Hindawi Publishing Corporation
Advances in Multimedia
Volume 2013, Article ID 281593, 17 pages
http://dx.doi.org/10.1155/2013/281593

Research Article

Efficient Region-of-Interest Scalable Video Coding with Adaptive Bit-Rate Control

Dan Grois and Ofer Hadar

Communication Systems Engineering Department, Ben-Gurion University of the Negev, P.O. Box 653, 84105 Beer-Sheva, Israel

Correspondence should be addressed to Dan Grois; grois@bgu.ac.il

Received 6 March 2013; Revised 23 July 2013; Accepted 29 August 2013

Academic Editor: Hermann Hellwagner

This work relates to the regions-of-interest (ROI) coding that is a desirable feature in future applications based on the scalable video coding, which is an extension of the H.264/MPEG-4 AVC standard. Due to the dramatic technological progress, there is a plurality of heterogeneous devices, which can be used for viewing a variety of video content. Devices such as smartphones and tablets are mostly resource-limited devices, which make it difficult to display high-quality content. Usually, the displayed video content contains one or more ROI(s), which should be adaptively selected from the preencoded scalable video bitstream. Thus, an efficient scalable ROI video coding scheme is proposed in this work, thereby enabling the extraction of the desired regions-of-interest and the adaptive setting of the desirable ROI location, size, and resolution. In addition, an adaptive bit-rate control is provided for the region-of-interest scalable video coding. The performance of the presented techniques is demonstrated and compared with the joint scalable video model reference software (JSVM 9.19), thereby showing significant bit-rate savings as a tradeoff for the relatively low PSNR degradation.

1. Introduction

Recently, significant changes have taken places in the content distribution network industry. The availability of cheaper and more powerful devices (such as smartphones and tablets, which have the ability to play, create, and transmit video content on various mobile networks) places unprecedented demands for high capacity and low-latency communications paths. The reduction of cost of digital video cameras, along with development of user-generated video sites (e.g., Vimeo, YouTube), has stimulated the new user-generated content sector. Growing premium content coupled with advanced video technologies, such as the Internet TV, will replace conventional technologies (e.g., cable or satellite TV) in the near future [1]. In this context, high-definition, highly interactive networked media applications pose challenges to network operators. The variety of end-user devices with different capabilities, ranging from smartphones with relatively small displays and restricted processing power to high-end PCs with high-definition displays, has stimulated a significant interest in effective technologies for providing video content in various spatial formats, employing limited computational complexity resources and operating under low bit-rates [2].

Much of the attention in the field of video adaptation is currently directed to the scalable video coding (SVC) extension [3] of the H.264/MPEG-4 AVC standard [4], since the bitstream scalability for video is a desirable feature for many multimedia applications. The above-mentioned extension, which was officially issued as a scalable video coding standard in 2007, is further discussed in Section 1.1.

1.1. Scalable Video Coding Extension of the H.264/MPEG-4 AVC Standard. The need for the scalability actually arises from the need for various spatial formats (depending on the particular end-user devices), bit-rates, and power [2]. To fulfill these requirements, it would be beneficial to simultaneously transmit or store video in a variety of spatial/temporal resolutions and qualities, leading to the video bitstream scalability.

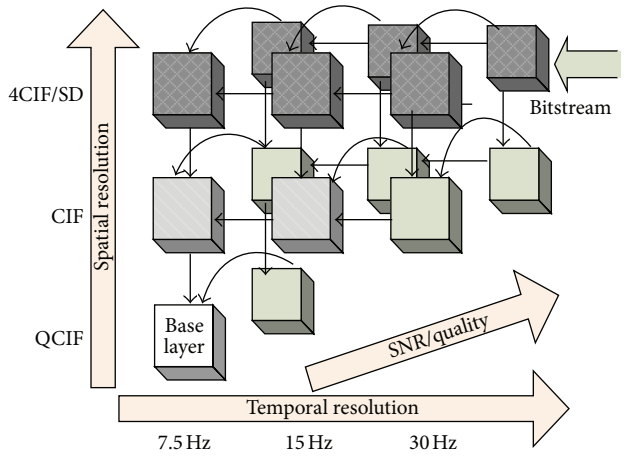


Figure 1: Schematic representation of the SVC bitstream: the spatial and/or temporal resolution is increased with the increase of a layer index (the base-layer (Layer 0), Layer 1, Layer 2, etc.), while the Base-Layer has the lowest bitstream resolution [3].

Figure 2: Defining ROIs with different spatial resolutions (e.g., CIF, SD/4CIF, 720 p resolutions) to be provided within a Scalable Video Coding stream.

Major requirements for the scalable video coding are to enable encoding of a high-quality video bitstream that contains one or more subset bitstreams, each of which can be transmitted and decoded to provide video services with lower temporal or spatial resolutions or to provide reduced reliability, while retaining reconstruction quality that is highly relative to the rate of the subset bitstreams. Therefore, the SVC standard provides functionalities, such as the spatial, temporal, and SNR (quality) scalability [3]. These functionalities lead to an enhancement of the video transmission and storage applications. The SVC standard has achieved significant improvements in coding efficiency with an increased degree of supported scalability relative to the scalable profiles of prior video coding standards, such as MPEG-2 [2].

One of the main purposes of the SVC-based systems is to enable encoding of a high-quality video bitstream that contains one or more subset bitstreams [3, 5]. Each of such subset bitstreams can be transmitted and decoded to provide video services with varying spatial resolutions (e.g., QCIF, CIF, 4CIF/SD, 720 p, 1080 p), temporal resolutions (e.g., 15 Hz, 30 Hz), fidelity (SNR/quality) resolutions, or any combination thereof, as schematically presented in Figure 1. It should be noted that the reconstruction quality is also highly relative to the bit-rate of the subset bitstreams, according to various end-user heterogeneous devices (e.g., smartphones, tablets, laptops, personal computers, etc.).

The SVC standard also supports a region-of-interest (ROI) scalability, in addition to the main scalability types, such as the temporal, spatial, and quality scalabilities, as already noted above [3, 6]. The ROI is a desirable feature in many future scalable video coding applications, for example, with which a cell phone user may require extracting only the ROI and tracking it smoothly. At the same time, other users having a larger mobile device screen can be suggested to extract other ROI in order to receive better video stream resolution. Thus, to fulfill these requirements, it would be beneficial to simultaneously transmit or store a video stream in a variety of ROIs, each of which can be selected as, for example, presented in Figure 2.

The dramatically growing number of video applications, programs, and distributed content (e.g., the Internet TV, Mobile TV, video conferences, news, user-generated videos, etc.) requires significant progress in the adaptation of scalable video coding to support these new demanding services. Due to a variety of heterogeneous end-user devices having different spatial and temporal resolution, the further significant optimization of scalable video encoders is also required. One use case for which there is likely to be a significant demand is the joint content-adaptive and user-adaptive scalable video coding which is based on video content analysis in the compressed domain, considering a variety of different user devices (a variety of decoders having different computational capabilities) to view any desired video content being adapted for each user device.

The challenges considered above are relatively new problems that have not been solved in the field. Nevertheless, some approaches described in the literature are relevant and can be used in an attempt to find a solution. There are two main methods to solve the ROI detection and tracking problem: (a) in the pixel domain approach and (b) in the compressed domain approach. First, the pixel domain approach [7–11] is, generally, more accurate than the compressed domain approach but has relatively high computational complexity and requires further additional computational resources for decoding compressed video streams. In [7], the authors demonstrate face detection using a skin-color model. The model in [7] exploits the fact that the people skin-colors have larger difference in the brightness/intensity and not in color. In [8], a fast method for detection faces based on simple features is proposed. The authors introduced a novel image representation, called the "integral image" which allows relatively quick computation of predefined features. In addition, the authors use the AdaBoost algorithm [9] that selects a relatively small number of critical visual features from a larger set and then yields extremely efficient classifiers. A visual attention system, inspired by the behavior and the

neuronal architecture of the early primate visual system, is presented in [10]. Multiscale image features are combined into a single topographical saliency map. The features are based on color, intensity, and orientation. One of the most popular methods for tracking a region in pixel domain is based on color histogram [11]. The algorithm searches, in the current frame, a region with a color histogram similar to the histogram of the ROI from the previous frame. Second, the compressed domain approach exploits DCT coefficients, motion vectors [12–14], or mode decisions instead of original pixel data as resources, in order to reduce computational complexity of ROI detection and tracking. The encoded data is not credible enough or insufficient to detect and track moving objects. In general, the compressed domain algorithms include two methods, such as the clustering method and the filtering method. The clustering-based methods attempt to perform grouping and merging all blocks into several regions according to their spatial or temporal similarity. Then, these regions are merged with each other or classified as background or foreground. The most advanced clustering-based method, which handles the H.264/MPEG-4 AVC standard, is the region growing approach, in which several seed fragments grow spatially and temporally by merging similar neighboring fragments [15]. On the other hand, the filtering-based methods extract foreground regions by filtering blocks, which are expected to belong to background or by classifying all blocks into foreground and background. Then, the foreground region is split into several object parts through a clustering procedure. Currently, there are three algorithms, which handle H.264/AVC compressed videos: the Markovian Random Field-based (MRF-based) algorithm [16], the dissimilarity minimization algorithm [17], and the Probabilistic Data Association Filtering (PDAF) algorithm [18].

Following the brief overview above, the region-of-interest coding in SVC is presented and explained in detail in the next Section 1.2.

1.2. Region-of-Interest Coding in SVC. Region-of-interest (ROI) coding is a very desirable feature in future SVC-based applications, especially in applications employed over limited-bandwidth networks. However, the H.264/MPEG-4 AVC standard does not explicitly teach how to perform the ROI coding.

The authors of this paper evaluate the ROI coding by using various techniques supported in the H.264/MPEG-4 AVC standard [19] and its scalable video coding extension [20–23]. Some of these techniques include a quantization step size control at slice and macroblock levels and are related to the concept of slice grouping, also known as the flexible macroblock ordering (FMO). The tradeoff and effectiveness of the six fixed FMO types are analyzed, for example, in [24], which proposes an adaptive FMO-type selection strategy for different video scenes and applications. Also, [25] handles the ROI-based fine granular scalability (FGS) coding, in which a user at the decoder side requires to receive better decoded quality ROIs, while the preencoded scalable bitstream is truncated. In addition, [26] presents the ROI-based spatial scalability scheme, concerning two main issues: overlapping regions between ROIs and providing different ROIs resolutions. However, [26] follows the concept of slice grouping of H.264/MPEG-4 AVC, considering the following two solutions to improve the coding efficiency: (a) supporting different spatial resolutions for various ROIs by introducing a concept of virtual layers and (b) avoiding duplicate coding of overlapping regions in multiple ROIs by encoding the overlapping regions in such a manner that the corresponding encoded regions can be independently decoded. Further, [27] presents ROI-based coarse granular scalability (CGS), using a perceptual ROI technique to generate a number of quality profiles, and in turn, to realize the CGS. According to [27], the proposed ROI-based compression achieves better perceptual quality and improves coding efficiency. Moreover, [12] relates to extracting the ROIs (i.e., of an original bitstream by introducing a description-driven content adaptation framework). According to [12], two methods for the ROI extraction are implemented: (a) a removal of non-ROI portions of a bitstream and (b) a replacement of the coded background with corresponding placeholder slices. As a result, bitstreams that are adapted by this ROI extraction process have a significantly lower bit-rate than their original versions. While this has, in general, a profound impact on the decoded video sequence quality, this impact is marginal in case of a fixed camera and static background. This observation may lead to new opportunities in the domain of video surveillance or video conferencing. According to [12], in addition to the bandwidth decrease, the adaptation process has a positive effect on the decoder due to the relatively simple processing of placeholder slices, thereby increasing the decoding speed.

In this work, a novel dynamically adjustable and scalable ROI video coding scheme is suggested, enabling to adaptively and efficiently define the desirable ROI location, size, resolution, and bit-rate. This further enables to provide in Section 4 an efficient ROI scalable video coding scheme with an adaptive bit-rate control, enabling to adaptively change the ROI visual quality and amount of bits allocated for each ROI, while considering the above constraints and considering various predefined settings (e.g., user's display spatial resolution, etc.).

This paper is organized as follows: in Section 2, an approach for efficiently extracting and obtaining the desired ROI scalability is presented. In Section 2.1, a method of cropping ROI from the original image and performing the inter-layer prediction is evaluated while also evaluating the ROI scalability by using the flexible macroblock ordering method in Section 2.2. Further, a proposed adaptive bit-rate control for the ROI scalable video coding is presented in Section 3. The experimental results are discussed in Section 4, and conclusions are provided in Section 5.

2. Extracting the Desired ROI Scalability according to User-Predefined Settings

In this section, an approach for efficiently extracting and obtaining the desired region-of-interest scalability of the SVC-based video stream is presented, thereby enabling to provide an efficient adaptive bit-rate control (presented in

Section 3). In such a way, the ROI scalable video coding can be used for various purposes, such as for streaming to heterogeneous end-user devices over either wireline or wireless networks and for broadcasting high-definition (HD) video content.

It should be noted that automatically determining and tracking the ROI region can be performed according to many conventional efficient techniques and algorithms [15–18]. Therefore, this issue is out of the scope of this paper. Hereinafter, it is supposed that the ROI region is appropriately determined prior to the extraction, based on various state-of-the-art methods, such as visual attention models [10, 28].

For extracting the desired ROI from each frame of the video sequence, the background region (e.g., which is defined as a particular FMO slice) is removed, and then the header information within the H.264/MPEG-4 AVC parameter sets is updated accordingly to exclude the removed data. It should be noted that the H.264/MPEG-4 AVC parameter sets contain higher-level parameters. Generally, the H.264/MPEG-4 AVC specification includes two types of parameter sets: a sequence parameter set (SPS), which applies to a series of consecutive coded video pictures (i.e., coded video sequence) and stay constant for the duration of the sequence and a picture parameter set (PPS), which applies to the decoding of one or more individual pictures within the coded video sequence and stay constant for the duration of each picture. This means that an active sequence parameter set remains unchanged throughout a coded video sequence, and an active picture parameter set remains unchanged within a coded picture. The sequence and picture parameter set structures contain information such as picture size, optional coding modes employed, and macroblock to slice group map. Thus, the PPS is used for decoding its video coding layer (VCL) data, while each slice references its corresponding picture parameter set.

The information regarding a particular ROI can be determined within the parameter set by first determining the physical location of the ROI within a picture (e.g., by determining upper-left and bottom-right macroblocks of a rectangular ROI).

In Section 2.1, a method of cropping ROI from the original image and performing the inter-layer prediction is evaluated.

2.1. Cropping the ROI from the Original Image and Performing SVC Inter-layer Prediction. According to the first evaluated method for the ROI scalable video coding, and in order to enable the obtaining of a high-quality ROI on devices with relatively small displays, the ROI is first cropped from the original image and then is used as a base-layer/Layer 0 during the SVC encoding process, which is schematically illustrated in Figure 3.

In addition, an inter-layer prediction can be performed in similar sections of each frame/picture, that is, defined as cropping areas. As a result, by using the inter-layer prediction for encoding of the three SVC layers (i.e., providing a different spatial resolution for each SVC layer, such as the QCIF, CIF, and 4CIF/SD resolution, resp.), a significantly low bit-rate overhead can be achieved. It should be noted that prior to cropping the image, the location of a cropping area in the next layer of the image is determined (e.g., first in Layer 1, and then in Layer 2, as shown in Figure 3). For this, an extended spatial scalability (ESS) method [29] can be employed.

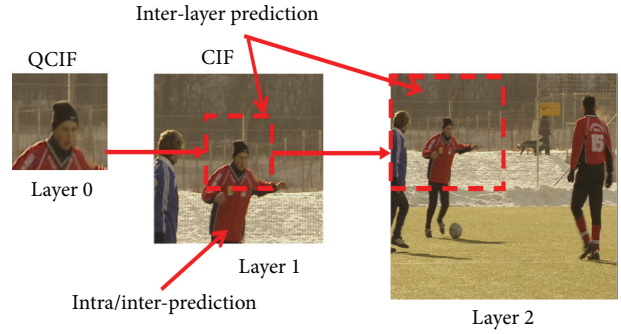


Figure 3: Example of the ROI dynamic adjustment and scalability (e.g., for mobile devices with different spatial resolution) by using a cropping method.

Figures 4(a) and 4(b) present additional examples of the ROI dynamic adjustment and scalability for end-user devices with various spatial resolutions. According to Figure 4(a), for example, a "box" in the man's hand has significant importance to users, and therefore, such a "box" is defined as the ROI. On the contrary, according to Figure 4(b), it is more important to see the man's face and man's lips clearly, which can be especially useful for deaf people.

Tables 1–3 present typical rate-distortion (R-D) experimental results for different cropping spatial resolutions of the *"SOCCER"* video sequence while using the inter-layer prediction. It should be noted that in this particular example, the coding parameters are as follows: the frame rate is 30 fps; the overall length is 300 frames; the GOP size is 16; the quantization parameters (QPs) are set to 22, 26, 30, and 34. The cropping was performed by selecting the ROI at the upper-left corner of each frame.

It is clearly seen from these tables that there is significantly low bit-rate overhead when compared to the single layer coding, which is especially important for transmitting over limited-bandwidth networks (such as wireless networks).

Table 1 presents the R-D experimental results for the two-layer coding versus single layer coding, by providing different spatial resolutions for each of the above-mentioned two layers. Specifically, the QCIF-resolution video sequence is defined as the SVC base-layer/Layer 0, and the CIF-resolution video sequence is defined as the SVC enhancement Layer 1. In addition, as already noted above, the QCIF-resolution video sequence is cropped from the CIF-resolution video sequence, and the single layer coding refers to the CIF-resolution video coding. Also, the SVC inter-layer prediction is performed, as already discussed with regard to Figure 4.

As illustrated in Table 1, the bit-rate overhead for encoding two SVC layers by using the inter-layer prediction is very low and is only up to 7%. Also, Table 2 presents the R-D experimental results for the two-layer coding versus single layer coding, while this time defining the CIF-resolution video sequence as the SVC base-layer/Layer 0, and the 4CIF/SD-resolution video sequence as the SVC enhancement Layer 1.

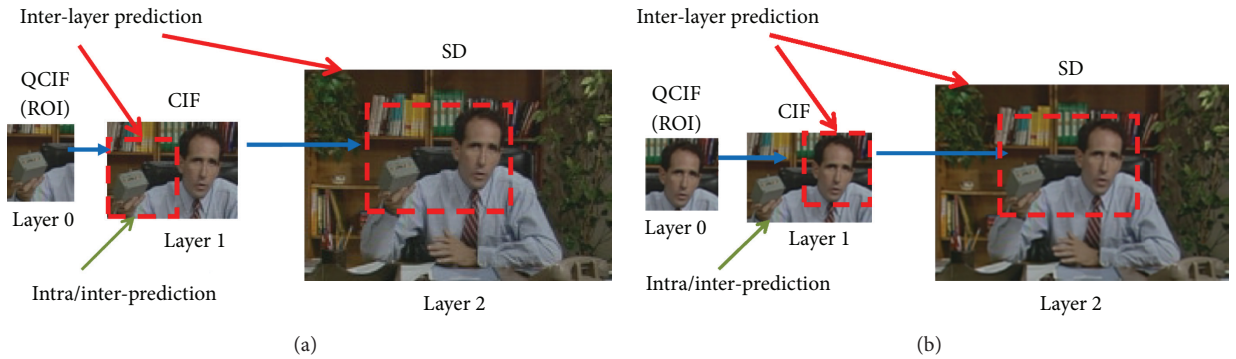


Figure 4: Additional examples of the ROI dynamic adjustment and scalability by using a cropping method.

Table 1: Two-layer (QCIF-CIF) spatial SVC coding versus single layer (CIF) coding ("*SOCCER*" video sequence, 30 fps, 300 frames, GOP size is 16).

Quantization parameters	Single layer (CIF)		QCIF-CIF		Bit-rate overhead (%)
	PSNR [dB]	Bit-rate [kb/s]	PSNR [dB]	Bit-rate [kb/s]	
22	40.9	1636.8	40.9	1713.5	**4.5**
26	38.6	917.2	38.6	968.8	**5.3**
30	36.5	544.0	36.5	578.1	**5.9**
34	34.4	332.9	34.4	357.5	**6.9**

The single layer coding in Table 2 refers to the SD-resolution video coding, and all other test conditions are exactly the same.

As it is observed from Table 2, the bit-rate overhead is only about 6%, which is very similar to the bit-rate overhead in Table 1. Further, Table 3 presents the R-D experimental results for the three-layer coding versus single layer coding, while defining the QCIF-resolution video sequence as the SVC base-layer/layer 0, the CIF-resolution video sequence as the SVC enhancement Layer 1, and the 4CIF/SD-resolution video sequence as the SVC enhancement Layer 2. Also, the QCIF and CIF-resolution video sequences are cropped accordingly, as in the previous tables. As it is seen from Table 3, the bit-rate overhead varies between 4.7% and 7.9%, when compared to the single layer coding, which is in line with the outcomes presented in Tables 1 and 2.

To summarize the evaluation results presented in this section, it is clearly seen that when using the SVC inter-layer prediction along with the cropping technique, the bit-rate overhead is very small and is less than 10%, when compared to the single layer coding.

In Section 2.2, a flexible macroblock ordering (FMO) method is presented for improving the region-of-interest SVC scalability.

2.2. ROI Scalability by Using the Flexible Macroblock Ordering. The second evaluated method refers to the ROI scalable video coding by using the flexible macroblock ordering (FMO) technique [4]. One of the basic elements of the H.264 video sequence is a slice, which contains a group of macroblocks. Each picture can be subdivided into one or more slices and each slice can be provided with increased importance as the basic spatial segment, which can be encoded independently from its neighbors [30–33] (the slice coding is one of the techniques used in H.264 for transmission). Usually, slices are provided in a raster scan order with continuously ascending addresses; on the other hand, the FMO is an advanced tool of H.264 that defines the information of slice groups and enables to assign different macroblocks to slice groups, according to several predefined patterns (types), as schematically presented in Figure 5.

Each slice of each picture/frame is independently intra-predicted, and the macroblock order within a slice must be in the ascending order. In H.264/MPEG-4 AVC standard, the FMO technique relates to seven slice group map types (Type 0 to Type 6), six of them are predefined fixed macroblock mapping types (interleaved, dispersed, foreground, box-out, raster scan and wipe-out), which can be specified through the Picture Parameter Set (PPS), and the last one is a custom type, which allows the full flexibility of assigning macroblocks to any slice group.

The ROI can be defined as a separate slice in the FMO Type 2 [24], which enables defining slices of rectangular regions, and then the whole sequence can be encoded accordingly, while making it possible to define more than one ROI region (these definitions should be performed in the SVC configuration files, according to the JSVM reference software manual (JSVM 9.19) [22]).

As noted above, for the ROI scalable video coding, the FMO Type 2 (Figure 5) can be used, where each ROI is represented by a separate rectangular region and is encoded as a separate slice.

Table 4 presents experimental results for the four-layer spatial scalable video coding (with the FMO technique enabled) versus six-layer spatial scalable video coding of the *"SOCCER"* video sequence. It should be noted that the coding

Table 2: Three-layer (CIF-SD) spatial SVC coding versus single layer (4CIF/SD) coding ("*SOCCER*" video sequence, 30 fps, 300 frames, GOP size is 16).

Quantization parameters	Single layer (4CIF/SD)		CIF-SD		Bit-rate overhead (%)
	PSNR [dB]	Bit-rate [kb/s]	PSNR [dB]	Bit-rate [kb/s]	
22	41.0	5663.3	40.9	5870.7	**3.5**
26	38.8	3054.9	38.7	3190.6	**4.3**
30	36.8	1770.2	36.7	1860.2	**4.8**
34	34.8	1071.3	34.7	1137.0	**5.8**

Table 3: Three-layer (QCIF-CIF-SD) spatial scalability coding versus single layer (4CIF/SD) coding ("*SOCCER*" video sequence, 30 fps, 300 frames, GOP size is 16).

Quantization parameters	Single layer (4CIF/SD)		QCIF-CIF-SD		Bit-rate overhead (%)
	PSNR [dB]	Bit-rate [kb/s]	PSNR [dB]	Bit-rate [kb/s]	
22	41.0	5663.3	41.0	5940.6	**4.7**
26	38.8	3054.9	38.8	3248.1	**6.0**
30	36.8	1770.2	36.8	1894.9	**6.6**
34	34.8	1071.3	34.8	1163.6	**7.9**

parameters are exactly the same as mentioned in Section 2.1. Also, the above-mentioned four SVC layers are represented by one CIF-resolution layer and three 4CIF/SD-resolution layers having the CIF-resolution ROI in an upper-left corner of the image. By such a way, the ROI quality in each layer can be controlled according to the user's needs (for simplicity, in these experiments, the ROI quality in each layer is the same as the quality of the background, that is, the ROI QP is equal to the QP of the background (non-ROI) region). On the other hand, the above-mentioned six layers are represented by three CIF-resolution layers and three 4CIF/SD-resolution layers.

According to Table 4, there are significant bit-rate savings of up to 10% by using the FMO technique. Further, Table 5 presents R-D experimental results for the high-definition (HD) video sequence *"STOCKHOLM"* (as presented in Figure 6) by using four-layer scalable video coding versus eight-layer scalable video coding. The above-mentioned four layers are represented by one 640 × 360 resolution layer and three 720 p resolution layers, which in turn have a ROI predefined in the upper-left corner of each frame by using the FMO technique. The ROIs resolutions are CIF and SD, respectively (for simplicity, the ROI quality in each layer is the same as the quality of the background; i.e., the ROI QP is equal to the QP of the background (non-ROI) region). On the other hand, the above-mentioned eight layers contain two CIF-resolution layers, three SD-resolution layers, and three 720 p resolution layers, which have different quantization parameters varying from 32 to 36 with an interval of 2.

As it is seen from Table 5, in this case the bit-rate savings are up to 33.5% by using the FMO technique. Further, Table 6 presents the R-D experimental results for the same *"STOCKHOLM"* video sequence by using four-layer scalable video coding versus six-layer scalable video coding. The above-mentioned four layers are represented by one 640×360 resolution layer and three 720 p resolution layers (similarly to the settings of Table 4), which in turn have a ROI predefined in the upper-left corner of each frame by using the FMO technique. The ROIs resolutions are CIF and SD, respectively (for simplicity, the ROI quality in each layer is the same as the quality of the background; i.e., the ROI QP is equal to the QP of the background (non-ROI) region). On the other hand, the above-mentioned six layers contain three SD-resolution layers and three 720 p resolution layers, which have different quantization parameters varying from 32 to 36 with an interval of 2.

According to Table 6, when using the FMO technique in this case, the bit-rate savings reach 39%. Also, it is noted that when perfuming the SVC coding of higher spatial resolutions (such as the 720 p resolution), the achieved bit-rate savings are much more significant.

In Section 3, a proposed adaptive bit-rate control for the ROI scalable video coding is presented.

3. Adaptive Bit-Rate Control ROI Scalable Video Coding

The bit-rate control is crucial in providing desired compression bit-rates for the H.264/MPEG-4 AVC video applications, and especially for the SVC-based applications [3, 34].

The bit-rate control has been intensively studied in the recent video coding standards, such as MPEG-2, MPEG-4, and H.264/MPEG-4 AVC [2, 35]. According to the existing single layer rate control schemes, the encoder employs the rate control as a way to control varying bit-rate characteristics of the coded bitstream. Generally, there are two objectives of the bit-rate control for the single layer video coding: one is to meet the bandwidth that is provided by the network, and another is to produce high-quality decoded pictures [36]. Thus, the inputs of the bit-rate control scheme are (a) the given bandwidth; (b) the statistics of the video sequence (e.g., including the mean squared error (MSE) or mean absolute difference (MAD) of colocated pixels/macroblocks in consecutive frames prior to performing the quantization);

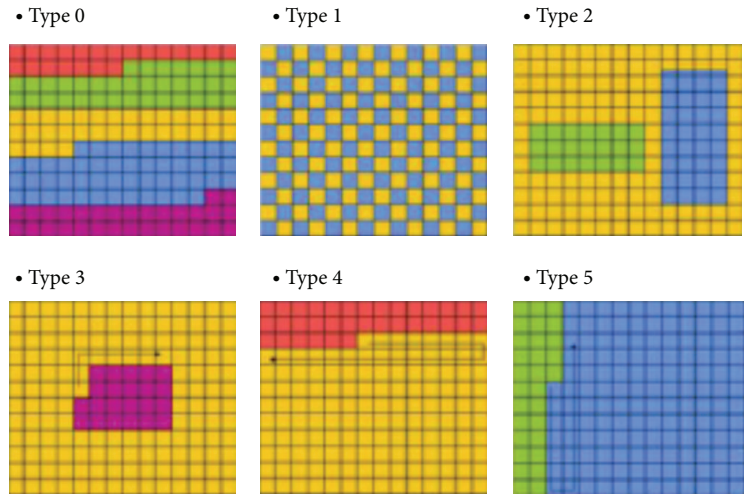


Figure 5: Six fixed types of the FMO (interleaved, dispersed, foreground, box out, raster scan, and wipe out), while each color represents a slice group [24].

Table 4: FMO: four-layer spatial scalability coding versus six-layer coding ("*SOCCER*" video sequence, 30 fps, 300 frames, GOP size 16).

Quantization parameters	Four layers (CIF and three SD layers) by using the FMO		Six layers (three CIF layers and three SD layers)		Bit-rate savings (%)
	PSNR [dB]	Bit-rate [kb/s]	PSNR [dB]	Bit-rate [kb/s]	
32	36.0	2140.1	36.0	2290.1	**6.6**
34	35.1	1549.4	35.1	1680.1	**7.8**
36	34.0	1140.1	34.0	1279.4	**10.9**

Figure 6: The 720 p resolution frame of the "STOCKHOLM" video sequence, which presents two ROIs having CIF and SD resolutions.

and (c) a header of each predefined unit (e.g., a basic unit which contains one or more macroblocks (MBs), or which is a frame or a slice). In turn, the outputs are a quantization parameter (QP) for the quantization process and another QP for the rate-distortion optimization (RDO) process of each basic unit, while these two quantization parameters, in the single layer video coding, are usually equal in order to maximize the coding efficiency [37–39].

In the current JSVM reference software [22], there is no rate control mechanism, besides the base-layer rate control, which does not consider the enhancement layers. The target bit-rate for each SVC layer is achieved by coding each layer with a fixed QP, which is determined by a logarithmic search [22, 40]. Of course, this is very inefficient and very time consuming. For solving this problem, only a few works have been published during the last years, thereby trying to provide an efficient SVC-based rate control mechanism. However, none of them handles scalable bit-rate control for the region-of-interest (ROI) coding. For example, in [41], the rate distortion optimization (RDO) involved in the step of encoding temporal subband pictures is only implemented on low-pass subband pictures, and rate control is independently applied to each spatial layer. Furthermore, for the temporal subband pictures obtained from the motion compensation temporal filtering (MCTF), the target bit allocation and quantization parameter selection inside a GOP makes a full use of the hierarchical relations inheritance from the MCTF. In addition, [40] proposes a switched model to predict the MAD of the residual texture from the available MAD information of the previous frame in the same layer and the same frame in its base-layer. Further, [42] describes a constant quality variable bit-rate (VBR) control algorithm for multiple layer coding. According to [42], a target quality is achieved by specifying memory capabilities and the bit-rate limitations of the storage device. In the more recent work [43], the joint optimization of layers in the layered video coding is investigated. However, as already mentioned above, there is currently no efficient bit-rate control scheme for the ROI scalable video coding.

Below, a method for the efficient ROI scalable video coding is presented. This method employs the FMO technique, which is evaluated in detail in Section 2.2. According to the proposed method, a bit-rate close to the target bit-rate is

Table 5: FMO: four-layer coding versus eight-layer coding ("*STOCKHOLM*", 30 fps, 96 frames, GOP size is 8).

Quantization parameters	Four layers (one 640 × 360 layer, and three 720p layers) by using the FMO		Eight layers (two CIF layers, three SD layers, and three 720p layers)		Bit-rate savings (%)
	PSNR [dB]	Bit-rate [kb/s]	PSNR [dB]	Bit-rate [kb/s]	
32	34.5	2566.2	34.5	3237.0	**20.7**
34	33.9	1730.2	33.9	2359.1	**26.7**
36	33.3	1170.0	33.3	1759.0	**33.5**

Table 6: FMO: four-layer coding versus six-layer coding ("*STOCKHOLM*", 30 fps, 96 frames, GOP size 8).

Quantization parameters	Four layers (640 × 360, and three HD layers) by using the FMO		Six layers (three SD layers, and three HD layers)		Bit-rate savings (%)
	PSNR [dB]	Bit-rate [kb/s]	PSNR [dB]	Bit-rate [kb/s]	
32	34.5	2566.2	34.5	3330.7	**19.3**
34	33.9	1730.2	33.9	2412.0	**29.7**
36	33.3	1170.0	33.3	1784.0	**39.9**

achieved, as further presented in Section 4. Also, the desirable ROI quality is achieved in terms of the peak signal-to-noise ratio (PSNR), while adaptively varying the background region quality according to the overall bit-rate.

In order to provide the different visual presentation quality to at least one ROI and to the background region (or other less important regions), each frame is divided into at least two slices: one slice is used for defining the ROI and one additional slice is used for defining the background region, for which fewer bits should be allocated. If more than one ROI is used, then the frame is divided into a larger number of slices, such that for each ROI a separate slice is used.

The method for performing the adaptive ROI SVC bit-rate control for each SVC layer is as follows:

(a) Compute the number of target bits for the current GOP and after that for each frame (of each SVC layer) within the above GOP by using a hypothetical reference decoder (HRD) [44]. The number of target bits $\mathrm{TB}_t^{\mathrm{SVC\ Layer}}(i)$ for frame i should be a weighted combination of the remaining bits for encoding the remaining frames (within the current GOP), and the target bits which were allocated for frame i; this is formulated as follows:

$$\mathrm{TB}_t^{\mathrm{SVC\ Layer}}(i) = \beta \cdot \widetilde{\mathrm{TB}}_r^{\mathrm{SVC\ Layer}}(i) + (1-\beta) \cdot \widehat{\mathrm{TB}}_t^{\mathrm{SVC\ Layer}}(i), \tag{1}$$

where β is a weight coefficient; $\widetilde{\mathrm{TB}}_r^{\mathrm{SVC\ Layer}}(i)$ is a number of remaining bits for encoding the current frame i in each SVC layer; and $\widehat{\mathrm{TB}}_t^{\mathrm{SVC\ Layer}}(i)$ is a number of target bits allocated for frame i in the current GOP of each SVC layer. $\widetilde{\mathrm{TB}}_r^{\mathrm{SVC\ Layer}}(i)$ and $\widehat{\mathrm{TB}}_t^{\mathrm{SVC\ Layer}}(i)$ are represented by the following expressions:

$$\widetilde{\mathrm{TB}}_r^{\mathrm{SVC\ Layer}}(i) = \frac{\mathrm{TB}_r^{\mathrm{SVC\ Layer}}}{F_r},$$
$$\widehat{\mathrm{TB}}_t^{\mathrm{SVC\ Layer}}(i) = \frac{\left(\mathrm{TB}_r^{\mathrm{SVC\ Layer}}/F_r\right) \cdot \sigma^2(i)}{\left(1/(F-F_r)\right) \sum_{j=1}^{F-F_r} \sigma^2(j)}, \tag{2}$$

where F_r is a number of remaining frames; F is the total number of frames in the current GOP; $\mathrm{TB}_r^{\mathrm{SVC\ Layer}}$ is the number of remaining bits for encoding the remaining frames; $\sigma(i)$ is the predicted MAD of the current frame i; and $\sigma(j)$ is the actual MAD of the previous frame j. It should be noted that upon determining a number of target bits $\mathrm{TB}_t^{\mathrm{SVC\ Layer}}(i)$ for frame (i), the bits for each region within the frame (i.e., one or more ROIs and background) are allocated according to the desired ROI QP, which is preset by a user. Of course, if better visual quality is required at the particular ROI, then more bits are allocated to that region, and vice-versa.

(b) Allocate the remaining bits to all noncoded macroblocks (MBs) for each predefined slice in the current frame of the particular SVC layer.

(c) Estimate the MAD for the current macroblock in the current slice by a linear prediction model [35, 45] using the actual MAD of the macroblocks in the colocated position of the previous slices (in the previous frames) within the same SVC layer and the MAD of neighbor macroblocks in the current slice. Suppose that the predicted MAD of current basic unit in the current frame and the actual MAD of basic unit in the colocated position of previous frame

Table 7: Test conditions for performing the tests/evaluation.

Settings	All tested video sequences
QP settings	
Base-layer (Layer 0), which only contains the ROI	40
Enhancement layer (Layer 1), which contains both the ROI and the background region	Is adaptively varied from 20 to 40, depending on the ROI QP: for ROI QP = 20, the background region QP is varied from 20 to 40; for ROI QP = 37, and the background region QP is varied from 37 to 40. The ROI was selected at the middle of the frame for all sequences.
Spatial resolution	
Base-layer (Layer 0)	CIF (352 × 288)
Enhancement layer (Layer 1)	4CIF/SD (704 × 576)
Frame rate	
Base-layer (Layer 0)	30 fps
Enhancement layer (Layer 1)	30 fps
Coding options used	Motion vector (MV) search range is 16; number of coded frames is 100 to 300; fast search is ON; number of reference frames is 1; GOP size is 16; GOP type is IPPP.
CODEC	JSVM 9.19

are denoted by $\mathrm{MAD}_{\mathrm{CUR}}$ and $\mathrm{MAD}_{\mathrm{PRED}}$, respectively. The linear prediction model is then given by

$$\mathrm{MAD}_{\mathrm{CUR}} = B_1 * \mathrm{MAD}_{\mathrm{PRED}} + B_2, \tag{3}$$

where B_1 and B_2 are corresponding coefficients, which are calculated regressively. The initial values B_1 and B_2 can be set, for example, to 0.85 and 0.15, respectively, based on empirical measurements. These coefficients are updated after coding each macroblock.

(d) Estimate a set of groups of coding modes (e.g., modes such as Inter-Search16 × 8, Inter-Search8 × 16, Inter-Search8 × 8, Inter-Search8 × 4, Inter-Search4 × 8, Inter-Search4 × 4 modes, and the like) of the current macroblock in the current frame within the above SVC layer by using the actual group of coding modes for the macroblocks in the colocated positions of the previous frame(s) and the actual group of coding modes of neighbor macroblocks in the current frame.

(e) Compute the corresponding QPs by using, for example, a quadratic model of [46, 47].

(f) Perform the rate-distortion optimization for each MB by using the QPs derived from the above Step (e) [45, 48].

(g) Adaptively adjust the QPs (increase/decrease the QPs by a predefined quantization step size), according to the current overall bit-rate.

In Figure 7, a block-diagram for performing the proposed adaptive bit-rate control for the scalable video coding is presented. For simplicity, only two layers are shown—base-layer (Layer 0) and enhancement layer (Layer 1). The block-diagram of Figure 7 contains the SVC adaptive bit-rate controller, which continuously receives data regarding the current buffer occupancy, actual bit-rate, and quantization parameters.

It should be noted that Step (f) above can be performed by using a method [45, 48] for determining an optimal coding mode for encoding each macroblock. According to [45, 48], the RDO for each macroblock is performed for selecting an optimal coding mode by minimizing the Lagrangian function as follows:

$$\begin{aligned} J\left(\mathrm{orig}, \mathrm{rec}, \mathrm{MODE} \mid \lambda_{\mathrm{MODE}}\right) &= D\left(\mathrm{orig}, \mathrm{rec}, \mathrm{MODE} \mid \mathrm{QP}\right) \\ &\quad + \lambda_{\mathrm{MODE}} \cdot R\left(\mathrm{orig}, \mathrm{rec}, \mathrm{MODE} \mid \mathrm{QP}\right), \end{aligned} \tag{4}$$

where the distortion $D(\mathrm{orig}, \mathrm{rec}, \mathrm{MODE} \mid \mathrm{QP})$ can be the sum of squared differences (SSD) or the sum of absolute differences (SAD) between the original block (orig) and the reconstructed block (rec); QP is the macroblock quantization parameter; MODE is a mode selected from the set of available prediction modes; $R(\mathrm{orig}, \mathrm{rec}, \mathrm{MODE} \mid \mathrm{QP})$ is the number of bits associated with selecting MODE; and λ_{MODE} is a Lagrangian multiplier for the mode decision [49].

According to a buffer occupancy constraint, due to the finite reference SVC buffer size, the buffer at each SVC layer should not be full or empty (overloaded or underloaded, resp.). The formulation of the optimal buffer control (for controlling the buffer occupancy for each SVC layer) can be given by

$$\min\left\{\sum_{i=1}^{N} e\left(i\right)\right\}, \quad \text{subject to } B_{\max}^{\mathrm{Layer}} \geq B^{\mathrm{Layer}}\left(i\right) \geq 0, \quad \text{for } i = 1, 2, \ldots, N, \tag{5}$$

where $e(i)$ is a distortion for basic unit i (the basic unit can contain one or more macroblocks, or to be a frame or a slice);

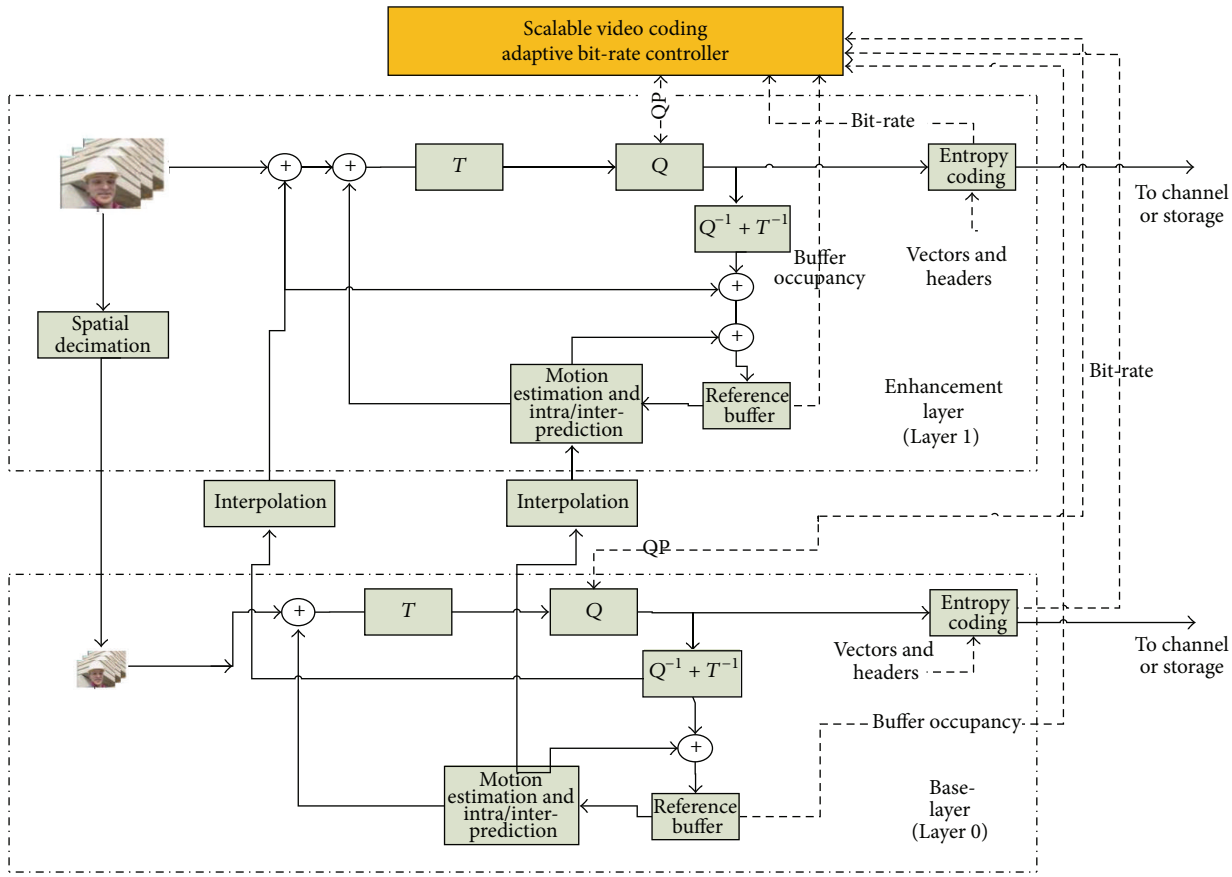


FIGURE 7: The block-diagram for performing the presented adaptive spatial bit-rate control for the scalable video coding (for simplicity, only two layers—Layer 0 and Layer 1—are presented).

$B^{\text{Layer}}(i)$ is a buffer size for basic unit i; and $B^{\text{Layer}}_{\max}$ is the maximal buffer size. The state of the buffer occupancy can be defined as

$$B^{\text{Layer}}(i+1) = B^{\text{Layer}}(i) + r^{\text{Layer}}(i) - r^{\text{Layer}}_{\text{out}}, \tag{6}$$

where $r^{\text{Layer}}(i)$ is the buffer input bit-rate with regard to each SVC layer, and $r^{\text{Layer}}_{\text{out}}$ is the output bit-rate of buffer contents. The optimal buffer control approach is related to the following optimal bit allocation formulation:

$$\min\left\{\sum_{i=1}^{N} e(i)\right\}, \quad \text{subject to} \sum_{i=1}^{n} r^{\text{Layer}}(i) \le R^{\text{Layer}}, \quad \text{for } i = 1, 2, \ldots, N, \tag{7}$$

where R^{Layer} is a target bit-rate for each SVC layer. The optimal buffer control approach is further schematically presented in Figure 8.

In order to overcome the buffer control drawbacks and overcoming buffer size limitations, preventing underflow/overflow of the buffer, and significantly decreasing the buffer delay, the computational complexity (such as a number of CPU clocks) and bits of each basic unit within a video sequence can be dynamically allocated, according to its predicted MAD. In turn, the optimal buffer control problem (5) can be solved by implementing the C-R-D analysis of [49–52] for each SVC layer.

For simplicity, in this paper, the experimental results for the bit-rate control of only two layers are shown: base-layer (Layer 0) and enhancement layer (Layer 1), while the ROI region is provided in both Layer 0 and Layer 1, and the background region is provided only in Layer 1, as illustrated in Figure 9. According to the presented adaptive bit-rate control method, different initial quantization parameters (QPs) are predefined for each layer: for example, for the whole Layer 0 an initial quantization parameter can be set to be equal to 40; on the other hand, for the ROI region provided in Layer 1, an initial quantization parameter can be set to be equal to 20, and then the QP of the remaining background region in Layer 1 is determined adaptively by the proposed bit-rate control. It should be noted that the detailed test conditions are further provided in Table 7.

By such a way, the desired quality of the region-of-interest can be obtained, according to the overall network bandwidth (either constant or variable bandwidth).

As a result, by encoding the video sequence with different QPs, the optimal presentation quality of the predefined ROI

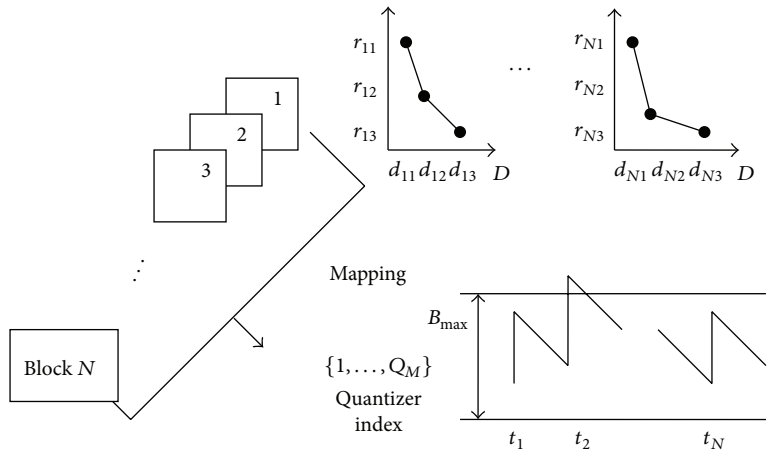


Figure 8: (a) Each block $(1 \cdots N)$ in the sequence has different R-D characteristics (for a given set of quantizers $(1 \cdots Q_M)$ for blocks in the sequence, R-D (rate-distortion) points (r_{N1}, r_{N2}, r_{N3} and d_{N1}, d_{N2}, d_N, etc.) can be obtained to form composite characteristics); and (b) R at t_2 is not a feasible solution to the selected maximum buffer size $B_{\max}$.

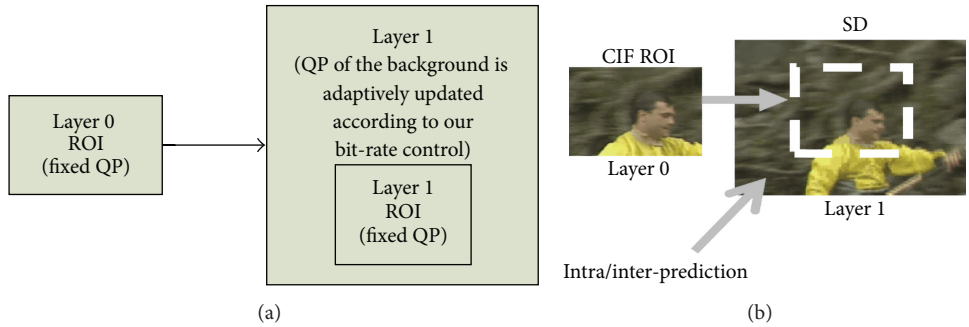


Figure 9: (a) Defining two or more layers with corresponding QPs. The QP of the background region in Layer 1 is determined adaptively by the proposed bit-rate control; (b) CIF ROI is used as base-layer (Layer 0), and 4CIF (SD) is used as an enhancement layer (Layer 1). The intra/inter-prediction is used for reducing the overall bit-rate.

region is obtained, thereby enabling to reduce the quality of the background, as presented, for example, in Figure 10 for the *"SOCCER"* video sequence (25 fps, SD resolution).

Further, the detailed experimental results are presented in Section 4.

4. Experimental Results

In this work, six conventional test sequences were evaluated, including *"PARKRUN," "HIELDS," "CREW," "CITY," "HARBOR," "ICE,"* and *"SOCCER,"* while all tests/experiments were conducted in similar test conditions, as presented in Table 7.

The used test platform is Intel Core 2 Duo CPU, 2.33 GHz, 2 GB RAM with Windows XP Professional operating system, version 2002, and Service Pack 3.

It should be noted that the performed experimental results mainly refer to the spatial SVC scalability evaluation and the quality SVC scalability evaluation [3] (i.e., the medium grained scalability (MGS) and coarse grain scalability (CGS) evaluation), which are presented in Sections 4.1 and 4.2, respectively.

4.1. Spatial Scalability Evaluation. Figure 11 illustrates a frame (number 90) of the *"PARKRUN"* video sequence, which contains the ROI region in the middle of the frame—the man with an umbrella. The quantization parameter of the background region is determined adaptively in order to achieve optimal video presentation quality (depending on the particular video content and on the desired ROI QP value, which is preset by a user). As it is seen from Figure 11(b), the QP of the background region is much higher than the QP of the ROI region (particularly, for this example, the ROI QP is equal to 20, and the QP of the non-ROI region (i.e., of the background region) is about 40).

Further, Figure 12 presents another frame (number 4) of the *"SHIELDS"* video sequence, which also contains the ROI region in the middle of the frame—a man's head and hand pointing to the shields. The quantization parameter of the background region is determined adaptively according to the adaptive bit-rate control. Similarly to Figure 11(b), in Figure 12(b) the QP of the background region is much higher than the QP of the ROI region (particularly, the ROI QP is equal to 20, and the QP of the non-ROI region is about 40).

Figure 10: The *"SOCCER"* video sequence (SD-resolution, 25 fps), which contains the ROI region in the upper-left corner.

(a)

(b)

Figure 11: The "PARKRUN" video sequence containing the ROI region in the middle of the frame, the man with an umbrella (the quantization parameter of the background region is determined adaptively); (a) the original frame; and (b) the compressed frame with the higher-quality ROI region.

Table 8 presents experimental results for the proposed bit-rate control scheme for various video sequences: "CITY," "CREW," "HARBOR," "ICE," and "SOCCER." According to the experimental results presented in Table 8, the QP of Layer 0 is equal to 40, and the QP of the ROI in Layer 1 is equal to 37, while the QP of the background of Layer 1 is determined adaptively. According to Table 8, the actual bit-rate is close to the target bit-rate. Also, it should be noted that the degradation of the PSNR values for the ROI is relatively low, especially when considering very significant bit-rate savings by using the proposed bit-rate control scheme. For example, for the "CITY" video sequence, by using the conventional base-layer rate control of JSVM 9.19 [22], the ROI PSNR value is 31.7 dB, while it is around 30 dB when using the proposed bit-rate control scheme. On the other hand, the bit-rate savings are very significant: the bit rate is reduced from 713.8 kb/s to about 445 kb/s, which is a decrease of about 38%. Very similar conclusions are made for all tested video sequences, which are presented in Table 8.

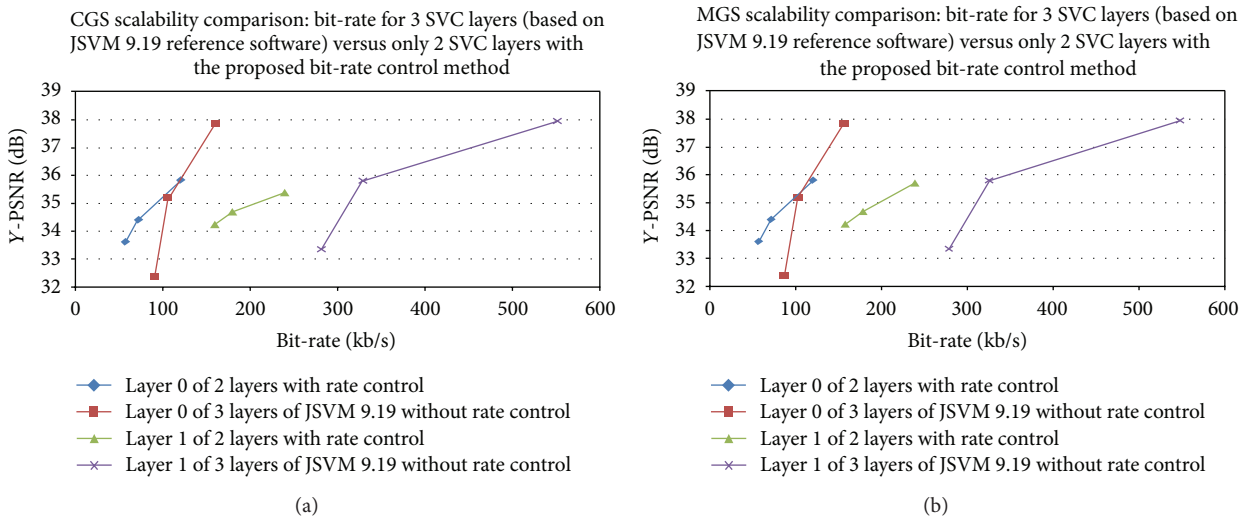
Also, Table 9 presents additional experimental results for the bit-rate control scheme for "CREW," "SHIELDS," "PARKRUN," and "SOCCER" video sequences. According to the experimental results presented in Table 9, the QP of Layer 0 is equal to 40, and the QP of the ROI in Layer 1 is equal to 20, while the QP of the background of Layer 1 is determined adaptively.

According to Table 9, the actual bit-rate is very close to the target bit-rate. Also, similarly to Table 8, the degradation of the PSNR values for the ROI is relatively low, especially when considering very significant bit-rate savings by using the proposed bit-rate control scheme. For example, for the "PARKRUN" video sequence, by using the conventional base-layer rate control of JSVM 9.19, the ROI PSNR value is 28.1 dB, while it is around 24 dB when using the proposed bit-rate control scheme. On the other hand, the bit-rate savings are very significant: the bit rate is reduced from 1435.2 kb/s to about 701.5 kb/s, which is a decrease of about 51%. Analogically to Table 8, the very similar conclusions are made for all tested video sequences.

4.2. Quality Scalability Evaluation: Medium Grained Scalability (MGS) and Coarse Grain Scalability (CGS). According to [3], the quality scalability can be considered as a special case of the spatial scalability with identical picture sizes and different QPs for the SVC base-layer and enhancement layers. This case is usually referred to as the coarse grain scalability (CGS), in which the same inter-layer prediction mechanisms as for the spatial scalable coding are employed, but without using upsampling operations and the interlayer deblocking for intra-coded reference layer macroblocks [3]. In addition, for increasing the flexibility of bit stream adaptation and error robustness, and also for improving the coding efficiency for bit streams that have to provide a variety of bit rates, a variation of the CGS approach, which is also referred to as the medium grain scalability (MGS), is provided as a part of the SVC standard [3]. The differences to the CGS concept are a modified high-level signaling, which allows the switching between different MGS layers in any access unit,

(a) (b)

Figure 12: The "*SHIELDS*" video sequence containing the ROI region, a man's head and hand pointing to the shields (the quantization parameter of the background region can be determined adaptively); (a) the original frame; and (b) the compressed frame with the higher-quality ROI region.



(a) (b)

Figure 13: Quality scalability comparison of the JSVM 9.19 reference software versus the proposed bit-rate control method (three layers of JSVM 9.19 versus two layers of the proposed rate control scheme; "*STOCKHOLM*" video sequence). The QP values for each layer are 30, 35, and 40. (a) CGS comparison; and (b) MGS comparison.

and the so-called key picture concept, which allows the adjustment of a suitable tradeoff between drift and enhancement layer coding efficiency for hierarchical prediction structures [3].

In this section, the performance of the JSVM 9.19 reference software [22] is compared with the proposed adaptive bit-rate control method. In the JSVM software, three quality layers of QCIF, CIF, and SD resolutions are used (the test conditions are specified in Table 7). On the other hand, in the proposed adaptive bit-rate control scheme, only two quality layers of CIF and SD-resolutions are used. In turn, the PSNR quality and bit-rate of the base-layer (Layers 0) and enhancement layer (Layers 1) of the JSVM 9.19 was compared with the same of the proposed bit-rate control scheme. As is seen from Figure 13, for both CGS and MGS scalability types, the proposed adaptive bit-rate control scheme provides, for each layer, much better results in terms of the bit-rate, while the PSNR degradation (when it exists) is relatively low.

Thus, according to Figures 13(a) and 13(b), by using the proposed rate control scheme, the bit-rate is reduced almost twice, while the PSNR degradation is at average only about 2 dB. Also, as it is observed from Figure 13, the CGS and MGS performance is relatively similar for both the proposed bit-rate control scheme and JSVM reference software. Further, Figure 14 presents a comparison of the CGS versus MGS quality types according to the proposed adaptive bit-rate control scheme.

As seen from Figure 14, the MGS and CGS of the base-layer (Layer 0) provide equal results, and the MGS of the enhancement layer (Layer 1) provides better results in terms

Table 8: Bit-rate control experimental results for "CITY," "CREW," "HARBOR," "ICE," and "SOCCER" video sequences (ROI QP in Layer 1 is equal to 37; the rest is determined by the proposed bit-rate control scheme; the detailed test conditions are provided in Table 7).

Video sequence	Layers				
	Target bit-rate for Layer 1 with the proposed bit-rate control	Actual bit-rate: Layer 1 with the proposed bit-rate control (ROI QP = 37, the rest by the proposed rate control)		Actual bit-rate of Layer 0 with JSVM 9.19 bit-rate control (QP = 40)	
	Bit-rate [kb/s]	Bit-rate [kb/s]	Average PSNR [dB]	Bit-rate [kb/s]	Average PSNR [dB]
CITY	400	406.0	29.7	713.8	31.7
	450	445.0	30.1		
	500	464.5	30.1		
	550	471.4	30.4		
CREW	400	437.0	31.8	897.7	34.5
	450	446.5	31.9		
	500	455.5	31.9		
	550	455.5	31.9		
HARBOR	600	598.4	28.0	1628.7	31.1
	650	683.7	28.5		
	700	728.0	28.7		
	750	729.0	28.7		
ICE	350	383.1	34.6	542.2	37.3
	400	386.7	35.0		
	450	391.3	35.2		
	500	391.3	35.2		
SOCCER	600	557.2	28.7	1119.2	32.0
	650	559.6	28.7		

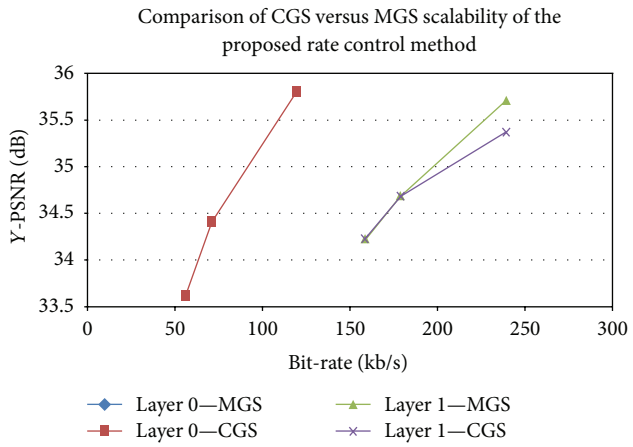


Figure 14: The CGS versus MGS comparison of the proposed adaptive bit-rate control method. The tested QP values for each layer are 30, 35, and 40.

of the PSNR, when compared to the JSVM reference software [22].

5. Conclusions

In this paper, an efficient scalable ROI video coding scheme was presented, enabling to extract the desired regions-of-interest and adaptively set the desirable ROI location, size, and resolution. Two methods for the scalable ROI video coding have been evaluated, that is, the ROI cropping method and the FMO method for the ROI coding, introducing a significantly low bit-rate overhead and very significant savings in bit-rate, respectively. In addition, an efficient adaptive bit-rate control for the ROI scalable video coding was presented, which, in turn, enables to obtain the desired high-quality region-of-interest and achieve significant overall bit-rate savings, while the average PSNR degradation for each frame (including the ROI) is relatively low. The performance of the presented techniques was demonstrated and compared with the SVC reference software (JSVM 9.19), thereby showing significant improvements in terms of the bit-rate savings as a tradeoff of the relatively low PSNR degradation.

Conflict of Interests

The authors do not have any direct financial relations with the commercial entities mentioned in this paper, which might lead to a conflict of interests.

Acknowledgments

This work was partially supported by the NEGEV consortium, MAGNET Program of the Israeli Chief Scientist,

Table 9: Bit-rate control experimental results for "CREW," "SHIELDS," "PARKRUN," and "SOCCER" video sequences (ROI QP in Layer 1 is equal to 20; the rest is determined by the proposed bit-rate control scheme; the detailed test conditions are provided in Table 7).

Video sequence	Layers				
	Target bit-rate for Layer 1 with the proposed bit-rate control	Actual bit-rate: Layer 1 with the proposed bit-rate control (ROI QP = 20, the rest by the proposed rate control)		Actual bit-rate of Layer 0 with JSVM 9.19 bit-rate control (QP = 40)	
	Bit-rate [kb/s]	Bit-rate [kb/s]	Average PSNR [dB]	Bit-rate [kb/s]	Average PSNR [dB]
CREW	1600	1691.4	30.1	2195.0	35.0
	1700	1691.4	30.1		
SHIELDS	5000	6393.1	37.8	6969.0	38.3
	6000	6399.6	38.2		
PARKRUN	700	701.5	24.0	1435.2	28.1
	750	714.9	24.1		
	800	717.4	24.2		
	850	843.8	25.1		
SOCCER	2300	2473.9	28.1	4105.9	34.1
	2500	2478.4	28.2		

Israeli Ministry of Trade and Industry, under Grant 85265610. The authors gratefully thank Igor Medvetsky, Ran Dubin, Aviad Hadarian, and Evgeny Kaminsky for their assistance in evaluation and testing.

References

[1] T. Spangler, "Lured by online video, digital broadcasts, more cable TV customers are cutting their service," 2008, http://www.multichannel.com/article/85964-Cover_Story_Breaking_Free.php.

[2] D. Grois, E. Kaminsky, and O. Hadar, "Optimization methods for H. 264/AVC video coding," in *The Handbook of MPEG Applications: Standards in Practice*, chapter 7, pp. 175–204, John Wiley and Sons, 2011.

[3] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103–1120, 2007.

[4] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[5] D. Grois and O. Hadar, "Recent trends in online mutimedia education for heterogeneous end-user devices based on Scalable Video Coding," in *Proceedings of the IEEE Global Engineering Education Conference (EDUCON '13)*, pp. 1141–1146, March 2013.

[6] D. Grois, E. Kaminsky, and O. Hadar, "Dynamically adjustable and scalable ROI video coding," in *Proceedings of the IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB '10)*, Shanghai, China, March 2010.

[7] M.-J. Chen, M.-C. Chi, C.-T. Hsu, and J.-W. Chen, "ROI Video coding based on H.263+ with robust skin-color detection technique," *IEEE Transactions on Consumer Electronics*, vol. 49, no. 3, pp. 724–730, 2003.

[8] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. I511–I518, December 2001.

[9] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Proceedings of the Computational Learning Theory (Eurocolt '95)*, pp. 23–37, Springer, 1995.

[10] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[11] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '00)*, pp. 142–149, Hilton Head Island, South Carolina, June 2000.

[12] P. Lambert, D. de Schrijver, D. van Deursen, W. de Neve, Y. Dhondt, and R. van de Walle, "A real-time content adaptation framework for exploiting ROI scalability in H.264/AVC," *Advanced Concepts for Intelligent Vision Systems*, vol. 4179, pp. 442–453, 2006.

[13] F. Manerba, J. Benois-Pineau, R. Leonardi, and B. Mansencal, "Multiple moving object detection for fast video content description in compressed domain," *Eurasip Journal on Advances in Signal Processing*, vol. 2008, Article ID 231930, 15 pages, 2008.

[14] C. Käs and H. Nicolas, "Compressed domain indexing of scalable H.264/SVC streams," *Signal Processing*, vol. 24, no. 6, pp. 484–498, 2009.

[15] C. Hanfeng, Z. Yiqiang, and Q. Feihu, "Rapid object tracking on compressed video," in *Proceedings of the 2nd IEEE Pacific Rim Conference on Multimedia*, pp. 1066–1071, October 2001.

[16] W. Zeng, J. Du, W. Gao, and Q. Huang, "Robust moving object segmentation on H.264/AVC compressed video using the block-based MRF model," *Real-Time Imaging*, vol. 11, no. 4, pp. 290–299, 2005.

[17] W. You, M. S. H. Sabirin, and M. Kim, "Moving object tracking in H. 264/AVC bitstream," in *Multimedia Content Analysis and*

Mining, vol. 4577, pp. 483–492, Springer, Heldelberg, Germany, 2007.

[18] V. Thilak and C. D. Creusere, "Tracking of extended size targets in H. 264 compressed video using the probabilistic data association filter," in *Real-Time Image and Video Processing (EUSIPCO '04)*, Proceedings of SPIE, pp. 281–284, September 2004.

[19] H. 264/AVC, Draft ITU-T Rec. and Final Draft Intl. Std. of Joint Video Spec. (H. 264/AVC), Joint Video Team, Doc. JVT-G050, 2003.

[20] "Applications and requirement for scalable video coding," JVT ISO/IEC JTC1/SC29/WG11 Doc. N6880, Hong-Kong, China, 2005.

[21] T. Wiegand et al., "ISO/IEC, 14496-10:200X/Amd. 3 Part 10: Advanced Video Coding—AMENDMENT 3: Scalable Video Coding Joint Draft ITU-T Rec. H. 264/ISO/IEC, 14496-10/Amd. 3 Scalable video coding," Joint Video Team Doc. JVT-X201, July 2007.

[22] JSVM, "JSVM Software Manual," Ver. JSVM 9. 19 (CVS tag: JSVM_9_19), 2009.

[23] T. Wiegand, G. Sullivan, J. Reichel, H. Schwarz, and M. Wien, "Joint draft 8 of SVC amendment," ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q. 6 9 (JVT-U201), Hangzhou, China, October 2006.

[24] H. Chen, Z. Han, R. Hu, and R. Ruan, "Adaptive FMO selection strategy for error resilient H.264 coding," in *Proceedings of the International Conference on Audio, Language and Image Processing (ICALIP '08)*, pp. 868–872, Shanghai, China, July 2008.

[25] Z. Lu et al., "CE8: ROI-based scalable video coding," JVT-O308, Busan, Korea, April 2005.

[26] T. C. Thang et al., "Spatial scalability of multiple ROIs in surveillance video," JVT-O037, Busan, Korea, April, 2005.

[27] Z. Lu, "Perceptual region-of-interest (ROI) based Scalable Video Coding," JVT-O056, Busan, Korea, April 2005.

[28] M. Shoaib and A. Cai, "Efficient residual prediction with error concealment in extended spatial scalability," in *Proceedings of the 9th Annual Wireless Telecommunications Symposium (WTS '10)*, April 2010.

[29] E. Francois and J. Vieron, "Extended spatial scalability: a generalization of spatial scalability for non dyadic configurations," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '06)*, pp. 169–172, October 2006.

[30] Y. Hu, D. Rajan, and L.-T. Chia, "Detection of visual attention regions in images using robust subspace analysis," *Journal of Visual Communication and Image Representation*, vol. 19, no. 3, pp. 199–216, 2008.

[31] L. Liu, S. Zhang, X. Ye, and Y. Zhang, "Error resilience schemes of H.264/AVC for 3G conversational video services," in *Proceedings of the 5th International Conference on Computer and Information Technology (CIT '05)*, pp. 657–661, Binghamton, New York, USA, September 2005.

[32] O. Ndili and T. Ogunfunmi, "On the performance of a 3D flexible macroblock ordering for H.264/AVC," in *Proceedings of the International Conference on Consumer Electronics (ICCE '06)*, pp. 37–38, January 2006.

[33] H. K. Arachchi, W. A. C. Fernando, S. Panchadcharam, and W. A. R. J. Weerakkody, "Unequal error protection technique for ROI based H.264 video coding," in *Proceedings of the Canadian Conference on Electrical and Computer Engineering (CCECE '06)*, pp. 2033–2036, Ottawa, Canada, May 2006.

[34] D. Grois, E. Kaminsky, and O. Hadar, "Adaptive bit-rate control for region-of-interest scalable video coding," in *Proceedings of the 26th Convention of Electrical and Electronics Engineers in Israel (IEEEI '10)*, pp. 761–765, Eilat, Israel, November 2010.

[35] Z. Li, F. Pan, K. P. Lim, G. Feng, X. Lin, and S. Rahardja, "Adaptive basic unit layer rate control for JVT," Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q. 6) Doc. JVT-G012, Pattaya, Thailand, 2003.

[36] Z. G. Li, W. Yao, S. Rahardja, and S. Xie, "New framework for encoder optimization of scalable video coding," in *Proceedings of the IEEE Workshop on Signal Processing Systems (SiPS '07)*, pp. 527–532, October 2007.

[37] D. Grois and O. Hadar, "Recent advances in Region-of-Interest coding," in *Recent Advances on Video Coding*, J. Del Ser Lorente, Ed., pp. 49–76, 2011.

[38] D. Grois and O. Hadar, "Advances in Region-of-Interest video and image processing," in *Multimedia Networking and Coding*, R. A. Farrugia and C. J. Debono, Eds., pp. 76–123, IGI Global, 2012.

[39] D. Grois and O. Hadar, "Region-of-Interest processing and coding techniques: overview of recent trends and directions," in *Intelligent Multimedia Technologies For Networking Applications: Techniques and Tools*, D. Kanellopoulos, Ed., pp. 126–155, IGI Global, 2013.

[40] Y. Liu, Z. G. Li, and Y. C. Soh, "Rate control of H.264/AVC scalable extension," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 1, pp. 116–121, 2008.

[41] L. Xu, S. Ma, D. Zhao, and W. Gao, "Rate control for scalable video model," in *Proceedings of the Visual Communications and Image Processing Conference*, pp. 525–534, July 2005.

[42] T. Anselmo and D. Alfonso, "Constant Quality Variable Bit-Rate control for SVC," in *Proceedings of the 11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS '10)*, April 2010.

[43] H. Roodaki, H. R. Rabiee, and M. Ghanbari, "Rate-distortion optimization of scalable video codecs," *Signal Processing*, vol. 25, no. 4, pp. 276–286, 2010.

[44] J. Ribas-Corbera, P. A. Chou, and S. L. Regunathan, "A generalized hypothetical reference decoder for H.264/AVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 674–687, 2003.

[45] K.-P. Lim, G. Sullivan, and T. Wiegand, "Text description of joint model reference encoding methods and decoding concealment methods," *Study of ISO/IEC, 14496-10 and ISO/IEC, 14496-5/ AMD6 and Study of ITU-T Rec. H. 264 and ITU-T Rec. H. 2. 64. 2, in Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG*, Doc. JVT-O079, Busan, Korea, April 2005.

[46] T. Chiang and Y.-Q. Zhang, "A new rate control scheme using quadratic rate distortion model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no. 1, pp. 246–250, 1997.

[47] E. Kaminsky, D. Grois, and O. Hadar, "Dynamic computational complexity and bit allocation for optimizing H.264/AVC video compression," *Journal of Visual Communication and Image Representation*, vol. 19, no. 1, pp. 56–74, 2008.

[48] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-constrained coder control and comparison of video coding standards," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 688–703, 2003.

[49] D. Grois, E. Kaminsky, and O. Hadar, "Buffer control in H.264/AVC applications by implementing dynamic

complexity-rate-distortion analysis," in *Proceedings of the IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB '09)*, May 2009.

[50] D. Grois, E. Kaminsky, and O. Hadar, "ROI adaptive scalable video coding for limited bandwidth wireless networks," in *Proceedings of the IFIP Wireless Days (WD '10)*, Venice, Italy, October 2010.

[51] D. Grois and O. Hadar, "Efficient adaptive bit-rate control for Scalable Video Coding by using computational complexity-rate-distortion analysis," in *Proceedings of the IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB '11)*, Nuremberg, Germany, June 2011.

[52] D. Grois and O. Hadar, "Complexity-aware adaptive spatial pre-processing for ROI scalable video coding with dynamic transition region," in *Proceedings of the 18th IEEE International Conference on Image Processing (ICIP '11)*, pp. 741–744, Brussels, Belgium, September 2011.