

Quality Scalability in H.264/AVC Video Coding

Ville-Pekka Limnell^{*a}, Dong Tian^a, Miska M. Hannuksela^b, Moncef Gabbouj^c

^aTampere Institute of Signal Processing, Tampere, Finland

^bNokia Research Center, Tampere, Finland

^cTampere University of Technology, Tampere, Finland

ABSTRACT

It is well-known that the problem of addressing heterogeneous networks in multicast can be solved by simultaneous transmission of multiple bitstreams of different bitrates and by layered encoding. This paper analyzes the use of H.264/AVC video coding in simulcast and for layered encoding. The sub-sequence feature of H.264/AVC enables hierarchical temporal scalability, which allows disposal of reference pictures from a coded bitstream without affecting the decoding of the remaining stream. In this paper we extend the scope of the H.264/AVC sub-sequence coding technique to quality scalability. The resulting quality scalable coding technique is similar to conventional coarse-granularity quality scalability but fully compatible with the H.264/AVC standard. It is found that the proposed method drops bitrate consumption in the core network compared to simulcast up to 20 %. However, the bitrate required for enhanced-quality reception for scalably coded bitstreams is considerably higher than that of non-scalable bitstreams.

Keywords: H.264/AVC, sub-sequence, scalable video, quality scalability, SNR scalability

1. INTRODUCTION

Scalable video coding is desirable in heterogeneous and error-prone environments for various reasons. For example, scalable coding helps streaming servers avoid congestions in network by allowing the server to reduce the bitrate of bitstreams whilst still transmitting a useable bitstream. One application for scalability is to improve error resilience in transport systems that allow different qualities of service. For example, the essential information could be delivered through a channel with high error protection. Scalability can also be used to enable different quality representations depending on playback devices processing power. Devices with better processing power can decode and display the full quality version, whereas devices having lower processing power decode the lower quality version.

There are three conventional types of scalability: temporal, quality and spatial. Temporal scalability enables adjustment of picture rate. This is commonly carried out with either disposable pictures or disposable sub-sequences, which are explained later on. Picture rate adjustment is then simply done by removing these disposable parts from the coded sequence thus lowering the frame rate. In conventional quality scalability, also known as SNR scalability, an enhancement layer is achieved with pictures having finer quantizers than the particular picture in the lower reference layer. In coarse-granularity quality scalability, pictures in enhancement layers may be used as prediction references and therefore all the enhancement layer pictures in a group of pictures typically have to be disposed as a unit. In fine-granularity scalability, the use of enhancement layer pictures as prediction sources is limited and therefore finer steps of bitrate can be achieved compared to coarse-granularity scalability. Finally, spatial scalability is used for creation of multi-resolution bitstreams to meet different display requirements or constraints and is very similar to SNR scalability. A spatial enhancement layer enables recovery of coding loss between an up-sampled version of the reconstructed layer used as a reference by the enhancement layer and a higher resolution version of the original picture.

The H.264/AVC coding standard, published as ITU-T Recommendation H.264 and ISO/IEC International Standard 14496-10, contains an extensive quantity of features, out of which we review the ones that are essential for the presented quality scalability method. H.264/AVC allows the use of multiple reference pictures for motion compensation, i.e. there is a reference picture buffer containing multiple decoded pictures from which an encoder can select a reference picture for inter prediction on block basis. In addition to reference pictures, which are stored to the reference picture buffer, the H.264/AVC features non-reference pictures, which cannot be used as prediction source for inter prediction. In contrast to earlier standards in which “disposable” pictures were always B pictures, non-reference pictures in H.264/AVC can be of

* ville-pekka.limnell@tut.fi; phone +358 3 3115 3880

any coding type. Decoupling of decoding and output order of pictures not only enables conventional B-picture-like temporal scalability in H.264/AVC but also facilitates a hierarchical temporal scalability scheme referred to as sub-sequences. Sub-sequences are used to construct a layered bitstream in which each enhancement layer contains sub-sequences and each sub-sequence contains a number of reference and/or non-reference pictures. Decoded pictures are buffered for two reasons, for prediction references and for reordering of decoded pictures from decoding order to output order. In order to minimize memory consumption, the H.264/AVC standard specifies a decoded picture buffer that unifies the processing for both needs and reduces memory consumption by maintaining only one copy of any decoded picture.

In this paper, we apply scalable H.264/AVC video coding in IP-based multicast and broadcast systems, and hence the key characteristics of these systems for video processing are reviewed in the following. Most of these multicast and broadcast systems utilize IP multicast in the core network to address heterogeneous downlink networks and to avoid unnecessary network traffic. Two solutions are commonly mentioned in the literature to cope with access links of different bitrates. First, in simulcast, multiple independent streams of different bitrates but originating from the same source sequence are sent simultaneously. Network routers can forward only those bitstreams that are supported by downlink. One problem with simulcast is that multiple copies of the same source sequence may result into considerable raise in the bitrate in the core network. Another solution to tackle network heterogeneity is layered IP multicast transmission. Each layer of a scalably coded bitstream is sent in its own multicast group, and receivers can subscribe to as many groups as they are capable of receiving or processing.

As reviewed above, H.264/AVC supports temporal scalability but does not contain features for quality and spatial scalability. In this work we present a novel technique enabling quality scalability with H.264/AVC. The proposal does not require any changes in the H.264/AVC syntax or decoding process and is therefore fully compatible with existing H.264/AVC decoder implementations. We use the sub-sequence feature to code multiple quality versions of the same uncompressed picture sequence within the same coded bitstream. Each quality version resides in its own sub-sequence layer, and sub-sequence layers form a hierarchical prediction structure. Coded pictures that originate from the same uncompressed picture share the same display timestamp and therefore only the last received and decoded picture having a particular display timestamp is displayed. The proposed method allows including differently quantized versions of the same uncompressed picture sequence into one coded video stream.

The Joint Video Team (JVT) of ITU-T Video Coding Experts Group (VCEG) and ISO/IEC Moving Picture Experts Group (MPEG) is in the process of specifying the Scalable Video Coding (SVC) standard². The base layer coding of the standard is chosen to be fully compatible with H.264/AVC. The goals of the standard include fine- and coarse-granularity quality scalability, spatial scalability, and improved temporal scalability. It is noted that in the method presented in this paper also the enhancement layer coding is compatible with H.264/AVC, and therefore the presented method enables quality scalability with fewer changes in codec implementations compared to the upcoming SVC standard.

This paper is constructed as follows. Section 2 reviews the proposed quality scalable coding method we present. The simulation results are presented in Section 3. Finally, in Section 4, we conclude our work.

2. PROPOSED QUALITY SCALABLE CODING

As the proposed coding method is based on the use of the sub-sequence coding tool, we review the sub-sequence method in Section 2.1. Then, Section 2.2 introduces how the proposed method utilizes sub-sequences to achieve quality scalability.

2.1. Sub-Sequences and Sub-Sequence Layers in H.264/AVC

A sub-sequence consists of a number of inter-dependent pictures that can be disposed without any disturbance to any other sub-sequence in any lower sub-sequence layer. When a sub-sequence in the highest enhancement layer is disposed, the remaining bitstream remains valid. The H.264/AVC design includes a single decoded picture buffer and reference picture buffer regardless of the number of sub-sequence layers. Moreover, one of the design goals of the sub-sequence feature was to make the decoding process unaware of the hierarchical nature of the bitstream. Consequently, no layer number is signaled in the integral part of the bitstream or plays a role in the decoding process. In contrast, the `frame_num` syntax element is incremented by 1 per each reference frame, and receivers infer disposed reference frames from gaps in `frame_num` value and insert a “non-existing” frame to the reference picture buffer as if the frame was received and

decoded. This procedure guarantees that the contents of and the picture order in the reference picture picture buffer remain unchanged when it comes to the remaining reference pictures.

Sub-sequence layers are hierarchically arranged based on their dependency on each other. The base layer (layer 0) is independently decodable. Sub-sequence layer 1 depends on some data in layer 0 i.e. to correctly decode a picture in sub-sequence layer 1 one needs to have decoded all referred pictures in sub-sequence layer 0, whether in a direct manner or not. Generally, correct decoding of sub-sequence layer N requires decoding of layers from 0 to N-1. It may be possible to arrange pictures to sub-sequences and sub-sequence layers in multiple ways. However, within one such arrangement, each coded picture belongs to exactly one sub-sequence, and each sub-sequence belongs to exactly one sub-sequence layer in any subsequence layer.

The sub-sequence coding method was reviewed in more details in ¹. The paper also analyzed the compression efficiency of certain coded picture patterns (e.g. so-called IBBP) in several picture rates, picture sizes, and sequences, with and without bi-prediction (i.e. conventional B pictures). The paper concluded that temporally scalable H.264/AVC coding outperforms non-scalable coding in terms of compression efficiency measured with average luma PSNR. Furthermore, it was found that the tested sub-sequence picture pattern provided similar or better compression performance than the use of non-reference pictures only when bi-prediction was not in use. Moreover, the simulations showed that sub-sequences outperformed the conventional IBBP coding pattern in compression efficiency when bi-prediction was in use.

2.2. Proposed Quality Scalable Coding Using Sub-Sequences

In this sub-section, we apply the sub-sequence coding scheme for quality scalability. In conventional coarse quality scalability, such as in Annex O of H.263, an enhancement layer picture can be upward-predicted from a temporally corresponding lower layer picture, forward-predicted from the previous reference picture in the same layer, or bi-predicted from both of these pictures. We create a similar coding arrangement in the proposed H.264/AVC-based coding method explained in the following.

Whereas each picture is normally coded only once in the sub-sequence coding method for temporal scalability, in the proposed technique each source picture is coded multiple times with different quantization parameters (QPs). The first picture coded from a source picture is the base layer representation of the picture coded with the coarsest quantization step. The second picture coded from the source picture resides in the first enhancement layer. It can be predicted upwards from the corresponding base layer picture or forwards from the previous reference picture in the first enhancement layer. In addition to these conventional prediction sources, the picture can be predicted from any reference picture in the same or lower layer provided that the reference picture is included in the reference picture buffer. The number of pictures coded from the same source picture is equal to the number quality scalability layers.

In the decoding end, the decoder processes a quality scalable H.264/AVC identically to any valid bitstream. As the H.264/AVC standard disallows equal output timestamps except for the two fields of a frame, an H.264/AVC decoder outputs all decoded pictures. Therefore, the output timestamps of the pictures are carried outside the integral H.264/AVC bitstream by some other means, such as RTP timestamps. In addition, the decoding terminal must have an additional rendering buffer, which maintains only the latest decoded picture of a single output timestamp.

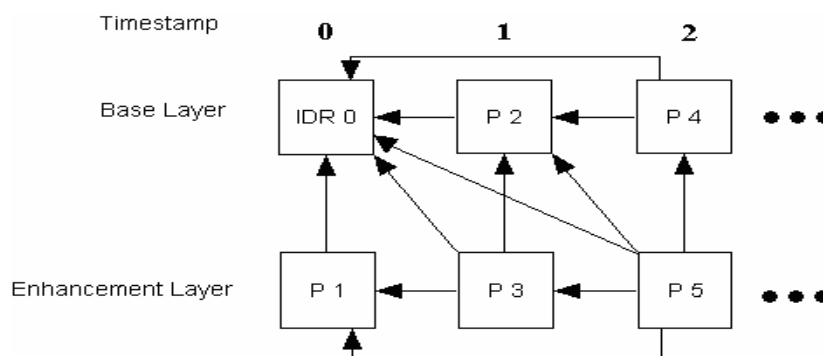


Figure 1: Example of sub-sequences: Quality scalable bitstream (The numbers behind the picture types in the figure indicates the encoding order and the number of reference frames is 5.)

Figure 1 illustrates an example of quality scalable coding with two layers using sub-sequences techniques. Each original picture is coded twice. For instance, the picture 0 is coded into IDR0 and P1 in the bitstream. The numbers associated with the picture coding types indicate the encoding /decoding order. The number of reference frames in the reference frame buffer is assumed to be 5. The coding procedure is as follows: Each original picture is input to the H.264 /AVC encoder twice. The reconstructed pictures from the first encoding are treated as base layer and the second encoding produces the enhancement layer. It is proposed to use a coarser QP for the first coding and a finer one for the second. The reconstructed pictures, including both of the reconstructed pictures from the same original picture, are put into the reference picture buffer immediately after they are coded. The encoder has to be aware that only the reconstructed pictures marked as base layer can be used as reference pictures for the other pictures in the base layer. The arrows in figure 1 show the valid reference pictures in the scalable coding scheme. Note that no changes are required in the H.264 /AVC decoder for the selective use of reference pictures. The reference picture buffer operates normally, e.g. the oldest reference picture will be marked as non-reference by the “sliding window” operation.

3. SIMULATIONS

We benchmarked our proposed quality scalable H.264/AVC against traditional simulcast in multicast streaming. We compared the bitrate and picture quality of the following cases

- 1) Resource consumption in the core network. The scalably coded bitstream compared to all streams in the simulcast. This case characterizes the required bitrate in the core network, in which no selection of forwarded bitstreams according to the available downlink bitrates can be made. Moreover, this case is applicable to broadcast systems offering multiple quality versions of the same bitstream.
- 2) Basic quality reception. The base layer of the scalable bitstream compared to the lowest bitrate stream in the simulcast. This case characterizes the subjective quality difference when the recipients get only the stream or the layer of the lowest transmitted quality.
- 3) Enhanced quality reception. The scalably coded bitstream compared to the highest bitrate stream in the simulcast. This case characterizes the subjective quality versus required bitrate when the recipients receive, decode, and play the bitstream of the highest transmitted quality.

This section is organized as follows.

3.1. Simulation Conditions

To simplify the tests we used two layers of quality scalability and compared it to normal non-scalable coding. In order to get comparable results, we turned bitrate control off in the encoder and used a constant quantizer throughout a non-scalable bitstream or within one scalability layer. To obtain an understanding on the achievable bitrate scalability, we coded two sets of scalable bitstreams: in the first one we used quantization parameter difference of 4 between the base layer and the enhancement layer, and in the second one the QP difference was 8. As a summary, for each tested original sequence, we created the following three sets of coded bitstreams:

- 1) Non-scalable bitstreams, QPs 16, 20, 24, 28, and 32
- 2) Scalable bitstreams, the following pairs of QPs for the base layer and enhancement layer: (24,16), (28,20), (32,24), (36,28)
- 3) Scalable bitstreams, the following pairs of QPs for the base layer and enhancement layer: (20,16), (24,20), (28,24), (32,28)

H.264/AVC Baseline Profile was used in the simulations, which causes that no bi-prediction was applied. The coding parameters for both scalable and non-scalable bitstreams were selected according to the appropriate Levels of H.264/AVC. For example, the size of the decoded picture buffer was selected according to level 1 (QCIF) and level 2 (CIF): we used 4 reference frames for QCIF sequences and 6 for CIF sequences. Rate-distortion optimized mode selection was turned on in the encoder, giving a similar compression performance as the Joint Model (JM) encoder².

Bjontegaard delta bitrate³ is a way of finding numerical averages between rate distortion curves as part of presentation of results. This is a more compact and in some sense more accurate way of presenting data and is used in addition to the rate distortion plots.

3.2. Resource Consumption in the Core Network

There can be a significant save in core networks resource consumption when the stream is quality scaled. Two rate distortion curves are provided. Figure 2 and Figure 3 are rate distortion curves from Foreman sequence in QCIF format and Coast guard in CIF format, respectively. Foreman was encoded with difference of 8 in quantization parameter and Coastguards difference in quantization parameter was 4. A Bjontegaards delta bitrate in rate distortion is also provided in

Table 1. From Table 1 it can be seen that encoding coastguard with quality scalability can result in almost 20% decrease in resource consumption in the core network.

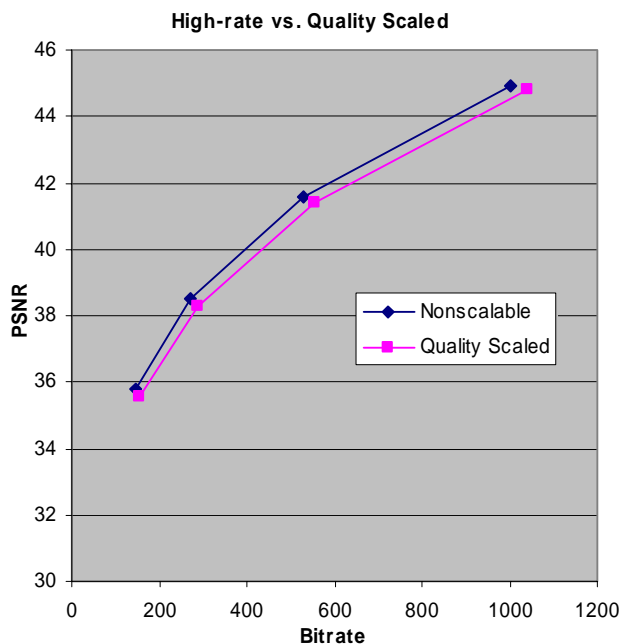


Figure 2: Rate-distortion curves for Foreman (QCIF) (nonscalable is normally encoded and contains the same quantization parameters as quality scaled both layers and Quality Scaled represents our quality scalable bitstreams both layers)

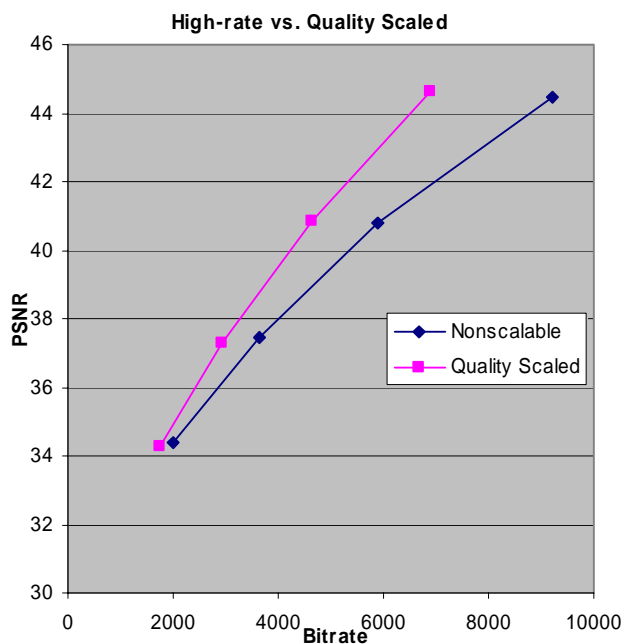


Figure 3: Rate-distortion curves for Coastguard (CIF) (nonscalable is normally encoded and contains the same quantization parameters as quality scaled both layers and Quality Scaled represents our quality scalable bitstreams both layers)

Table 1: Average rate-distortion differences (bitrate) (a: not scaled vs. quality scaled bitstreams both layers with 4 units' difference, b: not scaled vs. quality scaled bitstreams both layers with 8 units' difference. A positive value implies the former scheme outperforms the latter)

Sequences		a	b
QCIF	News	8,93	9,32
	Foreman	2,66	9,95
CIF	Paris	-14,08	-0,33
	Coastguard	-19,74	-4,96

3.3. Basic Quality Reception

Assuming the routers are IP-multicast-aware, there will be no unnecessary sending of layers or streams, which means either only low-rate stream/base layer is sent, or only high-rate stream/both base and enhancement layers are sent. We compared the rate-distortion performance of different coding schemes at full frame rate. The rate-distortion curves comparing low-rate non-scalable stream against quality scalable bitstreams base layer are shown in Figure 4 and Figure 5 for Foreman-sequence in QCIF-version and Paris-sequence in CIF-version, respectively. Bjontegaard delta bitrate was used to evaluate the average differences between rate-distortion curves. Table 2 contains the Bjontegaard delta bitrate values of the competitive pairs: Not scalable (normally encoded) vs. quality scalable bit-streams base layer with 4 units' difference in quantization parameters between layers and Not scalable vs. quality scalable bit-streams base layer with 8 units' difference. A positive value implies the former scheme outperforms the latter. It can be found that the compression performance of quality scalable bit-stream in this case is very close to that of normally encoded. The performance drop in the base layer coding compared to the non-scalable coding comes from the fewer available reference pictures.

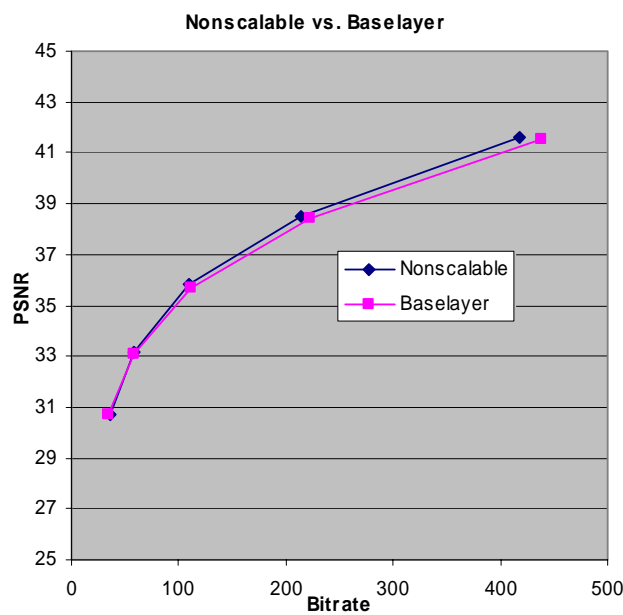


Figure 4: Rate-distortion curves for Foreman (QCIF) (nonscalable is normally encoded and baselayer represents our quality scalable bitstreams base layer)

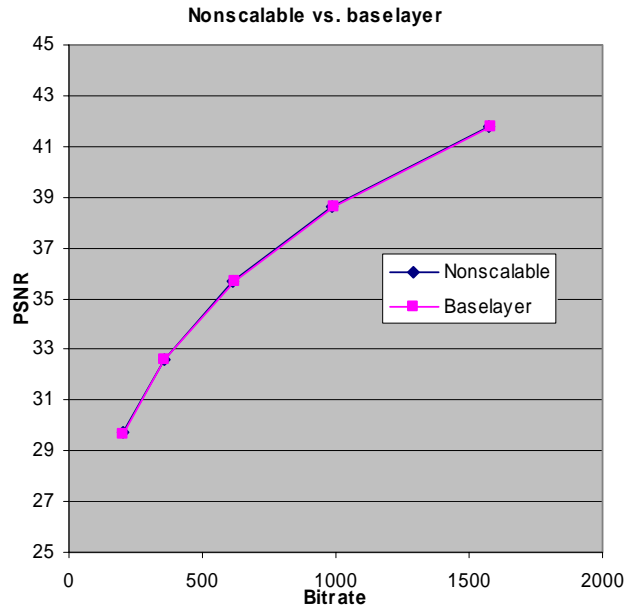


Figure 5: Rate-distortion curves for Paris (CIF) (nonscalable is normally encoded and baselayer represents our quality scalable bitstreams base layer)

Table 2: Average rate-distortion differences (bitrate) (a: not scaled vs. quality scaled bitstreams base layer with 4 units' difference, b: not scaled vs. quality scaled bitstreams base layer with 8 units' difference. A positive value implies the former scheme outperforms the latter)

Sequences		a	b
QCIF	News	0.19	0.02
	Foreman	4.70	2.28
CIF	Paris	0.72	0.94
	Coastguard	0.51	0.65

3.4. Enhanced Quality Reception

If only high-rate stream or both base and enhancement layers are sent, quality scalability performs quite poorly compared to normally encoded high-rate bitstream. Results from this scenario, where they are presented as Bjontegaard delta bitrates, can be seen in Table 3. Consequently, if the resource consumption in the core network is not an issue and if a majority of receivers are expected to receive an enhanced-quality stream, the results indicate that simulcast is preferred over layered coding. The rate distortion curves comparing high-rate stream against quality scaled bitstream with both layers are presented in Figure 6 and 7, where Figure 6 is Foreman in QCIF format and Figure 7 is Coastguard in CIF format. Quantization parameters of 16, 20, 24, 28, 32 and 36 were used in Foreman sequences non scalable bitstream and points in quality scaled stream are 24/16, 28/20, 32/24 and 36/28 where slash separates the quantization parameters between base and enhancement layers.

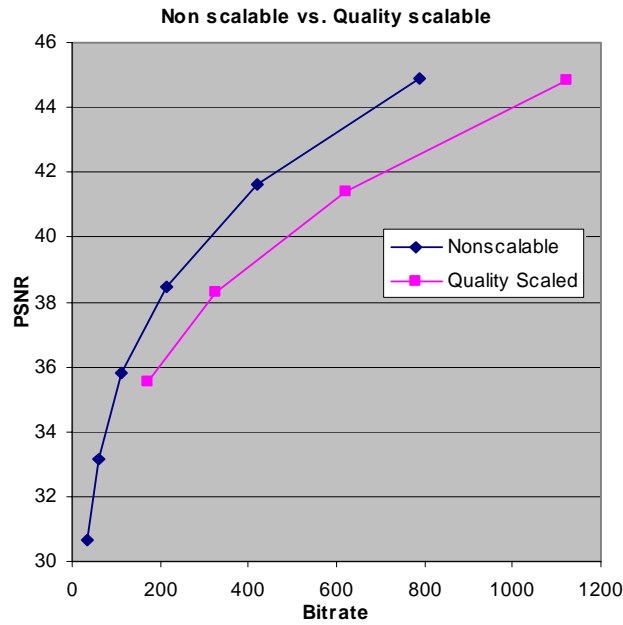


Figure 4: Rate distortion curves for Foreman (QCIF) (nonscalable is normally encoded with QPs from 16 to 36 and Quality Scaled represents our quality scalable bitstream coded with QP difference of 8)

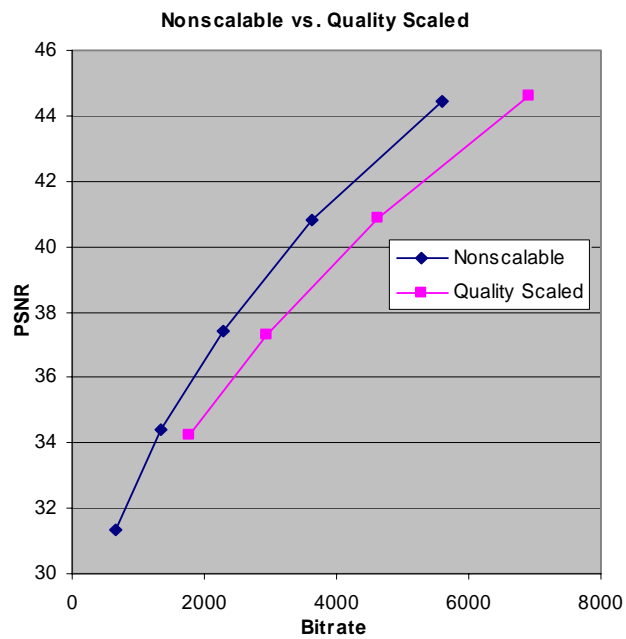


Figure 5: Rate distortion curves for Coastguard (CIF) (nonscalable is normally encoded with QPs from 16 to 32 and Quality Scaled represents our quality scalable bitstream coded with QP difference of 4)

Table3: Average rate-distortion differences (bitrate) (a: not scaled vs. quality scaled bitstreams both layers with 4 units' difference in quantization parameter, b: not scaled vs. quality scaled bitstreams both layers with 8 units' difference. Positive value implies the former scheme outperforms the latter)

Sequences		a	b
QCIF	News	85.20	67.10
	Foreman	59.08	46.66
CIF	Paris	44.31	45.49
	Coastguard	31.00	25.63

4. CONCLUSIONS

In this paper we proposed a technique for quality scalability in H.264/AVC video coding. The technique is based on the sub-sequence coding technique, which is modified such that one original picture is coded multiple times according to the number of layers. The versions of the same original picture share the same output timestamp and reside in the decoded picture buffer simultaneously, and therefore special care should be taken that only the version from the highest layer is rendered. Existing H.264/AVC decoders are able to decode bitstreams coded according to the proposed method, and the method can be applied to any video coding standards with a multiple-reference-picture buffer. It was shown that the proposed quality scalability method reduces the required total bitrate compared to simulcast up to 20 %. However, the simulation results also indicated that the bitrate required for enhanced-quality reception for scalably coded bitstreams is considerably higher than that of non-scalable bitstreams. Therefore, we see a big value in the ongoing work in the Joint Video Team for improved compression efficiency in quality and spatial scalability, which is going to result into the Scalable Video Coding (SVC) standard.

REFERENCES

1. D. Tian, M. M. Hannuksela, M. Gabbouj, "Sub-Sequence Video Coding For Improved Temporal Scalability", ISCAS 2005, Kobe Japan, 2005
2. J. Reichel, H. Schwarz, M. Wien (editors), "Scalable Video Coding - Working Draft 1", document JVT-N020, 2005
3. G. Bjontegaard, "Calculation of average PSNR differences between RD-curves", document VCEG-M33, USA, 2001