Quality Assessment of Multidimensional Video Scalability

Jong-Seok Lee, Yonsei University
Francesca De Simone and Touradj Ebrahimi, Ecole Polytechnique Fédérale de Lausanne (EPFL)
Naeem Ramzan and Ebroul Izquierdo, Queen Mary University of London

Abstract

Scalability is a powerful concept for adaptive video content delivery to many end users having heterogeneous and dynamic characteristics of networks and devices. In order to maximize users' quality of experience by selecting appropriate combinations of multiple scalability parameters, it is crucial to understand and model the relationship between multidimensional scalability and perceived quality. In this article, we address the latest advances in subjective and objective quality evaluation of multidimensional video scalability for optimal content distribution, present their applications, and discuss future trends and challenges.

Introduction

Video content delivery over networks has become popular due to the users' increasing demand for video content and remarkable evolution of network technologies such as videoconferencing, online surveillance, Internet Protocol television (IPTV), and streaming. The delivery chain includes two fundamental steps: lossy video compression and content distribution. The former aims at reducing the amount of data being transferred, while minimizing visibility of the resulting coding artifacts. The latter aims at providing contents to users effectively via transmission over physical networks. Since the ultimate users of the chain are human beings, both compression and distribution strategies should maximize the users' satisfaction in terms of visual quality, latency, and so on. Moreover, the same content needs to be delivered in different formats simultaneously to multiple users having different network conditions (e.g., bandwidth limitation) and end-user terminal characteristics (e.g., decoding and display capabilities).

In order to handle such a non-homogeneous and dynamic distribution scenario, scalability emerged in the field of video coding as an efficient alternative to simulcast encoding. From a single compressed scalable bitstream, a number of decodable video streams can be extracted, corresponding to various operating points in terms of spatial resolution (i.e., frame size), temporal resolution (i.e., frame rate), or reconstruction accuracy (i.e., signal-to-noise ratio [SNR] or quality) of the decoded video sequence. The bitstream description indicates the positions of the bitstream portions representing various combinations of scalability options. Thus, the extractor parses the bitstream and decides which portions to keep and which to discard. By leaving out parts of the stream, the bit rate can be adapted flexibly during transmission. The adapted bitstream is also scalable and thus can be fed into the extractor again if further adaptation is required (Fig. 1).

Scalable video compression evolved from two main branches of conventional non-scalable compression: block-based hybrid compression and wavelet-based motion-compensated compression techniques. They exploit the fact that there is statistical redundancy in any visual signal, such as spatial correlation among neighboring samples within an image frame and temporal correlation among samples in temporally adjacent frames, as well as the fact that the human visual system (HVS) can be thought of as a low-pass filter, so some information can be removed without human eyes noticing the loss (irrelevancy or psycho-visual redundancy reduction). Based on these principles, predictive and transform coding technologies have been developed. In predictive coding, information already sent or available (reference frame) is used to predict future values in the same frame (intra-prediction) or in other frames (inter-prediction or motion-compensated prediction), and the difference (prediction error) is quantized, losslessly encoded, and transmitted. In transform coding, the visual information is transformed from the spatial domain representation to a frequency representation, using transforms such as discrete cosine transform (DCT) or discrete wavelet transform (DWT). Then the transformed values (coefficients) are quantized to remove psycho-visual redundancy, and finally losslessly encoded to remove any statistical redundancy. Higher spatial frequencies are quantized more since HVS is less sensitive to them.

0163-6804/12/$25.00 © 2012 IEEE

Block-based hybrid compression is a combination of motion-compensated prediction to exploit temporal redundancy and intra-prediction and transform coding of prediction errors to reduce spatial redundancy. Each frame is partitioned into macroblocks that are coded using either intra-prediction or motion-compensated prediction. The former is for an intra-picture, that is, a frame whose macroblocks are all coded without referring to other pictures in the video sequence. In the latter, for each block, a displacement vector (i.e., the corresponding position of the block used for prediction in the reference image) is estimated by searching for the best matching reference block minimizing an error measure within a search range. Then only the motion information and prediction errors are encoded, where DCT is typically used for the latter.

Wavelet-based motion-compensated compression is based on the same principles as hybrid compression, but a full image transform (i.e., DWT) is used instead of block-based DCT; that is, motion vectors and DWT coefficients of the error after motion estimation are quantized and compressed.

In both these compression schemes, inevitable distortions appear in the decoded video due to the quantization process. Their visibility depends on the coding bit rate, and on the spatial detail and the amount of motion in the content. In general, quantization of block-wise high spatial frequency DCT coefficients results in the effects of a quantization error in any DCT coefficient to propagate without attenuation throughout the data block, which causes blurring, blockiness, and ringing artifacts [1]. However, wavelet-based compression does not result in blockiness artifacts, since it is based on full image transform and short time localization. Instead, blurring, pattern aliasing, ringing, and temporal fluctuation noise tend to be introduced because perfect reconstruction is impossible due to quantization in the transform domain [1]. These distortions are also present in the scalable version of the block-based or wavelet-based compression scheme.

For content distribution using a scalable video stream, it is necessary to select an appropriate combination of scalability options for given transmission and receiver conditions. Particularly, adjusting scalability options may introduce artifacts when the network resources are not sufficient, and different scalability options cause different spatial and temporal artifacts. Frame rate reduction results in jerkiness, frame size reduction causes blurring when upscaling is done for a fixed viewing window, and frame quality reduction typically results in blockiness for block-based encoding or increased ringing and loss of fine details for wavelet-based encoding. Figure 2 shows examples of blurring and blockiness resulting from different scalability options.

Thus, an optimization problem needs to be addressed: what is the optimal combination of the scalability options so as to maximize the quality of experience (QoE) of the delivered content for each target transmission and receiver condition? This is a challenging issue because of the complicated nature of the human judgment of visual quality, which can be thought of as the result of three levels of interaction with the media: a sensorial interaction, which refers to the simple conscious experience associated to a stimulus (i.e., the first contact between the human organism and the stimulus); a perceptual interaction, which is the conscious experience of the visual content carried by the stimulus; and an emotional interaction [2]. Knowledge about human sensorial, perceptual, and emotional mechanisms derives from many disciplines such as biological, psychophysical, and sociological studies, and is still far from complete. Particularly regarding human quality perception of video scalability, it is difficult to understand and model it in a multidimensional space involving spatial, temporal, and quality variations (and their corresponding artifacts) as well as application- and content-dependent expectations of users. Moreover, expressing and comparing quality across scalability dimensions on a unified scale is not straightforward.

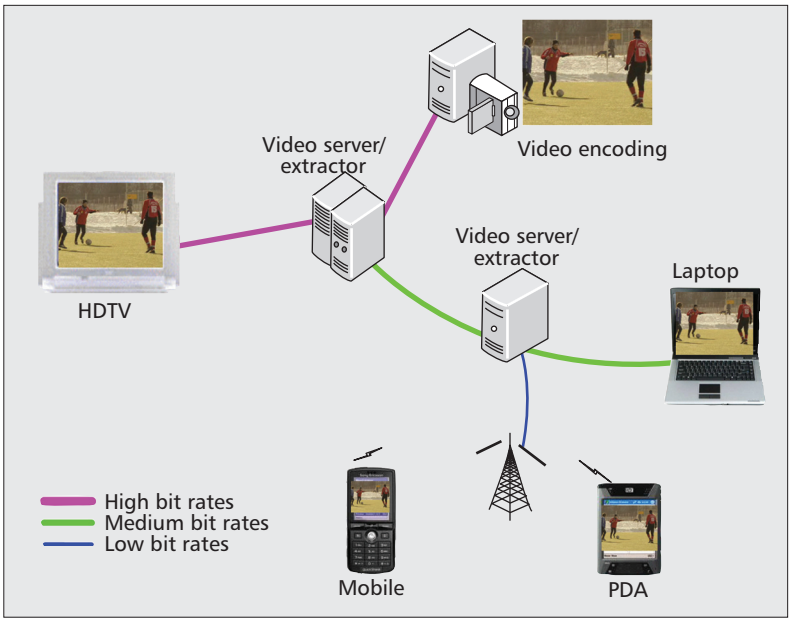


Figure 1. *Typical scalable video distribution chain and types of scalabilities by going to lower-rate decoding.*

The goal of this article is to provide an overview of the existing studies trying to overcome this challenge. Studies of human quality perception of video scalability and related modeling efforts are reviewed. It is important to highlight that different studies are usually based on different conditions including contents, codecs, bit rates, scalability ranges, subjective test environments, and so on, which are mentioned. Inconsistent results partially due to such differences between different studies are also discussed.

Basic Concepts

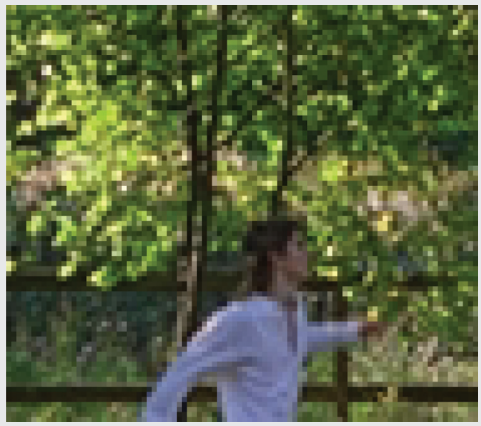
Subjective Quality Assessment

As human subjects usually act as end users of digital content, subjective tests are performed to measure the perceived quality in the context of multimedia services and applications.

Although subjective quality assessment provides highly informative and reliable results, it is usually expensive and time-consuming. Furthermore, it cannot be applied for real-time in-service quality evaluation.

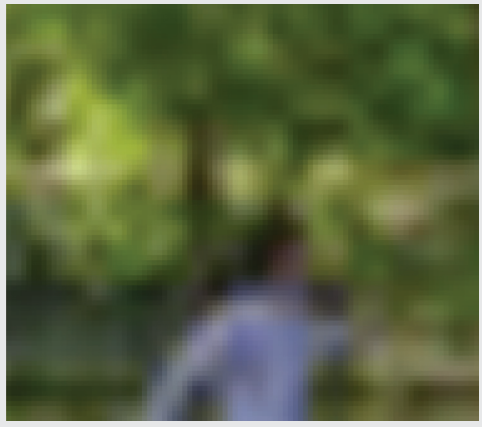
Subjective experiments have to be carried out with scientific rigor. They must be conducted in controlled environments with a significant number of subjects by following a methodology suitable for the given test objective, in order to ensure reproducibility and reliability of the results. Also, the test material needs to be carefully selected, including diverse contents, if possible, spanning all quality levels evenly. International standards provide guidelines for subjective test activities (e.g., International Telecommunication Union Radiocommunication Sector [ITU-R] BT.500-11). Three main categories of the stimulus presentation and rating procedure are defined: double stimulus methods (sequential presentation of each pair of a reference and a test stimuli), single stimulus methods (presentation and rating of the test stimuli only), and stimulus comparison methods (presentation of pairs of stimuli and rating of relative quality). The rating scale can be either continuous or discrete, and either numerical or categorical. Finally, the rating can be done after each stimulus presentation for assessing the overall quality, or continuously during presentation for assessing temporal quality variations. Examples of frequently used test methodologies are: single or double stimulus continuous quality scale (SSCQS or DSCQS), where a test video sequence or a pair of reference and test video sequences are played and rated on a continuous scale spanning {"bad," "good," "fair," "poor," "excellent"}; double stimulus impairment scale (DSIS), which is similar to DSCQS, but only the test sequence is assessed in comparison to the unimpaired reference on a discrete impairment scale {"imperceptible," "perceptible but not annoying," "slightly annoying," "annoying," "very annoying"}; stimulus comparison (SC), where two stimuli are shown and the relative quality of one against the other is assessed, for example, {"worse," "same," "better"}.

General guidelines for processing subjective test results are provided in ITU-R BT.500-11. First, the scores are screened to detect and exclude outliers whose ratings significantly deviate from the panel behavior. Then the scores of the subjects for each stimulus are summarized by two representative measures, the mean opinion score (MOS) or differential mean opinion score (DMOS) for single or double stimulus methods, respectively, and corresponding confidence interval. More thorough analysis of the results require appropriate statistical tools according to the properties of the data.



(a)

(b)



(c)

Figure 2. *Example of visual distortions resulting from different combinations of scalability options for scalable video coding (SVC), the scalable extension of the latest H.264/AVC: a) the original video frame; b) blockiness in the video frame encoded with SVC at a low bit rate; c) blurring in the video frame encoded with SVC at 1/16 of the original resolution and upscaled to the original resolution for presentation to the user.*

Objective Quality Assessment

Although subjective quality assessment provides highly informative and reliable results, it is usually expensive and time-consuming. Furthermore, it cannot be applied to real-time in-service quality evaluation. Therefore, objective quality metrics have been developed to predict outcomes of subjective evaluation.

Objective metrics can be classified into three categories according to the availability of the original (reference) signal besides the test signal: full reference (FR) metrics, when the original signal is accessible; reduced reference (RR) metrics, when description or parameters of the original signal are available; no reference (NR) metrics, when the original signal is not available. FR metrics can be used in offline scenarios for designing and optimizing video processing algorithms as replacements of or in conjunction with

Ref.	Scalability[1]	Codec	# subjects	# content	Methodology	Stimuli[2]
[5] Exp. 1	T, R	H.263+	19	5	DSCQS	S: 320 × 192; T: 7.5, 15, 30 Hz; R: 5 quantization parameters (QPs)
[6]	T, R	MCSBC	31	128	DSIS	S: CIF; T: 7.5, 15, 30 Hz; B: 50-1000 kb/s
[7]	T, R	MPEG-2	28	5	SC	S: 720 × 576; T: 25 Hz; 1 or 5 s frame dropping or quality loss
[5] Exp. 2	S, R	H.263+	19	5	DSCQS	S: 50, 75, 100% of 320 × 192; T: 30 Hz; R: 5 QPs
[8]	S, T	MPEG-4	120	4	SC	S: 40–100% of QCIF (upscaling); T: 5, 7, 11, 17, 25 Hz
[9]	S, T, R	H.263, H.264/AVC	20	5	DSIS	S: QCIF, CIF (upscaling); T: 7.5, 15, 30 Hz; B: 24-382 kb/s
[10]	S, T, R	SVC, wavelet-based scalable coding	16	3	SC	S: 1/16, 1/4, 1/1 of 1280 × 720 (upscaling); T: 6.25, 12.5, 25, 50 Hz; B: 358-4108 kb/s

[1] S: spatial dimension; T: temporal dimension; R: SNR dimension; B: bit rate.
[2] QCIF: 176 × 144; CIF: 352 × 288; upscaling: upscaling of small resolutions to the maximum.

Table 1. *Summary of subjective quality assessment studies on multidimensional scalability.*

subjective tests. On the other hand, RR and NR metrics are useful for in-service quality monitoring to adapt transmission and coding strategies to bandwidth fluctuations and packet losses.

Objective quality metrics can also be categorized into data metrics, which measure the fidelity of the signal without considering its content, and picture metrics, which treat the signal as the visual information it contains [3]. Examples of the former are mean square error (MSE) and peak signal-to-noise ratio (PSNR) in the pixel domain, and packet loss rate and bit error rate in the network domain. The latter is based on models of HVS or performs extraction and analysis of particular features in the data. More thorough reviews of objective visual quality metrics can be found in [3].

Subjective Quality Assessment of Scalability

In many existing applications of video scalability, traditional quality metrics such as PSNR or MSE have been used. However, they are not well correlated to the subjective quality that end users perceive, which brought the necessity of conducting subjective quality assessment of scalability and understanding human perception of different scalability options. This section reviews previous studies on this issue and draws a general conclusion from their results. Note that they differ in several factors including contents, codecs, scalability ranges, and subjective testing environments; thus, care must be taken in comparing their results.

Among the scalability dimensions, subjective assessment of temporal scalability has been conducted the most extensively. The central question is: what is the minimum acceptable frame rate? As a general conclusion of the related studies, the threshold of subjective acceptability seems to be around 15Hz, but its exact value varies with content, application, viewers, and so on [4].

More important, the trade-off among scalability dimensions and their relative importance in the viewpoint of perceived quality have been investigated recently. Table 1 summarizes representative studies considering multidimensional scalability, which are explained below. It is worth noting that when different spatial resolutions are considered, a small frame size is usually upsampled to the original full size, assuming a fixed size of viewing window.

The relationship between the temporal and SNR scalabilities has been investigated significantly. Traditionally, it is believed that a high frame rate is more important than high frame quality for content containing fast motion. Thus, reduction of the frame rate decreases subjective quality only slightly for slow motion content [5]. However, other studies showed results contradicting this belief. In [6], it was shown that the

The central question is: what is the minimum acceptable frame rate? As a general conclusion of the related studies, the threshold of subjective acceptability seems to be around 15Hz, but its exact value varies with content, application, viewers, etc.

preference of frame rate against frame quality varies according to the bit rate condition, i.e., frame rates of 7.5, 15, and 30 Hz were most preferred for low, middle, and high bit rate ranges, respectively. The boundaries between the three bit rate ranges were higher for complex scenes, which implies that reaching a certain satisfiable level of frame quality has priority over increasing the frame rate under a limited bit rate budget. A more thorough content dependence of the frame rate preference was investigated in [7]. For fast camera and background motions, a high frame rate was more important than high frame quality in order to prevent jerkiness that is easily detectable and disturbing. However, when the foreground motion is fast (e.g., soccer), frame quality preservation by dropping frames was preferred to frame quality degradation. It should be noted that the conclusion in [5] was not based on analysis for fixed bit rate conditions, which explains its inconsistency with those in [6,7].

The trade-off relation between the spatial and temporal dimensions was investigated in [8]. For each fixed bit rate condition, sequences with different combinations of spatial and temporal resolutions were produced, and their relative subjective preferences were obtained. The results showed that for fast motion content, the frame rate is more important than the frame size.

The relation of the spatial resolution and frame quality was considered in [5]. It was shown that a small frame size with high frame quality is preferable to a large size with low quality when the bit rate is not sufficiently high. In this work, smaller spatial resolutions were not upscaled to the original size; thus, it is not straightforward to compare these results with those in other studies using spatial upscaling.

Subjective quality assessment for all three scalability dimensions has been considered only recently [9, 10]. In [9], an extensive subjective experiment was conducted for low-bit-rate videos. The results showed that for a fixed bit rate, the frame size should be kept low, while a low (high) frame rate is preferable for fast (slow) motion content, which supports the aforementioned results of [6, 7]. When the frame rate is high (e.g., 30 Hz) while the frame size is small for low-bit-rate conditions, improvement in the SNR dimension is usually the most efficient to enhance perceived quality rather than improvement in the spatial dimension. Similarly, when the frame size is large (e.g., CIF) while the frame rate is low, perceived quality is enhanced more efficiently by improving picture quality in the SNR dimension than by increasing the frame rate.

Unlike other studies, the study [10] employed two scalable codecs, SVC and a wavelet-based codec, which facilitates investigation of codec-dependent results. In addition, wide ranges of spatial and temporal resolutions (up to high definition [HD] at 50 Hz) and bit rate conditions (up to 4 Mb/s) were considered. The results can be summarized as follows (Fig. 3). For fixed frame sizes, the frame rate and quality are compared, where the frame quality is governed by coding artifacts (Fig. 3a). The results showed that a higher frame rate was always preferred against better quality. The frame rate also

Bit rate	SVC	Wavelet-based
Low	T > R	-
High	T > R	T > R

(a)

Bit rate	SVC	Wavelet-based
Low	T < S	-
High	T > S	T > S

(b)

Bit rate	SVC	Wavelet-based
Low	-	S > R
High	S < R	S > R

(c)

Figure 3. *Summary of the subjective evaluation results for SVC and a wavelet-based scalable codec [10]. The bit rate conditions are divided into "low" and "high" conditions with thresholds of about 700 kb/s and 900 kb/s for the two codecs, respectively.* A > B *means an improvement in dimension* A *is preferable to an improvement in dimension* B. S, T *and* R *indicates the spatial, temporal, and SNR dimensions, respectively: a) when the frame size is fixed; b) when the frame size and rate vary simultaneously; c) when the frame rate is fixed.*

appeared to be important when the frame rate and size varied simultaneously (Fig. 3b). An exception was observed when the bit rate was low, which indicates that for a low-bit-rate condition, enhancing the frame quality through increase of the frame size was more important than increasing the frame rate. For fixed frame rates (Fig. 3c), the comparison is between the frame size, related to the amount of blurring artifacts, and the frame quality, affected by coding artifacts. For SVC, a better quality was preferred over a larger frame size, while a larger size was more important in the wavelet-based codec. This is because the dominant coding artifacts of the two codecs are fundamentally different (i.e., blockiness in SVC vs. blurring in the wavelet-based codec), and blurring is usually less annoying than blockiness.

In [11], subjective assessment of time-varying quality switching was conducted for an adaptive video transmission scenario using layered encoding. It was shown that the frequency and amplitudes of variations should be kept as small as possible in order to keep high perceived quality.

Although it is not easy to directly compare the aforementioned studies, their results can be roughly summarized as follows. The trade-off is basically between frame rate and frame quality, considering that low spatial resolutions are upscaled to a maximum resolution; thus, the frame quality is affected by both coding artifacts and blurring due to upscaling. It seems that there is a bit rate threshold (or, more accurately, "gray zone") at which the preference or optimal choice of scalability options is switched. Below the threshold, enhancing the frame quality has the priority by improvement in either the SNR or

Ref.	Scalability[1]	Considered ranges[2]	Codec	Content-dependence	Formula[3]
[12]	T, R	S: 130 × 192 T: 7.5–30 Hz	H.263+	Motion information	$\mathrm{PSNR} + \alpha_1 m_{\alpha_2} (f_{\max} - f)$
[13]	T, R	S: QCIF, CIF T: 3.75–30 Hz	SVC	Model parameters	$Q_{\max}\left(1-\frac{1}{1+e^{(\beta_1 \mathrm{PNSR}-\beta_2)}}\right)\left(\frac{1-e^{-\beta_3}\frac{f}{f_{\max}}}{1-e^{-\beta_3}}\right)$
[14]	S, T, R	S: QCIF-CIF T: 7.5–30 Hz	SVC	MPEG-7 motion activity	$\gamma_1 \mathrm{PSNR} + \gamma_2 M(f_{\max} - f) + \frac{\gamma_3}{1+e^{-\gamma_4 (h-h_0)}} + \gamma_5$
[15]	S, T, R	S: QCIF-CIF T: 7.5–30 Hz	SVC	MPEG-7 motion activity, edge histogram	$\delta_1\left(\frac{\mathrm{PNSR}-20}{25}\right)^{\delta_2-\delta_3 S} + \delta_4 M(f_{\max} - f) + \delta_5 e^{-0.5\left(\frac{h-h_{\max}}{\delta_6-\delta_7 S}\right)^2} + \delta_8$

[1] S: spatial dimension; T: temporal dimension; R: SNR dimension.
[2] QCIF: 176 × 144; CIF: 352 × 288.
[3] f: frame rate; $f_{\max}$: maximum frame rate; h: image height; hmax: maximum image height; h_0: mean of the minimum and maximum image heights; m: average magnitude of the top 25% largest motion vectors; M: MPEG-7 motion activity; S: MPEG-7 edge histogram; $\{\alpha_i\}$; $\{\beta_i\}$; $\{\gamma_i\}$; $\{\delta_i\}$: model parameters.

Table 2. *Summary of FR objective quality metrics developed for multidimensional scalability. Note that the contents as well as scalability ranges and codecs for metric development are different.*

spatial dimension. Above the threshold, the frame quality reaches a certain satisfactory level; thus, the frame rate becomes more critical for perceived quality. The threshold is mainly dependent on the content characteristics in such a way that it is higher for content containing faster motion because more bits are needed to achieve an acceptable level of quality for this kind of content, but it is also affected by the encoder type, viewing environment, user expectation, and so on.

The content dependence of the preferred scalability selection is a common finding in most of the studies. Especially, the spatial and temporal complexities have been frequently used to account for content-dependent features. Therefore, content features have been considered as important in developing objective quality metrics for scalability, as shown in the following section.

Objective Quality Assessment of Scalability

The objective metrics proposed specifically for the assessment of video scalability are mostly FR or NR metrics. In general, these metrics can be expressed as functions of frame size, frame rate, bit rate, encoding parameters, content characteristics (e.g., motion, texture), type of artifacts, and so on.

Additionally, since different scalability options may cause different spatial and temporal artifacts, there has been effort to combine metrics designed to measure single spatial or temporal artifacts into a unified model, considering the cross-modal influence of different artifacts on the overall quality.

In the following, we review the relevant works in the field, focusing on the metrics specifically designed for assessment of video scalability.

Full-Reference Quality Metrics

Table 2 summarizes the most relevant FR quality metrics developed in literature. They are usually defined as functions of the temporal and spatial resolutions of the source and test sequences, and the MSE or PSNR computed between the source and test sequences. In these works, subjective results, produced to study the effect of different scalability options on perceived quality, are used as training data to determine the functional forms and parameters of the metrics.

In [12], a quality metric considering the temporal and SNR dimensions was proposed. Subjective experiments for combinations of frame rates and quantization parameters were performed. Although monotonicity of PSNR with respect to subjective quality was observed, an offset was also found between them, which was large for large frame rate reduction and fast motion scenes. Thus, a correction offset for PSNR was defined, depending on the frame rate of the test sequence and the level of motion in the content. The metric recorded higher correlation coefficients with the subjective ratings than PSNR.

The metric proposed in [13] is also based on the observation that PSNR is not an accurate

Objective metrics considering content features and HVS characteristics can be employed in determining the compression parameters in the spatial, temporal, and SNR dimensions for non-scalable video sequences and layer structures of scalable video sequences.

predictor of subjective quality when joint temporal and SNR adjustments are adopted. However, this metric models the modulated relationship of PSNR and the frame rate as a product instead of their additive relationship assumed in [12]. The metric's superiority over the one in [12] and two jerkiness metrics was demonstrated in terms of correlation coefficient with respect to subjective scores. It was also demonstrated that the model parameters can be estimated from content-dependent spatial and temporal features such as frame difference, motion direction, and Gabor texture features.

Inspired by [12], the metric proposed in [14] for the three-dimensional scalability is expressed as a weighted sum of three terms: PSNR, the motion activity-modulated effect of the frame rate, and the effect of the frame size. The third term accounts for the observation that the perceived quality increases with the increasing spatial resolution of stimuli.

This method was further refined in [15], where the spatial complexity of contents is additionally considered. Normalization of PSNR was done to reflect the reduced influence of coding artifacts on the perceived quality for contents with high spatial complexity and the saturation of perceived quality for PSNR values over 45 dB. In addition, the normalized PSNR was further weighted by a spatial complexity measure in order to account for the fact that the quality degradation due to reduction of the spatial resolution is severe for contents having high spatial complexity. The effectiveness of the metric was demonstrated on 81 sequences different from those for model parameter estimation, from which it was shown to outperform PSNR and the metrics given in [12,14].

One of the limitations of the aforementioned metrics is that their design relied on specific sets of subjective data, which may narrow their applicability and reliability for assessment of different kinds of data (e.g., contents, encoding parameters, and spatial and temporal resolutions).

Apart from the aforementioned metrics, other FR metrics developed for different applications could be used at least for one of the scalability dimensions. For example, for a fixed frame rate, existing metrics for assessment of spatial distortion could be used [3]. However, their performance has not been evaluated for assessment of video scalability.

Reduced- and No-Reference Quality Metrics

Only a few RR and NR metrics have been proposed to directly target quality assessment of video scalability and consider both spatial and temporal artifacts.

In [16], a low-complexity NR algorithm, called quality impairment score, was proposed to assess video quality under different spatial, temporal, and SNR combinations. The metric is a weighted product of three factors, a blur metric, a blockiness metric, and a jerkiness metric, each of which is averaged over 30 frames. The weighting values were determined heuristically based on the observation that the blockiness metric has larger fluctuations than the other two terms. By assuming upsampling of sequences having lower spatial or temporal resolutions up to CIF at 30 Hz, each factor measures the quality in the spatial, SNR, and temporal scalability dimensions, respectively.

An NR and an RR metric considering content-dependence were proposed in [17]. In the NR metric, the motion information in the content is measured at the receiver side and used together with the frame rate to obtain a quality index. In the RR metric, the content is first classified to one of five categories and the class-dependent model parameters are used to obtain the quality index from the bit rate and frame rate.

As in the case of the FR metrics, these RR and NR metrics have a limitation in that they have been tested only over a small set of subjective benchmark data.

Applications

Applications of video scalability in multimedia content distribution can benefit from its subjective and objective quality assessment results (Fig. 4). Some examples are briefly discussed below.

Quality evaluation of scalability can be used to optimize video compression strategies. Objective metrics considering content features and HVS characteristics can be employed in determining the compression parameters in the spatial, temporal, and SNR dimensions for non-scalable video sequences and layer structures of scalable video sequences.

Environments such as wireless, mobile, and peer-to-peer (P2P) networks usually involve many heterogeneous users and have time-varying channel conditions in terms of available bandwidth capacity and end-to-end resources. In such cases, quality assessment results can be used for an optimal content delivery strategy that maximizes the quality perceived by users (determining suitable scalability operation points maximizing QoE, quality-aware unequal error protection, etc.).

Video scalability is also enviable for surveillance applications, where video sources not only need to be scrutinized on various devices ranging from videophones or mobiles to HD monitors, but also need to be stored and archived. Parts of a scalable bitstream that do not degrade quality significantly can be deleted after some cessation time so that only the rest is kept for long-term storage/archive.

Conclusion

We have presented a review of the state-of-the-art studies of subjective and objective quality assessment and applications of multidimensional video scalability. A significant amount of work has been done, which led to profound understanding of the challenging QoE issues with respect to video scalability. However, there are also limitations in the studies, which are left for future work.

It would be necessary to conduct more realistic field studies on diverse end-user terminals rather than laboratory environments by using diverse types of contents and to consider other

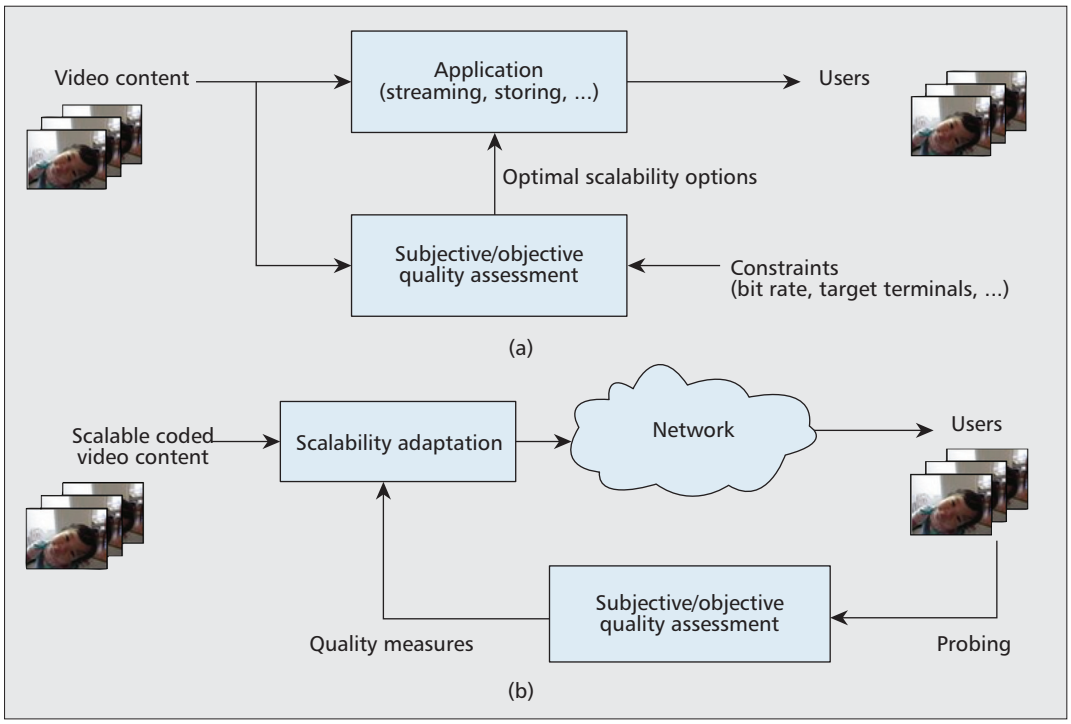


Figure 4. *Two use cases of quality assessment of video scalability: a) off-line selection of optimal combinations of scalability dimensions for content production; b) in-service quality monitoring for adaptive content distribution.*

factors affecting quality perception such as application- and context-specific users' expectation (e.g., attended viewing of video lectures vs. free viewing of movies), transmission errors (e.g., packet loss, jitter, delay), accompanied audio channels, and so on.

As for objective assessment in the context of video scalability, most of the metrics strongly rely on subjective data used for metric design, and systematic comparison of the metrics has rarely been conducted on common validation databases, which would be desirable to consider in the future. Moreover, they were developed mostly for traditional standard definition (SD) contents, so considering the increasing importance of HD content distribution, evaluation of them and development of new metrics on HD sequences will be necessary. Especially, common datasets for video scalability are crucial for evaluating different data analysis methods, developing objective metrics, and conducting cross-evaluation for benchmarking, as in other quality assessment research fields. To our best knowledge, the dataset[1] presented in [10] is the only publicly available database for quality assessment of video scalability. It contains video sequences produced by SVC and wavelet-based coding from three HD contents, ranging between 300 kb/s and 4 Mb/s, and corresponding subjective ratings given by 16 subjects through SSCQS and SC tests. It would be necessary to encourage further publication of open datasets of multidimensional scalability for diverse conditions and applications.

Bitstream-based objective quality evaluation that does not require full decoding of bitstreams is extremely rare in scalable coding. Since it is usually less computationally expensive than that using decoded images (those in Table 2) and thus effective for real-time applications, developing bitstream-based metrics is a promising research topic. Considering post-processing techniques compensating for reduced scalability options would also be interesting for quality assessment of codec-specific scalability options and their compensation.

Other important future trends and challenges in the field include high-level content analysis for QoE measurement (e.g., existence and sizes of faces, existence of text or subtitles, and camera motion), comparative study on QoE over a wide range of display devices (from mobile to HD or even larger displays), and consideration of scalability in 3D videos.

Other important future trends and challenges in the field include high-level content analysis for QoE measurement, comparative study on QoE over a wide range of display devices, and consideration of scalability in 3D videos.

Acknowledgments

This work was supported in part by the Ministry of Knowledge Economy, Korea, under the IT Consilience Creative Program (NIPA-2010-C1515-1001-0001), in part by the Yonsei University Research Fund, in part by the EU funded project SARACEN, and in part by COST IC1003 Qualinet.

References

[1] H. R. Wu and K. R. Rao, *Digital Video Image Quality and Perceptual Coding*, CRC Press, 2005.
[2] F. Pereira, "A Triple User Characterization Model for Video Adaptation and Quality of Experience Evaluation," *Proc. Int'l. Wksp. Multimedia Signal Processing*, Shanghai, China, 2005, pp. 1–4.
[3] S. Winkler and P. Mohandas, "The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics," *IEEE Trans. Broadcast.*, vol. 54, no. 3, Sept. 2008, pp. 660–68.
[4] J. Y. C. Chen and J. E. Thropp, "Review of Low Frame Rate Effects on Human Performance," *IEEE Trans. Sys., Man, Cybern. A*, vol. 37, no. 6, Nov. 2007, pp. 1063–76.

[1] *http://mmspg.epfl.ch/svd*

[5] D. Wang *et al.*, "Towards Optimal Rate Control: A Study of the Impact of Spatial Resolution, Frame Rate, and Quantization on Subjective Video Quality and bit Rate," *Proc. SPIE*, 2003, vol. 5150, pp. 198–209.
[6] Y. Wang *et al.*, "Classification-Based Multidimensional Adaptation Prediction for Scalable Video Coding Using Subjective Quality Evaluation," *IEEE Trans. Circuits Syst. Video Tech.*, vol. 15, no. 10, 2005, pp. 1270–79.
[7] N. van den Ende, H. de Hesselle, and L. Meesters, "Towards Content-Aware Coding: User Study," *Proc. European Conf. Interactive TV*, Amsterdam, The Netherlands, 2007, pp. 185–94.
[8] N. Cranley, P. Perry, and L. Murphy, "User Perception of Adapting Video Quality," *Int'l. J. Human-Computer Studies*, vol. 64, 2006, pp. 637–47.
[9] G. Zhai *et al.*, "Cross-Dimensional Perceptual Quality Assessment for Low Bit-Rate Videos," *IEEE Trans. Multimedia*, vol. 10, no.7, Nov. 2008, pp. 1316–24.
[10] J.-S. Lee, F. De Simone, and T. Ebrahimi, "Subjective Quality Evaluation Via Paired Comparison: Application to Scalable Video Coding," *IEEE Trans. Multimedia*, vol. 13, no. 5, Oct. 2011, pp. 882–93.
[11] M. Zink, J. Schmitt, and R. Steinmetz, "Layer-Encoded Video in Scalable Adaptive Streaming," *IEEE Trans. Multimedia*, vol. 7, no. 1, Feb. 2005, pp. 75–84.
[12] R. Feghali *et al.*, "Video Quality Metric for Bit Rate Control Via Joint Adjustment of Quantization and Frame Rate," *IEEE Trans. Broadcast.*, vol. 53, no. 1, Mar. 2007, pp. 441–46.
[13] Y.-F. Ou *et al.*, "Perceptual Quality Assessment of Video Considering Both Frame Rate and Quantization Artifacts," *IEEE Trans. Circuits Syst. Video Tech.*, vol. 21, no. 3, Mar. 2011, pp. 286–98.
[14] C. S. Kim *et al.*, "Measuring Video Quality on Full Scalability of H.264/AVC Scalable Video Coding," *IEICE Trans. Commun.*, vol. E91-B, no. 5, May 2008, pp. 1269–78.
[15] H. Sohn *et al.*, "Full Reference Video Quality Metric for Fully Scalable and Mobile SVC Content," *IEEE Trans. Broadcast.*, vol. 56, no. 3, Sept. 2010, pp. 269–80.
[16] G. Zhai *et al.*, "Three-Dimensional Scalable Video Adaptation Via User-End Perceptual Quality Assessment," *IEEE Trans. Broadcast.*, vol. 54, no. 3, Sept. 2008, pp. 719–27.
[17] M. Ries, O. Nemethova, and M. Rupp, "Video Quality Estimation for Mobile H.264/AVC Video Streaming," *J. Commun.*, vol. 3, no. 1, Jan. 2008, pp. 41–50.

BIOGRAPHIES

JONG-SEOK LEE (jong-seok.lee@yonsei.ac.kr) received his Ph.D. in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST) in 2006, where he was a postdoctoral researcher and an adjunct professor until 2008. He worked at Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, as a research scientist from 2008 to 2011. He is an assistant professor in the School of Integrated Technology at Yonsei University, Korea. His research interests include multimedia quality assessment, multimedia signal processing, and multimodal human-computer interaction.

FRANCESCA DE SIMONE received her B.Sc. and M.S. degrees in electronic engineering from Università degli Studi Roma Tre, Italy, in 2004 and 2006, respectively. Since May 2007 she is a research assistant in the Multimedia Signal Processing Group at EPFL, Switzerland, where, since March 2008, she is pursuing a Ph.D. degree. Her research interests include subjective and objective multimedia quality assessment and image and video compression.

NAEEM RAMZAN is a postdoctoral researcher at the Multimedia and Vision Group, Queen Mary University of London, United Kingdom. His research activities focus around multimedia search and retrieval, image and video coding, scalable video coding, surveillance-centric coding, and multimedia transmission over wireless and P2P networks. Currently, he is a senior researcher and core member of technical coordination team in the EU funded projects SARACEN and CUBRIK. He is the author or co-author of more than 40 research publications.

EBROUL IZQUIERDO [SM], Ph.D., M.Sc., CEng, FIET, MBMVA, is Chair of Multimedia and Computer Vision and head of the Multimedia and Vision Group in the School of Electronic Engineering and Computer Science at Queen Mary University of London. He received his Ph.D. from Humboldt University, Germany. He has been a senior researcher at the Heinrich-Hertz Institute for Communication Technology, Germany, and the Department of Electronic Systems Engineering of the University of Essex, United Kingdom. He holds several patents in multimedia signal processing and has published over 400 technical papers including chapters in books.

TOURADJ EBRAHIMI received his Ph.D. degree in electrical engineering from EPFL, Switzerland, in 1992. He was a research engineering at the Coorporate Research Laboratories of Sony Corporation, Japan, in 1993. In 1994, he served as a research consultant at AT&T Bell Laboratories in the United States. He is a professor heading the Multimedia Signal Processing Group at EPFL, where he is involved with various aspects of digital video and multimedia applications.