Subjective Quality Evaluation via Paired Comparison: Application to Scalable Video Coding

Jong-Seok Lee, *Member, IEEE*, Francesca De Simone, and Touradj Ebrahimi, *Member, IEEE*

***Abstract*—Scalable video coding is a powerful solution for content delivery in many interactive multimedia services due to its adaptability to varying terminal and network constraints. In order to successfully exploit such adaptability, it is necessary to understand users' preference among various scalability options and consequently develop an optimal bit rate adaptation strategy. In this paper, we present a study of subjective quality assessment of scalable video coding, which investigates the influence of the combination of scalability options on perceived quality with the goal of providing guidelines for an adaptive strategy that selects the optimal combination for a given bandwidth constraint. In particular, the study is based on paired comparison of stimuli that is suitable for our goal due to its simplicity and easiness. We propose a new method, called Paired Evaluation via Analysis of Reliability (PEAR), which analyzes paired comparison results and produces not only quality scores but also intuitive measures of confidence of the scores for significance analysis. Results and analysis of extensive subjective tests for two different scalable video codecs and high definition contents are described, from which general consistent conclusions are drawn. The video and subjective data used in the paper are publicly available to the research community.[1]**

***Index Terms*—Bradley-Terry model, content distribution, multimedia quality assessment, paired comparison, scalable video coding, subjective test.**

I. Introduction

MULTIMEDIA streaming over networks has become a popular application in these days. A wide range of interactive multimedia services such as video conferencing, interactive internet protocol television (IPTV), surveillance, e-learning, and other real-time multimedia content distribution rely on multimedia streaming techniques. An important issue for success of such applications is to deal with heterogeneous and dynamic end-user characteristics including decoding and display capabilities and network bandwidth limitations.

Scalable video coding offers an efficient solution for applications where content needs to be transmitted to many clients with different computational power [1]. In addition, the bit rate adaptability inherent in scalable video coding enables adaptive content distribution based on changes in network conditions. In general, scalable video coding offers three dimensions in scalability, namely, temporal, spatial, and quality (or signal-to-noise ratio). Parts of an encoded bit stream can be skipped in order to reduce the bit rate and quality during transmission over resource-constrained networks. The frame rate may be lowered (temporal scalability), the spatial resolution may be reduced (spatial scalability), and the quality of each image frame may be reduced (quality scalability). In order to effectively exploit such adaptability, it is necessary to have an intelligent strategy that determines the priority among the scalability options and select the best combination for a given bit rate condition, so that the quality of experience of the delivered content to the end-user is maximized.

Subjective quality assessment is a crucial step to develop such a strategy and reach the goal of optimal content delivery because it is the most accurate and reliable way to measure perceived multimedia quality and perform quantitative quality evaluation and comparison. When one conducts a subjective test, the methodology of the test needs to be determined according to the objectives of the test, the quality levels or artifacts in the stimuli, the number of stimuli, etc. There have been efforts to standardize test methodologies in order to ensure reliable and reproducible results [2].

Paired comparison is one of the standard methodologies for subjective multimedia quality assessment, where pair-wise comparisons of stimuli are repeatedly conducted. In comparison to other methods such as single stimulus and double stimulus methods, it has the advantage of simplicity, i.e., instead of assigning a score in a discrete or continuous scale to each stimulus, subjects only need to provide preference between each pair of stimuli. Especially, when the quality difference between the stimuli is not easily noticeable or multiple modalities of quality variations (e.g., spatial resolution, temporal resolution, or multiple sources of visual artifacts) are involved, the paired comparison method can be effectively used with improved reliability [3]–[5].

The results of a paired comparison test appear as a winning frequency matrix representing the frequencies that each stimulus is preferred against the other stimuli involved. Once the matrix is obtained from the comparison results collected from a sufficient number of subjects, they need to be translated into quality scores of the stimuli. In general, this process is not as straightforward as other test methods such as single stimulus methods directly producing quality scores called mean opinion scores (MOS) or double stimulus methods giving differential mean

Manuscript received September 01, 2010; revised February 28, 2011; accepted May 09, 2011. Date of publication May 23, 2011; date of current version September 16, 2011. This work was supported in part by the European Community's Seventh Framework Programme (FP7/2007-2011) under grant agreement no. 216444 (PetaMedia), and in part by the Swiss NCCR Interactive Multimodal Information Management (IM2). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jin Li.

The authors are with the Multimeida Signal Processing Group, Institute of Electrical Engineering, Swiss Federal Institute of Technology Lausanne (EPFL), CH-1015 Lausanne, Switzerland (e-mail: jong-seok.lee@epfl.ch; francesca.desimone@epfl.ch; touradj.ebrahimi@epfl.ch).

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

Digital Object Identifier 10.1109/TMM.2011.2157333

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

opinion scores (DMOS) between references and test stimuli. Converting the matrix to quality scores becomes even more difficult when the rating scale includes ties as well as preferences.

Besides the quality scores, information regarding the reliability of the scores also plays an important role in presenting and analyzing the results. Even if the quality score of a stimulus is higher than that of another one, the statistical significance of the quality difference should be further examined in order to make proper judgment about the result. In the single stimulus or double stimulus methods, it is common to compute the confidence interval of each quality score (i.e., MOS or DMOS) for a significance level of α, which represents that, if the same test is repeated for a large number of times and a confidence interval is obtained each time, $100 \cdot (1 - \alpha)\%$ of the intervals will contain the true value [2]. However, obtaining confidence information in paired comparison-based tests has not been sufficiently studied in the existing work. Overall, the paired comparison method has potential for subjective tests but, in comparison to other test methodologies, its theoretical and practical frameworks have not been investigated sufficiently in the field of multimedia quality assessment.

This paper presents an extensive and thorough study of evaluating quality of video sequences produced by scalable video coding for optimal content delivery over resource-constrained networks, which is performed by a novel method of interpreting paired comparison results for subjective quality assessment. The problem of quality evaluation of video sequences having different temporal, spatial, and quality scalability combinations can be effectively dealt with by paired comparison because providing preference between two sequences is much easier than assigning scores to such stimuli for human observers.

First, we propose a new method to convert comparison results into quality scores and confidence information, called Paired Evaluation via Analysis of Reliability (PEAR). The most important feature of the proposed method is its ability to produce intuitive measures of uncertainty (or conversely, reliability) underlying in the comparison results so that the significance of quality difference of considered stimuli can be judged easily. Then, we present the study where the paired comparison test method and the proposed analysis technique are applied to evaluate quality of video sequences produced by scalable video coding. The goal of the study is to obtain guidelines for a general bit rate adaptation strategy for scalable video coding. We present extensive subjective quality evaluation tests to reveal the relationship among the temporal, spatial, and quality scalability options in terms of perceived quality for different contents and codecs. The detailed description of the experiments shows how the paired comparison method can be effectively used for this problem and how the complete procedure of paired comparison can be performed based on the proposed analysis method. Moreover, we demonstrate the effectiveness of the paired comparison method by comparing the results to those obtained from the popular single stimulus continuous quality scale (SSCQS) method.

Although there have been efforts to analyze the effects of spatial and temporal scaling on perceived quality in simulcast encoded video sequences, the trade-off between the three scalability dimensions in scalable video coding has not been extensively studied. In [6], the trade-off between quality and temporal options in MPEG-4 fine-grained scalability coding was studied and it was concluded that the quality has the priority until it reaches a satisfactory level. Wang *et al.* [7] tried to find the optimal temporal resolution over a range of bandwidths through subjective quality evaluation for motion-compensated wavelet/subband video coding. In [8], it was shown that there exists an optimal adaptation trajectory in the space of possible combinations of spatial and temporal resolutions for MPEG-4 video sequences. The work presented in [9] provides subjective quality test results of low bit rate video sequences encoded by H.263 and H.264/AVC by considering different dimensions such as spatial resolution, temporal resolution, and bit rate.

In comparison to the above studies, the present work has distinct contributions as follows. First, while most of the existing studies only considered standard definition sequences with relatively low bit rates (up to about 1 Mbps) and low frame rates (up to 25 fps), we consider high definition (HD) video sequences having high bit rates (up to about 4 Mbps) and high frame rates (up to 50 fps) that are of interest in today's approaches. Second, various factors affecting the perceived quality are examined in our work for complete analysis, such as content type, scalable video codec, spatial resolution, temporal resolution, and quality option. On the other hand, many existing works considered only some of the options possible (e.g., [6], [7]). Third, the previous work [8], [9] simulated combinations of the three scalability options by using non-scalable codecs, whereas we use two popular scalable video codecs for analysis and examine general agreement between them in the results.

The rest of the paper is organized as follows. The definition of the problem to be solved in this paper is given in Section II. Section III reviews the existing theories of the paired comparison method. The proposed PEAR technique for analyzing comparison results is presented in Section IV along with numerical examples illustrating the results of the existing and the proposed methods. Section V presents the experiments for evaluation of scalable video coding, where the complete procedure of the proposed paired comparison analysis is provided and general guidelines for adaptive scalability are described. Finally, conclusions are given in Section VI.

II. Problem Definition

As stated in the introduction, scalable video coding provides efficient scalable representation of video content through flexible scalability in three dimensions, i.e., spatial, temporal, and quality scalability. A bit stream produced by a scalable video encoder contains a layered structure of several combinations of the three scalability options, from which a video sequence of a particular desired combination can be extracted. Scalable video coding is a suitable solution for adaptive content delivery to end-users having various preferences, terminal capabilities, and network conditions. For example, when low and high bit rate streams of the same content need to be transmitted simultaneously to different clients, necessary layers in a scalable bit stream can be sent to them, respectively, instead of preparing and sending two separate single-layer bit streams. Also, scalable video coding can work adaptively when the bit rate needs to be adjusted according to the varying network bandwidth limitations.

In a scalable bit stream, there may exist multiple different scalability combinations having the same or similar bit rates. Thus, it is necessary to have a policy that selects the best combinations of the three scalability options in terms of quality of experience. This requires quality comparison of layers having different combinations for each bit rate condition, from which guidelines for such an intelligent strategy are made.

While assessing individual video sequences in a pre-defined rating scale is traditionally popular for subjective multimedia quality evaluation, paired comparison is more appropriate and efficient for the goal of our quality evaluation study. First, it is necessary to perform evaluation of video stimuli across different modalities (e.g., spatial and temporal resolutions), which is not easy for human viewers to assign a quality score value in a continuous or discrete scale. On the other hand, paired comparison provides simplicity and easiness of evaluation because viewers only need to indicate which is better in terms of perceived quality between two stimuli. Second, in order to have complete comparison results of stimuli considered, one needs to perform comparison for all possible pairs; when there are many stimuli to be evaluated in a test, the number of comparisons may become too large, which is usual in many subjective tests. However, there are often only a few (typically less than 10) cases that have the same or similar bit rate and need to be compared, and thus the number of comparisons is kept reasonably small.

Therefore, the problem dealt with in our work is to investigate the perceived quality of different scalability combinations having (nearly) the same bit rates in scalable bit streams through paired comparison experiments, which will give insight for smart bit rate adaptation strategies for content distribution in networks.

III. Paired Comparison

A. Procedure

The task of a subject in paired comparison experiments is to provide an index of the relation between two stimuli presented. The judgment can be done either categorically or continuously. In the categorical judgment, the subject selects one of a set of categories defined in semantic forms. The simplest comparison scale for this is $\{\text{'better'}, \text{'worse'}\}$. If the tie between two stimuli is allowed, the scale becomes $\{\text{'better'}, \text{'same'}, \text{'worse'}\}$. A more subdivided scale can also be used, e.g., $\{\text{'much better'}, \text{'better'}, \text{'slightly better'}, \text{'same'}, \text{'slightly worse'}, \text{'worse'}, \text{'much worse'}\}$, as suggested in [2]. In the continuous judgment, the subject assigns each relation to a value on a continuous scale, e.g., $[-100,100]$.

When N stimuli are evaluated, the total number of comparisons ("rounds") is $\binom{N}{2}$. When N becomes large, the number of pairs to be compared may become too large to be feasible. In such cases, it is possible to use techniques that perform comparison only for some of the pairs and obtain estimated quality scores [10].

The paired comparison method can be also used complementarily as a secondary evaluation process in order to analyze further the stimuli that obtained similar quality scores during a primary evaluation run [2].

B. Analysis of Comparison Results Without Ties

When two discrete grades ("better" and "worse") are used, the results obtained from M subjects can be summarized by w_{ij}, i, $j = 1, 2, \ldots, N$, representing the frequency that stimulus i wins over stimulus j, where $w_{ij} + w_{ji} = M$ and $w_{ii} = 0$.

One of the most popular methods converting the winning frequencies to continuous-scale quality scores is to use the Bradley-Terry (BT) model [11]. The probability of choosing stimulus i against stimulus j is represented as

$$P_{ij} = \frac{w_{ij}}{M} = \frac{\pi_i}{\pi_i + \pi_j} \tag{1}$$

where π_i is a positive-valued parameter satisfying $\sum_{i=1}^{N} \pi_i = 1$, which can be considered as the quality score for stimulus i.

The parameter π_i can be estimated by maximizing the log-likelihood function given by

$$L(\pi_1, \pi_2, \ldots, \pi_N) = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} P_{ij} \log\left(\frac{\pi_i}{\pi_i + \pi_j}\right). \tag{2}$$

As a special case, the solution for $N = 2$ is easily obtained as

$$\pi_1 = P_{12} \tag{3}$$

$$\pi_2 = P_{21}. \tag{4}$$

C. Analysis of Comparison Results Having Ties

When ties are allowed in paired comparison, additional parameters, t_{ij}, i, $j = 1, 2, \ldots, N$, are used to summarize the results. t_{ij} represents the frequency that stimulus i and stimulus j are in a tie. In this case, we have $w_{ij} + w_{ji} + t_{ij} = M$.

The probability that stimulus i and stimulus j are in a tie can be calculated from the results as

$$P_{i=j} = \frac{t_{ij}}{M}. \tag{5}$$

1) Equal Division: A simple solution to deal with ties is to treat a tie as a half way between the other two options [12]. Thus, the number of ties is divided by two, each of which is added to the winning frequencies:

$$w_{ij}^{\text{new}} = w_{ij} + \frac{t_{ij}}{2} \tag{6}$$

$$w_{ji}^{\text{new}} = w_{ji} + \frac{t_{ij}}{2}. \tag{7}$$

Then, the BT model can be used to convert w_{ij}^{new} into quality scores.

This method has an advantage of avoiding complications of introducing extra parameters to account for the ties.

2) Rao and Kupper's Method: In [13], Rao and Kupper extended the BT model to consider ties as follows:

$$P_{ij} = \frac{\pi_i}{\pi_i + \theta\pi_j} \tag{8}$$

$$P_{ji} = \frac{\pi_j}{\theta\pi_i + \pi_j} \tag{9}$$

$$P_{i=j} = \frac{\pi_i\pi_j(\theta^2 - 1)}{(\pi_i + \theta\pi_j)(\theta\pi_i + \pi_j)} \tag{10}$$

where $\theta > 1$ is a threshold parameter related to the minimum difference of measurement to distinguish between two stimuli. When $\theta = 1$, the model becomes the original BT model. The parameters π_i and θ are estimated by maximizing the likelihood function that is similar to (2). When $N = 2$, the solution is written as

$$\pi_1 = \frac{\theta P_{12}}{1 + (\theta - 1)P_{12}} \tag{11}$$

$$\pi_2 = \frac{\theta P_{21}}{1 + (\theta - 1)P_{21}} \tag{12}$$

$$\theta = \sqrt{1 + \frac{P_{1=2}}{P_{12}P_{21}}}. \tag{13}$$

3) Davidson's Method: By assuming that the probability of no preference is proportional to the geometric mean of the probabilities of preferences, Davidson [14] proposed another extension of the BT model for incorporating ties:

$$P_{ij} = \frac{\pi_i}{\pi_i + \pi_j + \nu\sqrt{\pi_i \pi_j}} \tag{14}$$

$$P_{ji} = \frac{\pi_j}{\pi_i + \pi_j + \nu\sqrt{\pi_i \pi_j}} \tag{15}$$

$$P_{i=j} = \frac{\nu\sqrt{\pi_i \pi_j}}{\pi_i + \pi_j + \nu\sqrt{\pi_i \pi_j}} \tag{16}$$

where $1/\nu > 0$ can be considered as a discrimination index. $\nu = 0$ yields the BT model. Again, the maximum likelihood estimation is used to obtain the parameters. The solution for $N = 2$ can be written as

$$\pi_1 = \frac{P_{12}}{P_{12} + P_{21}} \tag{17}$$

$$\pi_2 = \frac{P_{21}}{P_{12} + P_{21}} \tag{18}$$

$$\nu = \frac{P_{1=2}}{\sqrt{P_{12}P_{21}}}. \tag{19}$$

IV. Proposed Analysis Method

We propose the PEAR method that gives an intuitive measure of confidence information so that the significance of quality difference can be easily checked. In our method, it is assumed that the ties contain information of uncertainty of the comparison results. The uncertainty may come from severe ambiguity in choosing one of the two given stimuli or unreliability of the ratings provided by the subjects. In any case, existence of the uncertainty in the results implies that the obtained quality scores are not definite but may vary within confidence intervals that will be defined in this section. Such information is crucial in judging whether difference between quality scores of stimuli is statistically significant.

The methods presented in the previous section also reflect the uncertainty underlying in the ties in a way that the scores and the auxiliary parameters (θ and ν) change according to the number of ties. However, such changes do not give intuitive measures of disclosing the significance of the results. Our proposed method produces the confidence interval for the score of a stimulus, which directly indicates whether difference of quality scores is significant. The upper and lower bounds of the confidence interval are obtained by considering the extreme cases where (a part of) the ties of two stimuli supposedly belong to one of the two preference options. Thus, significance of quality difference can be easily checked by examining the amount of overlap of confidence intervals.

A. Method

In our method, two additional parameter sets, π_i^- and π_i^+, are defined as

$$\pi_i^- = \pi_i - \Delta\pi_i^- \tag{20}$$

$$\pi_i^+ = \pi_i + \Delta\pi_i^+ \tag{21}$$

which indicate the lower and upper bounds of π_i, respectively, by considering the uncertainty underlying in the ties. $\Delta\pi_i^-$ (or $\Delta\pi_i^+$) represents the difference between the nominal value of π_i and π_i^- (or π_i^+). Therefore, the interval given by $\left[\pi_i^-, \pi_i^+\right]$ plays the role of the confidence interval for the quality score π_i.

The additional parameters are computed as follows. Once the nominal values of π_i are obtained by using the BT model without considering ties, $\Delta\pi_i^-$ and $\Delta\pi_i^+$ are estimated by solving the following equations:

$$P_{ij}^- = \frac{\pi_i^-}{\pi_i^- + \pi_j^+} = \frac{\pi_i - \Delta\pi_i^-}{\left(\pi_i - \Delta\pi_i^-\right) + \left(\pi_j + \Delta\pi_j^+\right)} \tag{22}$$

$$P_{ij}^+ = \frac{\pi_i^+}{\pi_i^+ + \pi_j^-} = \frac{\pi_i + \Delta\pi_i^+}{\left(\pi_i + \Delta\pi_i^+\right) + \left(\pi_j - \Delta\pi_j^-\right)}. \tag{23}$$

P_{ij}^- and P_{ij}^+ in the above equations can be calculated from the subjective ratings:

$$P_{ij}^- = \frac{w_{ij} + (1-\beta)\cdot t_{ij}}{M} \tag{24}$$

$$P_{ij}^+ = \frac{w_{ij} + \beta \cdot t_{ij}}{M}. \tag{25}$$

The additional parameter $\beta (0 < \beta \le 1)$ can be regarded as the proportion of the ties explaining uncertainty. When $\beta = 1$, all ties are considered to be uncertain, which can be used in most cases. If information regarding the reliability of the scores is available, the value of β can be set to be less than unity accordingly. This parameter may be regarded as analogous to the significance level (α) used in obtaining confidence intervals of MOS or DMOS, in the sense that both of them are related to reliability information and control widths of confidence intervals.

The log-likelihood function to be maximized to obtain $\Delta\pi_i^-$ and $\Delta\pi_i^+$ is written as

$$\begin{aligned} L_a &\left(\Delta\pi_1^-, \ldots, \Delta\pi_N^-, \Delta\pi_1^+, \ldots, \Delta\pi_N^+\right) \\ &= \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \ne i}}^{N} \left\{ P_{ij}^- \log\left(\frac{\pi_i - \Delta\pi_i^-}{\pi_i - \Delta\pi_i^- + \pi_j + \Delta\pi_j^+}\right) \right. \\ &\qquad \left. + P_{ij}^+ \log\left(\frac{\pi_i + \Delta\pi_i^+}{\pi_i + \Delta\pi_i^+ + \pi_j - \Delta\pi_j^-}\right)\right\}. \end{aligned} \tag{26}$$

When $N = 2$, by using $\Delta\pi_i^- = \Delta\pi_j^+$, the solution is given by

$$\Delta\pi_1^- = \Delta\pi_2^+ = \pi_1 - P_{12}^- = P_{21}^+ - \pi_2 \tag{27}$$

$$\Delta\pi_1^+ = \Delta\pi_2^- = P_{12}^+ - \pi_1 = \pi_2 - P_{21}^-. \tag{28}$$

B. Numerical Examples

The following numerical examples are given to effectively illustrate the procedures and results of the aforementioned methods of analyzing paired comparison results.

Let us consider two cases with the following subjective ratings for $N = 2$ and $M = 18$:

- Case I: $w_{12} = 4, w_{21} = 12, t_{12} = 2$
- Case II: $w_{12} = 1, w_{21} = 3, t_{12} = 14$

Note that in both cases, the winning frequency of stimulus 2 is three times that of stimulus 1 but case II has much more ties.

If the ties are divided equally and spread to other preferences, we obtain the scores as

$$\begin{aligned}\pi_1 =&0.278, \qquad \pi_2 = 0.722\\ &\text{and}\\ \pi_1 =&0.444, \qquad \pi_2 = 0.556\end{aligned}$$

for each case, respectively. More ties in case II lead to scores that are more similar.

By using (11)–(13), the Rao and Kupper's method produces

$$\pi_1 = 0.274, \quad \pi_2 = 0.726, \quad \theta = 1.323$$

for case I, and

$$\pi_1 = 0.352, \quad \pi_2 = 0.648, \quad \theta = 9.220$$

for case II.

Using the Davidson's method in (17)–(19), the parameters are estimated as

$$\begin{aligned}\pi_1 =&0.250, \quad \pi_2 = 0.750, \quad \nu = 0.289\\ &\text{and}\\ \pi_1 =&0.250, \quad \pi_2 = 0.750, \quad \nu = 8.083\end{aligned}$$

for each case, respectively.

In both methods, the auxiliary parameters, i.e., θ and ν, account for the uncertainty in the ties to some extent. In addition, when there are more ties, the Rao and Kupper's method yields the scores that are less different from each other.

In the proposed method, the nominal scores are obtained as

$$\pi_1 = 0.250, \quad \pi_2 = 0.750$$

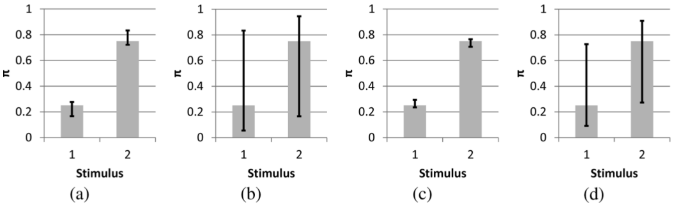
in both cases. By setting $\beta = 1$, we obtain the auxiliary parameters representing the confidence intervals as

$$\Delta\pi_1^- = \Delta\pi_2^+ = 0.028, \quad \Delta\pi_1^+ = \Delta\pi_2^- = 0.083$$

for case I, and

$$\Delta\pi_1^- = \Delta\pi_2^+ = 0.194, \quad \Delta\pi_1^+ = \Delta\pi_2^- = 0.583$$

for case II. Fig. 1 shows these results. One can clearly see that, while the nominal quality scores are the same in the two cases, more ties increase the confidence interval in case II. The confidence intervals are asymmetric around the nominal quality scores because separate parameters, i.e., $\Delta\pi_i^-$ and $\Delta\pi_i^+$, are used to define them. In this way, the confidence intervals obtained in the proposed method have direct implications about the uncertainty of the comparison results; overlapping confidence intervals of two stimuli indicate that there exists a possibility that the dominance appearing in the quality scores may be inverted and thus the quality difference is not significant. If one looks at only the quality scores, one may wrongly judge that stimulus 2 has better quality when compared to stimulus 1 even in case II. Although the auxiliary parameters (θ and ν) in the previous methods tend to become large according to the number of ties, our method provides a more effective and intuitive way to examine the reliability and confidence of the results.



(a) (b) (c) (d)

Fig. 1. Quality scores and confidence intervals calculated by the proposed method for (a) case I ($\beta = 1$), (b) case II ($\beta = 1$), (c) case I ($\beta = 0.5$), and (d) case II ($\beta = 0.5$).

The level of uncertainty is adjusted by varying the value of β, which results in changes of the widths of the confidence intervals. Such changes may cause different conclusions, e.g., two quality levels that are not significantly different for $\beta = 1$ due to the overlapping confidence intervals may become significantly different for $\beta < 1$.

Analysis using asymmetric confidence intervals in our method is particularly useful when more than two stimuli need to be compared. Each of the lower and upper bounds of the confidence interval for a stimulus can be used separately to compare the quality of the stimulus with that of another stimulus having a lower or higher quality score, whereas additional statistical tests such as t-tests need to be repeatedly conducted for results obtained from a single or double stimulus experiment.

V. Experiments

In this section, our subjective quality evaluation of scalable video coding based on the proposed analysis technique is presented. We employed two different scalable video codecs in order to examine the impact of encoding schemes on the subjective evaluation results. In addition, we conducted both conventional SSCQS and paired comparison tests in order to show the effectiveness of the latter over the former in terms of suitability for our goal due to better discriminability of quality scores.

First, details of the performed subjective tests such as codecs, test material, test environments and procedures, and subjective data processing are provided. Then, the test results and in-depth analysis based on the proposed analysis technique are given to draw important findings from the experiments.

A. Scalable Video Codecs

The subjective quality of the sequences produced by an encoder is highly dependent on the algorithms used in the encoder. In addition, the performance of an encoder usually varies according to contents and bit rate conditions. In order to investigate the quality variation due to the scalable video encoding schemes, two different scalable video codecs are employed, namely, scalable extension of H.264/AVC and wavelet-based scalable video coding [15]. The former is a standardized discrete cosine transform-based scalable video coding, while the latter is a popular alternative using wavelet transform.

1) Scalable Extension of H.264/AVC: The latest H.264/AVC standard provides a fully scalable extension, SVC, which achieves significant compression gain and complexity reduction for scalability in comparison to previous video coding standards [16]. It reuses the key features of H.264/AVC and employs new techniques to provide temporal, spatial, and quality scalability with minor increase of bit rate when compared to the single-layer H.264/AVC.

The scalable bit stream coded by SVC is organized by a base layer and several enhancement layers. Hierarchical prediction structures enable temporal scalability. Spatial scalability is achieved via the multi-layer coding approach with additional inter-layer prediction. The coarse-grain quality scalability (CGS) and medium-grain quality scalability (MGS) provide quality scalability. CGS is achieved by re-quantization of the residual signal in the enhancement layer, while MGS is enabled by distributing the transform coefficients of a slice into different network abstraction layer (NAL) units. All these three scalabilities are combined into one bit stream from which different operational points can be extracted.

In our experiments, we used the reference software JSVM 9.18 [17]. The default cascading of the quantization parameters over the temporal levels was disabled. CGS was used to support quality scalability for each of the spatial layers. Each spatial layer has a quality base layer and a CGS quality enhancement layer. Hierarchical B pictures were employed to enable five temporal layers.

2) Wavelet-Based Scalable Video Coding: Apart from the technology used in SVC, there has been also a great amount of research on wavelet-based scalable video coding (WSVC). It has been shown that recent WSVC systems perform well, especially when fine-grain quality scalability is required [18], [19].

WSVC typically operates as follows. First, the input video sequence undergoes spatio-temporal decomposition based on a wavelet transform, which provides the basis of spatial and temporal scalability. Then, motion estimation is performed to obtain motion information, and wavelet coefficients are computed on the texture information remaining after motion estimation. The wavelet coefficients and the motion vectors are compressed to remove redundancy. Finally, the compressed information is organized in the bit stream in a layered manner.

Specifically, the WSVC method in [19] was used in our experiments. The bit stream was encoded by using five temporal layers, three spatial layers, and several quality layers. The group-of-pictures (GOP) size of each sequence was set to 32. In contrast to SVC, the employed WSVC provides a rate control scheme so that it is possible to produce a bit stream containing different scalability combinations having the same bit rate.

B. Test Material

Three HD sequences of duration of 10 s were used, i.e., DucksTakeOff, IntoTree, and ParkJoy [20]. Fig. 2 shows example frames of the sequences. They have distinct spatial and temporal complexity as depicted in Fig. 3, which shows the spatial information (SI) and temporal information (TI) indexes computed for the luminance component of each content [21]. It is observed that IntoTree has relatively small SI and TI values, while ParkJoy shows large values for both measures. DucksTakeOff has a large SI index but a small TI index.

Fig. 2. Example frame images of the content used, namely, DucksTakeOff, IntoTree, and ParkJoy.

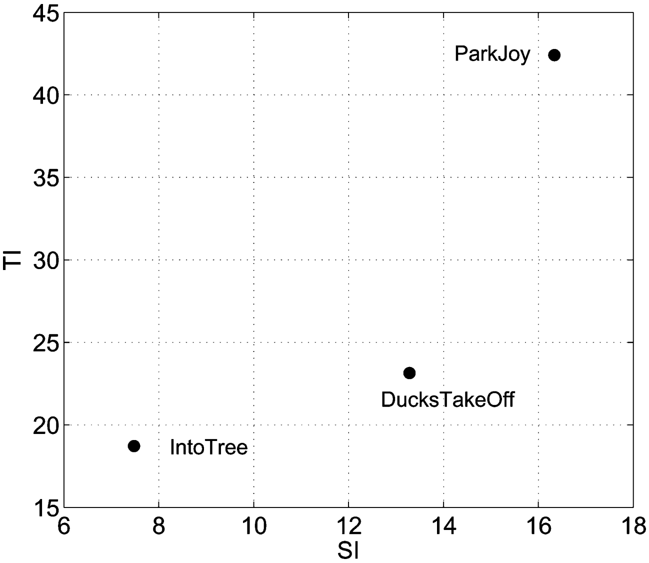


Fig. 3. Spatial information (SI) versus temporal information (TI) indexes of the selected contents.

The original raw sequences having a spatial resolution of 1280 × 720 and a temporal frequency of 50 fps were encoded by using the two codecs. Various layers of different combinations of spatial resolution, temporal resolution, and frame quality were extracted from the encoded bit streams.

Among the layers in the SVC and WSVC bit streams, some of them were selected for the experiments. Since our goal is to compare the perceived quality of different scalability options for each given bit rate condition, we first identified the bit rate conditions that are common to multiple layers. When layers having exactly the same bit rate were not available, we selected layers having similar bit rate values. Then, some of the bit rate conditions were discarded in order to keep the total duration of the subjective test reasonable, while the whole range of the bit rate was maintained and diverse quality levels and scalability options were included in the experiments. As a result, four to six bit rate conditions were selected for each content as shown in Tables I and II for each codec, respectively. The spatial resolutions $(W \times H)$ varying from 320 × 180, and 640 × 360, to 1280 × 720. The frame rates (F) are 6.25, 12.5, 25, and 50 fps. The frame quality is expressed in terms of the pixel bit rate (B_p) defined by

$$B_p = \frac{B}{H \times W \times F} \tag{29}$$

TABLE I
SELECTED COMPARISON SETS COMPOSED OF MULTIPLE LAYERS HAVING (NEARLY) THE SAME BIT RATES FROM THE BIT STREAMS ENCODED BY SVC. EACH LAYER IS SHOWN AS $(B, W \times H, F, B_p)$, WHERE B, $W \times H$, F, AND B_p ARE THE BIT RATE IN KBPS, SPATIAL RESOLUTION, TEMPORAL RESOLUTION, AND PIXEL BIT RATE IN BPS, RESPECTIVELY

	DucksTakeOff
1	(358, 320×180, 6.25, 0.50), (365, 320×180, 12.5, 1.01)
2	(533, 320×180, 12.5, 0.74), (536, 640×360, 6.25, 0.37)
3	(638, 1280×720, 6.25, 0.11), (642, 640×360, 6.25, 0.45)
4	(753, 1280×720, 6.25, 0.13), (790, 640×360, 12.5, 0.27)
5	(926, 1280×720, 12.5, 0.08), (971, 640×360, 12.5, 0.03)
6	(1542, 1280×720, 25, 0.07), (1552, 640×360, 25, 0.27)
	IntoTree
1	(508, 320×180, 12.5, 0.71), (528, 640×360, 6.25, 0.37)
2	(1527, 1280×720, 12.5, 0.13), (1550, 640×360, 25, 0.27)
3	(1932, 1280×720, 6.25, 0.34), (1960, 1280×720, 25, 0.09)
4	(2350, 1280×720, 12.5, 0.20), (2447, 1280×720, 50, 0.05)
	ParkJoy
1	(344, 320×180, 12.5, 0.48), (365, 320×180, 6.25, 1.01)
2	(509, 320×180, 12.5, 0.71), (531, 640×360, 6.25, 0.37)
3	(1542, 1280×720, 6.25, 0.27), (1556, 640×360, 25, 0.27)
4	(4062, 1280×720, 50, 0.09), (4108, 1280×720, 25, 0.18)

where B is the bit rate.

We assumed a fixed frame size equal to the maximum resolution (i.e., 1280×720) for all stimuli, considering applications such as video streaming in video sharing websites where video clips are usually shown with the same frame size. Therefore, sequences having smaller resolutions than 1280×720 were upsampled accordingly for presentation by using a bilinear filter.

C. Test Environments

The test environment is intended to assure the reproducibility of the subjective test results by avoiding involuntary influences of any external factors. Thus, it is important to fix features of the viewing environment such as general viewing conditions and crucial features of the used monitor.

The tests were performed in the space dedicated to professional subjective tests in our laboratory. The test room was equipped with an LCD monitor receiving input from a high performance server that was capable of playing HD raw content in real time. Detailed information about the monitor used is shown in Table III. The ambient lighting consisted of neon lamps with 6500 K color temperature and the walls color was gray 128, as recommended in [2]. Each subject was sitting in front of the monitor at a distance 2–3 times the height of the stimuli.

D. Test Procedures

TABLE II
SELECTED COMPARISON SETS CONTAINING MULTIPLE LAYERS HAVING (NEARLY) THE SAME BIT RATES FROM THE BIT STREAMS ENCODED BY WSVC. EACH LAYER IS SHOWN AS $(B, W \times H, F, B_p)$, WHERE B, $W \times H$, F, AND B_p ARE THE BIT RATE IN KBPS, SPATIAL RESOLUTION, TEMPORAL RESOLUTION, AND PIXEL BIT RATE IN BPS, RESPECTIVELY

	DucksTakeOff
1	(520, 640×360, 6.25, 0.36), (544, 320×180, 6.25, 0.51)
2	(768, 320×180, 12.5, 1.07), (768, 640×360, 12.5, 0.27)
3	(1024, 320×180, 12.5, 1.42), (1024, 640×360, 6.25, 0.71) (1024, 640×360, 12.5, 0.36), (1024, 640×360, 25, 0.18) (1024, 1280×720, 6.25, 0.18), (1024, 1280×720, 12.5, 0.09)
4	(3048, 1280×720, 6.25, 0.53), (3048, 1280×720, 12.5, 0.26) (3048, 1280×720, 25, 0.13), (3048, 1280×720, 50, 0.07)
	IntoTree
1	(520, 320×180, 6.25, 1.48), (520, 640×360, 6.25, 0.37)
2	(768, 320×180, 12.5, 1.07), (768, 640×360, 12.5, 0.27)
3	(1024, 320×180, 12.5, 1.42), (1024, 640×360, 6.25, 0.71) (1024, 640×360, 12.5, 0.36), (1024, 640×360, 25, 0.18) (1024, 1280×720, 6.25, 0.18), (1024, 1280×720, 12.5, 0.09)
4	(3048, 1280×720, 6.25, 0.53), (3048, 1280×720, 12.5, 0.26) (3048, 1280×720, 25, 0.13), (3048, 1280×720, 50, 0.07)
	ParkJoy
1	(520, 320×180, 6.25, 1.44), (520, 640×360, 6.25, 0.36)
2	(768, 320×180, 12.5, 1.07), (768, 640×360, 12.5, 0.27)
3	(1024, 320×180, 12.5, 1.42), (1024, 640×360, 6.25, 0.71) (1024, 640×360, 12.5, 0.36), (1024, 640×360, 25, 0.18) (1024, 1280×720, 6.25, 0.18), (1024, 1280×720, 12.5, 0.09)
4	(3048, 1280×720, 6.25, 0.53), (3048, 1280×720, 12.5, 0.26) (3048, 1280×720, 25, 0.13), (3048, 1280×720, 50, 0.07)

TABLE III
DETAILS OF THE MONITOR USED FOR THE TESTS

LCD monitor	Eizo CG301W
Diagonal size	30 inches
Resolution	2560×1600 (native)
Calibration tool	EyeOne Display 2
Gamut	sRGB
White point	D65
Brightness	120 cd/m^2
Black level	minimum
Response time	6 ms
Desktop window background	gray 128

Two subjective test methodologies were used in our experiments, namely, SSCQS and paired comparison. The two tests were conducted separately and their results are reported in order to demonstrate the effectiveness of the paired comparison method using the proposed analysis technique in comparison to the popular SSCQS methodology.

1) SSCQS: Prior to the test sessions, a training session was held, where the test methodology was described to the subject by using a set of training stimuli that were different from the test stimuli. A subject watched each stimulus played once for 10 s and had 5 s to vote on a score sheet. The rating scale used was a continuous scale ranging from 0 to 100. Adjective descriptions of each range of the scale ("excellent", "good", "fair", "poor", and "bad") were also indicated next to numeric scores. The presentation order was randomized and measures were taken to avoid playing the same content consecutively. The stimuli presentation was divided into two separate sessions so that the duration of each test session remains reasonable. Sixteen subjects (11 males and 5 females) participated in the tests. They reported normal or corrected-to-normal vision.

2) Paired Comparison: A pair of test stimuli in the same comparison set (each set in Tables I and II) were played side-by-side time-synchronously. All possible pairs in each comparison set were used for comparison in order to obtain complete winning frequency matrices. Since the monitor used has a native resolution of 2560×1600, two video sequences could fit in the horizontal space of the display. Each subject was asked to choose which stimulus had better quality between the two presented stimuli and to mark the answer between "left", "same", and "right" on the score sheet. Each pair of stimuli was played in loop in order to allow the subject to watch the stimuli as many times as he/she wanted for careful analysis. Each subject had two separate test sessions, each of which contained about 50 pairs of stimuli each. As in SSCQS, a training session took place before the test sessions. Another sixteen subjects (11 males and 5 females) having normal or corrected-to-normal vision participated in the tests.

E. Subjective Data Processing

Once subjective data are gathered from subjects, the data need to be processed to compute the final quality scores and confidence intervals for evaluation. The subjective data processing consisted of outlier detection and calculation of quality scores and confidence intervals, which are explained below for each test methodology.

1) SSCQS: Screening of subjects in SSCQS was performed by following the guidelines described in [2]. For each stimulus, it is tested whether the distribution of the ratings of all subjects is normal. If the kurtosis coefficient is between 2 and 4, the distribution is regarded as normal. Then, for each subject, the two counters, P and Q, are calculated by checking whether the score of the subject for a stimulus is far from the average score for all subjects. In other words, if the score by the subject for the stimulus is larger (or less) than the average score over all subjects plus (or minus) r times the standard deviation, P (or Q) is increased by 1. The value of r was chosen to be 2 for normally distributed and $\sqrt{20}$ for non-normally distributed scores. Finally, if the proportion of $P+Q$ among the whole test of a subject was larger than 5% and $|P-Q|/(P+Q)$ less than 30%, the subject was considered as an outlier and his/her ratings were discarded. In our case, no subject was found as an outlier.

The MOS was computed for each stimulus by averaging the scores of all subjects for the stimulus. The 95% confidence interval (δ) of the stimulus was computed by using the Student's t-distribution:

$$\delta = t(0.975, M-1) \cdot \frac{\sigma}{\sqrt{M}} \tag{30}$$

where σ is the standard deviation of the scores over all subjects for the stimulus and $t(0.975, M-1)$ is the t-value from a two-tailed Student's t-distribution with the significance level (5%) and $M-1$ degrees of freedom.

2) Paired Comparison: Outlier detection in paired comparison was conducted differently from SSCQS. An evidence that a particular subject was not capable of making reliable judgment can be detected by checking the transitivity appearing the paired comparison results of the subject. If a subject preferred stimulus 1 against stimulus 2, and stimulus 2 against stimulus 3, then stimulus 1 must be preferred against stimulus 3 by the same subject in order to satisfy the transitivity rule. If stimulus 3 is preferred against stimulus 1, the three stimulus forms a circular triad, which indicates a violation of the transitivity rule. Therefore, when the number of such circular triads in the results of the subject is large, we can conclude the subject's judgment is not reliable and thus he/she is regarded as an outlier. While previous methods only considered binary preference choices [3], [22], we present a method modified to accommodate ties in the results as follows.

Let $i \rightarrow j$ mean that stimulus i is preferred to j by a subject, and let $i \leftrightarrow j$ indicate a tie between the two stimuli. For each triad containing stimuli i, j, and k, the following cases are considered as circular triads:

$$i \rightarrow j \wedge j \rightarrow k \wedge (k \rightarrow i \vee k \leftrightarrow i) \tag{31}$$

$$i \rightarrow j \wedge j \leftrightarrow k \wedge k \rightarrow i \tag{32}$$

$$i \leftrightarrow j \wedge j \rightarrow k \wedge k \rightarrow i. \tag{33}$$

The transitivity satisfaction rate (or consistency rate) is defined by

$$\xi = 1 - \frac{C}{T} \tag{34}$$

where C is the number of circular triads and T is that of all possible triads. If ξ for a subject is less than a threshold, the subject is considered as an outlier.

For each subject, we calculated ξ for the comparison sets containing at least three layers, i.e., the last two comparison sets of each content in Table II. It turned out that, in our experiments, all subjects' transitivity satisfaction rates were above 0.96, which indicates high reliability of their ratings. Therefore, it was concluded that there were no outlier in the paired comparison experiments.

The quality scores and confidence intervals were obtained from the paired comparison results by using the proposed method described in Section IV. For analysis, we normalized the scores of the instances in a comparison set so that the maximum score becomes 100. The confidence intervals were also normalized accordingly.

F. Results and Analysis

The quality scores and the confidence intervals of the two test methods are shown in Figs. 4–9. The compared scalability combinations are written above each figure as $(B, W \times H, F, B_p)$, where B, $W \times H$, F, and B_p are the bit rate in kbps, spatial resolution, temporal resolution, and pixel bit rate in bps, respectively, as in Tables I and II. In the case of SSCQS, a

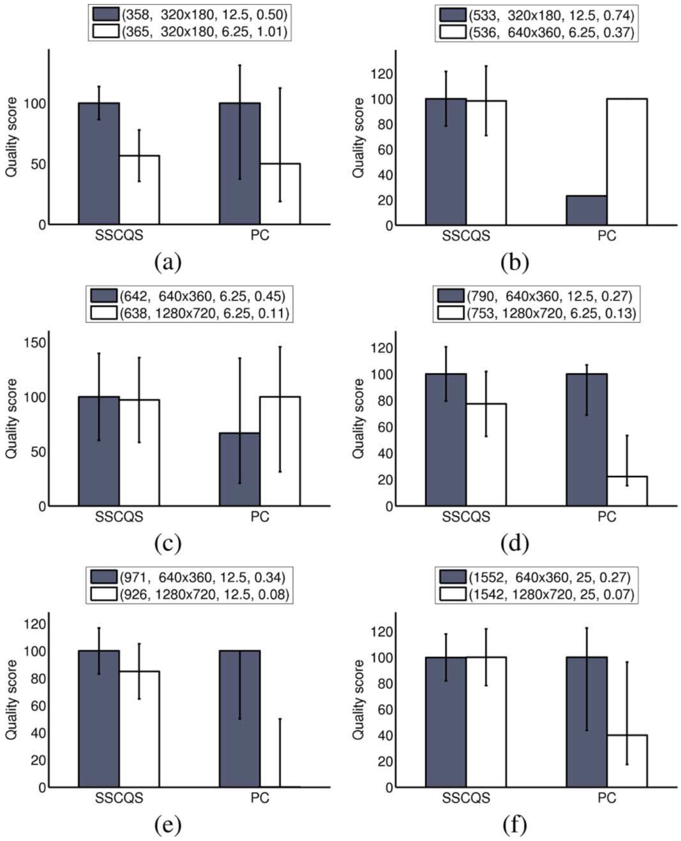


Fig. 4. Results of the subjective tests for DucksTakeOff encoded by SVC. In the case of SSCQS, the hypothesis test showed that the difference of the two MOS values was significant ($p < 0.05$) only in (a).

two-sample two-tailed t-test was additionally conducted for the stimuli having the highest and the second highest MOS values in each comparison set in order to determine if the difference between the two MOS values was statistically significant. The null hypothesis is that the two score sets are from a normal distribution with equal means, while the alternative hypothesis is that the two means are not equal. The cases where the null hypothesis was rejected at a significance level of 0.05 are mentioned in the captions of the figures.

When SSCQS and paired comparison are compared, the preference in each comparison set is consistent. The major difference between the two test methods is that the paired comparison results show large, significant differences among the layers in each comparison set, which indicates better discriminability of paired comparison for the given quality evaluation task in comparison to SSCQS. Since the layers in a comparison set have similar bit rate values, their quality tends to be similar through the SSCQS test containing sequences over the whole quality range, even though the layers are from different combinations of spatial, temporal, and quality scalabilities. The large discriminability of paired comparison is effective in our scenario where we want to find the best scalability combination for a given bit rate condition. Hereafter, the analysis is mainly based on the results of paired comparison.

Figs. 4–6 show the results of SVC. When a pair of sequences have the same spatial resolution, the one having a larger frame rate is preferred [Figs. 5(c)–(d) and 6(a)]. The last pair of ParkJoy [Fig. 6(d)] is an exception, which is thought to be because the high TI index of the content makes the subjects insensitive to the change of the frame rate higher than 25 fps and the frame quality is better in the case of 25 fps (i.e., a higher pixel bit rate).

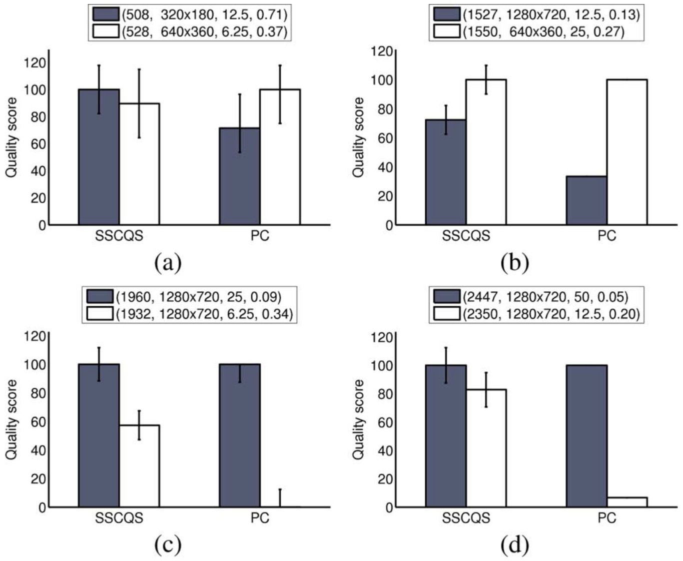


Fig. 5. Results of the subjective tests for IntoTree encoded by SVC. In the case of SSCQS, the hypothesis test showed that the difference of the two MOS values was significant ($p < 0.05$) in (b), (c), and (d).

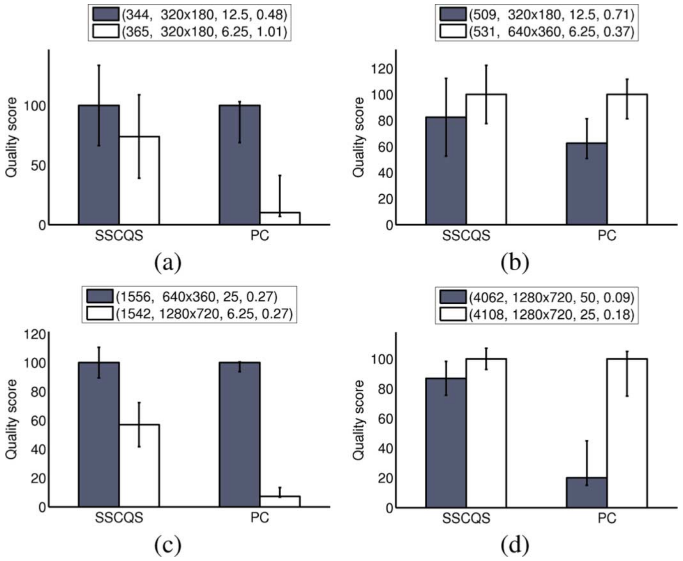


Fig. 6. Results of the subjective tests for ParkJoy encoded by SVC. In the case of SSCQS, the hypothesis test showed that the difference of the two MOS values was significant ($p < 0.05$) in (c) and (d).

Figs. 4(b) and (d), 5(b), and 6(b)–(c) show the results of the cases where the combination of the frame rate and frame resolution is different from each other and the quality difference between the two stimuli is significant in paired comparison. It is observed that, when the bit rate is small and the two resolutions 320×180 and 640×360 are compared, the latter is always preferred even though the frame rate is lower, which is because of the stronger interpolation effect in the smaller resolution [Figs. 4(b) and 6(b)]. However, when the bit rate becomes

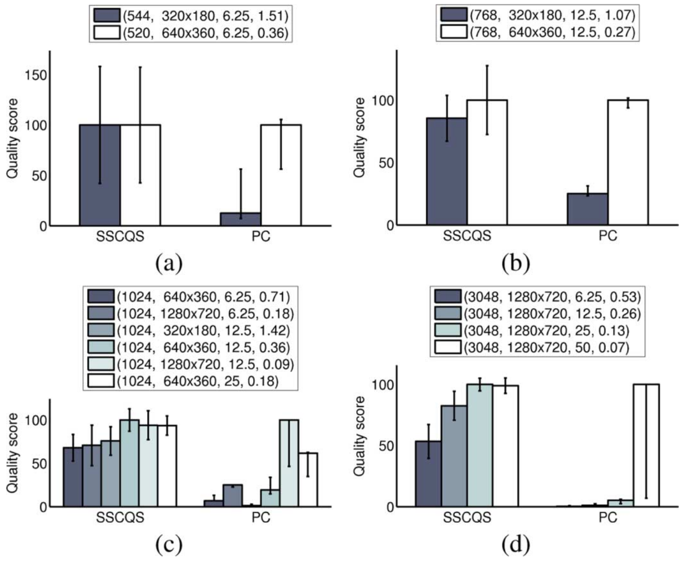


Fig. 7. Results of the subjective tests for DucksTakeOff encoded by WSVC. In the case of SSCQS, the hypothesis test showed that the difference of the two highest MOS values was not significant in these four comparison sets.

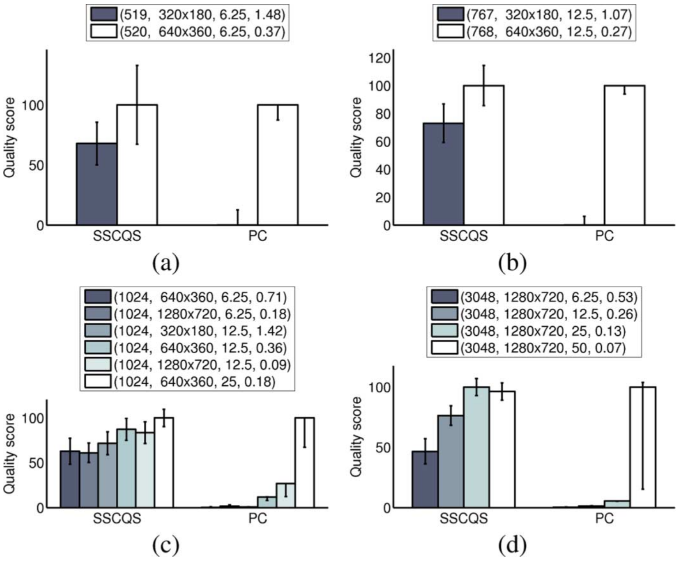


Fig. 8. Results of the subjective tests for IntoTree encoded by WSVC. In the case of SSCQS, the hypothesis test showed that the difference of the two highest MOS values was significant $(p < 0.05)$ in (c).

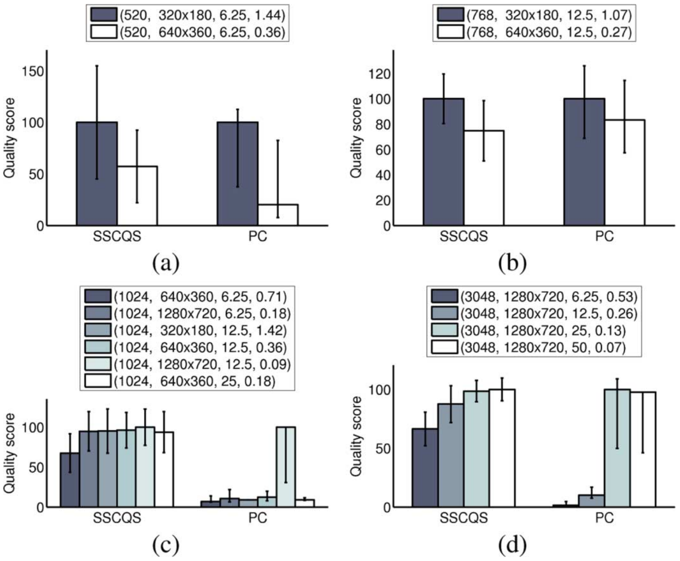


Fig. 9. Results of the subjective tests for ParkJoy encoded by WSVC. In the case of SSCQS, the hypothesis test showed that the difference of the two highest MOS values was not significant in these four comparison sets.

large, a high frame rate is more important than a high spatial resolution (i.e., 1280×720), as shown in Figs. 4(d), 5(b), and 6(c).

Fig. 4(c), (e), and (f) compares the cases of the same frame rate but different resolutions (1280×720 and 640×360). While the first and last cases show statistically insignificant results, Fig. 4(e) shows that the resolution of 640×360 is preferred because the frame quality is not good enough and the blocking artifacts are observed in the case of the resolution 1280×720.

In Figs. 7–9, the results of WSVC for each content are shown. For the comparison sets whose bit rates are less than 1024 kbps, the comparison is made between the two spatial resolutions, i.e., 320×180 and 640×360, for the same frame rates [Figs. 7(a)–(b), 8(a)–(b), and 9(a)–(b)]. For DucksTakeOff and IntoTree, a larger spatial resolution is preferred against a smaller one because the blurring artifact is stronger in the smaller resolution [Figs. 7(a)–(b) and 8(a)–(b)]. In these cases, low frame rates such as 6.25 fps and 12.5 fps are not sufficient for the content containing relatively fast visual motion. Thus, the subjects may prefer the more strongly blurred scene in the low resolution, which partially compensates for the jerky motion in the sequences having low frame rates.

Figs. 7(c), 8(c), and 9(c) compare the results when the bit rate is 1024 kbps. It is observed that the two combinations of the spatial and temporal resolutions, $(W \times H, F) = (320 \times 180, 12.5)$ and $(W \times H, F) = (640 \times 360, 6.25)$, show the worst quality due to the strong blurring effect caused by interpolation and the lowest temporal resolution, respectively. The combination of $(W \times H, F) = (1280 \times 720, 6.25)$ also shows low quality scores because of the lowest frame rate. For DucksTakeOff and ParkJoy, the combination of $(W \times H, F) = (1280 \times 720, 12.5)$ has the best quality, whereas $(W \times H, F) = (640 \times 360, 25)$ was evaluated as the best for IntoTree. Such content-dependence can be explained by the fact that IntoTree has a small SI index and thus spatial blurring does not degrade the quality as much as in the other two contents.

For the 3048 kbps bit rate condition, all the layers have the same spatial resolution and thus the comparison is performed among different combinations of the frame rate and frame quality. The results indicate that a larger frame rate is always preferred [Figs. 7(d), 8(d), and 9(d)]. For ParkJoy, quality difference between the cases of 25 fps and 50 fps is not statistically significant, which is in line with the exceptional result for SVC in Fig. 6(d).

Interestingly, the pixel bit rate does not have clear relationship with the perceived quality in both codecs, whereas the previous work [9] indicated that a higher quality score is usually marked for a higher pixel bit rate. Rather, the spatial and temporal resolutions seem to be more important for the perceived quality in our case. We think that such difference stems from different applications considered, i.e., low bit rate video sequences in [9] and HD sequences in our work.

Overall, the following conclusions can be made based on the above observations. First, when the bit rate is low and only layers having small spatial resolutions are available, a larger spatial resolution obtaining the lowest acceptable frame quality without strong blurring is preferable. In the case of SVC, this observation was valid even when the frame rate decreases at the cost of the increase of the spatial resolution. Second, for high bit rate conditions, acceptable frame quality is achieved and thus a high frame rate acquired by the decrease of the pixel bit rate becomes important for subjective quality. Here, the separation of low and high bit rate values appeared between 800 kbps and 1024 kbps for the case of WSVC; it may be estimated as about 700 kbps for SVC, which remains inconclusive due to limited availability of diverse layers having the same bit rate values. Third, the content is an important factor influencing the perceptual quality of different scalability combinations, which can be described by the SI and TI indexes to some extent. Fourth, while each encoder produces scalable bit streams of different structures and thus the encoder type affects the results significantly, the aforementioned overall tendency remains quite consistent across the two codecs.

VI. Conclusion

In this paper, we have presented a subjective quality evaluation study of scalable video coding for adaptive content delivery. The evaluation was performed via a paired comparison methodology, for which we have proposed the new PEAR method that converts comparison results to quality scores and facilitates performing intuitive significance analysis of the scores. It was demonstrated that the proposed paired comparison test methodology can be effectively used for comparing different scalability options and developing guidelines for an adaptive strategy of exploiting scalability in comparison to the conventional SSCQS method. The analysis of the test results showed that the priority between the spatial resolution and frame rate depends on the bit rate condition and content type, which was considerably consistent in the two codecs under test. For low bit rate conditions, the spatial resolution was important for perceived quality, whereas for higher bit rate conditions, a high frame rate was preferable.

While we applied the proposed paired comparison methodology for evaluation of scalable video coding, it can also be used to solve diverse multimedia quality evaluation problems. Particularly, it is suitable for online crowdsourcing from which many subjective data can be collected more easily when compared to MOS-based methodologies such as SSCQS. The rating procedure of paired comparison is simple so that training of subjects can be performed easily. In addition, the reliability of each subject's ratings can be judged independently in our methodology, where as other subjective data are required for outlier detection in MOS-based methodologies, as shown in Section V-E.

It will be necessary to perform further validation of the conclusions made in our work by using more diverse contents. In the future, it would be desirable to develop an adaptive decision strategy of scalability options for resource-constrained networks by taking into account the drawn observations in this paper as guidelines. Previously, there have been some work of proposing such strategies (e.g., [8], [23]), which are not usually based on thorough subjective evaluation across all three scalability dimensions. Moreover, it is useful to develop objective quality measures of video sequences generated by scalable video coding based on the results reported here, which can be used in adaptive strategies for real-time automatic quality measurement and monitoring of delivered content.

References

[1] J.-R. Ohm, "Advances in scalable video coding," *Proc. IEEE*, vol. 93, no. 1, pp. 42–56, Jan. 2005.

[2] Methodology for the Subjective Assessment of the Quality of Television Pictures, Geneva, Switzerland, 2002, Recommendation ITU-R BT.500-11.

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

[4] S. Voran and A. Catellier, "Gradient ascent paired-comparison subjective quality testing," in *Proc. Int. Workshop Quality of Multimedia Experience*, San Diego, CA, 2009, pp. 133–138.

[5] R. A. Doherty, A. C. Younkin, and P. J. Corriveau, "Paired comparison analysis for frame rate conversion algorithms," in *Proc. Int. Workshop Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, AZ, Jan. 2009.

[6] R. Rajendran, M. van der Schaar, and S.-F. Chang, "FGS+: Optimizing the joint SNR-temporal video quality in MPEG-4 fine grained scalable coding," in *Proc. Int. Symp. Circuits and Systems*, 2002, pp. 445–448.

[7] Y. Wang, S.-F. Chang, and A. C. Lou, "Subjective preference of spatio-temporal rate in video adaptation using multidimensional scalable coding," in *Proc. Int. Conf. Multimedia and Expo*, 2004, pp. 1119–1122.

[8] N. Cranley, P. Perry, and L. Murphy, "Optimum adaptation trajectories for streamed multimedia," *Multimedia Syst.*, vol. 10, no. 5, pp. 392–401, 2005.

[9] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, and M. Etoh, "Cross-dimensional perceptual quality assessment for low bit-rate videos," *IEEE Trans. Multimedia*, vol. 10, no. 7, pp. 1316–1324, Nov. 2008.

[10] D. A. Silverstein and J. E. Farrell, "Quantifying perceptual image quality," in *Proc. ST&T's Image Processing, Image Quality, Image Capture, Systems Conf.*, 1998, pp. 242–246.

[11] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, pp. 324–345, 1952.

[12] M. E. Glickman, "Parameter estimation in large dynamic paired comparison experiments," *J. Roy. Statist. Soc., Ser. C (Appl. Statist.)*, vol. 48, no. 3, pp. 377–394, 1999.

[13] P. V. Rao and L. L. Kupper, "Ties in paired-comparison experiments: A generalization of the Bradley-Terry model," *J. Amer. Statist. Assoc.*, vol. 62, no. 317, pp. 194–204, Mar. 1967.

[14] R. R. Davidson, "On extending the Bradley-Terry model to accommodate ties in paired comparison experiments," *J. Amer. Statist. Assoc.*, vol. 65, no. 329, pp. 317–328, Mar. 1970.

[15] J.-S. Lee, F. D. Simone, N. Ramzan, Z. Zhao, E. Kurutepe, T. Sikora, J. Ostermann, E. Izquierdo, and T. Ebrahimi, "Subjective evaluation of scalable video coding for content distribution," in *Proc. ACM Multimedia*, Firenze, Italy, Oct. 2010, pp. 65–72.

[16] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[17] J. Reichel, H. Schwarz, and M. Wien, Joint Scalable Video Model 11 (JSVM 11), Joint Video Team, 2007, doc. JVT-X202.

[18] N. Adami, A. Signoroni, and R. Leonardi, "State-of-the-art and trends in scalable video compression with wavelet-based approaches," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1238–1255, Sep. 2007.

[19] N. Ramzan, T. Zgaljic, and E. Izquierdo, "An efficient optimisation scheme for scalable surveillance centric video communications," *Signal Process.: Image Commun.*, vol. 24, no. 6, pp. 510–523, 2009.

[20] The SVT High Definition Multi Format Test Set. [Online]. Available: http://www.its.bldrdoc.gov/vqeg.

[21] Subjective Video Quality Assessment Methods for Multimedia Applications, 1999, Recommendation ITU-R P.910.

[22] M. G. Kendall and B. B. Smith, "On the method of paired comparisons," *Biometrika*, vol. 31, no. 3/4, pp. 324–345, Mar. 1940.

[23] W.-H. Peng, J. K. Zao, H.-T. Huang, T.-W. Wang, and L.-S. Huang, "A rate-distortion optimization model for SVC inter-layer encoding and bitstream extraction," *J. Vis. Commun. Image Represent.*, vol. 19, pp. 543–557, 2008.

Jong-Seok Lee (M'06) received the Ph.D. degree in electrical engineering and computer science in 2006 from KAIST, Daejeon, Korea.

He worked as a postdoctoral researcher and an adjunct professor at KAIST. He is now working as a research scientist in the Multimedia Signal Processing Group at Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland. His current research interests include audio-visual signal processing, multimedia quality assessment, and multimodal human-computer interaction. He is author or co-author of over 40 publications.

Dr. Lee is a member of the Multimedia Communication Technical Committee of the IEEE Communication Society.

Francesca De Simone was born in Italy in 1983. She received the B.S and M.S. degrees in electronic engineering from Università degli Studi Roma Tre, Rome, Italy, in 2004 and 2006, respectively. Since March 2008, she has been pursuing the Ph.D. degree.

From August 2006 until November 2006, she was a research assistant at the Institute of Signal Processing of Tampere University of Technology, Tampere, Finland. Since May 2007, she has been a research assistant at the Multimedia Signal Processing Group at Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland. Her current research interests include subjective and objective multimedia quality assessment and image and video compression.

Touradj Ebrahimi (M'92) received the M.Sc. and Ph.D. degrees, both in electrical engineering, from the Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland, in 1989 and 1992, respectively.

From 1989 to 1992, he was a research assistant at the Signal Processing Laboratory of EPFL. During the summer of 1990, he was a visiting researcher at the Signal and Image Processing Institute of the University of Southern California, Los Angeles. In 1993, he was a research engineer at the Corporate Research Laboratories of Sony Corporation in Tokyo, Japan. In 1994, he served as a research consultant at AT&T Bell Laboratories. He is currently a Professor heading the Multimedia Signal Processing Group at EPFL, where he is involved with various aspects of digital video and multimedia applications. He is author or co-author of over 100 papers and holds 10 patents.