporal indicator T*. While edginess indicator is a pure spatial indicator mostly for still images, the temporal indicator characterizes the motion of the video sequence. The fast-moving sequence will decrease visual sensitivity in details. Temporal indicator quantifies the temporal variability of the video sequence by calculating the correlation of the current frame (t) and the previous frame $(t-1)$.
  - *Chrominance indicator C*. Color errors in areas with saturated colors are less perceptible to the HVS. Chrominance indicator calculates the color saturation of the original and distorted videos.

Two cognitive models are further applied for pooling the above indicators from both spatial and temporal aspects:

- Spatial pooling. Errors on the edge are less disturbing than those in the central area, therefore, the edginess indicator and the chrominance indicator are given heavier weights in the center of the image and lighter weights on the bottom and top.

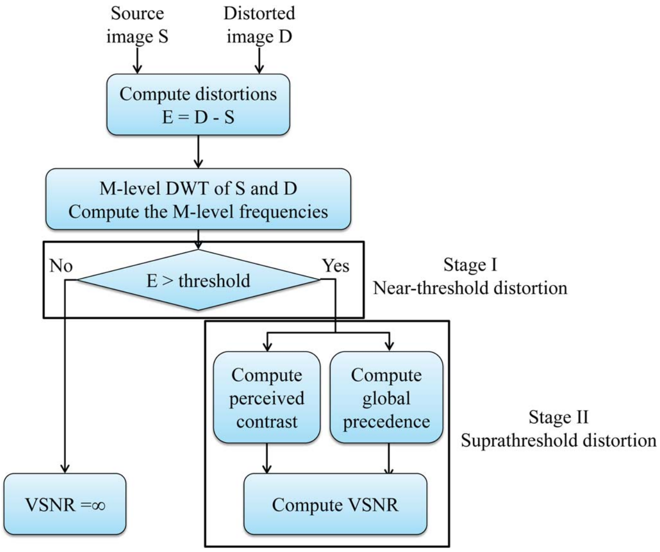

Source image S
Distorted image D
Compute distortions E = D - S
M-level DWT of S and D Compute the M-level frequencies
No
E > threshold
Yes
Stage I Near-threshold distortion
Compute perceived contrast
Compute global precedence
Stage II Suprathreshold distortion
VSNR =∞
Compute VSNR

Fig. 15. Flow of VSNR.

- Spatio-temporal pooling. HVS punishes more severe errors. Therefore, large local spatial and temporal errors are given heavier weights.

The final PVQM is the linear combination of the three indicators after aggregation.

$$PVQM = 3.95E + 0.74C - 0.78T - 0.4 \tag{13}$$

- Visual Signal-to-Noise Ratio (VSNR)
The VSNR determines near-threshold and suprathreshold distortions in two stages, as shown in Fig. 15. For preprocessing, the original image S and distorted image D are decomposed by M-level DWT to obtain two sets of $3M + 1$ subbands. Then, the assessment goes through two stages. In the first stage, near-threshold distortion is considered. Low-level HVS properties are used to determine whether the distortion is beyond the threshold: if not, the image is assessed to have perfect visual fidelity, thus $VSNR = \infty$; otherwise, the image will be put through the second stage. In the second stage, suprathreshold distortion is considered. Both low-level and mid-level HVS properties are used to compute the final VSNR value.

  - Stage I: Near-Threshold Distortion
  Whether an observer can detect the distortion depends on the spatial frequency of the image, which depends on the viewing conditions: the resolution of the display r and the viewing distance d. M-tuple frequency vector $f = [f_1, \ldots, f_m, \ldots, f_M]$ can be computed as

$$f_m = 2^{-m} rd \tan \frac{\pi}{180} \tag{14}$$

  To decide whether the distortions are visually perceptible, the contrast detection threshold for a particular frequency f is calculated as follows:

$$T_m = \frac{C(S_f)}{a_0 f^{a_2 \ln f + a_1}} \tag{15}$$

  in which $C(\cdot)$ is the root-mean-square (RMS) contrast function [84]; a_0, a_1, a_2 are parameters that can be obtained from experiment. If, for any subband f_m, the distortion contrast is less than the threshold T_m, assign $VSNR = \infty$ and the assessment process terminates. If, for a particular f_m the distortion contrast $C(E_m)$ exceeds the threshold T_m, Stage II is processed for further assessment.
  - Stage II: Suprathreshold Distortion
  The assessment of suprathreshold distortion is based on Global Precedence, a mid-level HVS property (see Section II). The principle of Global Precedence is, the HVS processes the image in a coarse-to-fine-grained manner: from the global structuring to the local details [85]. It is found in [86] that "structural distortion" that affects the global precedence is most perceptible; while additive white noise, which is uncorrelated with the image, is least perceptible. The global precedence-based VSNR is computed as

$$VSNR = 20 \log_{10} \frac{C(S)}{\alpha C(E) + (1-\alpha) GP / \sqrt{2}} \tag{16}$$

  in which $\alpha \in [0, 1]$ is to adjust the relative importance; $C(S)$ and $C(E)$ are the sum of $C(S_m)$ and $C(E_m)$, respectively; GP is the global precedence disruption given as follows:

$$GP = \sqrt{\sum_m \left[C^*(E_m) - C(E_m)\right]^2} \tag{17}$$

  in which $C^*(E_m)$ is the global-precedence preserving contrast.

- MOSp
MOSp is a simple and easy-to-compute metric, which is based on the linear relationship between MSE and subjective results.

$$MOSp = 1 - k \cdot MSE \tag{18}$$

in which k is the slope of the linear regression and the key element of MOSp. Due to the masking effect, distortions in highly detailed regions are less visible than those in low detailed regions. Therefore, k is calculated as follows:

$$k = 0.03585 \exp(-0.02439 * EdgeStrength) \tag{19}$$

in which $EdgeStrength$ is used to quantify the detail within a region.
- Attention-Driven Foveated Video Quality Metric (AFViQ)
AFViQ models the contrast sensitivity of the HVS based on the mechanisms of vision foveation and visual attention. The vision foveation refers to the fact that the HVS perceives different amount of detail, or resolution, across the area of view, with highest resolution at the point of fixation. The point of fixation is projected onto the center of the eye's retina, i.e., the fovea [99]. Different from existing quality metrics based on static foveated vision [99]–[101], AFViQ simulates the dynamic foveation by

TABLE VI
ENGINEERING APPROACH MODELS

CC: Correlation Coefficient; RMS: Root Mean Squared Error; OR: Outlier Ratio; SRCC: Spearman Rank-order Correlation Coefficient

	Year	Model Basis	Validation Database	Performance factors			
				CC	RMS	OR	SRCC
SSIM [33], [43], [76] [87]–[91]	2002	Structural similarity	VQEG Phase I FR-TV	0.821	N/A	0.644	0.833
LVQM [92]–[94]	2004	Blocking, content richness, masking	90 test video sequences	0.897	N/A	N/A	0.902
VIF [95] [96]	2005	Natural Scene Statistics	Subjective tests	0.950	5.025	0.013	0.950
Video VIF [97]	2005	Natural Scene Statistics and Video Motion	VQEG Phase I FR-TV	0.891	N/A	N/A	0.865
KVQM [98]	2008	Edginess, blockiness and blur	140 video sequences with H.263 and H.264/AVC	N/A	11.5	N/A	N/A
MOVIE [44]	2010	Spatial & temporal & spatio-temporal distortion	VQEG Phase I FR-TV	0.821	N/A	0.644	0.833

predicting video fixation based on eye movement. Given the traditional critical frequency f_c (beyond which the contrast change is imperceptible by the HVS) given in existing work [102], the adjusted critical frequency f_c' for a moving object is:

$$f_c' = f_c \frac{v_c}{|\cos\theta \cdot v_r| + v_c} \tag{20}$$

in which $v_c = 2$ deg/sec is the corner velocity, v_r is the difference between the velocity of the moving object and the eye movement, and θ is the retinal velocity direction. Moreover, the HVS has different attention towards different objects. The critical frequency of the different parts of the video can be adjusted by the attention map [103].

$$f_c'' = f_c' [\rho + (1 - \rho) AM] \tag{21}$$

in which AM is the attention map, $\rho \in [0, 1]$ is a control parameter. Then the contrast sensitivity for a given spatial frequency sf is:

$$CS(sf) = \begin{cases} f_c'', & f \le \hat{f} \\ 0, & f > \hat{f} \end{cases} \tag{22}$$

in which $\hat{f} = \min(f_c'', r/2)$, r is the effective display visual resolution [104].

The predicted perceived quality at the frame level is:

$$Q_{frame} = SD \cdot TD \tag{23}$$

in which SD is spatial distortion index and TD is temporal distortion index. Both SD and TD are a function of $CS(sf)$. Then the video sequence is partitioned into segments based on saccade duration, since the HVS has no visual detectability during the saccadic eye movement. The quality metric for $Q_{segment}$ is derived by a short-term spatial-temporal pooling. Finally, the overall quality metric for the entire video Q_{video} is derived by a long-term spatial-temporal pooling.

3) Engineering Approach: In this section, we first introduce engineering approach models which are based on modeling one or more video artefacts such as blockiness, edginess and blur; then we present a well-known NSS-based model; finally, we focus on an important branch of engineering approach models based on structural similarity. Table VI gives a summary of the engineering approach models.

Video Artefacts based Models:

- Low-Bitrate Video Quality Model (LVQM)
  Noting that pixel-wise error measurements (e.g., MSE, PSNR), used for TV types of video, are unsuitable for videos encoded at a low bitrate, LVQM is proposed and evaluated on QCIF and CIF videos encoded by MPEG-4 with bitrates ranging from 24 kbps to 384 kbps and frame rates ranging from 7.5 Hz to 30 Hz. LVQM incorporates three aspects:

  – Distortion-invisibility D. Subject to luminance masking, spatial-textural masking, and temporal masking, distortions below the detection threshold are deemed invisible. Distortions greater than the detection threshold are incorporated into D.
  – Block fidelity BF. At low bitrate, lossy block-based video compression will introduce distortions at block boundaries. Block fidelity computes the difference between the distorted video and the original video at block-boundaries.
  – Content richness fidelity RF. The HVS favors lively and colorful images. RF compares the content richness of the distorted video and the original video in terms of luminance occurrences.

  The final quality rating is:

$$LVQM = \frac{\sum_t D(t) \cdot BF(t) \cdot RF(t)}{N_t} \tag{24}$$

- KVQM
  KVQM metric is the linear combination of three factors:

  – F_{edge} quantifies the edge features, with the help of an edge detection algorithm and an edge boundary detection algorithm.
  – F_{block} quantifies the distortions at the block boundary, with the help of a block boundary detection algorithm.
  – F_{blur} quantifies blur distortion of the image, by calculating the differences of the average gradients of the distorted and original images.

  The flow of computing the KVQM is shown in Fig. 16. The edge detection algorithm extracts edge pixels, and the edge detection algorithm extracts pixels adjacent to the edge pixels, both from the original image, since the edges

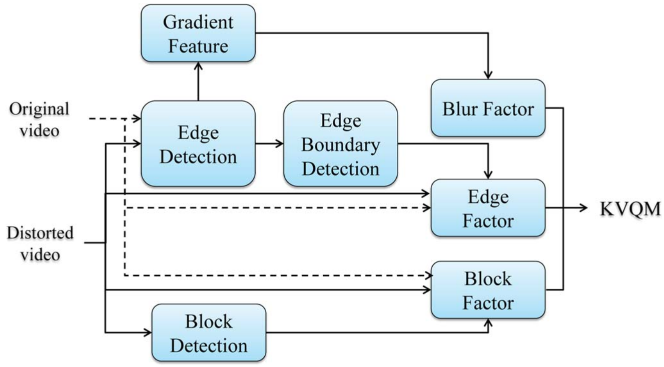


Fig. 16. Flow of KVQM.

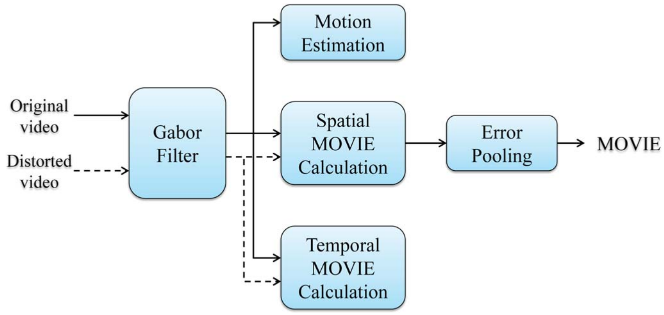


Fig. 17. Flow of MOVIE.

of the distorted image may suffer from blur or other degradation. The block boundary detection algorithm detects blockiness at block boundaries in the distorted image. The gradient feature is the difference between the average gradients of the original image and the distorted image. It quantifies the blur factor. Then KVQM is calculated as the weighted sum of the three factors.

$$KVQM = w_1 F_{edge} + w_2 F_{block} + w_3 F_{blur} + offset \quad (25)$$

in which w_1, w_2, and w_3 are the weights for each factor; $offset$ is the residual of the regression.

- MOtion-Based Video Integrity Evaluation (MOVIE)
  The MOVIE index assesses the video distortions not only separately in space domain and time domain, but also in space-time domain, characterizing the motion quality along the motion trajectories. Fig. 17 shows how to calculate the MOVIE index. The original video and distorted video signals first go through the Gabor filter to model the linear filtering function of the HVS. Let $i = (x, y, t)$ denote a spatio-temporal location; $R(i, k)$ denote the Gabor filtered original video signal, and $D(i, k)$ denote the Gabor filtered distorted video signal, in which $k = 1, 2, \ldots, K$ is the index of Gabor filters. The decomposed signals are then used to estimate motion and compute spatial and temporal MOVIE indexes.

  – Spatial MOVIE Index
  Local spatial movie index is computed for a reference location i_0, with N sample signals within a window centered at i_0:

$$Q_S(i_0) = 1 - \frac{PE_S(i_0)/K + E_{DC}(i_0)}{P+1} \quad (26)$$

  in which E_S is the error index of the Gabor sub-band and E_{DC} is the error index of the Gaussian sub-band. P is the scale of Gabor filters, K is the number of Gabor filters. $E_S(i_0, k)$ is calculated as

$$E_S(i_0) = \frac{1}{2N} \sum_k \sum_n \left[\frac{R(i_n, k) - D(i_n, k)}{C_1 + E(i_0, k)} \right]^2 \quad (27)$$

  in which $E(i_0, k)$ measures the local energy. $E_{DC}(i_0)$ is calculated in a similar manner.

  – Motion estimation
  Motion information is extracted from the original video based on the Fleet and Jepson algorithm [105]; and is used for the temporal MOVIE calculation.

  – Temporal MOVIE Index
  The idea of temporal MOVIE index is to compute a weighted sum of the Gabor filtered signals: if the distorted video has the same motion (speed and direction) as the original video, the weight is strongly positive, vice versa.

$$Q_T = 1 - \frac{1}{N} \sum_n \left(v_n^r - v_n^d\right)^2 \quad (28)$$

  in which v_n^r is the response of the original video to a mechanism that is tuned to its own motion, and v_n^d is the response of the distorted video to a mechanism that is tuned to the motion of the original video.

  – Error pooling
  Frame level spatial and temporal MOVIE is

$$F_S = \frac{\delta_{Q_s}}{\mu_{Q_s}}, \; F_T = \frac{\delta_{Q_T}}{\mu_{Q_T}} \quad (29)$$

  in which δ is the standard deviation and μ is the mean.

  The final MOVIE index is

$$MOVIE = \frac{1}{M} \sum_m F_S(t_m) \cdot \sqrt{\frac{1}{M} \sum_m F_T(t_m)} \quad (30)$$

  in which M is the number of frames.

NSS Based Models: Image and video are natural scenes, of which the statistical information is different from random signals. However, the compression artefacts will result in unnaturalness. Natural Scene Statistics models [107], [108], combined with distortion models, can better quantify the statistical information difference between the original and the distorted videos. Here, we introduce VIF, a widely cited NSS-based model.

- Video Visual Information Fidelity (VIF)
  Video VIF evaluates visual fidelity by comparing the information that can be extracted by the brain from the original video and the distorted video [97], as shown in Fig. 18. In the upper path in Fig. 18, the original video

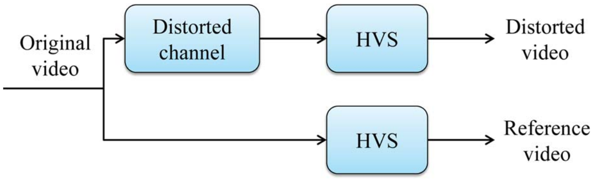


Fig. 18. Flow of video VIF.

first passes through the distortion channel, then passes through the HVS, resulting in the distorted video. In the lower path in Fig. 18, the original video directly passes through the HVS, resulting in the reference video. The quality of the video can be represented by the amount of information that the brain can extract from the video. Let $\mathcal{S}$ represent the original video, $\mathcal{D}$ represent the distorted video, $\mathcal{R}$ represent the reference video.

$$\begin{aligned} \mathcal{R} &= \mathcal{S} + \mathcal{N} \\ \mathcal{D} &= a\mathcal{S} + \mathcal{B} + \mathcal{N}' \end{aligned} \tag{31}$$

in which $\mathcal{N}$ and $\mathcal{N}'$ are the visual noises from the HVS channel, which can be approximated as additive white Gaussian noise. The response of the distortion channel is $a\mathcal{S} + \mathcal{B}$, in which $a = \{a_i, i \in I\}$ is a deterministic scalar gain (I represents all the spatiotemporal blocks), $\mathcal{B}$ is a stationary additive zero-mean Gaussian noise. This simple model is proved to be effective in modeling the noise (by $\mathcal{B}$) and blur (by a) effects in the distortion channel.

For one channel, the information that can be extracted from the reference and distorted video is as follows:

$$\begin{aligned} I_{\mathcal{R}} &= \frac{1}{2}\sum_{i \in I} \log_2\left(1 + \frac{s_i^2}{\delta_n^2}\right) \\ I_{\mathcal{D}} &= \frac{1}{2}\sum_{i \in I} \log_2\left(1 + \frac{a_i^2 s_i^2}{\delta_b^2 + \delta_n^2}\right) \end{aligned} \tag{32}$$

in which a_i is the distortion gain of the ith spatiotemporal block, s_i is the ith original spatiotemporal block, δ_b and δ_n are the variances of the distortion noise $\mathcal{B}$ and HVS noise $\mathcal{N}$ respectively.

The video VIF is defined as the information that can be extracted from the distorted video and that from the reference video of all channels.

$$VIF = \frac{\sum\limits_{\text{all channels}} I_{\mathcal{D}}}{\sum\limits_{\text{all channels}} I_{\mathcal{R}}} \tag{33}$$

Structural Similarity Based Models: The objective of the structural similarity based models is to measure the similarity (fidelity) between the original video and the distorted video, based on the knowledge of the transmitter, channel and the receiver [109]. Table VII shows the examples of widely-used structural similarity based models.

- Structural SIMilarity (SSIM)
  SSIM is first proposed in [76], then developed in [33], on the basis that HVS is highly developed to capture the "structure" of the image. Therefore, SSIM measures the "difference of structure" between the original image and the distorted image, by taking into consideration the following three factors: luminance, contrast and structure. The luminance and contrast are mostly affected by the illumination of the environment, while the structure is the intrinsic feature of the object. Let $x = \{x_i, i \in I\}$ and $y = \{y_i, i \in I\}$ denote the original and the distorted signals. I is the set of spatiotemporal blocks.

  – *Luminance* is represented by the mean of the signal. $\mu_x = \sum_i x_i, \mu_y = \sum_i y_i$. The luminance index is

  $$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}. \tag{34}$$

  in which C_1 is included to avoid near-zero denominator.

  – *Contrast* is represented by the standard deviation of the signal. $\delta_x = \sqrt{(x_i - \mu_x)^2/(I-1)}, \delta_y = \sqrt{(y_i - \mu_y)^2/(I-1)}$. Therefore, the contrast index is

  $$c(x, y) = \frac{2\delta_x\delta_y + C_2}{\delta_x^2 + \delta_y^2 + C_2} \tag{35}$$

  in which C_2 is included to avoid near-zero denominator.

  – *Structure*. The index to quantify the structural similarity is

  $$s(x, y) = \frac{\delta_{xy} + C_3}{\delta_x\delta_y + C_3} \tag{36}$$

  in which $\delta_{xy} = \sum_i (x_i - \mu_x)(y_i - \mu_y)/(I-1)$, C_3 is included to avoid near-zero denominator.

  In [76], when the SSIM was first proposed, the parameters C_1, C_2, and C_3 are excluded. But very soon they were added, because if $C_1 = C_2 = C_3 = 0$, the results become unstable when $\mu_x^2 + \mu_y^2$ or $\delta_x^2 + \delta_y^2$ are close to zero. The SSIM index is then calculated as

  $$SSIM(x, y) = [l(x, y)]^\alpha \, [c(x, y)]^\beta \, [s(x, y)]^\gamma \tag{37}$$

  in which α, β, and γ are constant parameters. The SSIM index has the following ideal properties:

  – Symmetric: $SSIM(x, y) = SSIM(y, x)$.
  – Bounded: $SSIM(x, y) \leq 1$.
  – Unique Maximal: $SSIM(x, y)$ is the maximum only when $x = y$.

  SSIM is calculated locally as in (37) for an 8×8 square window, which moves pixel-by-pixel to cover the whole image, resulting in an SSIM map. To avoid "blocking", the calculation of mean and standard deviation is weighted

TABLE VII
STRUCTURAL SIMILARITY BASED MODELS

CC: Correlation Coefficient; RMS: Root Mean Squared Error; OR: Outlier Ratio; SRCC: Spearman Rank-order Correlation Coefficient

	Year	Model Basis	Validation Database	Performance factors			
				CC	RMS	OR	SRCC
SSIM [76] [33]	2002	Luminance, Contrast, Structure	VQEG Phase I FR-TV	0.967	5.06	0.041	0.963
Multi-scale SSIM [43]	2003	Luminance, Contrast, Structure, Image details at different resolutions (Multi-scale)	LIVE JPEG/JPEG2000	0.969	4.91	1.16	0.966
Video SSIM [87]	2004	SSIM at the local region level, the frame level, and the sequence level	VQEG Phase I FR-TV	0.849	N/A	0.578	0.812
Spatial Weighted SSIM [89], [90]	2004	Minkowski SSIM , local quality- -weighted SSIM, information content-weighted SSIM	LIVE JPEG/JPEG2000	N/A	N/A	N/A	+0.0130[2]
Wang *et al.* [88]	2005	Structural and non-structural distortions (Luminance, Contrast, Gamma Distortion, Horizontal and Vertical translation)	N/A	N/A	N/A	N/A	N/A
Speed Weighted SSIM [106]	2007	Luminance, Contrast, Structure, Video Motion	VQEG Phase I FR-TV	N/A	N/A	N/A	0.8621

PF/FP-SSIM [91]	2009	Visual fixation-based weighting SSIM and Quality-based weighting SSIM	LIVE JPEG/JPEG2000	0.9664(SS)
				0.9554(MS)
					5.9383(SS)
					6.8245(MS)
						N/A	0.9402(SS)
							0.9469(MS)

by a circular-symmetric Gaussian weighted function $\mathbf{w} = \{w_1, w_2, \ldots, w_I\}$:

$$\mu_x = \sum_i w_i x_i$$
$$\delta_x = \sqrt{\sum_i w_i (x_i - \mu_x)^2}$$
$$\delta_{xy} = \sum_i w_i (x_i - \mu_x)(y_i - \mu_y) \tag{38}$$

The SSIM index for the whole image is

$$SSIM(X,Y) = \frac{1}{N}\sum_j SSIM(x_j, y_j) \tag{39}$$

in which N is the number of windows, and x_j, y_j are the signals at the jth window.

- Muti-Scale SSIM
Viewer's perceptibility of image details relies on the viewing conditions, such as the sampling density of the image, the distance between the viewer and the image, and the perceptual ability of the viewer's HVS. So, to choose the right scale on which to evaluate the perceptual quality is difficult. The single-scale SSIM is, therefore, extended to multi-scale SSIM [43], summing up the influence of each scale with different weights to account for their relative importance. Assume there are K intended scales. The original and distorted images are repeatedly processed by a low-pass filter, which downsamples the image by a factor of 2. The number of repetition is K. At the jth scale, the contrast index $c_j(x,y)$ and structure index $s_j(x,y)$ are computed; while the luminance index of the last iteration $l_K(x,y)$ is computed. Multi-scale SSIM is then calculated as:

$$SSIM(x,y) = [l_K(x,y)]^{\alpha_K}\, \Pi_{j=1}^{K} [c_j(x,y)]^{\beta_j} [s_j(x,y)]^{\gamma_j} \tag{40}$$

in which $\alpha_j, \beta_j, \gamma_j$ can be adjusted for different importance of each scale. In fact, the challenge of the method lies in determining the value of $\alpha_j, \beta_j, \gamma_j, j \in [1, K]$ and the number of scales K. One way is to refer to the contrast sensitivity function (CSF) of the HVS [110], another way is to calibrate the values via subjective test.

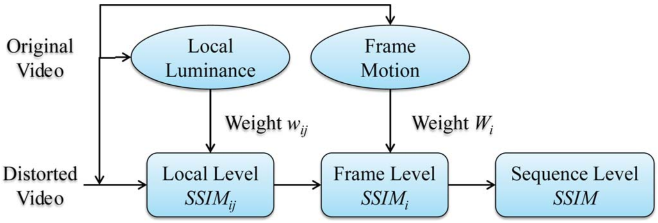


Fig. 19. Flow of video SSIM.

- Video SSIM
The SSIM for image is extended to SSIM for video sequence in [87]. The procedure of calculating the video SSIM is shown in Fig. 19.
  - *Local Level SSIM* is calculated for random sampled 8 × 8 windows in each frame, according to (37). The selection of windows is unlike that in image SSIM calculation, which exhausts all possible windows by moving pixel-by-pixel over the entire image. In video SSIM calculation, the number of sampled windows for each frame should consider both computational complexity and evaluation accuracy. Local SSIM for Y, Cb and Cr color components are calculated and then combined as (the jth window of the ith frame):

$$SSIM_{ij} = W^Y SSIM_{ij}^Y + W^{Cb} SSIM_{ij}^{Cb} + W^{Cr} SSIM_{ij}^{Cr} \tag{41}$$

  in which W^Y, W^{Cb}, and W^{Cr} are weights for Y, Cb and Cr color components.
  - *Frame Level SSIM* is calculated as the weighted sum of the local level SSIM. The weight given to each local level SSIM is based on its luminance. High weights are given to high-luminance regions as they are more likely to attract fixation. Frame level SSIM for the ith frame is:

$$SSIM_i = \frac{\sum_j w_{ij} SSIM_{ij}}{\sum_j w_{ij}} \tag{42}$$

in which the value of w_{ij} is determined as

$$w_{ij} = \begin{cases} 0, & \mu_x \leq 40 \\ (\mu_x - 40)/10, & 40 < \mu_x \leq 50 \\ 1, & \mu_x > 50 \end{cases} \tag{43}$$

in which μ_x is the mean of the Y components.

– *Sequence Level SSIM* is calculated as the weighted sum of the frame level SSIM. The weight given to each frame level SSIM is based on its motion with respect to the next frame. Low weights are given to large-motion frames as the experiments show that SSIM performs less stable with large-motion frames. A motion-related parameter M_i is defined as $M_i = \sum_j m_{ij}/(16N_i)$, in which m_{ij} is the motion vector of the jth window and N_i is the number of sampled windows in the ith frame. Sequence level SSIM is:

$$SSIM = \frac{\sum_i W_i SSIM_i}{\sum_i w_i} \tag{44}$$

in which the value of W_i is determined as

$$W_i = \begin{cases} \sum_j w_{ij}, & M_i \leq 0.8 \\ (3 - 2.5M_i)\sum_j w_{ij}, & 0.8 < M_i \leq 1.2 \\ 0, & M_i > 1.2 \end{cases} \tag{45}$$

- Spatial Weighted SSIM
  In stead of giving equal weight to local level SSIM in (39), three spatial weighting methods are proposed in [90].

  – *Minkowski weighting* gives high weights to windows with large distortions since the HVS is more sensitive towards poor quality. The Minkowski weighted SSIM is:

$$SSIM^{Minkowski} = \frac{1}{N}\sum_j SSIM_j^p \tag{46}$$

  in which p is the Minkowski power.

  – *Local quality weighting* also gives high weights to the windows with large distortions or poor qualities, but through a function of the local quality index, which is more flexible than the Minkowski weighting. The local quality weighted SSIM is:

$$SSIM^{Quality} = \frac{\sum_j f(SSIM_j)SSIM_j}{\sum_j f(SSIM_j)} \tag{47}$$

  in which $f(\cdot)$ is a (monotonic) function based on the local $SSIM_j$.

  – *Information content weighting* also gives high weights to the windows with large distortions or poor qualities, but through a function of the local quality index. The information content weighted SSIM is:

$$SSIM^{Information} = \frac{\sum_j g(x_j, y_j)SSIM_j}{\sum_j g(x_j, y_j)} \tag{48}$$

  in which $g(x_j, y_j)$ is a function of the signal of the original image x_j and the signal of the distorted image y_j. In [89], the weighting function $g(x_j, y_j)$ characterizes the local energy

$$g(\mathbf{x}, \mathbf{y}) = \delta_x^2 + \delta_y^2 + C \tag{49}$$

  in which C is included to account for near-zero $\delta_x^2 + \delta_y^2$. In [90], the weighting function $g(x_j, y_j)$ is defined based on the received information

$$g(\mathbf{x}, \mathbf{y}) = \log\left[\left(1 + \frac{\delta_x^2}{C}\right)\left(1 + \frac{\delta_y^2}{C}\right)\right] \tag{50}$$

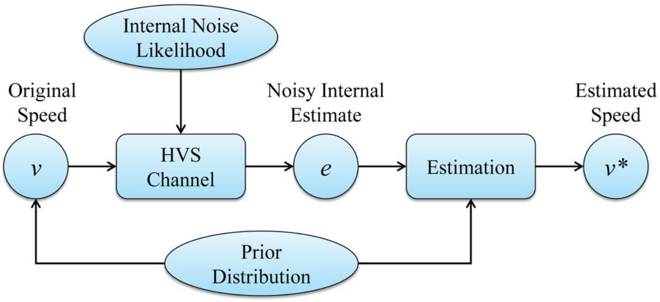


Fig. 20. Bayesian human visual speed perception model.

- Speed Weighted SSIM
  Different from a set of still images, the video sequence contains motion information, which is used to adjust the SSIM in [106]. The basis of the speed weighting adjustment is the Bayesian human visual speed perception model [111], as shown in Fig. 20. The original video passes through the noisy HVS channel, to get the noisy internal estimation of the motion, which is then combined with prior probability distribution of the speed of motion, to get the final estimated speed. Two kinds of speed are considered: v_g, the background speed, and v_m, the absolute speed subtracting v_g. v_m can be viewed as the motion of the moving object. The perception of the speed includes the following two aspects:

  – *Information Content*. High-speed motion acts as a surprisal for the human vision, and is likely to attract more attention. The prior probability distribution of v_m is assumed to be τ/v_m^{α} (τ and α are two positive constants). The information content is computed as the self-information of v_m.

$$I = \alpha \log_e v_m - \log_e \tau \tag{51}$$

  The information content increases with the speed of the object, which is reasonably true.

  – *Perception Uncertainty*. The perception uncertainty is determined by the noise in the HVS channel. As shown in Fig. 20, given the true speed (approximated by v_g), the likelihood of the internal noise e follows a log-normal distribution. The perception uncertainty is computed as the entropy of this likelihood function.

$$U = \log_e v_g + \beta \tag{52}$$

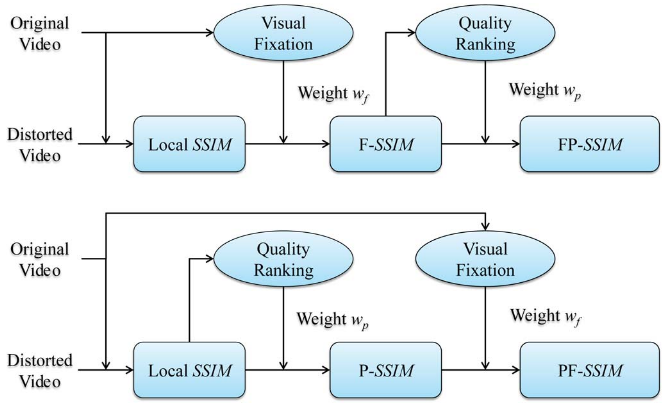


Fig. 21. Flow of PF/FP-SSIM.

in which β is a constant. The perception uncertainty increases with the background speed, meaning that the HVS channel cannot accurately process the video information, if the background motion is too fast. β decreases in video contrasts, meaning that high-contrast videos yield less uncertainty through the HVS channel.

Information content contributes to the importance of a visual stimulus, while the perception uncertainty reduces its importance. Hence, the speed-related weight is represented as $w = I - U$, and speed weighted SSIM is calculated as

$$SSIM^{speed} = \frac{\sum_x \sum_y \sum_t w(x,y,t) SSIM(x,y,t)}{\sum_x \sum_y \sum_t w(x,y,t)} \qquad (53)$$

in which $SSIM(x,y,t)$ is the SSIM index of the spatiotemporal region (x,y,t).

- PF/FP-SSIM
  PF/FP-SSIM is a combination of visual fixation weighted SSIM (P-SSIM) and quality weighted SSIM (F-SSIM) [91], as shown in Fig. 21. The weight for a local SSIM is determined by its visual importance.

  - *Visual fixation weighted SSIM (F-SSIM)*. The areas which attract most human attention and the eyes fix upon, are more important. For each image, ten fixation points are chosen according to the Gaze-Attentive Fixation Finding Engine (GAFFE) algorithm [112], then the fixation areas are determined by a 2-D Gaussian function. The pixels within the fixation areas are given weight $w_f > 1$, while other pixels are given weight $w_f = 1$. The F-SSIM of the jth window is obtained by:

$$F - SSIM_j = \frac{\sum_{x \in J} \sum_{y \in J} SSIM(x,y) w_f(x,y)}{\sum_{x \in J} \sum_{y \in J} w_f(x,y)} \qquad (54)$$

    For multi-scale SSIM, the number of fixation points and the size of fixation areas reduce with the scale level.
  - *Quality weighted SSIM (P-SSIM)*. The areas with "poor" quality are easier to capture attention than areas with "good" quality. Therefore, the "poor" quality areas hurt the perceptual quality more than the "good" quality areas improve the perceptual quality. Rank the quality of all windows according to their quality in ascending orders; then assign weight $w_p > 1$ to the lowest $p\%$ items, and assign weight $w_p = 1$ to others. In [91], $p = 6$ yields good results. For multi-scale SSIM, only the second scale image is given the weight w_p.
  - *PF/FP-SSIM*. The PF-SSIM is obtained by first applying the quality weighting to get P-SSIM, then visual fixation weighting to get FP-SSIM. The FP-SSIM is obtained by first applying the visual fixation weighting to get F-SSIM, then quality weighting to get PF-SSIM. F-SSIM and P-SSIM can also be computed separately.

  Unfortunately, the experiments show that only the P-SSIM gives significant improvements over the non-weighted SSIM [91].

B. Reduced Reference Model

We mainly introduce two kinds of reduced reference models: one is based on packet loss visibility (PLV), the other is based on natural scene statistics.

1) Packet Loss Visibility Based Model: Packet loss visibility based models indirectly measure the loss of video quality by measuring the visibility of the packet loss. The major problem is to classify what kind of packet loss is visible and what kind of packet loss is invisible. Therefore, different classification techniques and different packet types have been explored to improve the classification accuracy. Table VIII gives a summary of the packet loss visibility based RR models. Packet loss visibility based models usually process as follows. Firstly, subjective tests are conducted, in which assessors are asked whether they see artifacts in the displayed video. Then, classification algorithms (known as classifier) are applied to classify packet loss into visible or invisible classes, or the regression models are applied to predict the probability of packet loss visibility, using the subjective test results as the ground truth, and objective quality metrics as features.

In [113] and [114], the location of the packet loss and the content of the video are considered as the major factors that influence the visibility of the packet loss. The following objective quality metrics are specified to characterize the location of the packet loss and the content of the video:

- Content-independent factors
  - Temporal duration, that is, the number of frames affected by the packet loss. If the packet loss occurs in a B-frame, the influence will last only a single frame, however, if the packet loss occurs in an I-frame, the influence will last until the next I-frame.
  - Initial spatial extent, that is, the number of slices lost. Due to a single packet loss, the decoder may have to abandon one slice, double slices or the entire frame.
  - Vertical position, that is, the index of the topmost slice affected by the packet loss. In a frame, from the top to the bottom, slices are indexed from 0 to 29.

TABLE VIII
PACKET LOSS VISIBILITY BASED MODELS

	Year	Codec	Objective quality metrics	Analysis tool
[113]	2004	MPEG-2	Content-independent and content-dependent factors	CART
[114]	2004	MPEG-2	Same as [113]	GLM
[115]	2006	MPEG-2	More factors than [113]	GLM and CART
[116]	2006	H.264	[113] + Multiple packet loss factors	GLM
[117]	2007	MPEG-2 and H.264	[113] + SSIM-based factors	GLM
[118]	2007	MPEG-2 and H.264	[117]+camera motion and proximity to a scene change	Patient Rule Induction Method (PRIM)
[119]	2010	MPEG-2 and H.264	[118] + Group-of-picture structures	GLM

TABLE IX
NSS BASED RR MODELS

CC: Correlation Coefficient; RMS: Root Mean Squared Error; OR: Outlier Ratio; SRCC: Spearman Rank-order Correlation Coefficient

	Year	Model Basis	Validation Database	Performance factors			
				CC	RMS	OR	SRCC
WNISM [122] [123]	2005	KLD, wavelet-domain	LIVE database	0.8226	N/A	0.2311	0.8437
RRIQA [124]	2009	KLD, DNT-domain	LIVE database, Cornell-A57 [46]	0.9173		0.1069	0.9287
βW-SCM [125]	2010	Steerable pyramid decomposition Weibull distribution	LIVE database	0.8353			0.8391
RRED [126]	2011	Scaled entropy, wavelet-domain	LIVE database	N/A	N/A	N/A	0.9580
GSM model [127]	2012	KLD, Tetrolet-domain	Cornell-A57	0.70			0.74
RR-SSIM [128]	2012	SSIM, KLD, DNT-domain	LIVE database, Cornell-A57, IVC database [129], Toyama-MICT database [130], TID2008 [131], CSIQ database [132]	0.9194			0.9129

The location of the affected slice is considered since different regions of the picture capture different degrees of viewers' attention.

Content-independent factors do not rely on video content, and can be extracted from the distorted videos.

- Content-dependent factors:
  - Variance of motion and residual energy. These factors characterize the motion information of the video, which may mask the error and influence the visibility of the packet loss.
  - Initial Mean Square Error, is the mean square error per pixel between the decoded videos with and without packet loss, only considering the pixels in lost slices.

Content-dependent factors can be estimated with the help of reduced information of the original videos from the encoder.

In [113], tree-structured data analysis based on Classification And Regression Tree (CART) [120], is used to classify the visibility of the packet loss. However, using tree-structured data analysis is hard to distinguish the packet loss visibility near the threshold and far from the threshold. Therefore, in [114], a Generalized Linear Model (GLM) [121] is used to predict the probability that the packet loss is visible to the viewer. Also, in [114], two NR models are developed, in which the content-dependent factors are estimated from the distorted video. In [115], both CART and GLM are adopted and their performances are compared. More objective quality metrics are considered in [115], including: type of the frame in which packet loss occurs, the magnitude and the angle of the motion. [116] extends [115] in two ways: H.264 is considered in stead of MPEG-2; multiple packet loss is considered in stead of isolated packet loss. Multiple packet loss is considered because packet loss is usually bursty, and multiple packet loss may correlate with each other. More specifically, in [116], dual packet loss is considered, characterized by spatial and temporal separation of the two packet losses. In [117], SSIM is adapted for RR and NR models to predict the visibility of packet loss (SSIM is originally an FR model). In [118], scene-level factors, specifically camera motion and proximity of a scene cut, are considered, and the Patient Rule Induction Method (PRIM) is used to decide visibility of a packet loss. It is found that global camera motion will increase the packet loss visibility compared with a still camera, and packet loss near the scene cut is less invisible. In [119], different Group-of-Picture (GoP) structures (e.g., IBBP) are considered for prediction, and the model is applied to packet prioritization for the router to decide which packets to drop when the network is congested.

One of the problems of the PLV based models is that quality degradation is simply classified as visible or invisible, without further quantification of how severe the quality degradation is. PLV based models may be used for preliminary quality evaluation.

2) NSS Based Model: The NSS based models assume that the real-world image and video are natural scenes, whose statistical features will be disrupted by distortions. The comparison of the statistics of the original image and the distorted image can be used to quantify the quality degradation. Survey paper [133] offers a nice introduction of NSS based RR and NR models. Table IX gives a summary of the NSS based RR models. In this section, we introduce WNISM, recognized as the standard NSS based RR model proposed by [122].

Let $p(x)$ and $q(x)$ be the probability density functions of the wavelet coefficients in the same subband of the original image and distorted image respectively. According to the law of large numbers, the difference of log-likelihood between $p(x)$ and

$q(x)$ asymptotically approaches the Kullback-Leibler distance [134] between $p(x)$ and $q(x)$, denoted by $d(p\|q)$.

$$d(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} dx \tag{55}$$

While $q(x)$ can be easily extracted from the distorted image at the receiver, $p(x)$ should be extracted from the original image, and transmitting $p(x)$ as an RR feature is costly. Fortunately, it is found that $p(x)$ can be approximated by a 2-parameter generalized Gaussian density model (GGD) as:

$$p_m(x) = \frac{\beta}{2\alpha\Gamma\left(1/\beta\right)} e^{-(|x|/\alpha)^{\beta}} \tag{56}$$

where $\Gamma(\cdot)$ is the Gamma function. Also, the KLD between $p_m(x)$ and $p(x)$ is computed as

$$d(p_m\|p) = \int p_m(x) \log \frac{p_m(x)}{p(x)} dx \tag{57}$$

For each subband, based on the RR feature $\{\alpha, \beta, d(p_m\|p)\}$, the KLD between $p(x)$ and $q(x)$ can be approximated as

$$\hat{d}(p\|d) = d(p_m\|q) - d(p_m\|p) \tag{58}$$

in which $d(p_m\|q)$ can be calculated at the receiver side as

$$d(p_m\|q) = \int p_m(x) \log \frac{p_m(x)}{q(x)} dx \tag{59}$$

Finally, aggregate the distortions in all subbands and the overall distortion metric can be obtained as:

$$D = log_2 \left(1 + \frac{1}{D_0} \sum_{k=1}^{K} \left| \hat{d}^k(p^k\|q^k) \right| \right) \tag{60}$$

in which D_0 is a constant parameter; p^k and q^k are the probability density functions of the kth subband of the original image and distorted image respectively; and $\hat{d}^k$ is the KLD estimation between p^k and q^k.

In [123], the authors introduced the concept of quality-aware image, in which the RR information is encoded as invisible hidden messages. And after decoding, these hidden messages can help compute the quality metric. In [124], it is noted that linear image decomposition, such as wavelet transformation, cannot reduce statistical dependence between neuronal responses. Therefore, divisive normalization transform (DNT), a nonlinear decomposition, is leveraged as the image representation. Instead of using KLD, in [126], the quality metric is computed as the average difference between scaled entropies of wavelet coefficients of original image and distorted image. In [127], Tetrolet transform for both original image and distorted image is used to better characterize local geometric structures. Subbands are modeled by Gaussian Scale Mixture (GSM) to account for the statistical dependencies between tetrolet coefficients. In [125], coefficients with maximum amplitude, instead of all coefficients, are used to get the RR metric by fitting them with a Weibull distribution. In [128], an SSIM-like metric largely based on [124], [133] and structural similarity is developed.

Apart from the above mentioned PLV based and NSS based RR models, there are some other models that are worthy of noting. In [135], the blockiness and blurriness features are detected by harmonic amplitude analysis, and local harmonic strength values constitute the RR information for quality estimation. In [136], [137], the RR models are based on the HVS characteristics, more specifically, the contrast sensitivity function (CSP). The images are decomposed by contourlet transform in [136], and grouplet transform in [137]. The quality criterion C4 in [138] first models the HVS characteristics in respect of color perception, CSF, psychophysical subband decomposition and masking effect modeling; then extracts the structural similarity between the original image and distorted image to get the final RR metric.

C. No Reference Model

No reference model can meet the demand of real-time QoE monitor. However, it is hard to develop since there is no access to the original video. Therefore, much effort has been put on mapping the network statistics (e.g., packet loss rate, bandwidth), which can be obtained from simple measurement, and application-specific factors (e.g., encoding bitrate, packetization scheme), to the quality estimation. In this section, we introduce NR models according to Fig. 12. Note that the classification mostly depends on the major techniques or theory basis of the model, and may not be exclusive. In particular, the PLV based and NSS based NR models are the extensions of their RR counterparts; and the bitstream-layer, packet-layer, and hybrid models are based on the access of information of streamed videos. In this section, we will focus on the bitstream-layer, packet-layer, and hybrid models for streamed videos, since the PLV and NSS based models have already been explained in the previous session. Table X gives a summary of the NR models.

1) Bitstream-Layer Models: A survey of bitstream-based models is given in [139]. Now we introduce several typical bitstream-layer models.

- QANV-PA
  Apart from coding factors and video motion information, QANV-PA further consider the temporal information and the influence of packet loss.

  – Frame quality.
  QP parameter and spatial and temporal complexity of the nth frame are included in the frame quality:

$$Q_n = f(q_n) + (b_3 - f(q_n)) \left(\left(\frac{\delta_{S,n}}{a_1} \right)^{b_1} + \left(\frac{\delta_{T,n}}{a_2} \right)^{b_2} \right) \tag{61}$$

  in which $f(q_n)$ is a linear function of the QP parameter q_n, $\delta_{S,n}$, and $\delta_{T,n}$ are the spatial and temporal complexity, respectively, and a_1, a_2, b_1, b_2, b_3 are constant parameters.

  – Packet loss influence
  The degradation due to the packet loss is characterized by parameter p_n, which depends on the number of frames that are affected by the packet loss, and the

TABLE X
NO REFERENCE MODELS

CC: Correlation Coefficient; RMS: Root Mean Squared Error; OR: Outlier Ratio; SRCC: Spearman Rank-order Correlation Coefficient

	Year	Model Basis	Validation Database	Performance factors			
				CC	RMS	OR	SRCC
NSS based models							
BLIINDS [14]	2012	Natural scene statistics, DCT-domain	LIVE database				0.821
BLIINDS-II [15]	2012	Natural scene statistics, DCT-domain	LIVE database, TID2008	0.9232			0.9202
Bitstream-layer models							
QANV-PA [16]	2010	Quantization, packet loss and error propagation, temporal features of HVS	Standard test sequences	0.912	0.014	0.021	0.913
C-VQA [17]	2012	Quantization, motion, bit allocation	LIVE database	0.7927			0.7720
NR-blocky [18]	2012	Blockiness of a base layer, coding parameters of its enhancement layers	4 sequences	0.8719			
Packet-layer models							
V_q [19]	2008	Encoding betrate, packet loss rate	8 video sequences	0.968	0.287		
CAPL [20]	2012	Number of bits, frame type, temporal complexity, positions of lost packets	Standard test sequences	0.941	0.358	0.000	0.935
Hybrid models							
rPSNR [21]	2008	Codec, loss recovery technique, encoding bitrate, packetization	Video clips				
SVC quality model [22]	2008	Spatial and temporal information, network bandwidth determined SNR	10 video sequences				
Motion-based model [23]	2008	Bitrate, framework, content motion information	2 sets of video sequences	0.8190			
V_q [24]	2009	Encoding betrate, packet loss rate, spatial and temporal features	8 video sequences	0.96	0.37	0.36	
APM [25], [26]	2011	Startup delay, rebuffering, user-viewing activity	Flash videos				
UMTS quality model [27]	2012	Content type, sender bitrate, block error rate, mean burst length	LIVE database	0.912	0.014	0.021	0.913

temporal complexity $\delta_{T,n}$. Then, the quality metric becomes:

$$Q'_n = Q_n - p_n \tag{62}$$

- Temporal pooling
  The quality factors of the frames are integrated by temporal pooling.

$$QANV - PA = \frac{\sum_{n \in D} \left(Q_n^{F''} T_n\right)}{\sum_{n \in D} T_n} \tag{63}$$

in which D is the set of successfully decoded frames, T_n is the duration of the nth frame, and Q''_n is the contribution of the quality of the nth frame to the entire video.

$$Q''_n = Q'_n \left(a_4 + b_4 \delta'_{T,n} + c_4 \delta'_{T,n} log(T_n)\right) \tag{64}$$

in which $\delta'_{T,n} = \delta_{T,n} / \max(\delta_T)$ is the normalized temporal complexity.

- C-VQA
  Three factors: quantization parameter factor, motion factor and bit allocation factor, are calculated and then combined to form C-VQA.

  - Quantization parameter (QP) factor.
    The quantization process causes loss of temporal and spatial information. The higher the QP is, the more severe the quality degradation will be. The QP factor is computed as:

$$F_Q = (aC_n + b)^{c\overline{q}} \tag{65}$$

in which a, b, c are constants, $\overline{q}$ is the average QP over n consecutive frames, and C_n is the feature parameter of the n frames, including width, height and so on.

  - Motion factor.
    The motion factor accounts for the global motion consistency and local motion consistency. Global motion consistency M_g is calculated based on the variance of horizontal and vertical motion vector of moving objects (as opposed to stationary background). Local motion consistency M_l is calculated based on the absolute difference of motion factors between successive macro blocks. The motion factor is the combination of the above two motion factors.

$$F_m = M_g + M_l \tag{66}$$

  - Bit allocation factor
    Bitrate control is applied to streamed video because bitstream is restricted by limited bandwidth. The effectiveness of the bitrate control scheme is characterized by factor C_r, and the bit allocation factor is calculated as follows

$$F_B = V_B \times C_r \tag{67}$$

in which V_B is the variance of bit consumption of the macro blocks.

Finally, the C-VQA is a weighted sum of the QP factor, motion factor, and bit allocation factor:

$$C - VQA = \theta(\alpha\overline{F}_Q + \beta\overline{F}_M + \gamma\overline{F}_B + \eta) \tag{68}$$

in which $\overline{F}_Q, \overline{F}_M, \overline{F}_B$ are the average values over N frames.

2) Packet-Layer Models: The packet-layer models use only the information of the packet header for quality estimation, not depending on the information from the payload. For packets where the payload is encrypted, packet-layer models are more applicable.

- V_q
  V_q is a simple packet-layer model, which estimates the quality affected by the packet loss rate. Firstly, the video quality, when there is no packet loss, is estimated.

$$V_q|_{PL=0} = 1 + I_c \tag{69}$$

in which I_c is a function of the bitrate BR.

$$I_c = a_1 - \frac{a_1}{1 + (Br/a_2)^{a_3}} \tag{70}$$

in which a_1, a_2, a_3 are constant parameters.

When the packet loss rate PL is non-zero, the video quality is fitted by an exponential function

$$V_q = 1 + I_c \exp\left(-\frac{PL}{a_4}\right) \tag{71}$$

in which PL is the packet loss rate, a_4 is a constant.

- CARL
  CARL is developed based on the bitstream-layer model QANV-PA. However, due to a lack of payload information, the frame quality Q_n and temporal complexity $\delta_{T,n}$ are computed differently.

$$Q_n = 1 + a_1\left(1 - \left(\frac{R_n}{a_2\delta_{T,n} + b_2}\right)^{-b_1}\right) \tag{72}$$

in which a_1, a_2, b_1, b_2 are constant parameters, R_n is the average number of bit allocation for a frame in a Group of Pictures (GoP).

For packet-layer model, the motion vector, used to compute temporal complexity, is not available. Therefore, the temporal complexity is estimated as follows.

$$\delta_{T,n} = |R_{P,n}/R_{I,n} - a_3| \tag{73}$$

in which $R_{P,n}$ and $R_{I,n}$ are the average bit allocation for the P frame and I frame in a GoP respectively, a_3 is a constant. After calculating Q_n, the packet loss influence and temporal pooling process are similar to those of QANV-PA.

3) Hybrid Models:

- rPSNR
  rPSNR is a light-weight no reference model, focusing on the relationship between packet loss and QoE, while also considering video codec, loss recovery technique, encoding bitrate, packetization, and content characteristics. Video distortion (denoted by D) is measured through Mean Square Error (MSE), which is derived as a function of packet loss as follows:

$$D = P_e f(n) L D_1$$
$$PSNR = 10log_{10}\frac{255^2}{D} \tag{74}$$

in which P_e is the probability of packet loss event in the video streaming; $f(n)$ is the average number of slices affected by a loss event; L is the number of packets used for transmitting one frame; D_1 is total average distortion caused by losing a single slices. $f(n)$ is different for different codec. For example, in MPEG-2, once a packet loss is detected in a frame, the entire frame will be discarded, and replaced by the previously-decoded frame. However, in H.264, more sophisticated error-concealment is used. All slices will be decoded, and the slices affected by packet loss will be recovered using the corresponding slices in the previous slice and the motion information from other slices in the same frame. The estimation of D_1 depends on the error propagation resulting from loss of one slice due to coding dependencies between frames. P_e and $f(n)$ is network-dependent, and can be easily obtained from network statistics. L can be easily determined based on application configuration. However, D_1 is dependent on individual video characteristics and may not be efficiently estimated when considering real-time quality monitoring of a large number of video streams. To tackle this problem, we can compare the quality of the video transmitted over a path with that transmitted over a reference path. A reference path is a transmission path whose QoE is known beforehand. Usually, we can select the path which generates targeted QoE as the reference path, so that we know how much better or worse the actual path performs. Relative PSNR (rPSNR) is the difference between the monitored network path and the reference path.

$$rPSNR = PSNR - PSNR^0 \tag{75}$$

The resulting rPSNR is independent of D_1, and therefore, easy to compute.

- Application Performance Metrics (APM)
  APM characterizes the impact of rebuffering events on the QoE for HTTP video streaming service. Unlike traditional UDP-based video streaming, the HTTP over TCP video streaming does not suffer from frame loss. First, network QoS metrics, such as the round-trip time (RTT), packet loss rate, and bitrate(determined by bandwidth), are used to estimate the three APM metrics: startup delay, rebuffering time and rebuffering frequency. Then, the APM metrics are fed into the prediction model to get the estimated MOS value. Linear regression is performed for the APM values and MOS values obtained from subjective tests to get the QoE prediction model. The regression results show that the rebuffering frequency has the most significant influence on the QoE.

TABLE XI
USER-VIEWING ACTIVITIES

Activities	Viewer experience indication
Pause	Negative; need more buffering time
Resume	Unclear; continue playing after pause
Reload the page	Negative; the quality is bad, and it may help to allow the player to choose another server from the CDN
Switch to a lower video quality	Negative, lower quality video may be more smooth
Switch to a higher video quality	Positive, higher quality video may be supported by current speed
Full screen	Positive
Return to normal size	Negative
Minimize the video	Unclear, maybe the viewer just runs the video in the background
Forward	Unclear, maybe the video is not interesting
Backward	Unclear, maybe replay the buffered part can achieve a smoother playback
Frequent/infrequent mouse movement	Unclear, maybe frequent mouse movement indicate annoying QoS

In [26], the above APM model is refined by incorporating the influence of user-viewing activities and resort to logistic regression. It is observed that video impairment can trigger viewer interactive activities as listed in Table XI. Two major user-activity metrics, number of pause event and number of screen size reducing event, are put into the logistic regression model, along with the three APM metrics. The results show an improved explanatory power of the regression model.

- UMTS Quality Metric
  Video transmission over wireless network, more specifically the Universal Mobile Telecommunication System (UMTS) is considered in [27], taking into account the distortions caused by the transmission network. Subjective tests are first conducted for different combinations of sender bitrate (SBR), block error rate (BLER), mean burst length (MBL) and content type (CT). SBR reflects the distortion from the encoder; both BLER and MBL reflect the distortions from the transmission network; CT is the content type in terms of temporal and spatial features, identified by cluster analysis tool in [51]. Nonlinear regression on the subjective test results yields the following function:

$$MOS = \frac{\alpha + \beta \times \ln(SBR) + CT \times (\gamma + \delta * \ln(SBR))}{1 + (\eta \times (BLER) + \sigma(BLER)^2) \times MBL} \tag{76}$$

  in which α, β, γ, δ, η, and σ are regression parameters.

In Fig. 22, we show the timeline of all the major objective quality models introduced in this section. We can see several trends of the evolution of objective quality models.

- From FR models to NR models. As the need for real-time QoE monitoring and prediction becomes increasingly urgent, more and more NR models are being proposed. At the meantime, the FR models are further developed due to better understanding of HVS and other related areas.
- From image to streamed video. Previously, many models are first designed for image quality assessment, then extended to video quality assessment. The development of video streaming services motivates research on streamed video quality assessment depending on the information extracted from packet header or payload.

D. Performance Validation

The output of the objective quality model should be well correlated with the subjective results, which are regarded as the ground truth for user QoE. The Video Quality Expert Group (VQEG) gives a test plan [13] for validating objective quality models. The relationship between the output from the objective quality model and the results from the subjective test is usually estimated by a nonlinear regression function. It does not matter what form of nonlinear function is used as long as it is monotonic, applicable to a wide range of video content, and has minimum free parameters. Multiple forms of nonlinear functions will be tried to find the best-fitting one. Let VQR denote the output of the objective quality model; MOS_p denote the predicted MOS value by the regression function; MOS_{norm} denote the normalized output of the subjective test.

$$MOS_{norm} = \frac{MOS - MOS_{min}}{MOS_{max} - MOS_{min}} \tag{77}$$

Following are some of the most common-used nonlinear regression functions, fitted to data $[VQR, MOS_{norm}]$.

- Simplistic logistic function.

$$MOS_p = \frac{1}{1 + \exp C_0(VQR - C_1)} \tag{78}$$

  For ease of analysis, the above function can be transformed as the linear form $\log_e(1/MOS_p - 1) = C_0(VQR - C_1)$.
- Four-parameter cubic polynomial function

$$MOS_p = C_0 + C_1 \times VQR + C_2 \times VQR^2 + C3 \times VQR^3 \tag{79}$$

- "Inverse" four-parameter cubic polynomial function

$$VQR = C_0 + C_1 \times MOS_p + C_2 \times MOS_p^2 + C3 \times MOS_p^3 \tag{80}$$

- The 5-parameter logistic curve

$$DMOS_p(VQR) = A_0 + \frac{A_1 - A_0}{1 + A_4 \times (VQR + A_5)/A_3} \tag{81}$$

Apart from MOS, similar analysis can be performed on individual opinion scores (OS), and difference opinion scores (DOS). The performance of the objective quality model is evaluated from three aspects: prediction accuracy, monotonicity and consistency.

- *Prediction Accuracy* is represented by the Pearson linear correlation coefficient and root mean-square-error (MSE). The Pearson linear correlation coefficient between two variables X and Y is:

$$\rho_{X,Y} = \frac{E\left[(X - E(X))\,(Y - E(Y))\right]}{\sqrt{\left[E(X^2) - (E(X))^2\right]\left[E(Y^2) - (E(Y))^2\right]}} \tag{82}$$

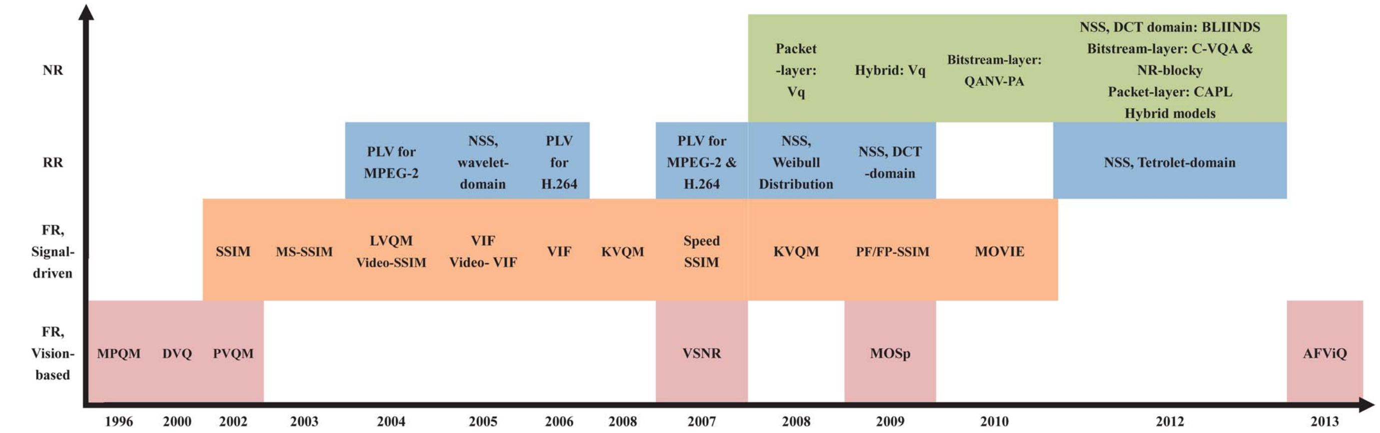


Fig. 22. Timeline of the objective quality models.

The Pearson linear correlation coefficient quantifies the correlation between two variables. It has the value in [-1,1], where -1 means total negative correlation, 0 means no correlation, and 1 means total positive correlation.

Root mean-square-error (MSE) is:

$$MSE = \frac{1}{N}\sum_i (MOS_p - MOS)^2 \tag{83}$$

- *Prediction Monotonicity* is represented by the Spearman rank order correlation coefficient.

  The Spearman rank order correlation coefficient characterizes how well one variable can be represented as a monotonic function of the other variable. One merit of the Spearman rank order correlation coefficient is that no knowledge of the relationship (e.g., linear, logistic) between the two variables is required (referred to as non-parametric). Assume that we have N raw samples (X, Y). The calculation of the Spearman rank order correlation coefficient is as follows:

  – Sort X and give rank number x_i to the ith sample, e.g., if in the 1st sample, the value of variable X is the 4th largest, then $x_1 = 4$;
  – Sort Y and give rank number y_i to the ith sample, e.g., if in the 1st sample, the value of variable Y is the 5th largest, then $y_1 = 5$;
  – The Spearman rank order correlation coefficient ρ is

$$\rho = 1 - \frac{6\sum_i (x_i - y_i)^2}{N(N^2-1)} \tag{84}$$

  The Spearman rank order correlation coefficient has the value in [-1,1], where -1 means X can be represented as a monotonically decreasing function of Y, 1 means X can be represented as a monotonically increasing function of Y.
- *Prediction Consistency* is represented by the outlier ratio.

$$\text{Outlier ratio} = \frac{\text{number of outliers}}{N} \tag{85}$$

  in which N is the total number of samples, and an outlier is a point for which $|MOS - MOS_p| > 2 *$ (Standard Error of MOS).

Furthermore, *wide application* and *computational complexity* are two other aspects to evaluate the objective quality model. It is ideal for the objective quality model to give relatively good prediction for a wide range of video content. However, there is no metric to evaluate the wide applicability of the model. Therefore, it is desirable to cover as many types of video content and test conditions as possible in the subjective test. It is recommended that at least 20 different video sequences should be included.

E. Objective Quality Model Projects and Standards

1) VQEG Projects: The Video Quality Experts Group (VQEG), established in 1997, with experts from ITU-T and ITU-R study groups, carried out a series of projects to validate objective quality models. Their work leads to inclusion of recommended objective quality models in International Telecommunication Union (ITU) standards for standard definition television and for multimedia applications [140]. Subjective test plan is given for laboratories to carry out subjective test. The resulting database is used for validating objective quality models' prediction power. Objective test plan is given to evaluate the submitted objective quality models with specified statistical techniques and evaluation metrics. The final report of each test summarizes the testing results as well as providing detailed description of the subjective evaluation procedure, the proposed objective quality models, the evaluation criteria and some discussion and comments. The subjective test sequences and corresponding scores are made accessible for researchers to validate their objective models. The validation test projects that have been accomplished by the VQEG is summarized in Table XII.

2) LIVE Project: The Laboratory for Image and Video Engineering (LIVE) at the University of Texas at Austin, led by Prof. Alan C. Bovik, establishes the LIVE Video Quality Database, due to two deficits of the existing VQEG Phase I FR-TV database [29]:

- VQEG database uses old-generation codec such as H.263 and MPEG-2, while the more advanced H.264/MPEG-4 Part 10 codec may exhibit different distortion patterns.

TABLE XII
VQEG COMPLETED VALIDATION TESTS

Project	Completion Date	Application Scope	Model type	Resulting ITU Recommendations
FRTV-I	June, 2000	SDTV	FR and NR	None
FRTV-II	August 25, 2003	SDTV	FR and NR	ITU-T Rec. J.144 (2004) ITU-R Rec. BT.1683 (2004)
MM-I	September 12, 2008	Multimedia model	FR and RR	ITU-T Rec. J.247 (2008) ITU-T Rec. J.246 (2008) ITU-R Rec. BT.1866 (2010) ITU-R Rec. BT.1867 (2010) ITU-T Rec J.340 (2010)
RRNR-TV	June 22, 2009	SDTV	RR and NR	ITU-T Rec. J.249, (2010) ITU-T Rec J.340 (2010)
HDTV-I	June 30, 2010	HDTV	FR, RR and NR	ITU-T Rec. J.341 (2011) ITU-T Rec. J.342 (2011)

- VQEG database subjective test results are skewed towards high user scores (it is ideal that user scores are uniformly distributed), suggesting that the processed video sequences have poor perceptual separation.

The LIVE database is publicly accessible with the aim to "enable researchers to evaluate the performance of quality assessment algorithms and contribute towards attaining the ultimate goal of objective quality assessment research—matching human perception" [141]. Table XIII summarizes the differences between the VQEG Phase I FR-TV database and the LIVE database. In the LIVE database, H.264 advanced video coding is used, and the wireless network distortion is simulated. There are ten source sequences provided by Boeing, with a diversity of motions, objects and people. The encoded source sequences are regarded as the original version as it is claimed that H.264 compression is visually lossless (with average PSNR greater than 45 dB). Each sequence is processed with a combination of 4 bitrates (500 kb/s, 1 Mb/s, 1.5 Mb/s, 2 Mb/s) and 4 packet loss rates (0.5%, 2%, 5%, 17%), resulting in 160 processed sequences. The subjective test results show that the DMOS value has good perceptual separation (i.e., the values are nearly uniformly distributed). Single stimulus continuous quality-scale (SSCQS) based on [57] is used as the subjective test method. To counteract the individual bias, the original video sequences are inserted in the testing sequence. Therefore, (score for the processed video)—(score for the original video) is regarded as an unbiased score. The use of continuous quality-scale also breaks the limitation of categorical quality-scale used in VQEG database. However, only 60 Hz refresh rate is considered (the VQEG Phase I FR-TV test includes both 50 Hz and 60 Hz).

TABLE XIII
DIFFERENCES BETWEEN LIVE DATABASE AND VQEG PHASE I FR-TV DATABASE

	LIVE database	VQEG FRTV-I database
Codec	H.264 and MPEG-2	H.263 and MPEG-2
Video format	50Hz	50Hz, 60Hz
Number of source video sequence	10	20
Encoding bitrate	500 kb/s, 1 Mb/s, 1.5 Mb/s, 2 Mb/s	768kb/s, 1.5Mb/s, 2Mb/s, 3Mb/s, 4.5Mb/s, 6Mb/s, 8Mb/s, 12Mb/s, 19Mb/s, 50Mb/s
Packet loss rate	0.5%, 2%, 5%, 17%	N/A
Subjective test method	SSCQS	DSCQS
Subjective test score	DMOS	DMOS

F. Discussion

Objective quality model has a wide range of applications, including equipment testing (e.g., codec evaluation), in-service network monitoring, and client-based quality measurement. However, in [142], the author points out seven challenges facing the current objective quality models. Interested readers can refer to the original paper for more details.[2]

- Insufficient knowledge of HVS and natural image. Most of the objective quality models only employ low-level HVS properties. Though VSNR leverages mid-level HVS property (global precedence), the modeling of higher level HVS property is far from complete. Another problem is that visual neurons have different responses to simple, controlled stimuli and to natural image. This may affect masking results, in particular, the contrast threshold. However, there is a lack of ground truth data of local contrast detection thresholds for natural images.
- Compound and suprathreshold distortions. Compound distortions refer to distortions that stimulate more than one channel of the HVS multichannel system; suprathreshold distortions refer to distortions that are obviously visible. Existing near-threshold distortion analysis focuses on the visual detectability of the distortion. However, it is found that visual detectability of the distortions may not accord with viewers' perception towards suprathreshold distortions [143], [144]. Therefore, models suitable for near-threshold distortions may not be able to be extended to account for suprathreshold distortions.
- Interaction of the distortion and the image. There are two different assumptions about the relationship between the distortion and the image. One is that the distorted image is a single stimulus ("overlay distortion"); the other is that the distorted image is a combination of two separate stimuli: the distortion added to the image (additive distortion). It is important to distinguish these two types of distortions.
- Interaction between distortions. One type of distortion may *mask* another type of distortion, known as cross-masking. To quantify the interaction between distortions and their effect on the image quality is needed.

[2] Though [142] limits the discussion to image quality assessment, the main points are still applicable to video quality assessment.

- Geometric changes. It is argued that current objective quality models are bad at dealing with geometric changes. For example, slight rotation of the objects has little impact on perceptual quality but will result in lower quality estimation by the objective quality models.
- Evaluation of enhanced image. Image enhancement such as noise reduction, color correction and white-balancing, may in turn make the original image seem like inferior. One way to evaluate enhanced image is to treat the original image as "distorted", and the enhanced image as "original"; then apply existing objective quality models. The feasibility of such method still needs to be verified.
- Efficiency. Efficiency concern includes running time and memory requirement.

Apart from the above challenges, we also have the following comments for the objective quality models.

- Full reference model is impossible to implement for real-time QoE prediction and monitoring, because of its complexity and the need to access the original video. Reduced reference model, though does not need the access to the original video, requires extra resources (e.g., a side channel) to transmit the extracted information of the original video. Psychophysical approach models that are based on the mechanisms of the HVS, though perform well with the subjective MOS scores, often have high complexity. Engineering approach models usually have lower complexity, and can be calibrated using the subjective test results.
- All of the existing objective quality models compared their predicted QoE with the MOS scores to evaluate their performance. The MOS scores are obtained from the subjective test, which is limited in test video types, number of human assessors, and test conditions. Therefore, objective quality models with predicted QoE close to one set of MOS scores of a particular subjective test, may not have the same good performance compared with another set of MOS scores obtained from a different subjective test.

V. DATA-DRIVEN QoE ANALYSIS

The dramatic development of video distribution over the Internet makes massive data available for analysis, and triggers a new research interest of data-driven QoE assessment. Commercial broadcast television corporations (e.g., FOX, NBC) and on-demand streaming video service providers (e.g., Netflix, Hulu) now provide millions of videos online. Improving user QoE is crucial to the service providers and network operators, since small changes in viewer behavior will lead to whopping changes in monetization opportunities due to huge viewer base over the Internet.

To begin with, we give a detailed description of a typical video viewing session, based on which we introduce the QoE and QoS metrics that are concerned by the current data-driven QoE-related works. Define a *viewer* as a specific identifiable user who watches video through the service of a provider; define a *view* as the event that a viewer watches a specific video; define a *visit* as the event that a viewer continually watches a series of videos from a specific website. Two visits are separated by a duration of inactivity for a time threshold. Fig. 23 shows a typical video watching session. A viewer initiates a video request, and the video player establishes the connection to the server. A certain amount of data has to be downloaded in the buffer before the video starts playing (*startup state*). During playing, the video player fetches the data in the buffer; and meanwhile, downloads more data from the server (*playing state*). If the rate of using the data exceeds the rate of downloading (e.g., due to poor connection), the buffer will be exhausted. In this case, the video player has to pause to fill its buffer to a certain level before start playing again (*rebuffer state*). The viewer therefore experiences interruptions during this period. During the video session, viewers may also have interactive actions such as pausing, fast-forwarding, rewinding, changing the resolution or changing the screen size. A view may end in four manners.

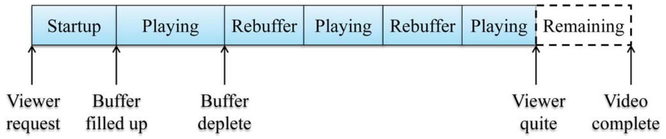


Fig. 23. A typical video watching session.

- *Abandoned view*. The viewer voluntarily quits during the startup state, and does not watch any of the video.
- *Aborted view*. The viewer watches a certain part of the video, but voluntarily quits during the playing state or rebuffer state before the video completes.
- *Failed view*. The requested video involuntarily ends due to failure of the server, the connection or the video content.
- *Complete view*. The view ends when the video is completely watched.

Except for the case of complete view, all other three cases may be a result of user dissatisfaction, which may be caused by poor video quality, user's lack of interest in the video content, or external interruption (e.g., mobile users on the train reaches destination). The following metrics are often used to represent user QoE by quantifying the user *engagement* for the video service:

- View-level metrics, which regard the engagement of each video viewing session.
  - *Viewing time per view*: the actual time that a user watches a video. Usually, the ratio of the viewing time to the total duration of the video is used as an indicator for user engagement.
  - *Abandoned view ratio*, the percentage of views that are voluntarily abandoned by the viewers during startup state.
- Viewer-level metrics, which regard the engagement of each viewer.
  - *Number of view*, the number of video clips a user watches within a certain time period on a certain website.
  - *Viewing time per visit*, the total length a user watches the video during a visit to a certain website.

TABLE XIV
MEASUREMENT STUDY ON USER BEHAVIOR RESEARCH

	Scenario	User behavior	QoS factors	External factors
[145]	Online VoD service	Viewing ratio distribution number of views per viewer	Rate of buffering	Video length, video popularity time of the day
[146]	YouTube	Number of unique users and request viewing time, viewer rating	HTTP request methods video bitrate, file size	Content type, video release and update time viewer geography, temporal features
[147]	P2P IPTV ystem	Number of peers, peer arrival and departure rate	TCP connection, video traffic characteristics	TV channel popularity, peer geography, temporal features
[148]	Live VoD system	Number of daily access, viewing time ratio		Video length, video popularity temporal features
[149]	YouTube	Resolution switch, video download ratio, viewing time ratio	Flow size, startup delay	Video length, size, bitrate and format, device type
[150]	Online VoD service	Viewing ratio distribution, seeks, Number of views per viewer	Rate of buffering	Content type, video popularity index
[151]	Mobile video service	Subjective rating	Video quality, bandwidth, startup delay, rate of buffering, RTT	

TABLE XV
DATA-DRIVEN VIDEO QOE RESEARCH

	QoE metrics	QoS metrics	External factors	Method	Data collection			
					Size (million views)	Source	Method	Duration
[48]	Viewing time, number of views, total viewing time	Average bitrate, startup delay, buffering time ratio, rate of buffering, rendered quality	Video length (short or long), VoD or live	Linear regression	300/ week	Conviva	Client-side instrumentation library	Fall, 2010
[30] [53]	Viewing time	Average bitrate, startup delay, buffering time ratio, rate of buffering	Video type, time of day, device, connectivity	Decision tree	40	Conviva	Client-side instrumentation library	Over 3 months
[152]	Viewing time, prob. of return, abandonment rate	Average bitrate, startup delay, rebuffer delay, number of failed views	N/A	QED	23	Worldwide	Akamai's client-side media analytics plug in	10 days

  - *Return rate*, the percentage of viewers who visit the same website again within a specified time period. The return rate indicates the possibility that a user will visit the video website in the future.
  - *Video rating*. Many video websites enable users to rate the video. For example, YouTube uses a scale of 0–5 "stars"; Youku and Tudou have a "Thumb-up" or "Thumb-down" choice.

These measurable QoE metrics are also directly related to the service providers' business objectives. For example, for advertisement-supported video service, if the viewing time is longer, more ads can be played to the viewers; for subscription-supported video service, better QoE can reduce the viewer churn rate.

While still considering the influential factors described in Section II, current data-driven QoE research is more focused on the following QoS metrics.

- *Startup delay*, also called join time. As shown in Fig. 23, join time is the time between the user requests the video and the video actually begins playing, during which the buffer is being loaded.
- *Rebuffering*. The encoded video stream is temporarily put in a buffer to be played back later. As shown in Fig. 23, when the buffer is depleted, the player pauses to rebuffer. There are two ways to quantify the rebuffering event.
  - *Rebuffering time ratio*, the ratio of the total time for rebuffering to the total viewing time.
  - *Number of rebuffering*. If the rebuffering happens quite frequently, but the time for each rebuffering is very short, the ratio of rebuffering is low, yet such intermittent playing may annoy the viewer. The number of rebuffering can characterize the frequency of the rebuffering event.
- *Average Bitrate* at which the video is rendered on the screen to the viewer. This rendered bitrate depends on the video encoding bitrate, network connectivity and the bitrate-switch heuristics employed by the media player.

In the rest of this section, we first introduce the earlier work of video measurement study on user behavior, as summarized in Table XIV, then we introduce three recent directions of data-driven QoE analysis, as summarized in Table XV.

A. *Measurement Study on User Behavior in Video Service*

Large-scale measurement studies have long been carried out to study general user behavior in various video services, including online VoD service [145], [150], Live VoD [148], P2P IPTV system [147], the YouTube traffic [146], [149], [153] and mobile video service [151]. A survey of user behavior in P2P video system is recently given by [154]. In this section, we first identify the general user behavior revealed by these measurement study, then introduce a decision theoretic approach to model user behavior.

1) General User Behavior: We discuss the following user behaviors that have been studied by many measurement studies.

- Early quitter phenomenon/video browsing. It is found that the most video sessions are terminated before completion [55], [148], [150]. More specifically, many viewers quit the video session within the first short period of time. One of the explanations for this early quitter phenomenon is that a viewer browses several videos before dedicating to a specific one which interests him. The video browsing behavior is intensively studied by [150]. It is found that viewers often use seeks (jump to a new part) to browse a video, and that the viewers are more likely to browse popular videos first due to recommendation. Another problem caused by the early quitter problem is that the downloaded video files will exceed the watched video files, resulting in data waste, which is found to be more severe for the player on the mobile device than the computer [149].
- Temporal user access pattern. It has been confirmed in many papers that user access has a clear and consistent daily or weekly pattern [55], [145], [146], [148]. The diurnal viewing pattern is also found in the P2P video system [147].
- Video quality metrics. Three video quality metrics, i.e., startup delay, rebuffer events, and encoding bitrate, are most-commonly characterized by their cumulative distribution function [145], [147], [149]. In particular, the impact of rebuffering time is studied in [151] by a subjective test like experiment. Each assessor watches preassigned videos with different bandwidth, quality and rebuffering time combinations, in a mobile context. Then, they are asked to answer questionnaires to express their experience. Finally, the relationship between the rebuffering time and viewers' acceptance of the video quality is fitted by a logistic regression model.
- Video popularity. It is found that the video popularity can be approximated by the Parento Principle, or 80–20 rule. That is to say, a few top videos account for most of the viewer accesses [55], [145], [148], which is usually compared with a Zipf-like distribution. It is found that the popular video list changes quite frequently [148]. As the video release time increases, the video popularity often drops. However, if later, a remake version appears or a certain event happens, the related video may have a surge in popularity [148].
- Flash crowd phenomenon. Normally, user arrival distribution is found to follow the Poisson distribution in [55]. Flash crowd refers to a burst of video access or request within a short period of time. It is usually triggered by special national or international events, for example, popular events in the Olympic Games [148], or Chinese spring festival gala show [147]. The flash crowd phenomenon will impose great pressure on the network due to huge amount of video traffic. One solution is to push related videos to multiple edge servers during such event.

2) Decision Theoretic User Behavior Model: In [155], a theoretic model based on decision network, an extension to the Bayesian network [156], is proposed to characterize user behavior. There are four types of nodes in the decision network.

- Chance nodes, also the bottom nodes. Chance nodes represent all random variables in the system, including all possible QoS parameters and external factors we introduce in Section II.
- Query nodes, the parents of chance nodes. Query nodes determine the current state, including four contexts: network context, service context, environment context and user behavior.
- Utility nodes, associated with each of the four types of query nodes, including network utility, service utility, environment utility and user behavior utility. Utility nodes specify the utility function in each context.
- Decision nodes, the top nodes. Decision nodes choose the optimal option according to predefined target, e.g., maximum QoE.

Firstly, the chance nodes are fed with evidence variables. After the values of the evidence variables are determined, the posterior probability distribution of the query nodes can be calculated. Then, the utility nodes figure out the utility for different options. Finally, the decision nodes choose the option which maximizes the QoE. The Bayesian network or decision network can be applied to estimate user departure time [156] or perceptual quality [155]. Further development and verification of such models are expected.

Measurement study can only give a general understanding of the user behavior in video service under different conditions. In order to monitor, predict and even control user QoE, we need more in-depth analysis.

B. Data-Driven QoE Analysis

1) Correlation and Linear Regression Based Analysis: In [48], a framework is built for identifying QoS metrics that have significant impact on user QoE for different video types; and quantifying such influence by linear regression. QoS metrics include startup delay, rebuffering and bitrate; QoE metrics include the viewing time ratio, number of views and total time of viewing. The data is collected at the client side via affiliated video websites, covering five influential content providers. Videos are classified as Long VoD, Short VoD and Live videos. The flow of the analysis is shown in Fig. 24.

- QoE-QoS Kendall Correlation
  The correlations between each QoS and QoE metrics are calculated to determine the magnitude and the direction of the influence of each QoS metric. The paper chooses Kendall correlation coefficient, a non-parametric rank correlation measurement to quantify the similarity between two random variables. Unlike Pearson correlation coefficient, which measures the linear dependence of two random variables, the Kendall correlation coefficient does not assume the relationship between the two variables. High absolute correlation value is regarded as an indicator

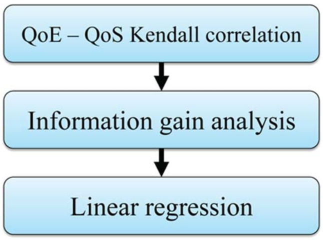


Fig. 24. Linear regression based QoS-QoE model.

for significant impact of the QoS metric on the QoE metric. Kendall correlation coefficient can be calculated as follows. Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ denote the joint observation of two random variables X and Y. Pair $(x_i, y_i), (x_j, y_j)$ is *concordant* if $x_i > x_j, y_i > y_j$ or $x_i < x_j, y_i < y_j$, otherwise, the pair is *discordant*. The case $x_i = x_j, y_i = y_j$ can be treated as concordant or discordant. The Kendall correlation can be calculated as:

$$\tau = \frac{N_{concordant} - N_{discordant}}{\frac{1}{2}n(n-1)} \tag{86}$$

The number of possible pairs of observation is $\frac{1}{2}n(n-1)$, so $\tau \in [-1, 1]$. If the ordering of X and Y is perfectly agreed, $\tau = 1$; If the ordering of X and Y is perfectly disagreed, $\tau = -1$; If X and Y are independent, $|\tau| \approx 0$.

- Information Gain Analysis
  The Kendall correlation coefficient cannot reveal the non-monotonic relationship between the QoS and QoE metrics. *Information gain* helps to get a more in-depth understanding of the QoS-QoE relationship by quantifying how the knowledge of a certain QoS metric decreases the uncertainty of the QoE metrics. Let X denote the QoE metric, and Y denote the QoS metric. The information gain for X, given Y is $[I(X) - I(X|Y)]/I(Y)$, in which $I(\cdot)$ is the entropy, a characterization of how much information is known of the random variable. Information gain can be calculated for not only an isolated QoS metric, but also the QoS metric combinations. High information gain is regarded as an indicator for significant impact of the QoS metric on the QoE metric.
- Linear Regression
  Linear regression based curve fitting is applied to the QoS-QoE pairs which are visually confirmed to have quasi-linear relationship. By observing the QoS-QoE curves, it is obvious that the relationship is not linear in the entire range. Therefore, linear regression is only applicable to a certain range of data.

The above analysis framework is applied for Long VoD, Short VoD and Live videos. There are two key findings. First, certain QoS metrics have high influence on one type of video, but low influence on other types of video. In other words, the influence of QoS metrics is content-dependent. Second, certain QoS metrics have low absolute correlation coefficient values, but high information gain. The possible reason is that the QoS-QoE relationship may be non-monotonic. Therefore, correlation analysis alone is not enough to decide the importance of QoS metrics.

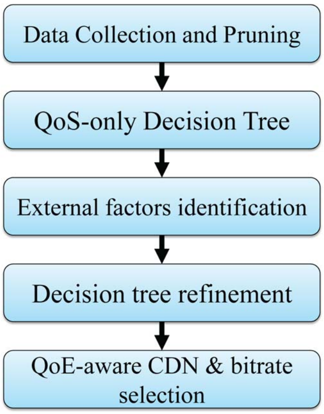

Data Collection and Pruning
QoS-only Decision Tree
External factors identification
Decision tree refinement
QoE-aware CDN & bitrate selection

Fig. 25. Decision-tree based QoE prediction model.

Though being a simple way to characterize the QoS-QoE relationship, the correlation and linear regression based analysis fail to deal with the following problems.

- Non-monotonic relationship between the QoS and QoE.
- Interdependence between QoS parameters. The linear regression requires that the QoS parameters are independent, which may not be true, e.g., it is shown that bitrate and startup delay are correlated [30], [53].
- External factors handling. There is a lack of analysis on external factors and their influence on user QoE.

2) Decision Tree Based QoE Prediction Model: To overcome the drawbacks of the linear regression and correlational analysis, in [30], [53], a decision-tree based QoE prediction model is developed based on 40 million video views collected on the video website *conviva.com*. Viewing time ratio is chosen as the QoE metric; startup delay, buffer events and average bitrate are chosen as the QoS metrics; external factors considered are video type (live or VoD), connectivity and so on. The analysis framework is shown in Fig. 25.

- Data Collection and Pruning
  Not only the QoE and QoS metrics are recorded, the viewer-specific parameters (e.g., video type, device type and time stamp) are also collected for external factors. The early-quitters who watch the video for a very brief time are eliminated from the data set to improve prediction accuracy.
- QoS-only Decision Tree Building
  Decision Tree model is a non-parametric model, which does not presume the QoS-QoE relationship (therefore can deal with non-monotonicity), and does not require the QoS metrics to be independent. In addition, it is simple but expressive enough to characterize QoS-QoE relationship and give relatively accurate predictions. First, each parameter is discretized because decision tree can only deal with discrete values. Then, the data set is separated into 10 groups. The model is trained 10 times. Each time, 9 groups are used for training and the remaining group for validation.
- External Factors Identification
  The impact of external factors is on three aspects: the QoE metrics, the QoS metrics and the QoS-QoE relationship. The impact on QoS and QoE metrics is identified by the

information gain; and the impact on QoS-QoE relationship is identified by the difference in decision tree structure and QoE-QoS curve. If an external factor has high information gain for a certain QoS metric or QoE metric, or makes the tree structure and/or QoE-QoS curve different, it is identified as an important external factor.

- Decision Tree Refinement
  After figuring out the important external factors, there are two ways to refine the QoS-only decision tree

  - *Add as an input* to build the decision tree. It is simple, but mixing the QoS metrics with external factors gives confusing guidance.
  - *Split the data* according to different external factors (or combinations, like VoD-TV). It will lead to a forest of decision trees. The curse of dimensionality may happen when the data is sparse.

  It is shown that splitting the data often gives better results than adding the external factor as an input.
- QoE-aware CDN & Bitrate Selection
  Brute force method is used to find the optimal Content Delivery Networks (CDN) and bitrate combination by feeding the (CDN, bitrate) pair and other QoS metrics and external factors into the QoE prediction model. The (CDN, bitrate) pair that yields the highest predicted QoE is optimal.

Though overcoming the drawbacks of the linear regression, the above decision tree based analysis framework still suffers from the following major problems:

- The final QoE prediction is a range rather than a value. Therefore, it cannot meet the need for fine-grained QoE prediction.
- The decision tree can only deal with discrete values. The way how the parameters are discretized may influence the performance of the model.

3) QED Based QoS-QoE Causality Analysis: To verify the existence of causal relationship between QoS and QoE, a QED-based model is built to identify the QoS metrics that have a significant causal effect on the QoE metrics, thus providing a guidance to service providers of which QoS metrics should be optimized [152]. Correlational relationship does not infer causal relationship, thus may lead to incorrect conclusions. For example, one can not conclude that high bitrate alone will result in longer viewing time, unless all the other factors (e.g., video popularity, buffering time) are accounted for. The authors only consider VoD videos, with a dataset of 23 million views from 6.7 million unique viewers, using cable, fiber, mobile and DSL as major connections. The QoE metrics under analysis are the abandonment rate, viewing time and return rate; and the QoS metrics are failures, startup delay, average bitrate and rebuffer delay.

To verify that a QoS metric X has a causal influence on the QoE metric Y, the most ideal method is through controlled test. In the test, two viewers with perfectly the same attributes but only differ in X are compared in terms of their resulting Y. Such controlled test is infeasible to implement for the video distribution service. But Quasi-Experimental Designs (QED) [157], can be leveraged to reveal the causal relationship from the observational data. The flow of the QED-based QoS-QoE causal relationship analytical framework is shown in Fig. 26.

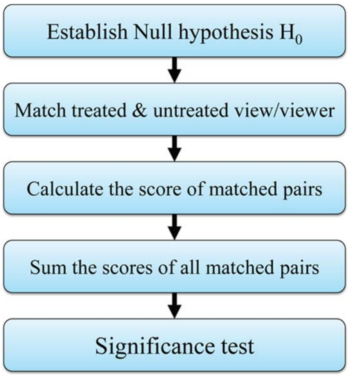

Establish Null hypothesis H_0
Match treated & untreated view/viewer
Calculate the score of matched pairs
Sum the scores of all matched pairs
Significance test

Fig. 26. QED-based QoS-QoE causal relationship analytical framework.

- Establish Null Hypothesis
  A null hypothesis usually takes the form "The QoS metric X has no impact on the QoE metric Y". The null hypothesis will be rejected if there is causal relationship between the QoS metric and the QoE metric.
- Match Treated and Untreated View/Viewer
  A view/viewer is treated if the view/viewer undergoes a certain "bad" QoS condition, e.g., a rebuffering time ratio more than $\alpha\%$. A view/viewer is untreated if the view/viewer undergoes a corresponding normal QoS condition, e.g., a rebuffering time ratio less than $\alpha\%$. Regarding a certain QoS metric, all the treated view/viewer form the treated set T, all the untreated view/viewer form the untreated set U. Then, for each $t \in T$, uniformly and randomly pick a $u \in U$, that is "identical" to t in every other aspects. (t, u) is a matched pair, and all matched pairs form the match set M.
- Calculate Scores for Matched Pairs
  For a matched pair (t, u) in M, if the QoE values conform to the hypothesis, (e.g., t has lower QoE value than u), the score of pair (t, u) is assigned 1; otherwise, the score of pair (t, u) is assigned -1. Other ways of assigning score value are also possible [152].
- Sum Up Scores
  The sum of the scores for all matched pairs is

$$\text{Sum of score} = \frac{\sum_{(u,t)\in M} \text{Score}(u, t)}{|M|} \tag{87}$$

- Significant Test
  A "p-value" based on sign test is calculated, which indicates the probability that the data conforms with the null hypothesis. If the "p-value" is small, the null hypothesis can be rejected with high confidence, corroborating the assumption that the QoS metric has a causal influence on the QoE metric.

Though verifying the causal relationship between QoS and QoE, the above framework does not quantify the QoS-QoE relationship. Hence, it cannot be used for QoE prediction, or providing instrumental guidance on how to achieve QoE-based video service optimization.

C. Discussion

After discussing the advantages and disadvantages of the existing models, we now identify the requirements of an ideal data-driven QoE analysis model:

- Requirement for QoE Metrics.
  - *Measurable*. Since the raw data is collected in the wild rather than in a controlled laboratory environment, the QoE metrics for large-scale data-driven analysis should be easy to measure and monitor in real-time. This is also true for the QoS metrics and the external factors.
  - *Informative*. The selected QoE metrics should be a good indication of user experience or engagement. It may be needed to verify the correlation between the measurable QoE metrics (such as viewing time ratio) and real subjective user QoE.
  - *Business fitting*. Ideally, the QoE metrics should be closely linked to the service providers' business objectives, e.g., contributing to the monetization of the advertisement-supported or subscription-supported video service.
- Requirement for QoS-QoE Model
  - *Reliable*. The model should give reliable QoE prediction, given the QoS parameters and external factors. Models that assume independency among QoS variables may be not be accurate, e.g., it is found that the bitrate and buffering are correlated [30].
  - *Expressive*. The model should be expressive enough to capture the complex and non-monotonic relationship between QoS and QoE. Regression models that preassign a certain relationship (linear, logistic, etc.) may be problematic.
  - *Real-time*. For the model to be able to conduct real-time QoE prediction, monitoring and even controlling, the computational complexity and storage requirement have to be acceptable.
  - *Scalable*. As the network and user experience evolves with time, the model should be able to readily take new variables, and give relatively accurate results.

VI. Applications of Video QoE Analysis Models

In this section, we introduce existing works which leverage video quality assessment models for video transmission optimization or network control.

A. Cross-Layer Video Transmission Optimization

QoE metrics evaluate the video quality from the users' perspective, which can provide the guideline for MAC/PHY level optimization. This is especially important for delivering video over wireless network, constrained by the limited bandwidth and unstable channel quality. There are two major concerns for the cross-layer video transmission optimization:

- *Reliable QoE prediction model*. Given the input of QoS parameters and external factors, the QoE prediction model should give reliable results, so that corresponding adaptation actions can be taken to improve user QoE. The process should be performed online to give real-time feedback.
- *Cross-layer timescale difference*. At the application level, the video source adaptation is at the timescale of one frame or one Group of pictures (GoP), which is much longer than the link adaptation at the PHY level. Furthermore, the channel condition variation is much faster than the video signal variation. Therefore, the application level video source adaptation may use the aggregated PHY level information, while the PHY level link adaptation uses the relatively coarse application level information.

Cross-layer video transmission optimization is studied in [158]–[160], using PSNR as the QoE metric. In [161], the authors propose a classification-based multi-dimensional video adaptation using the subjective test results, not practical for online network management. In [162], the authors propose an APP/MAC/PHY cross-layer video transmission optimization architecture. An online QoS-QoE mapping is developed to estimate the lower-bound of QoE value based on the packet error rate. Then, the QoS-QoE mapping is leveraged by the PHY level to perform unequal error protection to maximize the QoE. At the APP level, source rate is adapted based on channel condition and buffer state. In [163], the authors use slice loss visibility (SLV) model [115] to estimate the visual importance of video slices (a frame is divided into multiple slices, each of which consists of multiple macroblocks). The most important slices are allocated to the most reliable subbands of the OFDM channels.

B. QoE-Aware Congestion Control

Congestion control in the conventional TCP protocol, when applied to video traffic, may lead to long delay due to the following reasons:

- According to the TCP protocol, a lost packet will be retransmitted until it is successfully received, resulting in long delay and therefore poor QoE.
- The Additive Increase Multiplicative Decrease (AIMD) algorithm leads to fluctuated throughput over time, which will increase the delay, leading to user dissatisfaction.
- The congestion control is QoS-based while the video is more user-centric and QoE-based.

In order to design a video-friendly congestion control mechanism for the TCP protocol, Media-TCP is proposed in [164], which optimizes the congestion window size to maximize the long-term expected QoE. The distortion impact and delay deadline of each packet are considered, in order to provide differential services for different packet classes. Media-TCP is shown to improve the PSNR over the conventional TCP congestion control approaches. While Media-TCP is still QoS-based, a MOS-based congestion control for multimedia transmission is proposed in [165]. The MOS value is estimated in real time by the Microsoft Lync system, based on quantitative measurements such as packet loss, bit errors, packet delay and jitter. The

QoE-aware congestion window adaptation is then formulated as a Partially Observable Markov Decision Process (POMDP), and is solved by the online learning algorithm. Another way to mitigate the delay problem in video transmission, without modify the TCP protocol, is to use video-friendly application protocol such as Dynamic Adaptive Streaming over HTTP (DASH).

C. Video Transmission Over Wireless Network

Special attention has been paid to video transmission over wireless network because of two reasons. First, channel condition in wireless network is ever changing due to noise, interference, multipaths and the mobility of user devices. Second and more importantly, with the growing popularity of smartphones and tablets, mobile video traffic is expected to be dominant in the near future. There are two mainstream wireless networks: licensed cellular networks and unlicensed wireless local area networks (WLANs). While the cellular system has a centralized management, the WLAN, most of which based on IEEE 802.11 standards, operates in a distributed way, sharing the same spectrum with many other networks or systems without a centralized interference management. Thus, video transmission over WLAN is more challenging and attracts more research interests.

1) Interference Management: Rather than average video quality, it is found that viewers are sensitive to small regions of poor quality in the recent past (hysteresis effects) [166], [167]. Rapid change of channel condition and network throughput lead to variation in video quality, which contributes to poor QoE. Different from existing interference management schemes, which often target at reducing the interference power, in [168], the authors propose an interference shaping scheme, which spreads the received interference power in time to "smooth" the burstiness of interference. Though prioritizing real-time video traffic over best effort traffic, it is shown that the QoE improvement (quantified by MS-SSIM index) for the video users only leads to negligible decrease in QoE for best effort users (quantified by Weber-Fechner Law (WFL)-based web QoE modeling [169], [170]).

2) Admission Control: Admission control, or access control, of the IEEE 802.11 WLAN is generally contention based. To cater for different traffic types (real-time and non real-time), it is proposed to prioritize the video traffic, or split the contention time into real-time and non real-time traffic [171]. In [172], the authors use Pseudo-Subjective Quality Assessment PSQA as the QoE metric, and propose a QoE-aware real-time admission control mechanism to manage the network access of multiple users. In [173], the authors consider the reverse problem where a user has multiple network to choose from. Given the information provided by the access points (AP), the user estimates the overall QoE (represented by PSQA [174]) of the APs' existing users and chooses the AP with lower load.

3) Resource Allocation: Resource allocation concerns about how to allocate frequency, transmission time, or bandwidth to multiple users when a centralized scheduling is possible. In [175], a channel allocation scheme is proposed for cognitive radio (CR) network. The CR base station will allocate available channels to secondary users based on their QoE expectations. In [176], [177], the system adapts video configurations through transcoding to meet resource constraints, aiming to have the best possible quality (PSNR).

4) Multicast Rate Selection: In [178], the authors design a video multicast mechanism for multirate WLANs. The hierarchical video coders of the H.264 are combined with the multicast data rate selection: users with poor channel condition (low data rate) will receive only the Base Layer of the encoded video, while users with good channel condition (high data rate) will receive both the Base Layer and the Enhancement Layers. The mechanism is extended for compatibility with IEEE 802.11 standards in [179].

D. QoE-Aware Video Streaming

HTTP-based video streaming protocols have been developed to cater for video traffic. The representative protocols include HTTP Live Streaming (HLS) protocol and Dynamic Adaptive Streaming over HTTP (DASH), also known as MPEG-DASH. A video is divided into chunks of the same time duration, and each chunk is available in multiple quality levels (with different encoding bitrates). During the video session, the player can switch between video streams of the same video content but different bitrates. For instance, if the buffer is nearly empty, the player can select a low bitrate to quickly fill up the buffer to avoid interruption. Given the choice of different quality video streams, the remain issue is the *streaming strategy*, which specifies how to choose the "right" quality for each video chunk, in order to maximize QoE, subject to network conditions and buffer size. The intuition to achieve better QoE is to get higher quality, less frequent quality switch, and avoid video "freezing" (rebuffer). Single user adaptive video streaming is considered in [180], [181]. In [180], the wireless channel prediction information is assumed to be available to the video streaming application, which schedules the video chunks and chooses their quality at each time slot. The problem is formulated as an optimization problem to maximize quality and minimize rebuffering time. In [181], the number of quality switches is added in the utility function, and Markov Decision Process(MDP) is used to solve the optimization problem. Three MDP approaches are proposed, based on online or offline network bandwidth statistics. Multi-user adaptive video streaming is considered in [182], [183]. Different from single user scenario, multi-user scenario has to consider not only efficiency but also fairness among multiple users.

E. Media Player Buffer Design

The design of media player buffer is of great importance, since the rebuffering event has a major influence on user QoE. The buffer size will affect the startup delay and the rebuffering time. If the buffer size is large, the startup delay will be longer because more data has to be downloaded before the player starts playing. Nevertheless, during the playing state, fewer rebuffering events may happen, vice versa. In addition, it is found in [149] that most of the downloaded data in the buffer is useless because many users quit before the video completes.

This results in a huge waste of the bandwidth both for the Internet Service Providers (ISP) and the Content Delivery Network (CDN) operators. Predicting the fraction of videos that may be watched by the viewer will be a great help to avoid transferring excessive data.

VII. Future Direction

In this section, we present future directions of QoE-oriented video quality assessment.

A. Development of Data-Driven QoE Research

Data-driven QoE analysis is still at its infancy, and there is still great room for development.

- *New metric selection*. New metrics for representing QoE, QoS and external factors may come up as the network and the user expectations change with time. The selected QoE metrics should be a good indicator of user experience or engagement, and easy to track and monitor in real-time. Other aspects of user QoE are also interesting. For example, interactivity can be reflected by user behaviors such as pause, fast-forward and rewind. With abundant QoS metrics and external factors, it should be verified which QoS metrics and external factors have a significant impact on user QoE.
- *In-depth user expectation understanding*. Just as most objective quality models are designed based on HVS, theories on user expectation of Internet video service may be further advanced, for example, the user patience for waiting a video to start or restart; the user viewing habits at different time of a day or different days of a week.
- *Analysis tool development*. Many advanced analysis tools can be leveraged to give a more accurate QoE prediction. For example, deep learning algorithms can help extract important QoS and external factors that contribute to user QoE; better regression models can characterize complex QoS-QoE relationship.
- *Early-quitter phenomenon analysis*. A large number of viewers will first “skim” a few number of videos before devoting to watching a specific one or simply quit the website. The early-quitters may exhibit different behaviors from other viewers, e.g., their QoE may be more sensitive towards the video content (e.g., popularity), but less sensitive towards some QoS metrics (due to small QoS changes within a very short time). Other interesting observations also deserve further investigations.
- *Database establishment*. As consumer data is often hard to access and time-consuming to collect, a database that is available to the research community will be of great boost to the QoE-related research. So far, there is no such well-established database like the VQEG database and the LIVE database.

B. QoE-Based Video Transmission Optimization

Most of the previous video transmission optimization is QoS-oriented. As the goal changes from QoS-oriented to QoE-oriented, the optimization problem may be quite different. Though many existing video QoE-related applications have been discussed in Section VI, there are still more to be explored. The following may be some potential research directions.

- *QoE-aware multi-user video traffic scheduling*. This is especially needed for the scenario where multiple users share a bottleneck link. Since different users have different QoE expectations, scheduling can be performed based on user QoE sensitivity. In this way, higher aggregated user QoE may be achieved with limited network resources.
- *QoE-aware video streaming*. Built on the existing adaptive video streaming protocols (e.g., DASH and HLS), sophisticated streaming strategy (find the optimal quality for each video chunk) still needs further exploration. Future solutions must strike a balance between video quality, rebuffering time and quality switch frequency, while relying on relatively accurate channel capacity estimation. In the multi-user case, fairness is also a concern.
- *QoE-aware network management*. Once QoE degradation is detected, first and foremost, the causes should be identified (possibly through the QoE prediction model). If the cause is network-related, ISP and CDN operators may take corresponding actions. If the cause is due to external factors, there is no need for ISP and CDN operators to waste their resources, such as increase bandwidth or change edge servers. All the management decisions should be based on a comprehensive understanding of the QoS-QoE relationship.
- *QoE-aware traffic prioritization*. Video traffic often has larger packet size than other traffic, and the user patience for video service delay is often less than that for other services. Traffic prioritization based on different definitions of user QoE towards different services will be a matter of concern for future research directions.

C. QoE Evaluation in Emerging Technologies

1) 3D Video: There have been a huge number of research works on perceptual quality of 2D video, while the works on 3D video QoE are rather limited. The evaluation of 3D video QoE is challenging because additional factors such as depth perception, comfort levels and naturalness, have to be considered. There are two mainstream coding schemes for 3D video: Scalable Video Coding (SVC) and Multi-view Video Coding (MVC), see Table II. SVC is simulcast coding, where views are independently encoded with different SNR, temporal or spatial scalability. MVC exploits inter-view correlations, and sequential views are dependently encoded. Apart from different coding methods, 3D video can also leverage asymmetric coding. Asymmetric coding encodes the right and left views at different PSNR, spatial resolution or frame rate, being able to reduce the overall bitrate and required bandwidth for transmission. The performance of symmetric coding and asymmetric coding is compared in [184]–[186] via subjective test, based on which efficient asymmetric video encoding approaches are proposed. The influence of packet losses on the QoE of 3D video is studied in [187] using subjective test. The relationship between the DMOS results and the PSNR is characterized by a symmetrical logistic function. The future direction for 3D video QoE evaluation may be a study of the combination of scalable stereo coding, multi-view video coding and asymmetric coding.

2) Interactive Video: Interactive video services, or audio-visual communication services, include videotelephony, video conferencing, and online gaming. Unlike QoE metrics for conventional video services, in the interactive video services, interactivity measurement is of great importance, and should be incorporated in the QoE assessment. In [188], a conceptual framework is proposed to model, measure and evaluate QoE in the distributed interactive multimedia environments. In particular, cognitive perceptions (such as telepresence and perceived technology acceptance) and behavioral consequences (such as performance gains and technology adoption) are incorporated in the QoE metrics. A novel test methodology for QoE evaluation in the interactive video services is proposed in [189]. Conversational interactivity and perceived social presence are incorporated in the QoE metrics. Social presence is the "degree of salience of the other person in the (mediated) interaction and the consequent salience of the interpersonal relationships" [190]. An objective quality model for voice and video over IP (VVoIP) is built in [191], using network bandwidth, delay, jitter and loss to predict QoE. However, there is a lack of consideration for interactivity.

3) Ultra Definition Video: Ultra-high definition television (UHDTV) is standardized in the ITU-R Recommendation BT.2020 [192], aiming at providing users with advanced viewing experience beyond high definition TV. Various works have compared the performance of two common compression methods for UHDTV: High Efficiency Video Coding (HEVC) and H.264/MPEG-4 Part 10 or AVC (Advanced Video Coding). The results show that the HEVC generally outperforms the AVC, achieving higher MOS scores [193], [194] and higher PSNR [195]. However, there is a lack of study on understanding the human perception towards ultra definition video, and building the models to characterize the QoS-QoE relationship for ultra definition video.

4) New Transmission Network: With the rapid development of network technologies, it is desirable to evaluate the QoE of video transmission over different networks, such as mobile network, sensor network and vehicular network.

- *Mobile network*. The popularization of the smartphone has made the traffic of mobile media increase dramatically. The mobile video is characterized by its usage in dynamic and heterogeneous environment. According to the study of mobile TV in [196], the subjective test results in real contexts (e.g., wait in the train station, kill time in cafe or transit by bus) are different from those in the controlled lab. Therefore, it is proposed to evaluate QoE of mobile video in a Living Lab setting, where the viewers watch the pre-defined videos and perform evaluation tasks on mobile devices in real-life scenarios [197], [198].
- *Sensor network*. Wireless Multimedia Sensor Network (WMSN) refers to the sensor network that is able to retrieve, process, store and fuse multimedia information from the physical world [199]. WMSN can be applied for video surveillance, traffic control system, environmental monitoring and so on. However, WMSN faces challenges of resource constraints, channel capacity variation, video processing complexity as well as network management. QoS-provisioning system design for the WMSN has been widely explored [200], [201], but there is a lack of work on the QoE evaluation of such systems.
- *Vehicular network*. Vehicular communications include vehicle-to-vehicle, vehicle-to-infrastructure and vehicle-to-roadside wireless communications. Video transmission over vehicular networks is studied in [202]–[204], using PSNR or packet loss rate as evaluation metrics.

D. QoE-Based Internet Video Economics

The success of the advertisement-supported or subscription-supported revenue models is the major driven force for the fast development of Internet video. Improving user QoE is essential to maintain such revenue models. Therefore, creating a QoE-based economic analysis framework for Internet video will be of great interest.

Fig. 27 shows the general architecture of an Internet video transmission network. Video files are initially generated by the video content providers; then distributed by the Content Delivery Networks (CDN), often chosen by the content providers. After that, the video files are transmitted via wired or wireless network provided by the Internet Service Providers (ISP); and finally displayed on end users' devices by the media player. We can see that there are four major participants in the Internet video service ecosystem:

- *Video Content Provider*, e.g., YouTube, Netflix, and Comcast.
- *Content Delivery Network (CDN) Operator*, e.g., Akamai Technologies in the U.S. [205], ChinaCache in China, and StreamZilla in the Europe. CDN consists of large numbers of servers distributed across multiple ISPs' data centers close to the end users. CDN transports the videos from the content provider to servers at the "edge" of the internet, where the videos are cached and delivered to the end users with high quality.
- *Internet Service Providers (ISP)*, e.g., AT & T, Vodafone, and China Telecom. There are two major types of ISP: fixed-line operators who provide wired network access, and mobile network operators who provide wireless network access. Typical wireless networks include cellular network and WLAN (Wi-Fi); typical wired networks include cable, DSL and fiber.
- *Media Player Designer*, e.g., Adobe which designed Adobe Flash Player, Microsoft which designed Windows Media Player, and Apple which designed QuickTime.

The economic ties between these participants are as follows. The video content providers will choose and pay the CDN operators for delivering their videos. The CDN operators have to pay the ISPs for hosting CDN servers in the ISPs' data centers. Though most media players are free of charge, they benefit the designers by completing their products or services. Improving user QoE is of common interest to all participants, but different participants have different control parameters. For example, CDN operators can select the CDN servers; ISPs can decide

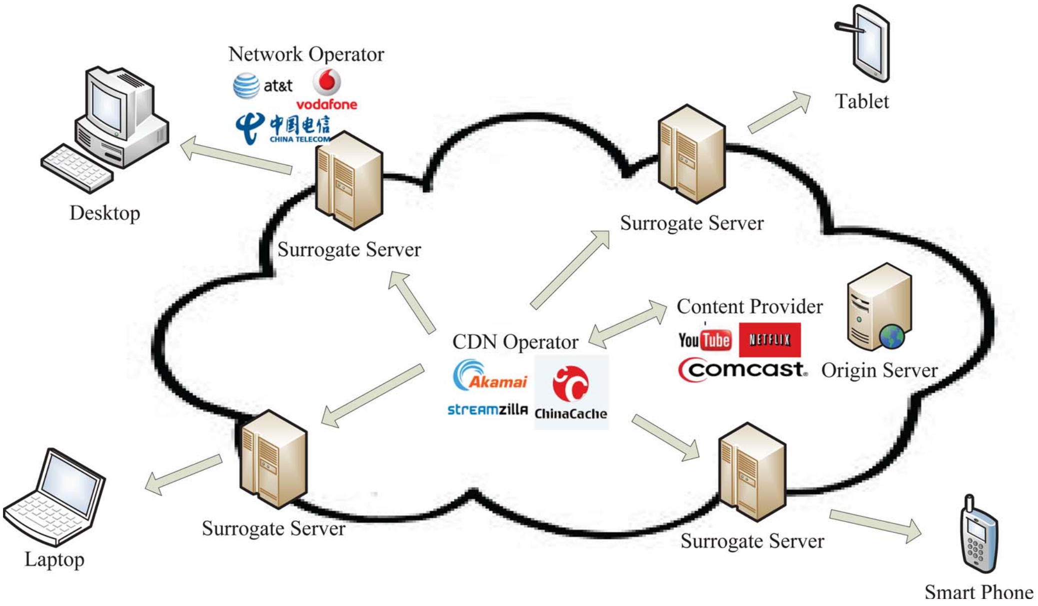


Fig. 27. Video delivery network.

the bandwidth. On the one hand, each participant can maximize his individual utility by choosing his own control strategy; on the other hand, all or some participants can cooperate with each other to maximize end user QoE or total utility. Future research on either direction is promising.

VIII. Conclusion

Video quality assessment has evolved from system-centric QoS-oriented to user-centric QoE-oriented. With the ever-increasing user demand for video service, developing reliable models that can monitor, predict and even control QoE is of great importance to the service providers and network operators. In this tutorial, we give a comprehensive review of the evolution of QoE-based video quality assessment methods: first the subjective test, then the objective quality model, and finally the data-driven analysis. We give detailed description of the state of art of each method. Subjective test is a direct way of measuring QoE, but has a great many of limitations. Objective quality model indirectly predicts QoE through objective metrics, but it relies heavily on the subjective test results. With growing popularity of video streaming over the Internet, large-scale data-driven QoE models have emerged, based on new QoE metrics and data mining techniques. We believe that this will be the research frontier, with many issues to be explored and resolved. We also identify other future research directions, such as QoE-based video transmission optimization and QoE-based Internet video economics.

References

[1] Cisco Visual Networking Index: Forecast and Methodology, 2012–2017, May 29, 2013.

[2] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013–2018, Feb. 5, 2014.

[3] G. J. Sullivan and T. Wiegand, “Rate-distortion optimization for video compression,” *IEEE Signal Process. Mag.*, vol. 15, no. 6, pp. 74–90, Nov. 1998.

[4] Z. Chen and K. N. Ngan, “Recent advances in rate control for video coding,” *Signal Process. Image Commun.*, vol. 22, no. 1, pp. 19–38, Jan. 2007.

[5] Y. Liu, Z. G. Li, and Y. C. Soh, “A novel rate control scheme for low delay video communication of H.264/AVC standard,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 1, pp. 68–78, Jan. 2007.

[6] S. Chong, S.-Q. Li, and J. Ghosh, “Predictive dynamic bandwidth allocation for efficient transport of real-time VBR video over ATM,” *IEEE J. Sel. Areas Commun.*, vol. 13, no. 1, pp. 12–23, Jan. 1995.

[7] A. M. Adas, “Using adaptive linear prediction to support real-time VBR video under RCBR network service model,” *IEEE/ACM Trans. Netw.*, vol. 6, no. 5, pp. 635–644, Oct. 1998.

[8] M. Wu, R. A. Joyce, H.-S. Wong, L. Guan, and S.-Y. Kung, “Dynamic resource allocation via video content and short-term traffic statistics,” *IEEE Trans. Multimedia*, vol. 3, no. 2, pp. 186–199, Jun. 2001.

[9] H. Luo and M.-L. Shyu, “Quality of service provision in mobile multimedia–A survey,” *Human-Centric Comput. Inf. Sci.*, vol. 1, no. 1, pp. 1–15, 2011.

[10] B. Wah, X. Su, and D. Lin, “A survey of error-concealment schemes for real-time audio and video transmissions over the Internet,” in *Proc. Int. Symp. Multimedia Softw. Eng.*, 2000, pp. 17–24.

[11] Q. Zhang, W. Zhu, and Y.-Q. Zhang, “End-to-end QoS for video delivery over wireless Internet,” *Proc. IEEE*, vol. 93, no. 1, pp. 123–134, Jan. 2005.

[12] B. Vandalore, W.-C. Feng, R. Jain, and S. Fahmy, “A survey of application layer techniques for adaptive streaming of multimedia,” *Real-Time Imag.*, vol. 7, no. 3, pp. 221–235, Jun. 2001.

[13] VQEG Objective Video Quality Model Test Plan, May 7–29, 1998.

[14] M. A. Saad, A. C. Bovik, and C. Charrier, “Blind image quality assessment: A natural scene statistics approach in the DCT domain,” *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[15] M. A. Saad and A. C. Bovik, “Blind quality assessment of videos using a model of natural scene statistics and motion coherency,” in *Conf. Rec. 46th ASILOMAR Signals, Syst. Comput.*, 2012, pp. 332–336.

[16] F. Yang, S. Wan, Q. Xie, and H. R. Wu, “No-reference quality assessment for networked video via primary analysis of bit stream,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1544–1554, Nov. 2010.

[17] X. Lin, H. Ma, L. Luo, and Y. Chen, “No-reference video quality assessment in the compressed domain,” *IEEE Trans. Consum. Electron.*, vol. 58, no. 2, pp. 505–512, May 2012.

[18] S.-O. Lee and D.-G. Sim, “Hybrid bitstream-based video quality assessment method for scalable video coding,” *Opt. Eng.*, vol. 51, no. 6, pp. 067403-1–067403-9, Jun. 2012.

[19] K. Yamagishi and T. Hayashi, “Parametric packet-layer model for monitoring video quality of IPTV services,” in *Proc. IEEE ICC*, 2008, pp. 110–114.

[20] F. Yang, J. Song, S. Wan, and H. R. Wu, "Content-adaptive packet-layer model for quality assessment of networked video services," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 672–683, Oct. 2012.

[21] S. Tao, J. Apostolopoulos, and R. Guérin, "Real-time monitoring of video quality in IP networks," *IEEE/ACM Trans. Netw.*, vol. 16, no. 5, pp. 1052–1065, Oct. 2008.

[22] G. Zhai, J. Cai, W. Lin, X. Yang, and W. Zhang, "Three dimensional scalable video adaptation via user-end perceptual quality assessment," *IEEE Trans. Broadcast.*, vol. 54, no. 3, pp. 719–727, Sep. 2008.

[23] M. Ries, O. Nemethova, and M. Rupp, "Video quality estimation for mobile H.264/AVC video streaming," *J. Commun.*, vol. 3, no. 1, pp. 41–50, Jan. 2008.

[24] K. Yamagishi, T. Kawano, and T. Hayashi, "Hybrid video-quality-estimation model for IPTV services," in *Proc. IEEE Global Telecommun. Conf.*, 2009, pp. 1–5.

[25] R. K. Mok, E. W. Chan, and R. K. Chang, "Measuring the quality of experience of http video streaming," in *Proc. IFIP/IEEE Int. Symp. IM Netw.*, 2011, pp. 485–492.

[26] R. K. Mok, E. W. Chan, X. Luo, and R. K. Chang, "Inferring the QoE of http video streaming from user-viewing activities," in *Proc. 1st ACM SIGCOMM Workshop Meas. Stack*, 2011, pp. 31–36.

[27] A. Khan, L. Sun, and E. Ifeachor, "QoE prediction model and its application in video quality adaptation over UMTS networks," *IEEE Trans. Multimedia*, vol. 14, no. 2, pp. 431–442, Apr. 2012.

[28] Video Quality Experts Group (VQEG), Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, 2000, M.G.T.P.

[29] A. K. Moorthy, K. Seshadrinathan, R. Soundararajan, and A. C. Bovik, "Wireless video quality assessment: A study of subjective scores and objective algorithms," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 4, pp. 587–599, Apr. 2010.

[30] A. Balachandran *et al.*, "Developing a predictive model of quality of experience for Internet video," in *Proc. ACM SIGCOMM*, 2013, pp. 339–350.

[31] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 165–182, Jun. 2011.

[32] W. Lin and C.-C. Jay Kuo, "Perceptual visual quality metrics: A survey," *J. Vis. Commun. Image Represent.*, vol. 22, no. 4, pp. 297–312, May 2011.

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[34] Y. Wang, "Survey of objective video quality measurements," EMC Corporation, Hopkinton, MA, USA, p. 39, 2006, vol. 1748.

[35] M. Yuen and H. Wu, "A survey of hybrid MC/DPCM/DCT video coding distortions," *Signal Process.*, vol. 70, no. 3, pp. 247–278, Nov. 1998.

[36] C. J. Van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatiotemporal model of the human visual system," in *Proc. Int. Soc. Opt. Photon. Electron. Imag.—Sci. Technol.*, 1996, pp. 450–461.

[37] A. Liu, W. Lin, M. Paul, C. Deng, and F. Zhang, "Just noticeable difference for images with decomposition model for separating edge and textured regions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1648–1652, Nov. 2010.

[38] W. Osberger, A. J. Maeder, and D. McLean, "A computational model of the human visual system for image quality assessment," in *Proc. DICTA*, 1997, vol. 97, pp. 337–342.

[39] W. Osberger, N. Bergmann, and A. Maeder, "An automatic image quality assessment technique incorporating higher level perceptual factors," in *Proc. Int. Conf. Image Process.*, 1998, pp. 414–418.

[40] S. Westen, R. Lagendijk, and J. Biemond, "Perceptual image quality based on a multiple channel HVS model," in *Proc. ICASSP*, 1995, vol. 4, pp. 2351–2354.

[41] F. Xiao, "DCT-based video quality evaluation," Final Proj. EE392J Stanford University, 2000.

[42] A. B. Watson, J. Hu, and J. F. McGowan, "Digital video quality metric based on human vision," *J. Electron. Imag.*, vol. 10, no. 1, pp. 20–29, Jan. 2001.

[43] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Conf. Rec. 37th Asilomar Signals, Syst. Comput.*, 2003, vol. 2, pp. 1398–1402.

[44] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[45] U. Ansorge, G. Francis, M. H. Herzog, and H. Öğmen, "Visual masking and the dynamics of human perception, cognition, and consciousness–A century of progress, a contemporary synthesis, and future directions," *Adv. Cogn. Psychol.*, vol. 3, no. 1/2, pp. 1–8, Jul. 2007.

[46] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, Sep. 2007.

[47] H. R. Wu and K. R. Rao, *Digital Video Image Quality and Perceptual Coding*. Boca Raton, FL, USA: CRC Press, 2005.

[48] F. Dobrian *et al.*, "Understanding the impact of video quality on user engagement," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. 362–373, Aug. 2011.

[49] P. Read and M.-P. Meyer, *Restoration of Motion Picture Film*. Oxford, U.K.: Butterworth-Heinemann, 2000.

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

[51] A. Khan, L. Sun, and E. Ifeachor, "Content clustering based video quality prediction model for mpeg4 video streaming over wireless networks," in *Proc. IEEE Int. Conf. Commun.*, 2009, pp. 1–5.

[52] M. Claypool and J. Tanner, "The effects of jitter on the perceptual quality of video," in *Proc. 7th ACM Int. Conf. Multimedia*, 1999, pp. 115–118.

[53] A. Balachandran *et al.*, "A quest for an Internet video quality-of-experience metric," in *Proc. 11th ACM Workshop Hot Topics Netw.*, 2012, pp. 97–102.

[54] H. Chen, S. Ng, and A. R. Rao, "Cultural differences in consumer impatience," *J. Market. Res.*, vol. 42, no. 3, pp. 291–301, Aug. 2005.

[55] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng, "Understanding user behavior in large-scale video-on-demand systems," *ACM SIGOPS Oper. Syst. Rev.*, vol. 40, no. 4, pp. 333–344, Oct. 2006.

[56] M. H. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," in *Proc. Int. Soc. Opt. Photon. Vis. Commun. Image Process.*, 2003, pp. 573–582.

[57] International Telecommunication Union, Methodology for subjective assessment of the quality of television picture, Geneva, Switzerland, ITU-R Rec. BT.500-10, 2000.

[58] International Telecommunication Union, Specifications and alignment procedures for setting of brightness and contrast of displays, Geneva, Switzerland, ITU-R Rec. BT.814-1, 1994.

[59] International Telecommunication Union, Specification of a signal for measurement of the contrast ratio of displays, Geneva, Switzerland, ITU-R Rec. BT.815-1, 1994.

[60] International Telecommunication Union, Subjective assessment of stereoscopic television pictures, Geneva, Switzerland, ITU-R Rec. BT.1438, 2000.

[61] J. Gutierrez, P. Perez, F. Jaureguizar, J. Cabrera, and N. García, "Validation of a novel approach to subjective quality evaluation of conventional and 3D broadcasted video services," in *Proc. 4th Int. Workshop QoMEX*, 2012, pp. 230–235.

[62] A. Sorokin and D. Forsyth, "Utility data annotation with Amazon Mechanical Turk," in *Proc. IEEE Compu. Vis. Pattern Recog. Workshops*, 2008, pp. 1–8.

[63] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, "A crowdsourceable QoE evaluation framework for multimedia content," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 491–500.

[64] Q. Xu, J. Xiong, Q. Huang, and Y. Yao, "Robust evaluation for quality of experience in crowdsourcing," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 43–52.

[65] Q. Xu *et al.*, "Hodgerank on random graphs for subjective video quality assessment," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 844–857, Jun. 2012.

[66] C. Wu, K. Chen, Y. Chang, and C. Lei, "Crowdsourcing multimedia QoE evaluation: A trusted framework," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1121–1137, Aug. 2013.

[67] R. G. Cole and J. H. Rosenbluth, "Voice over IP performance monitoring," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 31, no. 2, pp. 9–24, Apr. 2001.

[68] D. Hands, O. V. Barriac, and F. Telecom, "Standardization activities in the ITU for a QoE assessment of IPTV," *IEEE Commun. Mag.*, vol. 46, no. 2, pp. 78–84, Feb. 2008.

[69] S. Winkler, A. Sharma, and D. McNally, "Perceptual video quality and blockiness metrics for multimedia streaming applications," in *Proc. Int. Symp. Wireless Pers. Multimedia Commun.*, 2001, pp. 547–552.

[70] S. Olsson, M. Stroppiana, and J. Baina, "Objective methods for assessment of video quality: State of the art," *IEEE Trans. Broadcast.*, vol. 43, no. 4, pp. 487–495, Dec. 1997.

[71] A. Punchihewa, D. G. Bailey, and R. Hodgson, "A survey of coded image and video quality assessment," in *Proc. Image Vis. Comput. New Zealand*, 2003, pp. 326–331.

[72] U. Engelke and H.-J. Zepernick, "Perceptual-based quality metrics for image and video services: A survey," in *Proc. 3rd EuroNGI Netw. Conf.*, 2007, pp. 190–197.

[73] S. Winkler, *Digital Video Quality: Vision Models and Metrics*. Hoboken, NJ, USA: Wiley, 2005.

[74] International Telecommunication Union, Reference algorithm for computing Peak Signal to Noise Ratio (PSNR) of a video sequence with a constant delay, Geneva, Switzerland, ITU-T. J.340, 2009.

[75] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, Dec. 1995.

[76] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[77] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?" in *Proc. IEEE ICASSP*, 2002, vol. 4, pp. IV-3313–IV-3316.

[78] A. Schertz, *IRT Tektronix Investigation of Subjective and Objective Picture Quality for 2–10 Mbit-sec MPEG-2 Video*. Munchen, Germany: Institut für Rundfunktechnik GmbH, 1997.

[79] A. P. Hekstra *et al.*, "PVQM–A perceptual video quality measure," *Signal Process. Image Commun.*, vol. 17, no. 10, pp. 781–798, Nov. 2002.

[80] A. Bhat, I. Richardson, and S. Kannangara, "A new perceptual quality metric for compressed video," in *Proc. IEEE ICASSP*, 2009, pp. 933–936.

[81] J. You, T. Ebrahimi, and A. Perkis, "Attention driven foveated video quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 200–213, Jan. 2014.

[82] Video Quality Experts Group, Report on the Validation of Video Quality Models for High Definition Video Content, 2010.

[83] J. Mannos and D. Sakrison, "The effects of a visual fidelity criterion of the encoding of images," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 4, pp. 525–536, Jul. 1974.

[84] P. J. Bex and W. Makous, "Spatial frequency, phase, and the contrast of natural images," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 19, no. 6, pp. 1096–1106, Jun. 2002.

[85] D. Navon, "Forest before trees: The precedence of global features in vi sual perception," *Cogn. Psycholo.*, vol. 9, no. 3, pp. 353–383, Jul. 1977.

[86] D. M. Chandler, K. H. Lim, and S. S. Hemami, "Effects of spatial correlations and global precedence on the visual fidelity of distorted images," in *Proc. Int. Soc. Opt. Photon. Electron. Imag.*, 2006, pp. 60570F-1–60570F-15.

[87] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process. Image Commun.*, vol. 19, no. 2, pp. 121–132, Jan. 2004.

[88] Z. Wang and E. P. Simoncelli, "An adaptive linear system framework for image distortion analysis," in *Proc. IEEE ICIP*, 2005, vol. 3, pp. III-1160.1–III-1160.3.

[89] Z. Wang and E. P. Simoncelli, "Stimulus synthesis for efficient evaluation and refinement of perceptual image quality metrics," in *Proc. Int. Soc. Opt. Photon. Electron. Imag.*, 2004, pp. 99–108.

[90] Z. Wang and X. Shang, "Spatial pooling strategies for perceptual image quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, 2006, pp. 2945–2948.

[91] A. K. Moorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 193–201, Apr. 2009.

[92] E. Ong, X. Yang, W. Lin, Z. Lu, and S. Yao, "Video quality metric for low bitrate compressed videos," in *Proc. ICIP*, 2004, vol. 5, pp. 3531–3534.

[93] E. Ong, W. Lin, Z. Lu, and S. Yao, "Colour perceptual video quality metric," in *Proc. IEEE ICIP*, 2005, vol. 3, pp. III-1172.1–III-1172.5.

[94] P. Ndjiki-Nya, M. Barrado, and T. Wiegand, "Efficient full-reference assessment of image and video quality," in *Proc. IEEE ICIP*, 2007, vol. 2, pp. II-125–II-128.

[95] 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.

[96] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[97] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *Proc. 1st Int. Workshop Video Process. Qual. Metrics Consum. Electron.*, 2005, pp. 23–25.

[98] S.-O. Lee and D.-G. Sim, "New full-reference visual quality assessment based on human visual perception," in *Proc. ICCE Dig. Tech. Papers*, 2008, pp. 1–2.

[99] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video quality assessment," *IEEE Trans. Multimedia*, vol. 4, no. 1, pp. 129–132, Mar. 2002.

[100] Z. Wang, A. C. Bovik, L. Lu, and J. L. Kouloheris, "Foveated wavelet image quality index," in *Proc. Int. Soc. Opt. Photon.—Int. Symp. Opt. Sci. Technol.*, 2001, pp. 42–52.

[101] S. Rimac-Drlje, M. Vranješ, and D. Žagar, "Foveated mean squared error—A novel video quality metric," *Multimedia Tools Appl.*, vol. 49, no. 3, pp. 425–445, Sep. 2010.

[102] W. S. Geisler and J. S. Perry, "A real-time foveated multiresolution system for low-bandwidth video communication," in *Proc. SPIE*, 1998, pp. 294–305.

[103] J. You, A. Perkis, M. M. Hannuksela, and M. Gabbouj, "Perceptual quality assessment based on visual attention analysis," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 561–564.

[104] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Process.*, vol. 6, no. 8, pp. 1164–1175, Aug. 1997.

[105] D. J. Fleet and A. D. Jepson, "Computation of component image velocity from local phase information," *Int. J. Comput. Vis.*, vol. 5, no. 1, pp. 77–104, Aug. 1990.

[106] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 24, no. 12, pp. B61–B69, Dec. 2007.

[107] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu, "On advances in statistical modeling of natural images," *J. Math. Imag. Vis.*, vol. 18, no. 1, pp. 17–33, Jan. 2003.

[108] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Ann. Rev. Neurosci.*, vol. 24, no. 1, pp. 1193–1216, 2001.

[109] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[110] B. A. Wandell, *Foundations of Vision*. Sunderland, MA, USA: Sinauer Associates, 1995.

[111] A. A. Stocker and E. P. Simoncelli, "Noise characteristics and prior expectations in human visual speed perception," *Nat. Neurosci.*, vol. 9, no. 4, pp. 578–585, Apr. 2006.

[112] U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack, "GAFFE: A gaze-attentive fixation finding engine," *IEEE Trans. Image Process.*, vol. 17, no. 4, pp. 564–573, Apr. 2008.

[113] A. R. Reibman, S. Kanumuri, V. Vaishampayan, and P. C. Cosman, "Visibility of individual packet losses in MPEG-2 video," in *Proc. ICIP*, 2004, vol. 1, pp. 171–174.

[114] S. Kanumuri, P. Cosman, and A. R. Reibman, "A generalized linear model for MPEG-2 packet-loss visibility," in *Proc. 14th Int. PV Workshop*, 2004, pp. 1–9.

[115] 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.

[116] 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. IEEE Image Process.*, 2006, pp. 2245–2248.

[117] A. R. Reibman and D. Poole, "Characterizing packet-loss impairments in compressed video," in *Proc. IEEE ICIP*, 2007, vol. 5, pp. V-77–V-80.

[118] A. R. Reibman and D. Poole, "Predicting packet-loss visibility using scene characteristics," in *Proc. Packet Video*, 2007, pp. 308–317.

[119] T.-L. Lin *et al.*, "A versatile model for packet loss visibility and its application to packet prioritization," *IEEE Trans. Image Process.*, vol. 19, no. 3, pp. 722–735, Mar. 2010.

[120] L. Breiman, *Classification and Regression Trees*. Boca Raton, NJ, USA: CRC Press, 1993.

[121] P. MacCullagh and J. A. Nelder, *Generalized Linear Models*, vol. 37. Boca Raton, NJ, USA: CRC Press, 1989.

[122] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," in *Proc. Int. Soc. Opt. Photon. Electron. Imag.*, 2005, pp. 149–159.

[123] Z. Wang *et al.*, "Quality-aware images," *IEEE Trans. Image Process.*, vol. 15, no. 6, pp. 1680–1689, Jun. 2006.

[124] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 202–211, Apr. 2009.

[125] W. Xue and X. Mou, "Reduced reference image quality assessment based on Weibull statistics," in *Proc. 2nd Int. Workshop QoMEX*, 2010, pp. 1–6.

[126] R. Soundararajan and A. C. Bovik, "RRED indices: Reduced reference entropic differencing framework for image quality assessment," in *Proc. IEEE ICASSP*, 2011, pp. 1149–1152.

[127] A. A. Abdelouahad, M. El Hassouni, H. Cherifi, and D. Aboutajdine, "Image quality assessment measure based on natural image statistics in the tetrolet domain," in *Image and Signal Processing*. Berlin, Germany: Springer-Verlag, 2012, pp. 451–458.

[128] A. Rehman and Z. Wang, "Reduced-reference image quality assessment by structural similarity estimation," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3378–3389, Aug. 2012.

[129] P. Le Callet and F. Autrusseau, Subjective Quality Assessment IRCCYN/IVC Database, 2005.

[130] Y. Horita, K. Shibata, Y. Kawayoke, and Z. P. Sazzad, MICT Image Quality Evaluation Database, 2011.

[131] N. Ponomarenko *et al.*, "TID2008—A database for evaluation of full-reference visual quality assessment metrics," *Adv. Mod. Radioelectron.*, vol. 10, no. 4, pp. 30–45, 2009.

[132] E. C. Larson and D. Chandler, Categorical Image Quality (CSIQ) Database, 2010. [Online]. Available: http://vision.okstate.edu/csiq

[133] Z. Wang and A. C. Bovik, "Reduced-and no-reference image quality assessment," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 29–40, Nov. 2011.

[134] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.

[135] I. P. Gunawan and M. Ghanbari, "Reduced-reference picture quality estimation by using local harmonic amplitude information," in *Proc. London Commun. Symp.*, 2003, pp. 137–140.

[136] D. Tao, X. Li, W. Lu, and X. Gao, "Reduced-reference IQA in contourlet domain," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 6, pp. 1623–1627, Dec. 2009.

[137] A. Maalouf, M.-C. Larabi, and C. Fernandez-Maloigne, "A grouplet-based reduced reference image quality assessment," in *Proc. Int. Workshop QoMEx*, 2009, pp. 59–63.

[138] M. Carnec, P. Le Callet, and D. Barba, "Objective quality assessment of color images based on a generic perceptual reduced reference," *Signal Process. Image Commun.*, vol. 23, no. 4, pp. 239–256, Apr. 2008.

[139] F. Yang and S. Wan, "Bitstream-based quality assessment for networked video: A review," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 203–209, Nov. 2012.

[140] K. Brunnstrom, D. Hands, F. Speranza, and A. Webster, "VQEG validation and ITU standardization of objective perceptual video quality metrics [standards in a nutshell]," *IEEE Signal Process. Mag.*, vol. 26, no. 3, pp. 96–101, May 2009.

[141] R. S. A. K. Moorthy, K. Seshadrinathan, and A. C. Bovik, Live Wireless Video Quality Assessment Database, 2009. [Online]. Available: http://live.ece.utexas.edu/research/quality/live_wireless_video.html

[142] D. M. Chandler, "Seven challenges in image quality assessment: Past, present, and future research," *ISRN Signal Process.*, vol. 2013, pp. 905 685-1–905 685-53, 2013.

[143] T. N. Pappas, T. A. Michel, and R. O. Hinds, "Supra-threshold perceptual image coding," in *Proc. Int. Conf. Image Process.*, 1996, vol. 1, pp. 237–240.

[144] D. M. Chandler and S. S. Hemami, "Dynamic contrast-based quantization for lossy wavelet image compression," *IEEE Trans. Image Process.*, vol. 14, no. 4, pp. 397–410, Apr. 2005.

[145] M. Vilas, X. G. Pañeda, R. García, D. Melendi, and V. G. García, "User behavior analysis of a video-on-demand service with a wide variety of subjects and lengths," in *Proc. 31st EUROMICRO Conf. Softw. Eng. Adv. Appl.*, 2005, pp. 330–337, IEEE.

[146] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "Youtube traffic characterization: A view from the edge," in *Proc. 7th ACM SIGCOMM Conf. Internet Meas.*, 2007, pp. 15–28.

[147] X. Hei, C. Liang, J. Liang, Y. Liu, and K. W. Ross, "A measurement study of a large-scale P2P IPTV system," *IEEE Trans. Multimedia*, vol. 9, no. 8, pp. 1672–1687, Dec. 2007.

[148] H. Yin *et al.*, "Inside the bird's nest: Measurements of large-scale live vod from the 2008 olympics," in *Proc. 9th ACM SIGCOMM Conf. Internet Meas. Conf.*, 2009, pp. 442–455.

[149] A. Finamore, M. Mellia, M. M. Munafò, R. Torres, and S. G. Rao, "Youtube everywhere: Impact of device and infrastructure synergies on user experience," in *Proc. ACM SIGCOMM Conf. Internet Meas. Conf.*, 2011, pp. 345–360.

[150] L. Chen, Y. Zhou, and D. M. Chiu, "Video browsing—A study of user behavior in online VoD services," in *Proc. 22nd ICCCN*, 2013, pp. 1–7.

[151] T. De Pessemier, K. De Moor, W. Joseph, L. De Marez, and L. Martens, "Quantifying the influence of rebuffering interruptions on the user's quality of experience during mobile video watching," *IEEE Trans. Broadcast.*, vol. 59, no. 1, pp. 47–61, Mar. 2013.

[152] S. S. Krishnan and R. K. Sitaraman, "Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs," in *Proc. ACM Conf. Internet Meas. Conf.*, 2012, pp. 211–224.

[153] A. Balachandran, V. Sekar, A. Akella, and S. Seshan, "Analyzing the potential benefits of CDN augmentation strategies for Internet video workloads," in *Proc. Internet Meas. Conf.*, 2013, pp. 43–56.

[154] I. Ullah, G. Doyen, G. Bonnet, and D. Gaiti, "A survey and synthesis of user behavior measurements in P2P streaming systems," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 3, pp. 734–749, 2012.

[155] A. U. Mian, Z. Hu, and H. Tian, "A decision theoretic approach for in-service QoE estimation and prediction of P2P live video streaming systems based on user behavior modeling and context awareness," *JICS*, vol. 10, no. 11, pp. 3429–3436, 2013.

[156] I. Ullah, G. Doyen, G. Bonnet, and D. Gaiti, "User behavior anticipation in P2P live video streaming systems through a Bayesian network," in *Proc. IFIP/IEEE Int. Symp. IM Netw.*, 2011, pp. 337–344.

[157] W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA, USA: Houghton Mifflin, 2002.

[158] D. Wang, P. C. Cosman, and L. B. Milstein, "Cross layer resource allocation design for uplink video OFDMA wireless systems," in *Proc. IEEE GLOBECOM*, 2011, pp. 1–6.

[159] Y. P. Fallah, H. Mansour, S. Khan, P. Nasiopoulos, and H. M. Alnuweiri, "A link adaptation scheme for efficient transmission of H.264 scalable video over multirate WLANs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 7, pp. 875–887, Jul. 2008.

[160] Y. Zhang, W. Gao, Y. Lu, Q. Huang, and D. Zhao, "Joint source-channel rate-distortion optimization for H.264 video coding over error-prone networks," *IEEE Trans. Multimedia*, vol. 9, no. 3, pp. 445–454, Apr. 2007.

[161] Y. Wang, M. van der Schaar, S.-F. Chang, and A. C. Loui, "Classification-based multidimensional adaptation prediction for scalable video coding using subjective quality evaluation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 10, pp. 1270–1279, Oct. 2005.

[162] A. A. Khalek, C. Caramanis, and R. Heath, "A cross-layer design for perceptual optimization of H.264/SVC with unequal error protection," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 7, pp. 1157–1171, Aug. 2012.

[163] L. Toni, P. C. Cosman, and L. B. Milstein, "Channel coding optimization based on slice visibility for transmission of compressed video over OFDM channels," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 7, pp. 1172–1183, Aug. 2012.

[164] H.-P. Shiang and M. van der Schaar, "Media-TCP: A quality-centric TCP-friendly congestion control for multimedia transmission," *arXiv preprint*, vol. arXiv:0910.4186, 2009.

[165] O. Habachi, Y. Hu, M. van der Schaar, Y. Hayel, and F. Wu, "MOS-based congestion control for conversational services in wireless environments," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 7, pp. 1225–1236, Aug. 2012.

[166] M. A. Masry and S. S. Hemami, "A metric for continuous quality evaluation of compressed video with severe distortions," *Signal Process. Image Commun.*, vol. 19, no. 2, pp. 133–146, Feb. 2004.

[167] K. Seshadrinathan and A. C. Bovik, "Temporal hysteresis model of time varying subjective video quality," in *Proc. IEEE ICASSP*, 2011, pp. 1153–1156.

[168] S. Singh, J. G. Andrews, and G. de Veciana, "Interference shaping for improved quality of experience for real-time video streaming," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 7, pp. 1259–1269, Aug. 2012.

[169] E. Ibarrola, F. Liberal, I. Taboada, and R. Ortega, "Web QoE evaluation in multi-agent networks: Validation of ITU-TG.1030," in *Proc. 5th ICAS*, 2009, pp. 289–294.

[170] P. Reichl, S. Egger, R. Schatz, and A. D'Alconzo, "The logarithmic nature of QoE and the role of the Weber–Fechner law in QoE assessment," in *Proc. IEEE ICC*, 2010, pp. 1–5.

[171] S.-T. Sheu and T.-F. Sheu, "A bandwidth allocation/sharing/extension protocol for multimedia over IEEE 802.11 ad hoc wireless LANs," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 10, pp. 2065–2080, Oct. 2001.

[172] K. Piamrat, A. Ksentini, C. Viho, and J.-M. Bonnin, "QoE-aware admission control for multimedia applications in IEEE 802.11 wireless networks," in *Proc. IEEE 68th VTC—Fall*, 2008, pp. 1–5.

[173] K. Piamrat, A. Ksentini, C. Viho, and J.-M. Bonnin, "QoE-based network selection for multimedia users in IEEE 802.11 wireless networks," in *Proc. 33rd IEEE LCN*, 2008, pp. 388–394.

[174] G. Rubino, M. Varela, and J.-M. Bonnin, "Controlling multimedia QoS in the future home network using the PSQA metric," *Comput. J.*, vol. 49, no. 2, pp. 137–155, Mar. 2006.

[175] T. Jiang, H. Wang, and A. V. Vasilakos, "QoE-driven channel allocation schemes for multimedia transmission of priority-based secondary users over cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 7, pp. 1215–1224, Aug. 2012.

[176] J.-G. Kim, Y. Wang, and S.-F. Chang, "Content-adaptive utility-based video adaptation," in *Proc. ICME*, 2003, vol. 3, pp. III-281.1–III-281.4.

[177] Y. Wang, J.-G. Kim, and S.-F. Chang, "Content-based utility function prediction for real-time MPEG-4 video transcoding," in *Proc. ICIP*, 2003, vol. 1, pp. I-189–I-192.

[178] J. Villalón, P. Cuenca, L. Orozco-Barbosa, Y. Seok, and T. Turletti, "Cross-layer architecture for adaptive video multicast streaming over multirate wireless LANs," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 4, pp. 699–711, May 2007.

[179] M. A. Santos, J. Villalón, and L. Orozco-Barbosa, "A novel QoE-aware multicast mechanism for video communications over IEEE 802.11 WLANs," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 7, pp. 1205–1214, Aug. 2012.

[180] M. Draxler and H. Karl, "Cross-layer scheduling for multi-quality video streaming in cellular wireless networks," in *Proc. 9th IWCMC*, 2013, pp. 1181–1186.

[181] A. Bokani, M. Hassan, and S. Kanhere, "HTTP-based adaptive streaming for mobile clients using Markov decision process," in *Proc. 20th Int. PV Workshop (PV)*, 2013, pp. 1–8.

[182] J. Jiang, V. Sekar, and H. Zhang, "Improving fairness, efficiency, and stability in HTTP-based adaptive video streaming with festive," in *Proc. 8th Int. Conf. Emerging Netw. Exp. Technol.*, 2012, pp. 97–108.

[183] J. Chen, R. Mahindra, M. A. Khojastepour, S. Rangarajan, and M. Chiang, "A scheduling framework for adaptive video delivery over cellular networks," in *Proc. 19th Annu. Int. Conf. Mobile Comput. Netw.*, 2013, pp. 389–400.

[184] G. Saygili, C. G. Gurler, and A. M. Tekalp, "Evaluation of asymmetric stereo video coding and rate scaling for adaptive 3D video streaming," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 593–601, Jun. 2011.

[185] S. S. Savas, C. G. Gurler, and A. M. Tekalp, "Evaluation of adaptation methods for multi-view video," in *Proc. 19th IEEE ICIP*, 2012, pp. 2273–2276.

[186] C. Hewage *et al.*, "Quality evaluation of asymmetric compression for 3d surgery video," in *Proc. IEEE 15th Int. Conf. e-Healthcom Netw., Appl. Serv.*, Oct. 2013, pp. 680–684.

[187] C. T. Hewage, M. G. Martini, M. Brandas, and D. De Silva, "A study on the perceived quality of 3d video subject to packet losses," in *Proc. IEEE ICC Workshops*, 2013, pp. 662–666.

[188] W. Wu *et al.*, "Quality of experience in distributed interactive multimedia environments: Toward a theoretical framework," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 481–490.

[189] S. Egger, M. Ries, and P. Reichl, "Quality-of-experience beyond MOS: Experiences with a holistic user test methodology for interactive video services," in *Proc. 21st ITC Spec. Semin. Multimedia Appl.-Traffic, Perform. QoE*, 2010, pp. 13–18.

[190] J. Short, E. Williams, and B. Christie, *The Social Psychology of Telecommunications*. London, U.K.: Wiley, 1976.

[191] P. Calyam, E. Ekici, C.-G. Lee, M. Haffner, and N. Howes, "A 'gap-model' based framework for online vvoip QOE measurement," *J. Commun. Netw.*, vol. 9, no. 4, pp. 446–456, Dec. 2007.

[192] International Telecommunication Union, Parameter values for ultra-high definition television systems for production and international programme exchange, Geneva, Switzerland, ITU-R Rec. BT.20207, 2012.

[193] P. Hanhart, M. Rerabek, F. De Simone, and T. Ebrahimi, "Subjective quality evaluation of the upcoming HEVC video compression standard," in *Proc. Int. Soc. Opt. Photon. SPIE Opt. Eng. Appl.*, 2012, p. 84990V.

[194] M. Horowitz *et al.*, "Informal subjective quality comparison of video compression performance of the HEVC and H.264/MPEG-4 AVC standards for low-delay applications," in *Proc. SPIE Int. Soc. Opt. Photon. Opt. Eng. Appl.*, 2012, pp. 84990W-1–84990W-6.

[195] M. T. Pourazad, C. Doutre, M. Azimi, and P. Nasiopoulos, "HEVC: The new gold standard for video compression: How does HEVC compare with H.264/AVC?" *IEEE Consum. Electron. Mag.*, vol. 1, no. 3, pp. 36–46, Jul. 2012.

[196] S. Jumisko-Pyykkö and M. M. Hannuksela, "Does context matter in quality evaluation of mobile television?" in *Proc. 10th Int. Conf. Human Comput. Interaction Mobile Devices Serv.*, 2008, pp. 63–72.

[197] K. De Moor *et al.*, "Proposed framework for evaluating quality of experience in a mobile, testbed-oriented living lab setting," *Mobile Netw. Appl.*, vol. 15, no. 3, pp. 378–391, Jun. 2010.

[198] T. De Pessemier, K. De Moor, W. Joseph, L. De Marez, and L. Martens, "Quantifying subjective quality evaluations for mobile video watching in a semi-living lab context," *IEEE Trans. Broadcast.*, vol. 58, no. 4, pp. 580–589, Dec. 2012.

[199] I. F. Akyildiz, T. Melodia, and K. R. Chowdhury, "A survey on wireless multimedia sensor networks," *Comput. Netw.*, vol. 51, no. 4, pp. 921–960, Mar. 2007.

[200] S. Ehsan and B. Hamdaoui, "A survey on energy-efficient routing techniques with QoS assurances for wireless multimedia sensor networks," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 2, pp. 265–278, 2012.

[201] S. Pudlewski, A. Prasanna, and T. Melodia, "Compressed-sensing-enabled video streaming for wireless multimedia sensor networks," *IEEE Trans. Mobile Comput.*, vol. 11, no. 6, pp. 1060–1072, Jun. 2012.

[202] P. Bucciol, E. Masala, N. Kawaguchi, K. Takeda, and J. De Martin, "Performance evaluation of H.264 video streaming over inter-vehicular 802.11 ad hoc networks," in *Proc. IEEE 16th Int Symp. PIMRC*, 2005, vol. 3, pp. 1936–1940.

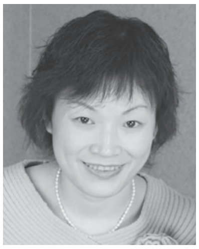
[203] F. Xie, K. A. Hua, W. Wang, and Y. H. Ho, "Performance study of live video streaming over highway vehicular ad hoc networks," in *Proc. IEEE 66th VTC*, 2007, pp. 2121–2125.

[204] I. Rozas-Ramallal, T. M. Fernandez-Carames, A. Dapena, and P. A. Cuenca-Castillo, "Improving performance of H.264/AVC transmissions over vehicular networks," in *Proc. IFIP/IEEE Int. Symp. IM Netw.*, 2013, pp. 1324–1327.

[205] E. Nygren, R. K. Sitaraman, and J. Sun, "The Akamai network: A platform for high-performance Internet applications," *ACM SIGOPS Oper. Syst. Rev.*, vol. 44, no. 3, pp. 2–19, Jul. 2010.

Yanjiao Chen received the B.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 2010. She is currently working toward the Ph.D. degree with The Hong Kong University of Science and Technology, Kowloon, Hong Kong. Her research interests include spectrum management for femtocell networks and network economics.

Kaishun Wu received the Ph.D. degree in computer science and engineering from The Hong Kong University of Science and Technology (HKUST), Kowloon, Hong Kong, in 2011. He is currently a Research Assistant Professor with the Fok Ying Tung Graduate School, HKUST. He is also currently with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. His research interests include wireless communication, mobile computing, wireless sensor networks, and data center networks.



Qian Zhang (F'11) received the B.S., M.S., and Ph.D. degrees from Wuhan University, Wuhan, China, in 1994, 1996, and 1999, respectively, all in computer science. In September 2005, he joined The Hong Kong University of Science and Technology, Kowloon, Hong Kong, where she is currently a Full Professor with the Department of Computer Science and Engineering. Before that, she was a Research Manager of the Wireless and Networking Group with Microsoft Research Asia, Beijing, China, from July 1999. He has authored or coauthored about 300 refereed papers in international leading journals and key conferences in the areas of wireless/Internet multimedia networking, wireless communications and networking, wireless sensor networks, and overlay networking. Her current research is on cognitive and cooperative networks, dynamic spectrum access and management, and wireless sensor networks. Dr. Zhang is a Fellow of the IEEE for her "contribution to the mobility and spectrum management of wireless networks and mobile communications." She has been a recipient of the MIT TR100 (MIT Technology Review) World Top Young Innovator Award. She was also a recipient of the Best Asia Pacific Young Researcher Award elected by the IEEE Communication Society in 2004.