# Subjective Evaluation of Scalable Video Coding for Content Distribution

Jong-Seok Lee
Multimedia Signal Processing Group
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015, Lausanne, Switzerland
jong-seok.lee@epfl.ch

Francesca De Simone
Multimedia Signal Processing Group
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
francesca.desimone@epfl.ch

Naeem Ramzan
School of Electronic Engineering and Computer Science
Queen Mary University of London
Mile End Road, London E1 4NS, UK
naeem.ramzan@elec.qmul.ac.uk

Zhijie Zhao
Institut für Informationsverarbeitung
Leibniz Universität Hannover
30167 Hannover, Germany
zhao@tnt.uni-hannover.de

Engin Kurutepe
Communication Systems Group
TU Berlin
Einsteinufer 17
10587 Berlin, Germany
kurutepe@nue.tu-berlin.de

Thomas Sikora
Communication Systems Group
TU Berlin
Einsteinufer 17
10587 Berlin, Germany
sikora@nue.tu-berlin.de

Jörn Ostermann
Institut für Informationsverarbeitung
Leibniz Universität Hannover
30167 Hannover, Germany
ostermann@tnt.uni-hannover.de

Ebroul Izquierdo
School of Electronic Engineering and Computer Science
Queen Mary University of London
Mile End Road, London E1 4NS, UK
ebroul.izquierdo@elec.qmul.ac.uk

Touradj Ebrahimi
Multimedia Signal Processing Group
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015, Lausanne, Switzerland
touradj.ebrahimi@epfl.ch

## ABSTRACT

This paper investigates the influence of the combination of the scalability parameters in scalable video coding (SVC) schemes on the subjective visual quality. We aim at providing guidelines for an adaptation strategy of SVC that can select the optimal scalability options for resource-constrained networks. Extensive subjective tests are conducted by using two different scalable video codecs and high definition contents. The results are analyzed with respect to five dimensions, namely, codec, content, spatial resolution, temporal resolution, and frame quality.

## Categories and Subject Descriptors

E.4 [**Coding and Information Theory**]: Data Compaction and Compression; H.1.2 [**Information Systems**]: User/Machine Systems—*Human information processing*

## General Terms

Measurement, performance

## Keywords

Scalable video coding, quality of experience, bit-rate adaptation, subjective test

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
*MM'10,* October 25–29, 2010, Firenze, Italy.
Copyright 2010 ACM 978-1-60558-933-6/10/10 ...$10.00.

## 1. INTRODUCTION

Scalable video coding (SVC) schemes offer an efficient alternative to simulcast encoding for applications where content needs to be transmitted to many non-homogeneous clients with different decoding and display capabilities. Moreover, the bit-rate adaptability inherent in the scalable codec designs provides a natural and efficient way of adaptive content distribution according to changes in network conditions.

In general, a scalable video sequence can be adapted in three dimensions, namely temporal, spatial, and quality dimensions, by leaving out parts of the encoded representation, thus reducing the bit-rate and the video quality during transmission. We define these dimensions of scalability as follows:

- Temporal scalability refers to the possibility of reducing the temporal resolution of the encoded video directly from the compressed bit-stream, i.e. the number of frames contained in one second of the video is reduced.
- Spatial scalability refers to the possibility of reducing the spatial resolution of the encoded video directly from the compressed bit-stream, i.e. the number of pixels in a video frame is reduced.
- Quality scalability, or commonly called signal-to-noise ratio (SNR) scalability, or fidelity scalability, refers to the possibility of reducing the quality of the encoded video. This is achieved by extracting and decoding coarsely quantized pixels from the compressed bit-stream.

By adjusting one or more of the scalability options, the SVC scheme allows flexibility and adaptability of video transmission over resource-constrained networks. In order to uti-

lize such adaptability efficiently, however, it is necessary to have a strategy of selecting a layer of an appropriate combination of the temporal, spatial and quality parameters among several layers contained in the SVC bit-stream. An intelligent strategy should maximize the subjective quality of experience of the end-user by determining the priority among the scalability dimensions and selecting the best combination of them for certain bit-rate conditions.

Although there exists some work investigating the subjective effects of spatial and temporal scaling in simulcast encoded video sequences, the trade-off between different dimensions of SVC has not been extensively studied. In [7], the authors presented a series of subjective comparisons of the scalable extension of H.264/MPEG-4 advanced video coding (AVC), i.e. H.264/SVC, and the single layer H.264 coding. The results show that H.264/SVC can provide reasonable scalability with no more than 10% additional rate relative to single layer H.264. Rajendran *et al.* [8] studied the trade-off between SNR and temporal scalability in MPEG-4 fine grained scalability (FGS) coding and concluded that the SNR-quality has priority until it reaches a satisfactory level. Wang *et al.* [14] investigated the optimal temporal resolution over a range of bandwidths through subjective quality evaluation for the motion-compensated wavelet/subband video coding (MCSBC). The work presented in [4] showed that, through subjective tests with MPEG-4 video sequences, there exists an optimal adaptation trajectory in the space of possible encodings (i.e. combinations of spatial and temporal resolutions), which maximizes visual quality. In [16], quality assessment of low bit-rate video sequences encoded by H.263 and H.264 was performed by considering different dimensions such as spatial resolution, temporal resolution, and bit-rate.

In this paper, we investigate the perceived visual quality of the scalable video sequences produced by two different scalable video codecs. The goal is to provide guidelines for a general bit-rate adaptation strategy for SVC that selects the optimal combination of the scalability dimensions according to a given bit-rate constraint. We conduct extensive subjective quality evaluation tests to reveal the relationship among the temporal, spatial and quality scalability options in enhancing the subjective quality for different contents and codecs. The distinct contributions of the work in comparison to the aforementioned prior researches are as follows. First, we include high definition (HD) sequences having high bit-rates (up to about 4 Mbps) and high frame rates (up to 50 fps) that are of interest in these days for on-line video content distribution such as streaming, while most of the existing studies only considered standard definition sequences with relatively low bit-rates (up to 1 Mbps) and low frame rates (up to 25 fps). Second, various factors affecting the quality are considered for complete analysis, i.e. three scalability dimensions, content type, and scalable video codec, whereas much of the existing work considers only some of them (e.g. [8, 14]). Third, while the previous work simulated different combinations of scalability by using non-scalable codecs (e.g. [4, 16]) or considered only one scalable codec (e.g. [8, 7, 14]), the results of two popular scalable video codecs are analyzed in this paper in order to understand whether there exists general agreement between different codecs.

The remainder of this paper is structured as follows. We first present the two different scalable video codecs under evaluation in Section 2 and describe our subjective test methodology in Section 3. The results are presented in Section 4 and the paper concludes in Section 5.

# 2. SCALABLE VIDEO CODING

The subjective quality of the sequences produced by a scalable video encoder may highly depend on the algorithms used for encoding. Consequently, the performance of an encoder may change significantly over diverse types of content and bit-rate conditions. Therefore, we employed two representative scalable video codecs, namely, H.264/SVC [13] and wavelet-based SVC (W-SVC) [2], in order to investigate the effect of the encoding scheme of SVC on the perceived quality. The former is a standardized discrete cosine transform-based SVC, while the latter is a popular alternative using a wavelet transform. In this section, the two scalable video codecs are briefly described.

## 2.1 Scalable Extension of H.264/AVC

The latest H.264/MPEG-4 AVC standard provides a fully scalable extension, H.264/SVC, which achieves significant compression gain and complexity reduction when scalability is sought, compared to the previous video coding standards [13]. According to evaluations done by MPEG, SVC based on H.264/AVC provided significantly better subjective quality than alternative scalable technologies at the time of standardisation. H.264/SVC reuses the key features of H.264/AVC and also employs some other new techniques to provide scalability and to improve coding efficiency. It provides temporal, spatial and quality scalability with a low increase of bit-rate relative to the single layer H.264/AVC.

The scalable bit-stream is organized into a base layer and one or several enhancement layers. Temporal scalability can be enabled by using hierarchical prediction structures. Spatial scalability is achieved using the multi-layer coding approach. Each layer corresponds to a supported spatial resolution. Within each spatial layer, single-layer coding techniques are employed. Moreover, inter-layer prediction mechanisms are utilized to further improve the coding efficiency. Quality scalability is provided using the coarse-grain quality scalability (CGS) and medium-grain quality scalability (MGS). CGS is achieved by requantization 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 can be combined into one scalable bit-stream that allows for extraction of different operation points of the video.

## 2.2 Wavelet-based Scalable Video Coding

Although a hybrid based technology was chosen for standardisation within MPEG, a great amount of research continued also on W-SVC. Several recent W-SVC systems (e.g. [9]) have shown a very good performance in different types of application scenarios, especially when fine granular quality scalability is required.

A typical W-SVC encoder is shown in Figure 1. First, the input video is subjected to a spatio-temporal (ST) decomposition, which is based on a wavelet transform. The purpose of the decomposition is to decorrelate the input video content and provide the basis for spatial and temporal scalability. The ST decomposition results in two distinctive types of data: wavelet coefficients representing the texture informa-

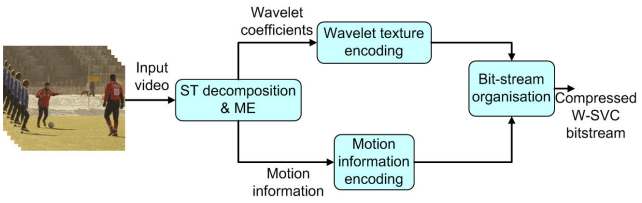


**Figure 1: Structure of wavelet-based scalable video encoder.**

tion remaining after the wavelet transform and motion information obtained from motion estimation (ME), which describe spatial displacements between blocks in neighbouring frames. Although the wavelet transform generally performs very well in the task of video content decorrelation, some amount of redundancies still remains between the wavelet coefficients after the decomposition. Moreover, a strong correlation also exists between motion vectors. For these reasons, further compression of the texture and motion vectors is performed. Texture coding is performed in conjunction with so-called embedded quantisation (bit-plane coding) in order to provide the basis for quality scalability. Finally, the resulting data are mapped into the scalable stream in the bit-stream organisation module, which creates a layered representation of the compressed data. This layered representation provides the basis for low-complexity adaptation of the compressed bit-steam.

## 3. SUBJECTIVE TESTS

In order to understand whether, for a fixed bit-rate, a reduction of spatial resolution, temporal resolution, or frame quality is the best choice to optimize the overall quality of the video sequence as perceived by the end user, pair-wise comparison tests were carried out. Particularly, the question that we tried to answer was: what are the optimal combinations of the spatial, temporal, and quality scalability dimensions for the best quality of experience for a set of fixed bit-rate conditions?

### 3.1 Test Material

We used three 10 second long HD sequences having different spatial and temporal complexity, namely, DucksTakeOff, IntoTree, and ParkJoy (Figure 2) [1]. Figure 3 shows the spatial information (SI) and temporal information (TI) indices on the luminance component of each content, as indicated in [11]. It is observed that IntoTree has small SI and TI values, while ParkJoy shows large values for both measures. DucksTakeOff has a large SI index but a small TI index.

The original raw sequences having a spatial resolution of 1280×720 and a temporal frequency of 50 Hz were encoded by using the SVC codecs described in Sections 2.1 and 2.2. Various layers of different combinations of spatial resolution, temporal resolution, and frame quality were extracted from the coded bit-streams[1].

First, H.264/SVC reference software JSVM 9.18 [12] was used to code the test sequences without rate control. The default cascading of the quantization parameters over the temporal levels was disabled. CGS was used to support quality

[1]The test sequences and subjective data used in this paper are available on request to the first author.

**Figure 2: Example frame images of the content used, namely, DucksTakeOff, IntoTree, and ParkJoy.**

scalability for each of the spatial layers. Each spatial layer has a quality base layer and a CGS quality enhancement layer. Only the first frame of each sequence was encoded as I frame. In addition, hierarchical B pictures were employed to enable five temporal layers.

Second, the W-SVC method in [9] was used to code the test material. The W-SVC 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.

Among the layers in the bit-streams, we selected some of them for the subjective tests as follows: As our goal is to compare the subjective quality of different scalability options for a given bit-rate condition, we first identified the bit-rate conditions that are common to multiple layers of different scalability options. In the cases of H.264/SVC and low bit-rate conditions of W-SVC, layers having exactly the same bit-rate were not available, in which 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 is covered and diverse quality levels and scalability options are included in the test.

As a result, four to six bit-rate conditions were selected for each content, as shown in Tables 1 and 2 for each codec, respectively. The spatial resolution varies among 320×180, 640×360, and 1280×720. The frame rates are 6.25, 12.5, 25, and 50 fps. In the tables, the frame quality is expressed as the pixel bit-rate ($B_p$) that is defined as

$$B_p = \frac{B}{H \cdot W \cdot F} \tag{1}$$

where $B$, $H$, $W$, and $F$ are the bit-rate, frame height, frame width, and frame rate, respectively. As can be seen in the tables, the comparison within a bit-rate condition is made

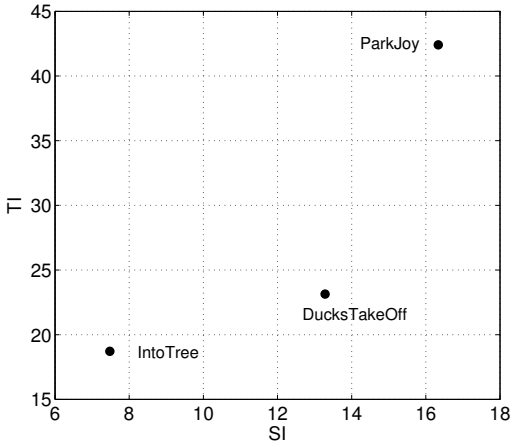


**Figure 3: Spatial information (SI) versus temporal information (TI) indices of the selected contents.**

**Figure 4: Space for the subjective tests.**

among two to six instances having different scalability options according to their availability in the SVC bit-streams. It should be noted that, due to the lack of accurate rate control schemes in the codecs, the layers in the SVC bit-streams produced by the two codecs do not have direct correspondence in terms of bit-rate and scalability options.

## 3.2 Test Environment and Subjects

The test room environment is intended to assure the reproducibility of the subjective test activity by avoiding the involuntary influence of any controllable external factors. Thus, it is important to fix some features of the viewing environment, regarding general viewing conditions and some crucial features of the used monitor.

The tests were performed at the premises of the Multimedia Signal Processing Group (MMSPG) laboratory at EPFL. The test room was equipped with a LCD monitor receiving input from a high performance server that was able to play HD raw content in real time. Detailed information regarding the monitor is shown in Table 3. The ambient lighting consisted of neon lamps with 6500 K color temperature and the wall color was grey 128, as recommended by [10]. One subject per time was sit in front of the screen at a distance about 2-3 times the height of the stimuli. A picture of the test room is shown in Figure 4.

## 3.3 Test Methodology

Several subjective quality evaluation methods can be used for comparison of different scalability combinations having

**Table 1: Selected comparison sets composed of multiple layers having (nearly) the same bit-rates from the bit-streams encoded by JSVM. 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) |

(nearly) the same bit-rate, among which the stimulus comparison (SC) method was used in this work [10]. The method consisted of pair-wise comparison between two stimuli, i.e. video sequences. The subject was asked to indicate which one has better quality. In our case, the stimuli being compared often showed small difference in quality, and the SC method enables subjects to compare the visual quality of such stimuli easily.

The subjective test proceeded as follows. A pair of test stimuli that are in a comparison set (each set in Tables 1 and 2) were played side by side in a time-synchronous manner. We considered all possible pair combinations in each comparison set in order to obtain relative quality scores of all the stimuli in the set (see Section 4.1). A fixed size of the viewing window that is equal to the maximum resolution of the stimuli (i.e. 1280×720) was assumed, considering applications such as video streaming in video sharing websites, where normally the video clips are shown with the same frame size. Thus, video sequences smaller than the resolution of 1280×720 were upsampled to match the fixed resolution by using the bilinear filter. Since a monitor having a native resolution of 2560 × 1600 pixels was used, two video sequences could fit in the horizontal space of the display. The desktop window background was set to grey 128. Each subject was asked to choose which stimulus had a better quality between the two presented, and write the answer on the answer sheet. The option "same" was also included in the possible choices. Each pair of stimuli was played in loop so that each subject could watch them as many times as she/he wanted in order to carefully analyze and rate them. To limit the duration of a test session, the stimuli presentation was divided into two separate sessions, each of which contained about 50 pairs of stimuli.

Prior to the test sessions, a training session took place, where the test methodology was described to the subject by using a set of training stimuli different from the test stimuli.

**Table 2: Selected comparison sets containing multiple layers having (nearly) the same bit-rates from the bit-streams encoded by W-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 | (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 | (384, 320×180, 6.25, 0.27), (384, 320×180, 6.25, 1.09) |
| 2 | (520, 320×180, 6.25, 1.48), (520, 640×360, 6.25, 0.37) |
| 3 | (768, 320×180, 12.5, 1.07), (768, 640×360, 12.5, 0.27) |
| 4 | (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) |
| 5 | (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) |

Sixteen subjects (11 men and 5 women) participated in the experiment. They reported normal or corrected to normal vision. The average subject age was 28.2 years old.

# 4. RESULTS

## 4.1 Subjective Data Processing

For a comparison set containing $n$ different scalability combinations, $R_1, R_2, ..., R_n$ (in our case $n$ varies from 2 to 6), there are $\binom{n}{2}$ pairs to be compared. The comparison results for the set can be summarized by a matrix of winning frequencies $\{c_{ij}\}$. By treating a tie as a half way between the two preference options, $c_{ij}$ is computed as [5]:

$$c_{ij} = 2 \times p_{ij} + q_{ij} \quad (2)$$

where $p_{ij}$ is the number of subjects who preferred $R_i$ over $R_j$ and $q_{ij}$ the number of subjects who rated them the same. Note that $c_{ij} + c_{ji} = 32$, which is two times the number of subjects. An example of the matrix for $n = 4$ is shown in Table 4.

In order to obtain continuous scale quality score values for $R_i$'s from the matrix of winning frequencies, we used the Bradley-Terry-Luce (BTL) model that is frequently applied for analysis of pair comparison data [3, 6]. In this model, the probability of choosing $R_i$ against $R_j$, $P_{ij}$, is represented as

$$P_{ij} = \frac{c_{ij}}{c_{ij} + c_{ji}} = \frac{\pi_i}{\pi_i + \pi_j} \quad (3)$$

**Table 3: 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 |

**Table 4: Example of pair comparison results for a bit-rate condition with $n = 4$**

| | $R_1$ | $R_2$ | $R_3$ | $R_4$ |
|---|---|---|---|---|
| $R_1$ | 0 | $c_{12}$ | $c_{13}$ | $c_{14}$ |
| $R_2$ | $c_{21}$ | 0 | $c_{23}$ | $c_{24}$ |
| $R_3$ | $c_{31}$ | $c_{32}$ | 0 | $c_{34}$ |
| $R_4$ | $c_{41}$ | $c_{42}$ | $c_{43}$ | 0 |

where $\pi_i$ is the quality score of $R_i$, which is referred to as the true rating or preference in literature. Every $\pi_i \geq 0$ and $\sum_i \pi_i = 1$. $\pi_i$'s can be estimated by the maximum likelihood estimation based on the emperical probability values $P_{ij}$'s. In addition, the confidence intervals for the maximum likelihood estimates of the scores can be obtained from the Hessian matrix of the log-likelihood function [15]. For analysis, we normalized the scores of the instances in a comparison set so that the maximum score becomes 100. Thus, comparison across the comparison sets (over different bit-rate conditions, contents, and codecs) is not valid.

## 4.2 Results and Analysis

In this section, the results of the tests and their analysis are presented. The quality scores obtained from the comparison tests and the confidence intervals are depicted in Figures 5 to 10. The compared scalability combinations are shown below 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 1 and 2.

Figures 5 to 7 show the results of H.264/SVC. When a pair of sequences have the same spatial resolution, the one having a larger frame rate is preferred (Figures 5(a), 6(c)-(d), and 7(a)). The last pair of ParkJoy, Figure 7(d), is an exception, which might be because the high TI index of the content (as shown in Figure 3) 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).

Figures 5(b), 5(d), 6(a)-(b), and 7(b)-(c) show the results of the cases where the combination of the frame rate and frame resolution is different from each other. 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 (Figures 5(b), 6(a), and 7(b)). However, when the bit-rate becomes large, a high frame rate is more important than a high spatial resolution (i.e. 1280×720), as shown in Figures 5(d), 6(b) and 7(c).

Figures 5(c), (e), and (f) compare the cases of the same

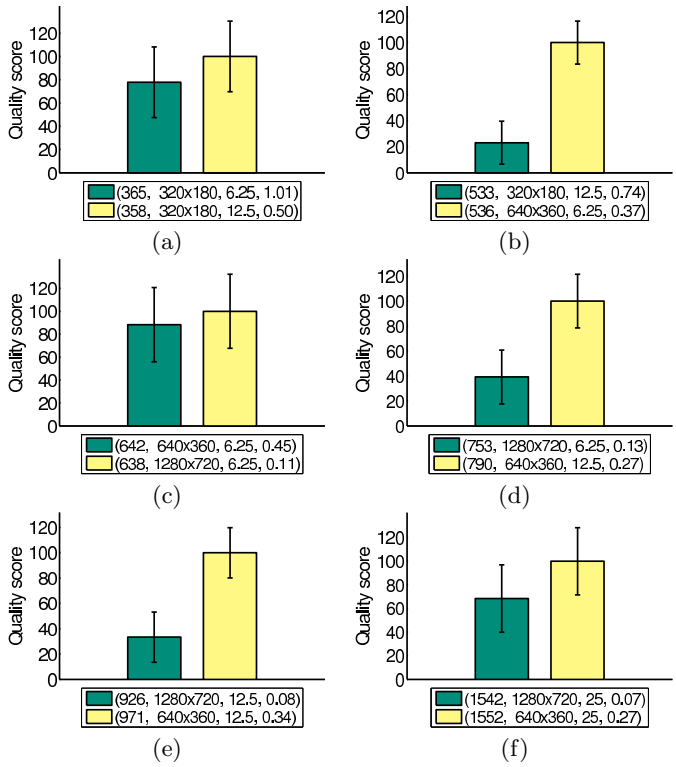


**Figure 5: Results of the subjective tests for DucksTakeOff encoded by JSVM.**

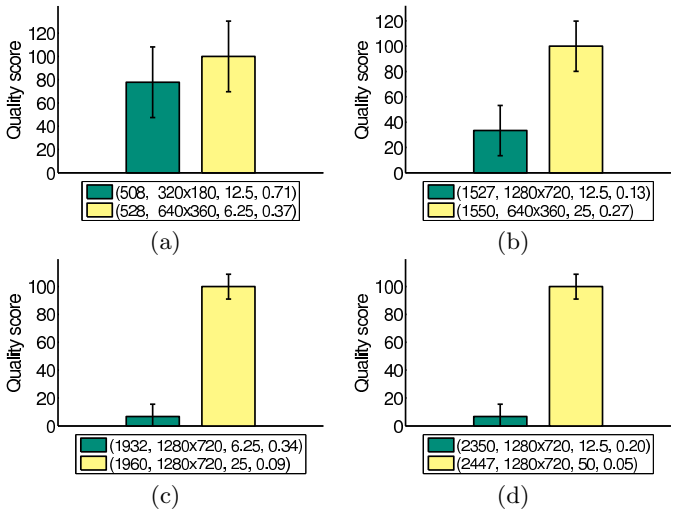


**Figure 6: Results of the subjective tests for IntoTree encoded by JSVM.**

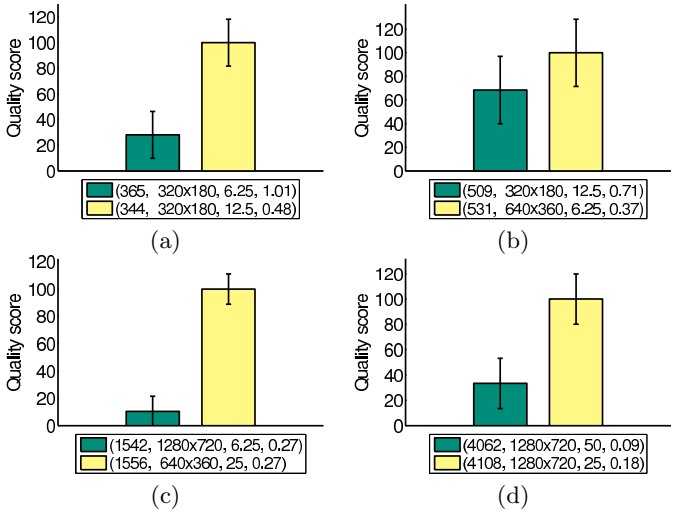


**Figure 7: Results of the subjective tests for ParkJoy encoded by JSVM.**

frame rate but different resolutions (1280×720 and 640×360). Except for the statistically insignificant case (Figure 5(c)), the resolution of 640×360 is preferred because the frame quality is not good enough and the blocking artifacts are observed in the cases of the resolution 1280×720.

In Figures 8 to 10, the results of W-SVC for each content are shown, respectively. For the comparison sets whose bit-rates are less than 1024 kbps, the comparison is made between the two spatial resolutions, 320×180 and 640×360, for the same frame rates (Figures 8(a)-(b), 9(a)-(c), and 10(a)-(b)). In most cases, a larger spatial resolution is preferred against a smaller one because the blurring effect is stronger in a smaller resolution (Figures 8(a)-(b) and 9(a)-(c)). However, ParkJoy shows different behavior, i.e. the smaller resolution is preferred (Figure 10(a)). In this case, a low frame rate such as 6.25 fps is not sufficient for the content containing relatively fast visual motion. Thus, the subjects prefer the more strongly blurred scene in the low resolution that partially compensates for the jerky motion in the low frame rate sequence. As for Figure 10(b), the difference of the two cases is not statistically significant.

Figures 8(c), 9(d), and 10(c) show the results when the bit-rate is 1024 kbps. In these cases, 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. For DucksTakeOff and ParkJoy, the combination of $(W \times H, F) = (1280 \times 720, 12.5)$ has the best quality, whereas $(W \times H, F) = (640 \times 360, 25)$ is 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 done among different combinations of the frame rate and the frame quality. The results show that a larger frame rate is always preferred (Figures 8(d), 9(e), and 10(d)). For ParkJoy, quality difference of the cases of 25 fps and 50 fps is not statistically significant, which is in line with the exceptional result for JSVM in Figure 7(d).

Interestingly, the pixel bit-rate does not have clear relationship with the perceived quality in both codecs, whereas the previous work in [16] indicated that a higher quality score is usually obtained for a higher pixel bit-rate. Rather, the spatial and temporal resolutions seem to be more important for the perceived quality in our case. The difference

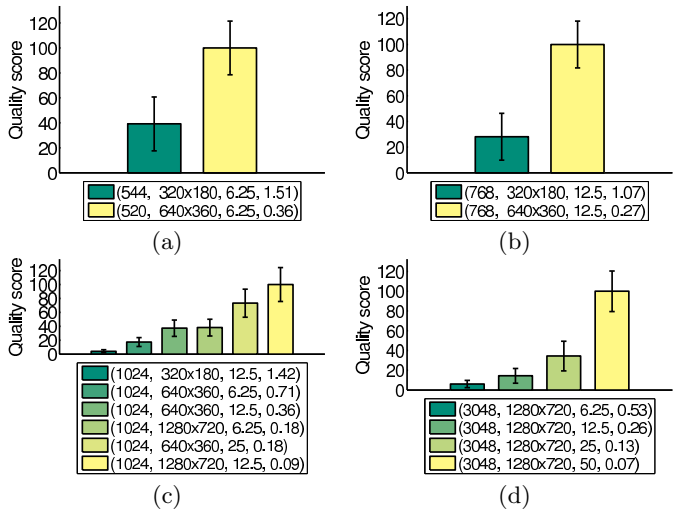


**Figure 8: Results of the subjective tests for Ducks-TakeOff encoded by W-SVC.**

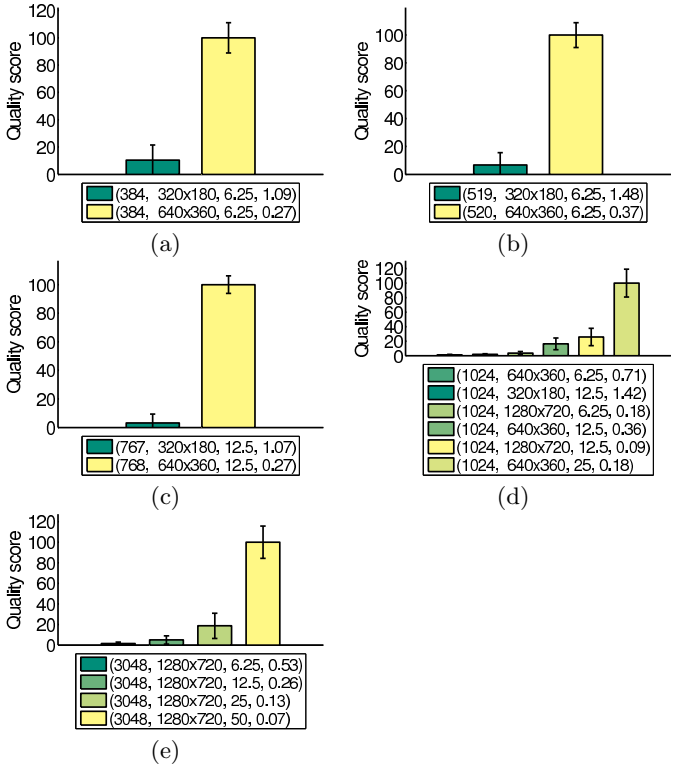


**Figure 9: Results of the subjective tests for IntoTree encoded by W-SVC.**

may stem from the applications considered, i.e. the low bit-rate video sequences in [16] and HD sequences in this paper.

Overall, the following observations can be drawn. First, when the bit-rate is small and only layers having small spatial resolutions are available, a larger spatial resolution is preferable in order to obtain the lowest acceptable frame quality without strong blurring. In the case of H.264/SVC, this observation was valid even when the frame rate decreases along with increase of the spatial resolution. Second, for large bit-rate conditions, acceptable frame quality is achieved and thus a high frame rate, which is obtained at the cost of the decreased pixel bit-rate, becomes important for better subjective quality. The threshold between the small and large bit-rate values appears as between about 800 kbps and 1 Mbps for the case of W-SVC; for H.264/SVC, it may be estimated as about 700 kbps, which remains inconclusive due to limited availability of diverse layers having the same bit-rate values. Third, the content is an important factor that influences the perceptual quality of different scalability combinations, which can be described by the SI and TI indices to some extent. Fourth, while the encoder type affects the quality evaluation results significantly because each encoder produces SVC bit-streams containing different structures, it is observed that the aforementioned overall tendency remains quite consistent across the two codecs.

# 5. CONCLUSIONS

In this paper, extensive subjective quality evaluation results on two scalable video codecs have been presented. By using the pair comparison scheme, the perceived quality of HD sequences was analyzed in five dimensions, i.e. codec, content, frame size, frame rate, and frame quality. The results showed that the preference between the spatial resolution and frame rate depends on the bit-rate condition and content type, which was quite consistent across the two codecs. It is expected that the drawn observations are useful as guidelines for an adaptive decision strategy of scalability options for resource-constrained networks. In our future work, we will work on developing objective quality measures of video sequences generated by SVC based on the subjective evaluation results reported in this paper.

# 6. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2011) under grant agreement no. 21644 (PetaMedia). Furthermore, the authors gratefully acknowledge the support of the Swiss National Foundation for Scientific Research in the framework of the NCCR Interactive Multimodal Information Management (IM2).

# 7. REFERENCES

[1] The SVT High Definition Multi Format Test Set. `http://www.its.bldrdoc.gov/vqeg`.

[2] N. Adami, A. Signoroni, and R. Leonardi. State-of-the-art and trends in scalable video compression with wavelet-based approaches. *IEEE Trans. Circuits and Systems for Video Technology*, 17(9):1238–1255, 2007.

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

[4] N. Cranley, P. Perry, and L. Murphy. Optimum adaptation trajectories for streamed multimedia. *Multimedia Systems*, 10(5):392–401, 2005.

[5] M. E. Glickman. Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 48(3):377–394, 1999.

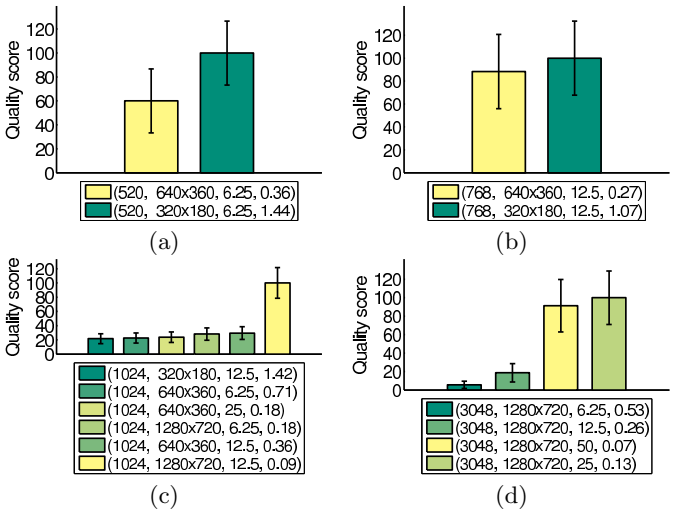


**Figure 10: Results of the subjective tests for ParkJoy encoded by W-SVC.**


[6] R. D. Luce. *Individual choice behavior: A theoretical analysis*. Wiley, New York, USA, 1959.

[7] T. Oelbaum, H. Schwarz, M. Wien, and T. Wiegand. Subjective performance evaluation of the SVC extension of H.264/AVC. In *Proc. 15th IEEE Int. Conf. Image Processing (ICIP)*, pages 2772–2775, Oct. 12–15, 2008.

[8] 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*, pages 445–448, 2002.

[9] N. Ramzan, T. Zgaljic, and E. Izquierdo. An efficient optimisation scheme for scalable surveillance centric video communications. *Signal Processing: Image Communication*, 24(6):510–523, 2009.

[10] Methodology for the subjective assessment of the quality of television pictures. Recommendation ITU-R BT.500-11, 2002.

[11] Subjective video quality assessment methods for multimedia applications. Recommendation ITU-R P.910, 1999.

[12] J. Reichel, H. Schwarz, and M. Wien. Joint scalable video model 11 (JSVM 11). *Joint Video Team, doc. JVT-X202*, Jul. 2007.

[13] H. Schwarz, D. Marpe, and T. Wiegand. Overview of the scalable video coding extension of the H.264/AVC standard. *IEEE Trans. Circuits and Systems for Video Technology*, 17(9):1103 – 1120, Sep. 2007.

[14] Y. Wang, S.-F. Chang, and A. C. Lou. Subjective preference of spatio-temporal rate in video adaptation using multi-dimensional scalable coding. In *Proc. Int. Conf. Multimedia and Expo (ICME)*, pages 1119–1122, 2004.

[15] F. Wickelmaier and C. Schmid. A Matlab function to estimate choice model parameters from paired-comparison data. *Behavior Research Methods, Instruments, and Computers*, 36(1):29–40, 2004.

[16] 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*, 10(7):1316–1324, Nov. 2008.