We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900
Open access books available

186,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

3

Recent Advances in Region-of-interest Video Coding

Dan Grois and Ofer Hadar
Ben-Gurion University of the Negev, Beer-Sheva, Israel

1. Introduction

Recently, the content distribution network industry has become exposed to significant changes. The advent of cheaper and more powerful mobile devices having the ability to play, create, and transmit video content and which maximize a number of multimedia content distributions on various mobile networks will place unprecedented demands on networks for high capacity, low-latency, and low-loss communications paths. The reduction of cost of digital video cameras along with development of user-generated video sites (e.g., iTunes™, Google™ Video and YouTube™) have stimulated the new user-generated content sector. Growing premium content coupled with advanced video technologies, such as the Internet TV, will replace in the near future conventional technologies (e.g., cable or satellite TV).

The relatively recent ITU-T H.264/AVC (ISO/IEC MPEG-4 Part 10) video coding standard (Wiegand & Sullivan, 2003), which was officially issued in 2003, has become a challenge for real-time video applications. Compared to others standards, it gains about 50% in bit rate, while providing the same visual quality. In addition to having all the advantages of MPEG-2, H.263 and MPEG-4, the H.264 video coding standard possesses a number of improvements, such as the content-adaptive-based arithmetic codec (CABAC), enhanced transform and quantization, prediction of "Intra" macroblocks (spatial prediction), and others. H.264 is designed for both constant bit rate (CBR) and variable bit rate (VBR) video coding, useful for transmitting video sequences over statistically multiplexed networks (e.g. asynchronous transfer mode (ATM), the Ethernet, or other Internet networks). This video coding standard can also be used at any bit rate range for various applications, varying from wireless video phones to high definition television (HDTV) and digital video broadcasting (DVB). In addition, H.264 provides significantly improved coding efficiency and greater functionality, such as rate scalability, “Intra” prediction and error resilience in comparison with its predecessors, MPEG-2 and H.263. However, H.264/AVC is much more complex in comparison to other coding standards and to achieve maximum quality encoding, high computational resources are required.

Due to the recent technological achievements and trends, the high-definition, highly interactive networked media applications pose challenges to network operators. The variety of end-user devices with different capabilities, ranging from cell phones with small screens and restricted processing power to high-end PCs with high-definition displays, have stimulated significant interest in effective technologies for video adaptation for spatial formats, consuming power and bit rate.

As a result, much of the attention in the field of video adaptation is currently directed to the Scalable Video Coding (SVC), which was standardized in 2007 as an extension of H.264/AVC (Schwarz et al., 2007), since the bit-stream scalability for video is currently a very desirable feature for many multimedia applications.
The need for the scalability arises from the need for spatial formats, bit rates or power (Wiegand & Sullivan, 2003). To fulfill these requirements, it would be beneficial to simultaneously transmit or store video in variety of spatial/temporal resolutions and qualities, leading to the video bit-stream scalability. 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 Scalable Video Coding provides important functionalities, such as the spatial, temporal and SNR (quality) scalability, thereby enabling the power adaptation. In turn, these functionalities lead to enhancements of video transmission and storage applications.
SVC has achieved significant improvements in coding efficiency comparing to the scalable profiles of prior video coding standards. Also, in addition to the temporal, spatial and quality scalabilities, the SVC supports the Region-of-Interest (ROI) scalability. The ROI is a desirable feature in many future scalable video coding applications, such as mobile device applications, which have to be adapted to be displayed on a relatively small screen (thus, a mobile device user may require to extract and track only a predefined Region-of-Interest within the displayed video). At the same time, other users having a larger mobile device screen may wish to extract other ROI(s) to receive greater video stream resolution. Therefore, to fulfill these requirements, it would be beneficial to simultaneously transmit or store a video stream in a variety of Regions-of-Interest (e.g., each Region-of-Interest having different spatial resolution, as illustrated in *Fig. 1*), as well to enable efficiently tracking the predefined Region-of-Interest.

Fig. 1. Defining ROIs with different spatial resolutions (e.g., CIF, SD/4CIF, 720p resolutions) to be provided within a Scalable Video Coding stream.

This chapter is organized as follows: in *Section 2*, the Region-of-Interest (ROI) detection and tracking is described in detail, while presenting the Pixel-Domain approach (*Section 2.1*) and Compressed-Domain approach (*Section 2.2*), and further presenting various models and techniques, such as the Visual Attention model (*Section 2.1.1*), Object Detection (*Section*

2.1.2), Face Detection (*Section 2.1.3*), Skin Detection (*Section 2.1.4*), etc.; in *Section 3,* the ROI Coding in H.264/SVC Standard is presented, including the ROI Scalability by Performing Cropping (*Section 3.1*) and the ROI Scalability by Using Flexible Macroblock Ordering (FMO) technique (*Section 3.2*); in *Section 4,* the bit-rate control for the ROI coding is presented; and Conclusions are provided in *Section 5*.

2. Region-of-interest detection and tracking

In order to successfully perform the ROI coding, it is important to accurately detect, and then correctly track, the desired Region-of-Interest. There are mainly two methods for the ROI detection and tracking: (a) the pixel-domain approach; and (b) the compressed-domain approach. The pixel-domain approach is more accurate compared to the compressed-domain approach, but it requires relatively high computational complexity resources. On the other hand, the compressed-domain approach does not consume many resources since it exploits the encoded information (such as DCT coefficients, motion vectors, macroblock types which are extracted in a compressed bitstream, etc.) (Manerba et al., 2008; Kas & Nicolas, 2009; Hanfeng et al., 2001; Zeng et al., 2005), but it results in a relatively poor performance. Also, for the same reason, the compressed-domain approach has significantly fast processing time and is adaptive to compressed videos. As a result, the compressed-domain approach is applicable mainly for simple scenarios.
Both the pixel-domain and compressed-domain approaches are explained in detail in the following *Sections 2.1 and 2.2*.

2.1 Pixel-domain approach

Generally, the main researches on object detection and tracking have been focused on the pixel domain approach since it can provide powerful capability of object tracking by using varyous technologies. The pixel-domain detection can be classified into the following types:

- Region-based methods. According to these methods, the object detection is performed according to ROI features, such as motion distribution and color histogram. The information with regard to the object colors can be especially useful when these colors are distinguishable from the image background or from other objects within the image (Vezhnevets, 2002).
- Feature-based methods (Shokurov et al., 2003). According to these methods, various motion parameters of feature points are calculated (the motion parameters are related to affine transformation information, which in turn contains rotation and 2D translation data).
- Contour-based methods. According to these methods, the shape and position of objects are detected by modeling the contour data (Wang et al., 2002).
- Template-based methods. According to these methods, the objects (such as faces) are detected by using predetermined templates (Schoepflin et al., 2001).

As mentioned above, phe pixel-domain approach 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. Therefore, the desired ROI can be predicted in a relatively accurate manner by defining various pixel-domain models, such as visual attention models, object detection models, face detection models, etc., as presented in detail in the following *Sub-Sections 2.1.1 to 2.1.4*.

2.1.1 Visual attention

The visual attention models refer to the ability of a human user to concentrate his/her attention on a specific region of an image/video. This involves selection of the sensory information by the primary visual cortex in the brain by using a number of characteristic, such as intensity, color, size, orientation in space, and the like (Hu et al., 2008). Actually, the visual attention models simulate the behavior of the Human Visual System (HVS), and in turn enable to detect the Region-of-Interest within the image/video, such as presented in *Fig. 2.*

Fig. 2. An example of concentrating the attention on a specific region of an image.

Several researches have been conducted with this regard in order to achieve better ROI detection performance, and in turn improve the ROI visual presentation quality. Thus, for example (Cheng et al., 2005) presents a framework for automatic video Region-of-Interest determination based on user attention model, while considering the three types of visual attention features, i.e. intensity, color and motion. The contrast-based intensity model is based on the fact that particular color pairs, such as red-green and blue-yellow possess high spatial and chromatic opposition; the same characteristics exist in high deference lighting or intensity pairs. Thus, according to (Cheng et al., 2005), the intensity, red-green color and blue-yellow color constant models should be included into the user attention representation module. Also, when there is more than one ROI within the frame (e.g., a number of football players), then a saliency map is used which shows the ability to characterize the visual attraction of the image/video. The saliency map is divided into n regions, and ROI is declared for each such region, thereby enabling to dynamically and automatically determine ROI for each frame-segment.

Further, (Sun et al., 2010) proposes a visual attention based approach to extract texts from complicated background in camera-based images. First, it applies the simplified visual attention model to highlight the region of interest (ROI) in an input image and to yield a map consisting of the ROIs. Second, an edge map of image containing the edge information of four directions is obtained by Sobel operators; character areas are detected by connected component analysis and merged into candidate text regions. Finally, the map consisting of the ROIs is employed to confirm the candidate text regions.

Further, other visual attention models have been recently proposed to improve the ROI visual presentation quality, such as (Engelke et al., 2009), which discusses two ways of obtaining subjective visual attention data that can be subsequently used to develop visual attention models based on the selective region-of-interest and visual fixation patterns; (Chen et al., 2010) discloses a model of the focus of attention for detecting the attended regions in video sequences by using the similarity between the adjacent frames, establishing the gray histogram, selecting the maximum similarity as predicable model, and finally obtaining a position of the focus of attention in the next fame; (Li et al., 2010) presents a three-stage method that combines the visual attention model with target detection by using the saliency map, covering the region of interest with blocks and measuring the similarity between the blocks and the template; (Kwon et al., 2010) shows a ROI based video preprocessor method that deals with the perceptual quality in a low-bit rate communication environment, further proposing three separated processes: the ROI detection, the image enhancement, and the boundary reduction in order to deliver better video quality at the videoconferencing application for use in a fixed camera and to be compatible as a preprocessor for the conventional video coding standards.
As seen from the above, the visual attention approach has recently become quite popular among researchers, and many improved techniques have been lately presented.

2.1.2 Object detection

Automatic object detection is one of the important steps in image processing and computer vision (Bhanu et al., 1997; Lin et al., 2005). The major task of object detection is to locate objects in images and extract the regions containing them (the extracted regions are ROIs). The quality of object detection is highly dependent on the effectiveness of the features used in the detection. Finding or designing appropriate features to capture the characteristics of objects and building the feature-based representation of objects are the key to the success of detection. Usually, it is not easy for human experts to figure out a set of features to characterize complex objects, and sometimes, simple features directly extracted from images may not be effective in object detection.
The ROI detection is especially useful for medical applications (Liu, 2006). Automatic detection of ROI in a complex image or video like endoscopic neurosurgery video, is an important task in many image and video processing applications such as image-guide surgery system, real-time patient monitoring system, and object-based video compression. In telemedical applications, object-based video coding is highly useful because it produces a good perceptual quality in a specified region, i.e., a region of interest (ROI), without requiring an excessive bandwidth. By using a dedicated video encoder, the ROI can be coded with more bits to obtain a much higher quality than that of the non-ROI which is coded with fewer bits.
In the last decade, various object detection techniques have been proposed. For example, (Han et al., 2008) presents a fully automated architecture for object-based ROI detection, based on the principle of discriminant saliency, which defines as salient the image regions of strongest response to a set of features that optimally discriminate the object class of interest from all the others. It consists of two stages, saliency detection and saliency validation. The first detects salient points, the second verifies the consistency of their geometric configuration with that of training examples. Both the saliency detector and the configuration model can be learned from cluttered images downloaded from the web.

Also, (Wang J. M. et al., 2008) describes a simple and novel algorithm for detecting foreground objects in video sequences using just two consecutive frames. The method is divided in three layers: sensory layer, perceptual layer, and memory layer (short-term memory in conceptual layer). In sensory layer, successive images are obtained from one fixed camera, and some early computer vision processing techniques are applied here to extract the image information, which are edges and inconsistent region. In perceptual layer, moving objects are extracted based on the information from the sensory layer, and may request the sensory layer support more detail. The detecting results are stored in the memory layer, and help the perceptual layer to detect the temporal static objects.
In addition, (Jeong, 2006) proposes an objectionable image detection system based on the ROI. The system proposed by (Jeong, 2006) excels in that ROI detection method is specialized in objectionable image detection. In addition, a novel feature consisting of weighted SCD based on ROI and skin color structure descriptor is presented for classifying objectionable image. Using the ROI detection method, (Jeong, 2006) can reduce the noisy information in image and extract more accurate features for classifying objectionable image.
Further, (Lin et al., 2005) uses genetic programming (GP) to synthesize composite operators and composite features from combinations of primitive operations and primitive features for object detection. The motivation for using GP is to overcome the human experts' limitations of focusing only on conventional combinations of primitive image processing operations in the feature synthesis. GP attempts many unconventional combinations that in some cases yield exceptionally good results. Compared to a traditional region-of-interest extraction algorithm, the composite operators learned by GP are more effective and efficient for object detection. Still further, (Kim & Wang, 2009) proposes a method for smoke detection in outdoor video sequences, which contains three steps. The first step is to decide whether the camera is moving or not. While the camera is moving, the authors skip the ensuing steps. Otherwise, the second step is to detect the areas of change in the current input frame against the background image and to locate regions of interest (ROIs) by connected component analysis. In the final step, the authors decide whether the detected ROI is smoke by using the k-temporal information of its color and shape extracted from the ROI.

2.1.3 Face detection

The face detection can be regarded as a specific case of object-class detection. In object-class detection, the task is to find the locations and sizes of all objects in an image that belong to a given class (such as pedestrians, cars, and the like). Also, the face detection can be regarded as a more general case of face localization. In face localization, the task is to find the locations and sizes of a known number of faces (usually one). In face detection, one does not have this additional information.
Early face-detection algorithms focused on the detection of frontal human faces, whereas recent face detection method aim to solve the more general and difficult problem of multi-view face detection. The face detection from an image video is considered to be a relatively difficult task due to a plurality of possible visual representations of the same face: the face scale, pose, location, orientation in space, varying lighting conditions, face emotional expression, and many others (e.g., as presented in *Fig. 3*). Therefore, in spite of the recent technological progress, this field still has many challenges and problems to be resolved.

Generally, the challenges associated with face detection can be attributed to the following factors (Yang et al., 2010):

- *Facial expression*. The appearance of faces is directly affected by a person's facial expression.
- *Pose*. The images of a face vary due to the relative camera-face pose (frontal, 45 degree, profile, upside down), and some facial features such as an eye or the nose may become partially or wholly occluded.
- *Occlusion*. Faces may be partially occluded by other objects. In an image with a group of people, some faces may partially occlude other faces.
- *Image orientation*. Face images directly vary for different rotations about the camera's optical axis.
- *Imaging conditions*. When the image is formed, factors such as lighting (spectra, source distribution and intensity) and camera characteristics (sensor response, lenses) affect the appearance of a face.
- *Presence or absence of structural components*. Facial features such as beards, mustaches, and glasses may or may not be present and there is a great deal of variability among these components including shape, color, and size.

Fig. 3. An example of a plurality of possible visual representations of the same face, which has an influence on the accurate face detection. Although the accuracy of face detection systems has dramatically increased during the last decade, such systems still have many challenges and problems to be resolved, such as varying lighting conditions, facial expression, presence or absence of structural components, etc.

During the last decade, many researchers around the world tried to improve the face detection and develop an efficient and accurate detection system. Such for example, (Mustafah et al., 2009) proposes a design of a face detection system for real-time high resolution smart camera, while making an emphasis on the problem of crowd surveillance where the static color camera is used to monitor a wide area of interest, and utilizing a background subtraction method to reduce the Region-of-Interest (ROI) to areas where the moving objects are located. Another work was performed by (Zhang et al., 2009), in which was presented a ROI based H.264 encoder for videophone with a hardware macroblock level face detector. The ROI definition module operates as a face detector in videophone, and it is embedded into the encoder to define the currently processed and encoded ROI macroclocks, while the encoding process is dynamically controlled according to the ROI (the encoding parameters vary according to ROI).

Further, other face detection techniques have been recently proposed to improve the face detection, such as: (Micheloni et al., 2005) presents an integrated surveillance system for the

outdoor security; (Qayyum & Javed, 2006) discloses a notch based face detection, tracking and facial feature localization system, which contains two phases: visual guidance and face/non-face classification; and (Sadykhov & Lamovsky, 2008) discloses a method for real-time face detection in 3D space.

2.1.4 Skin detection

The successful recognition of the skin ROI simplifies the further processing of such ROI. The main aim of traditional skin ROI detection schemes is to detect skin pixels in images, thereby generating skin areas. According to (Abdullah-Al-Wadud & Oksam, 2007), if ROI detection process misses a skin region or provides regions having lots of holes in it, then the reliability of applications significantly decreases. Therefore, it is important to maintain the efficiency of the human-computer interaction (HCI) based systems. In turn, (Abdullah-Al-Wadud & Oksam, 2007) presents an improved region-of-interest selection method for skin detection applications. This method can be applied in any explicit skin cluster classifier in any color space, while do not requiring any learning or training procedure. The proposed algorithm mainly operates on a grayscale image (DM), but the processing is based on color information. The scalar distance map contains the information of the vector image, thereby making this method relatively simple to implement.

Also, (Yuan & Mu, 2007) presents an ear detection method, which is based on skin-color and contour information, while introducing a modified Continuously Adaptive Mean Shift (CAMSHAFT) algorithm for rough and fast profile tracking. The aim for profile tracking is to locate the main skin-color region, such as the ROI that contains the ear. The CAMSHIFT algorithm is based on a robust non-parameter technique for climbing density gradients to find peak of probability distribution called the mean shift algorithm. The mean shift algorithm operates on probability distribution, so in order to track colored objects in video sequence, the color image data has to be represented as the color distribution first. According to (Yuan & Mu, 2007), the modified CAMSHIFT method is performed as follows:

- Generating the skin-color histogram on training set skin images.
- Setting the initial location of the 2D mean shift search window at a fixed poison in the first frame such as the center of the frame.
- Using the generated skin-color histogram to calculate the skin-color probability distribution of the 2D region centered at the area slightly larger than the mean shift window size.
- Calculating the zeroth moment (area of size) and mean location (the centroid).
- For the next frame, centering the search window at the calculated mean location and setting the window size using a function of the zeroth moment. Then the previous two steps are repeated.

In addition, (Chen et al., 2003) presents a video coding H263 based technique for robust skin-color detection, which is suitable for real time videoconferencing. According to (Chen et al., 2003), the ROIs are automatically selected by a robust skin-color detection which utilizes the Cr and RGB variance instead of the traditional skin color models, such as YCbCr, HSI, etc. The skin color model defined by Cr and RGB variance can choose the skin color region more accurately than other methods. The distortion weight parameter and variance at the macroblock layer are adjusted to control the qualities at different regions. As a result, the quality at the ROI is can significantly improved.

2.2 Compressed-domain detection

The conventional compressed domain algorithms exploit motion vectors or DCT coefficients instead of original pixel data as resources in order to reduce computational complexity of object detection and tracking (You, 2010).

In general, the compressed domain algorithms can be categorized as follows: the clustering-based methods and the filtering-based methods.

The clustering-based methods (Benzougar et al., 2001; Babu et al., 2004; Ji & Park, 2000; Jamrozik & Hayes, 2002) 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/AVC standard, is the region growing approach, in which several seed fragments grow spatially and temporally by merging similar neighboring fragments.

On the other hand, the filtering-based methods (Aggarwal et al., 2006; Zheg et al., 2005; You et al., 2007; You et al., 2009) 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 clustering procedure.

2.3 Region-of-interest tracking

Object tracking based on video sequence plays an important role in many modern vision applications such as intelligent surveillance, video compression, human-computer interfaces, sports analysis (Haritaoglu et al, 2000). When object is tracked with an active camera, traditional methods such as background subtraction, temporal differencing and optical flow may not work well due to the motion of camera, tremor of camera and the disturbance from background (Xiang, 2009).

Some researchers propose methods of tracking moving target with an active camera, yet most of their algorithms are too computationally complex due to their dependence on accurate mathematical model and motion model, and can't be applied to real-time tracking in presence of fast motion from the object or the active camera, irregular motion and un-calibrated camera. (Xiang, 2009) makes great effort to find a fast, computationally efficient algorithm, which can handle fast motion, and can smoothly follow-up track moving target with an active camera, by proposing a method for real-time follow-up tracking fast moving object with an active camera. (Xiang, 2009) focuses on the color-based Mean Shift algorithm which shows excellent performance both on computationally complexity and robustness.

(Wei & Zhou, 2010) presents a novel algorithm that uses the selective visual attention mechanisms to develop a reliable algorithm for objects tracking that can effectively deal with the relatively big influence by external interference in a-priori approaches. To extract the ROI, it makes use of the "local statistic" of the object. By integrating the image feature with state feature, the synergistic benefits can bring following obvious advantages:

- It doesn't use any a-priori knowledge about blobs and no heuristic assumptions must be provided;
- The computation of the model for a generic blob doesn't take a long processing time.

According to (Wei & Zhou, 2010), during the detection phase, there are some false-alarms in any actual image. To reduce the fictitious targets as much as possible, it needs to identify the extracted ROI, while the tracing target can be defined by the following characteristics:

- *The length of boundary of the tracing target in the ROI.*
- *Aspect ratio.* The length and the width of the target can be expressed by the two orthogonal axes of minimum enclosing rectangle. The radio between them is the aspect ratio.
- *Shape complexity.* The radio between the length of the boundary and the area.

The ROI, whose parameters accord with the above three features, can be considered as the ROI including the real- target.

Further, there are many other recent tracking methods, such as: (Mehmood, 2009) implements kernel tracking of density-based appearance models for real-time object tracking applications; (Wang et al., 2009) discloses a wireless, embedded smart camera system for cooperative object tracking and event detection; (Sun, Z. & Sun, J., 2008) presents an approach for detecting and tracking dynamic objects with complex topology from image sequences based on intensive restraint topology adaptive snake mode; (Wang & Zhu, 2008) presents a sensor platform with multi-modalities, consisting of a dual-panoramic peripheral vision system and a narrow field-of-view hyperspectral fovea; thus, only hyperspectal images in the ROI should be captured; (Liu et al., 2006) presents a new method that addresses several challenges in automatic detection of ROI of neurosurgical video for ROI coding, which is used for neurophysiological intraoperative monitoring (IOM) system. According to (Liu et al., 2006), the method is based on an object tracking technique with multivariate density estimation theory, combined with the shape information of the object, thereby by defining the ROIs for neurosurgical video, this method produces a smooth and convex emphasis region, within which surgical procedures are performed. (Abousleman, 2009) presents an automated region-of-interest-based video coding system for use in ultra-low-bandwidth applications.

3. Region-of-interest coding in H.264/SVC standard

Region-of-Interest (ROI) coding is a desirable feature in future applications of Scalable Video Coding (SVC), especially in applications for the wireless networks, which have a limited bandwidth. However, the H.264/AVC standard does not explicitly teach as how to perform the ROI coding.

The ROI coding is supported by various techniques in the H.264/AVC standard (Wiegand & Sullivan, 2003) and the SVC (Schwarz et al., 2007) extensions. Some of these techniques include quantization step size control at the slice and macroblock levels, and are related to the concept of slice grouping, also known as Flexible Macroblock Ordering (FMO). For example, (Lu et al., 2005a) 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 pre-encoded scalable bit-stream is truncated. (Lu et al., 2005a) presents a number of ROI enhancement quality layers to provide fine granular scalability. In addition, (Thang et al., 2005) presents ROI-based spatial scalability scheme, concerning two main issues: overlapped regions between ROIs and providing different ROIs resolutions. However, (Thang et al., 2005) follows the concept of slice grouping of H.264/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) enabling to avoid duplicate coding of overlapped regions in multiple ROIs by encoding the overlapped regions such that the corresponding encoded regions can be independently decoded. Further, (Lu et al., 2005b) 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 (Lu et al., 2005b), the proposed ROI based compression achieves better perceptual quality and improves coding efficiency. Moreover, (Lampert et al., 2006) relates to extracting the ROIs (i.e., of an original bit-stream by introducing a description-driven content adaptation framework. According to (Lampert et al., 2006), two methods for ROI extraction are implemented: (a) the removal of the non-ROI portions of a bit-stream; and (b) the replacement of coded background with corresponding placeholder slices. In turn, bit-streams 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 quality of the decoded video sequence, 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 (Lampert et al., 2006), in addition to the bandwidth decrease, the adaptation process has a positive effect on the decoder due to the relatively easy processing of placeholder slices, thereby increasing the decoding speed.
Below we present a novel dynamically adjustable and scalable ROI video coding scheme, enabling to adaptively and efficiently set the desirable ROI location, size, resolution and bit-rate, according to the network bandwidth (especially, if it is a wireless network in which the bandwidth is limited), power constraints of resource-limited systems (such as mobile devices/servers) where the low power consumption is required, and according to end-user resource-limited devices (such as mobile devices, PDAs, and the like), thereby effectively selecting best encoding scenarios suitable for most heterogonous and time-invariant end-user terminals (i.e., different users can be connected each time) and network bandwidths.
In the following *Sections 3.1* and *3.2*, different types of ROI scalability are presented: the ROI scalability by performing cropping and ROI scalability by employing the Flexible Macroblock Ordering (FMO) technique, respectively.

3.1 ROI scalability by performing cropping

According to the first method for the ROI video coding, and in order to enable obtaining a high-quality ROI on resource-limited devices (such as mobile devices), we crop the ROI from the original image and use it as a baselayer (or other low enhancement layers, such as Layer 1 or 2), as schematically illustrated in *Fig. 4* below (Grois et al., 2010a).

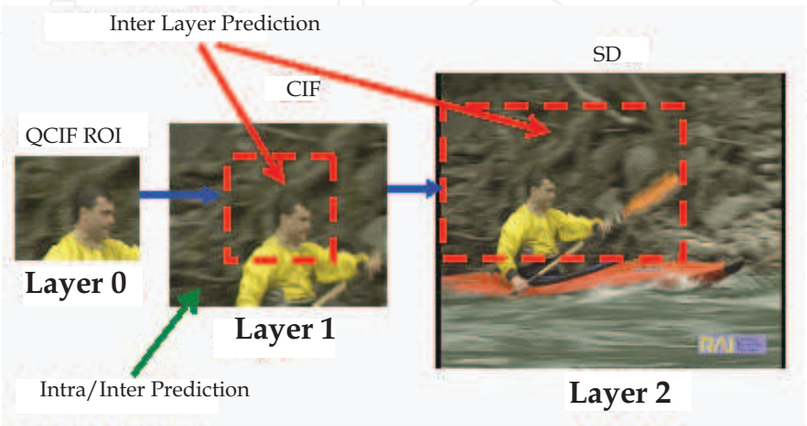


Fig. 4. The example of the ROI dynamic adjustment and scalability (e.g., for mobile devices with different spatial resolutions) by using a cropping method.

Then, we perform an Inter-layer prediction in the similar sections of the image, i.e., in the cropping areas. As a result, for example (*Fig. 4*), by using the Inter-layer prediction for the three-layer (QCIF-CIF-SD) coding (with the similar quantization parameter (QP) settings at each layer), we achieve the significantly low bit-rate overhead. Prior to cropping the image, we determine the location of a cropping area in the successive layer of the image (in Layer 1, and then in Layer 2, as shown on *Fig. 4*). For this, we employ an ESS (Extended Spatial Scalability) method (Shoaib & Anni, 2010). In addition, we define a GOP for the SVC as a group between two I/P frames, or any combination thereof. Thus, as shown for example in *Table 3*, for the "*SOCCER*" video sequence (30 fp/sec; 300 frames; GOP size 16; QPs varying from 22 to 34) we obtain the bit-rate overhead of only 4.7% to 7.9% compared to conventional single layer coding.

Tables 1 to *3* below present R-D (Rate-Distortion) experimental results for the variable-layer coding with different cropping spatial resolutions, while using the Inter/Intra-layer prediction. As it is clearly seen from these tables, there is significantly low bit–rate overhead, which is especially important for transmitting over limited-bandwidth networks (such as wireless networks). Particularly, the *Tables 1* below presents the R-D (Rate-Distortion) experimental results for the two-layer coding (QCIF-CIF) with the QCIF cropping versus the single layer coding.

Quantization Parameters	**Single layer**		**QCIF-CIF**		**Bit–Rate Overhead (%)**
	PSRN [dB]	**Bit-Rate[K/sec]**	**PSNR [dB]**	**Bit-Rate[K/sec]**	
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

Table 1. Two-layer (QCIF-CIF) spatial scalability coding vs. single layer coding ("*SOCCER*" video sequence, 30 fp/s, 300 frames, GOP size 16).

Also, the *Tables 2* below presents the R-D (Rate-Distortion) experimental results for the two-layer coding (CIF-SD) with the CIF cropping versus the single layer coding.

Quantization Parameters	**Single layer**		**CIF-SD**		**Bit–Rate Overhead (%)**
	PSRN [dB]	**Bit-Rate [K/sec]**	**PSNR [dB]**	**Bit-Rate [K/sec]**	
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 2. Three-layer (CIF-SD) spatial scalability coding vs. single layer coding ("*SOCCER*" video sequence, 30 fp/s, 300 frames, GOP size 16).

Further, the *Tables 3* below presents the R-D (Rate-Distortion) experimental results for the three-layer coding (QCIF-CIF-SD) with the QCIF-CIF cropping versus the single layer coding.

Quantization Parameters	Single layer		QCIF-CIF-SD		Bit–Rate Overhead (%)
	PSRN [dB]	Bit-Rate [K/sec]	PSNR [dB]	Bit-Rate [K/sec]	
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

Table 3. Three-layer (QCIF-CIF-SD) spatial scalability coding vs. single layer coding ("*SOCCER*" video sequence, 30 fp/s, 300 frames, GOP size 16).

As was mentioned above, it is clearly seen from the above experimental results that when using the Inter/Intra-layer prediction, the bit-rate overhead is very small and is much less than 10%.

3.2 ROI scalability by using flexible macroblock ordering

The second method refers to the ROI dynamic adjustment and scalability (Grois et al., 2010a) by using the FMO (Flexible Macroblock Ordering) in the scalable baseline profile (not for Layer 0, which is similar to the H.264/AVC baseline profile without the FMO).

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 (the slice coding is one of the techniques used in H.264 for transmission) (Chen et al., 2008; Liu et al., 2005; Ndili & Ogunfunmi, 2006; Kodikara et al., 2006). 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 employ different macroblocks to slice groups of mapping patterns.

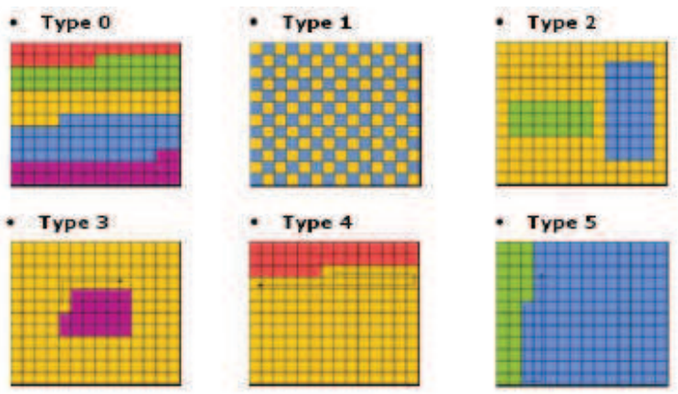


Fig. 5. Six fixed types of the FMO (interleaved, dispersed, foreground, box-out, raster scan and wipe-out), while each color represents a slice group).

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 standard, FMO consists of seven slice group map types (*Type 0* to *Type 6*), six of them are predefined fixed macroblock mapping types (as illustrated in Fig. 5: interleaved, dispersed, foreground, box-out, raster scan and

wipe-out), which can be specified through picture parameter setting (PPS), and the last one is a custom type, which allows the full flexibility of assigning macroblock to any slice group. The ROI can be defined as a separate slice in the FMO *Type 2* 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 regions (these definitions should be made in the SVC configuration files, according to the JSVM 9.19 reference software manual (JSVM, 2009).

For the Scalable Video Coding, we use the FMO *Type 2* above, where each ROI is represented by a separate rectangular region and is encoded as a separate slice. *Tables 4* presents experimental results for the four layers spatial scalability coding versus six layers coding of the "*SOCCER*" sequence (30 fp/s; 300 frames; GOP size is 16), where four layers are presented by one CIF layer and three SD layers having the CIF-resolution ROI in an upper-left corner of the image. In turn, the six layers are presented by three CIF layers (each layer is a crop from the SD resolution) and three 4CIF/SD layers.

Quantization Parameters	**Four Layers (CIF and three SD layers)**		**Six Layers (three CIF layers and three SD layers)**		**Bit-Rate Savings (%)**
	PSRN [dB]	**Bit-Rate [K/sec]**	**PSNR [dB]**	**Bit-Rate [K/sec]**	
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

Table 4. FMO: Four-layer spatial scalability coding vs. six-layer coding ("*SOCCER*" video sequence, 30 fp/s, 300 frames, GOP size 16).

It should be noted that each of the above three CIF layers (crops extracted from the SD resolution image) can be considered, for example, as a zoom of the image in a upper-left corner, as shown in *Fig. 6* below.

(a) (b)

Fig. 6. (a) the CIF crop (representing Layer 0, i.e. the base-layer) extracted from the SD resolution frame of the "*SOCCER*" sequence; (b) the corresponding HD resolution image, representing Layer 1 of the "*SOCCER*" sequence. The white dashed lines show the zoomed ROI.

Further, *Table 5* presents R-D (Rate-Distortion) experimental results for the HD (High Definition) video sequence "*STOCKHOLM*" (*Fig. 1*, 1280x720, 30 fp/sec, GOP size 8, 160 frames) by using four-layer coding (640x360 layer and three HD layers having two ROIs

(CIF and 4CIF/SD resolutions) in the upper left corner of the image) versus eight-layer coding (two CIF layers (scalable baseline profile without B frames), three 4CIF/SD layers, and three HD layers having different quantization parameters). The quantization parameters vary from 32 to 36 with a step size of 2.

Quantization Parameters	Four Layers (640x360, and three HD layers)		Eight Layers (two CIF layers, three SD layers, and three HD layers)		Bit–Rate Savings (%)
	PSRN [dB]	Bit-Rate [K/sec]	PSNR [dB]	Bit-Rate [K/sec]	
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 5. FMO: Four-layer coding vs. eight-layer coding ("*STOCKHOLM*", 30 fp/s, 96 frames, GOP size 8)

Further, *Table 6* below presents R-D (Rate-Distortion) experimental results for the HD video sequence "*STOCKHOLM*" by using four-layer coding (640x360 layer and three HD layers having two ROIs (CIF and SD resolution, respectively) in the upper-left corner of the image) versus six-layer coding (three CIF and three SD layers).

Quantization Parameters	Four Layers (640x360, and three HD layers)		Six Layers (three CIF layers and three SD layers)		Bit–Rate Savings (%)
	PSRN [dB]	Bit-Rate [K/sec]	PSNR [dB]	Bit-Rate [K/sec]	
32	34.5	2566.2	34.5	3237.0	19.3
34	33.9	1730.2	33.9	2359.1	29.7
36	33.3	1170.0	33.3	1759.0	39.9

Table 6. FMO: Four-layer coding vs. six-layer coding ("*STOCKHOLM*", 30 fp/s, 96 frames, GOP size 8)

As it is clearly observed from *Table 4* to *6* above, there are very significant bit-rate savings – up to 39%, when using the FMO techniques.

4. Bit-rate control for region-of-interest coding

The bit-rate control is crucial in providing desired compression bit rates for H264/AVC video applications, and especially for the Scalable Video Coding, which is the extension of H264/AVC.

The bit-rate control has been intensively studied in existing single layer coding standards, such as MPEG 2, MPEG 4, and H.264/AVC (Li et al., 2003). 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 bit-stream. 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 (Li et al., 2007). Thus, the inputs of the bit-rate control scheme are: the given bandwidth; usually, the

statistics of video sequence including Mean Squared Error (MSE); and a header of each predefined unit (e.g., a basic unit, macroblock, frame, 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.

In the current JSVM reference software (JSVM, 2009) there is no rate control mechanism, besides the base-layer rate control, which do not consider 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 (JSVM, 2009; Liu et al., 2008). Of course, this is very inefficient and much time-consuming. For solving this problem, only a few works have been published during the last years, trying to provide an efficient rate control mechanism for the SVC. However, none of them handles scalable bit-rate control for the Region-of-Interest (ROI) coding. Such, in (Xu et al., 2005) 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 make a full use of the hierarchical relations inheritance from the MCTF. In addition, (Liu et al., 2008) proposes a switched model to predict the mean absolute difference (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, (Anselmo & Alfonso, 2010) describes a constant quality variable bit-rate (VBR) control algorithm for multiple layer coding. According to 0 (Anselmo & Alfonso, 2010), the algorithm allows achieving a target quality by specifying memory capabilities and the bit-rate limitations of the storage device. In the more recent work (Roodaki et al., 2010), the joint optimization of layers in the layered video coding is investigated. The authors show that spatial scalability, like the SNR scalability, does benefit from joint optimization, though not being able to exploit the relation between the quantizer step sizes. However, as mentioned above, there is currently no efficient bit-rate control scheme for the ROI Scalable Video Coding.

Below, we present a method and system for the efficient ROI Scalable Video Coding, according to which we achieve a bit-rate that is very close to the target bit-rate, while being able to define the desirable ROI quality (in term of QP or Peak Signal-To-Noise Ratio (PSNR)) and while adaptively changing the background region quality (the background region excludes the ROI), according to the overall bit-rate.

In order to provide the different visual presentation quality to at least one ROI and to the background (or other less important region of the frame), we divide each frame to at least two slices, while one slice is used for defining the ROI and at least 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 on larger number of slices, such that for each ROI we use a separate slice.

The general proposed 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) ((Ribas-Corbera et al., 2003). The calculation should consider that each SVC layer

contains a number of predefined slices (the ROI slice, background slice, etc.), which should be encoded with different QPs.

b. Allocate the remaining bits to all non-coded macroblocks (MBs) for each predefined slice in the current frame of the particular SVC layer.
c. Estimate the MAD (Mean Absolute Difference) for the current macroblock in the current slice by a linear prediction model (Li et al., 2003; Lim et al., 2005) using the actual MAD of the macroblocks in the co-located position of the previous slices (in the previous frames) within the same SVC layer and the MAD of neighbor macroblocks in the current slice.
d. Estimate a set of groups of coding modes (e.g., modes such as Inter-Search16X8, Inter-Search8X16, Inter-Search8X8, Inter-Search8X4, Inter-Search4X8, Inter-Search4X4 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 co-located 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 a quadratic model (Chiang & Zhang, 1997; Kaminsky et al., 2008; Grois et al., 2010c).
f. Perform the Rate-Distortion Optimization (RDO) for each MB by using the QPs derived from the above step 5.
g. Adaptively adjust the QPs (increase/decrease the QPs by a predefined quantization step size) according to the current overall bit-rate.

In *Fig. 7* below, is presented a system for performing the proposed adaptive bit-rate control for the Scalable Video Coding (for simplicity, only two layers are shown – Base-Layer (Layer 0), and Enhancement Layer (Layer 1). The system contains the SVC adaptive bit-rate controller, which continuously receives data regarding the current buffer occupancy, actual bit-rate and quantization parameters (Grois et al., 2010b).

The step (f) above can be performed by using a method (Lim et al., 2005; Wiegand et al., 2003) for determining an optimal coding mode for encoding each macroblock. According to method (Lim et al., 2005; Wiegand et al., 2003), the RDO for each macroblock is performed for selecting an optimal coding mode by minimizing the Lagrangian function as follows:

$$J\left(orig, rec, MODE \middle| \lambda_{MODE}\right) = D\left(orig, rec, MODE \middle| QP\right) + \lambda_{MODE} \cdot R\left(orig, rec, MODE \middle| QP\right) \quad (1)$$

where the distortion $D\left(orig, rec, MODE \middle| QP\right)$ 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; $R\left(orig, rec, MODE \middle| QP\right)$ is the number of bits associated with selecting $MODE$; and λ_{MODE} is a Lagrangian multiplier for the mode decision (Lim et al., 2005).

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, respectively). The formulation of the optimal buffer control (for controlling the buffer occupancy for each SVC layer) can be given by:

$$\min_{\{r(i)\}}\{\sum_{i=1}^{N} e(i)\}, \qquad \text{subject to } B^{Layer}{}_{\max} \geq B^{Layer}(i) \geq 0 \quad (2)$$

for i = 1, 2, ..., N

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

$$B^{Layer}(i+1) = B^{Layer}(i) + r^{Layer}(i) - r_{out}^{Layer} \tag{3}$$

where $r^{Layer}(i)$ is the buffer input bit-rate with regard to each SVC layer and r_{out}^{Layer} is the output bit-rate of buffer contents.

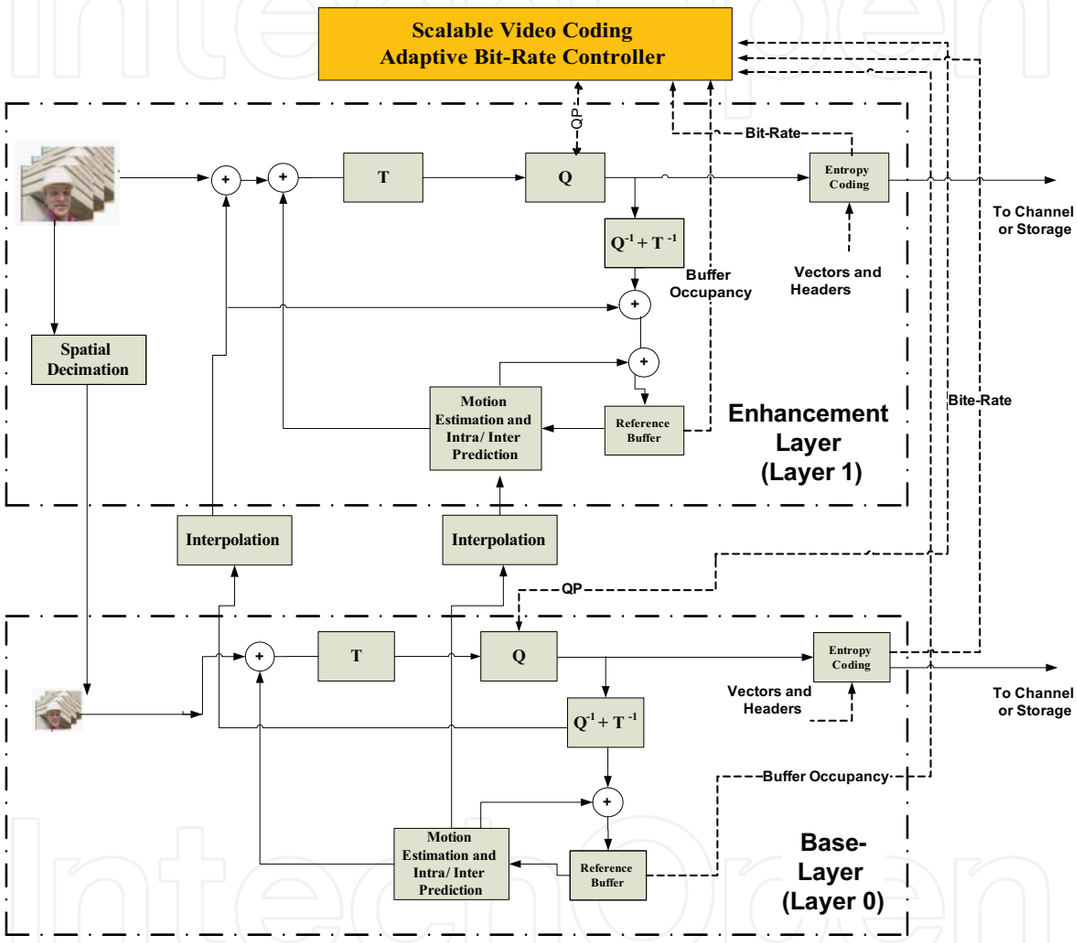


Fig. 7. The system for performing the presented adaptive spatial bit-rate control for the Scalable Video Coding (for simplicity, only two layers – Layer 0 and Layer 1 - are presented).

The optimal buffer control approach is related to the following optimal bit allocation formulation,

$$\min_{\{r(i)\}}\{\sum_{i=1}^{N} e(i)\} \text{ , subject to } \sum_{i=1}^{n} r^{Layer}(i) \le R^{Layer} \tag{4}$$

for i = 1, 2, ..., N
and is schematically presented in *Fig. 8* below.

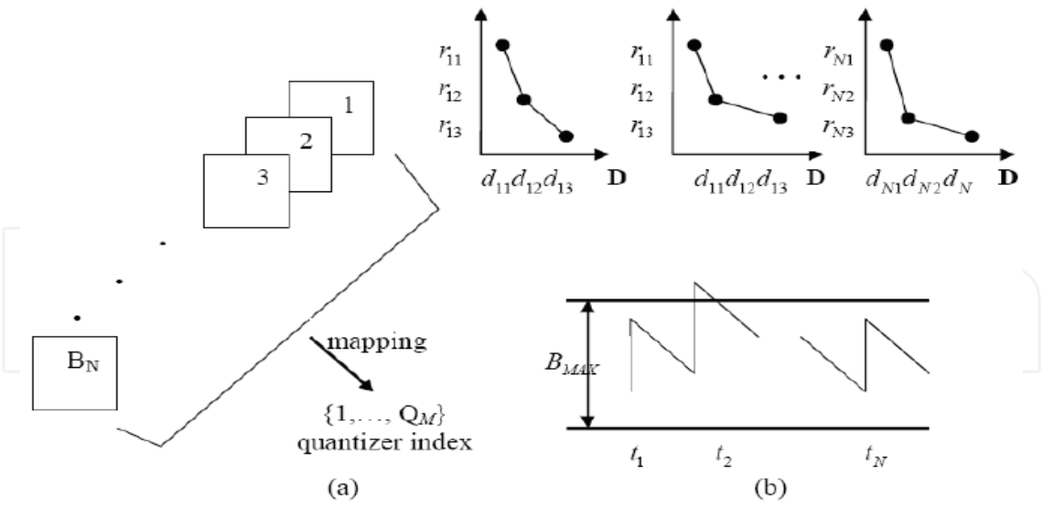


Fig. 8. (a) Each block (*1...*B_N) in the sequence has different R-D characteristics (for a given set of quantizers (*1...*Q_M) for blocks in the sequence, we can obtain R-D (Rate-Distortion) points (r_{N1}, r_{N2}, r_{N3} *and* d_{N1}, d_{N2}, d_N, etc.) to form composite characteristics); and (b) R at t_2 is not a feasible solution to the selected maximum buffer size B_{MAX}.

For overcoming 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 (2) can be solved by implementing the C-R-D analysis of (Grois et al., 2009) for each SVC layer.

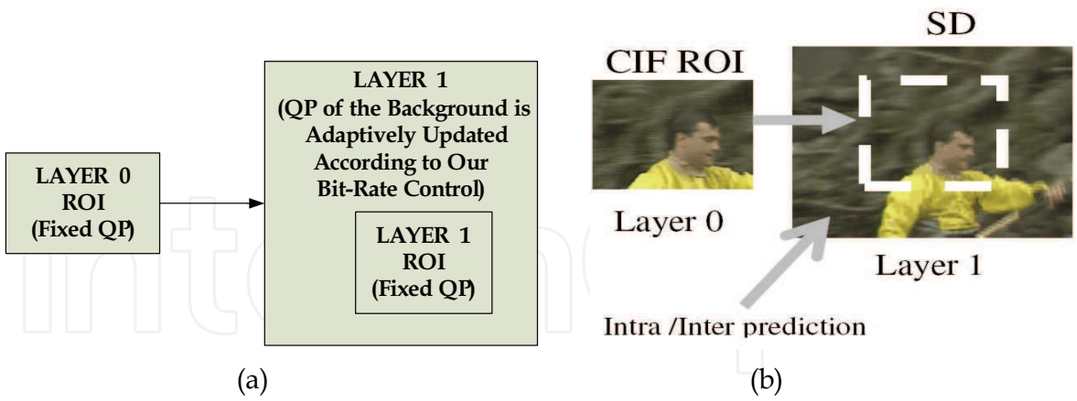


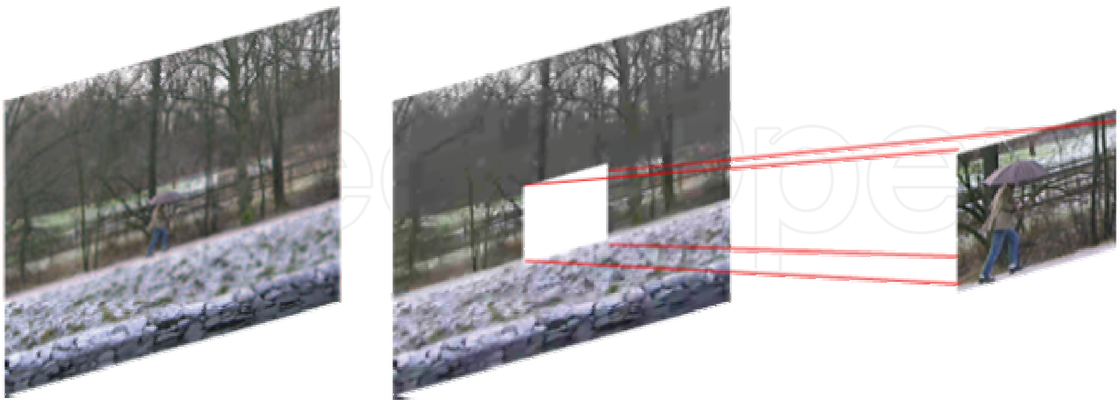
Fig. 9. (a) Defining two or more layers with corresponding QPs. The QP of the background region in Layer 1 is determined adaptively by our bit-rate control; (b) CIF ROI is used as Base Layer (Layer 0), and 4CIF (SD) is used as an Enhancement Layer (Layer 1). The Intra/ Inter-prediction is used for reducing the overall bit-rate.

For simplicity, in this section, we show results for the bit-rate control of two layers: 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 *Fig. 9*. According to the presented adaptive bit-rate control, we preset for each layer different

initial quantization parameters (QPs): e.g., for the whole Layer 0 we can define an initial quantization parameter to be equal to 40, and for the ROI region provided in Layer 1 we can define an initial quantization parameter to be equal to 20; and then the QP of the remaining background region in Layer 1 is determined adaptively by our bit-rate control. In such a way, we can obtain the desired quality of the Region-of-Interest, and as a result, of the remaining background region (or any other less important region) according to the overall network bandwidth (either constant or variable bandwidth).
As a result, by encoding the video sequence with different QPs, we enable obtaining the optimal presentation quality of the predefined ROI region and enable reducing the quality of the background, as presented for example, in *Fig. 10* ("*SOCCER*" video sequence, SD resolution).

Fig. 10. The "SOCCER" video sequence (SD 704x576, 25 fp/sec.) containing the ROI region in the upper-left corner.

Figs. 11 and *12* below illustrate sample frames of the "*PARKRUN*" video sequence, which contains the ROI region – the man with an umbrella. The quantization parameter of the background region can be determined adaptively in order to achieve optimal video presentation quality (as it is clearly seen from *Figs. 11(b)* and *12(b)*, the QP of the background region is much higher than the QP of the ROI region).



(a) (b)

Fig. 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 can be determined adaptively); (a) the original frame; and (b) the compressed frame with the higher-quality ROI region.

(a) (b)

Fig. 12. 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 can be determined adaptively); (a) the original frame; and (b) the compressed frame with the higher-quality ROI region.

Further, *Fig. 13* below shows another frame of the *"SHIELDS"* video sequence, which contains the ROI region – man's head and hand pointing to the shields. The quantization parameter of the background region can be determined adaptively according to the adaptive bit-rate control (as it is seen from *Fig. 13(b)*, the QP of the background region is much higher than the QP of the ROI region).

(a) (b)

Fig. 13. The "SHIELDS" video sequence containing the ROI region – 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.

The following *Table 7* presents experimental results for the bit-rate control operation for various video sequences ("CITY", "CREW", "HARBOR", "ICE", and "SOCCER"), along with the corresponding PSNR and bit-rate values. According to the conducted tests, 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 by our adaptive SVC bit-rate control scheme.

Video Sequence	Target Bit-Rate for Layer 1 with our Bit-Rate Control	Layers			
		Actual Bit-Rate: Layer 1 with our Bit-Rate Control (ROI QP=20, the rest by our Rate Control)		Actual Bit-Rate of Layer 0 with JSVM 9.19 Bit-Rate Control (QP=40)	
	Bit-Rate [K/sec]	Bit-Rate [K/sec]	Average PSNR [dB]	Bit-Rate [K/sec]	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	38.3		
PARKRUN	7000	7010.5	24.0	3435.2	28.1
	7500	7140.9	24.1		
	8000	7172.4	24.2		
	8500	8431.8	25.1		
SOCCER	2300	2473.9	28.1	2105.9	34.1
	2500	2478.4	28.2		

Table 7. 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 our bit-rate control).

5. Conclusions

In this chapter we have presented a comprehensive overview of recent developments in the area of Region-of-Interest Video Coding, making an emphasis on the ROI Scalable Video Coding field, which has become popular in the last couple of years due to standardization of the SVC in 2007, as an extension of H.264/AVC.

Also, we have presented our efficient novel scalable video coding schemes, enabling to adaptively set the desirable ROI location, size, resolution (e.g., the spatial resolution), ROI visual quality and amount of bits allocated for the ROI, and perform other predefined settings. According to these schemes, we achieve a significantly low bit–rate overhead and very significant savings in bit-rate, thereby enabling to provide an efficient adaptive bit-rate control for the ROI Scalable Video Coding, which was also presented in detail. In turn, the adaptive bit-rate control has enabled us to provide the high-quality video coding for the desired Region-of-Interest, while considering the overall available bandwidth, and other predefined parameters. The performance of the presented schemes was demonstrated and compared with the (Joint Scalable Video Model) JSVM reference software (JSVM 9.19), thereby showing a significant improvement in term of the PSNR values and bit-rate.

6. Acknowledgments

This work was supported by the NEGEV consortium, MAGNET Program of the Israeli Chief Scientist, Israeli Ministry of Trade and Industry under Grant 85265610. We thank Igor Medvetsky, Ran Dubin, Aviad Hadarian and Evgeny Kaminsky for their assistance in evaluation and testing.

7. References

Abdullah-Al-Wadud, M. & Oksam C. (2007). Region-of-Interest Selection for Skin Detection Based Applications, *Convergence Information Technology, 2007. International Conference on*, vol., no., pp.1999-2004, 21-23 Nov. 2007.

Abousleman, G.P. (2009). Target-tracking-based ultra-low-bit-rate video coding, *Military Communications Conference, 2009. MILCOM 2009. IEEE*, vol., no., pp.1-6, 18-21 Oct. 2009.

Aggarwal, A.; Biswas, S.; Singh, S.; Sural, S. & Majumdar, A. K. (2006). Object Tracking Using Background Subtraction and Motion Estimation in MPEG Videos, *ACCV 2006*, LNCS, vol. 3852, pp. 121-130, Springer, Heidelberg (2006).

Anselmo, T. & Alfonso, D., (2010). Constant Quality Variable Bit-Rate control for SVC, Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on, vol., no., pp.1-4, 12-14 April 2010.

Babu, R. V.; Ramakrishnan, K. R. & Srinivasan, S. H. (2004). Video object segmentation: A compressed domain approach, *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, No. 4, pp. 462–474, April 2004.

Bae T. M.; Thang T. C.; Kim D. Y.; Ro Y. M.; Kang J. W. & Kim J. G. (2006). Multiple region-of-interest support in scalable video coding," ETRI journal 2006, vol. 28, no. 2, pp. 239 – 242.

Benzougar, A.; Bouthemy, P. & Fablet, R. (2001). MRF-based moving object detection from MPEG coded video, *in Proc. IEEE Int. Conf. Image Processing*, 2001, vol. 3, pp.402-405.

Bhanu, B.; Dudgeon, D. E.; Zelnio, E. G.; Rosenfeld, A.; Casasent & D.; Reed, I. S. (1997). Guest Editorial Introduction To The Special Issue On Automatic Target Detection And Recognition, *Image Processing, IEEE Transactions on*, vol.6, no.1, pp.1-6, Jan 1997.

Bing L.; Mingui S.; Qiang L.; Kassam, A.; Ching-Chung Li & Sclabassi, R.J. (2006). Automatic Detection of Region of Interest Based on Object Tracking in Neurosurgical Video, *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, vol., no., pp.6273-6276, 17-18 Jan. 2006.

Chen, M.-J.; Chi, M.-C.; Hsu, C.-T. & Chen, J.-W. (2003). ROI video coding based on H.263+ with robust skin-color detection technique, *Consumer Electronics, 2003. ICCE. 2003 IEEE International Conference on*, vol., no., pp. 44- 45, 17-19 June 2003.

Chen, H.; Han, Z.; Hu, R. & Ruan, R. (2008). Adaptive FMO Selection Strategy for Error Resilient H.264 Coding, *Int. Conf. on Audio, Lang. and Image Proc.*, *ICALIP* 2008, Jul. 7-9, pp. 868-872, Shanghai, China.

Chen, Q.-H.; Xie X.-F.; Guo T.-J.; Shi L. & Wang X.-F (2010). The Study of ROI Detection Based on Visual Attention Mechanism, *Wireless Communications Networking and Mobile Computing (WiCOM), 2010 6th International Conference on*, vol., no., pp.1-4, 23-25 Sept. 2010.

Chiang, T. & Zhang, Y.-Q. (1997). A new rate control scheme using quadratic rate distortion model, *IEEE Trans. Circuit Syst. Video Technol*., vol. 7, no. 1, pp. 246–250, 1997.

Engelke, U.; Zepernick, H.-J. & Maeder, A. (2009). Visual attention modeling: Region-of-interest versus fixation patterns, *Picture Coding Symposium, 2009. PCS 2009*, vol., no., pp.1-4, 6-8 May 2009.

Grois, D.; Kaminsky, E. & Hadar, O. (2009). Buffer control in H.264/AVC applications by implementing dynamic complexity-rate-distortion analysis, *Broadband Multimedia Systems and Broadcasting, 2009. BMSB '09. IEEE International Symposium on*, pp.1-7, 13-15 May 2009.

Grois, D.; Kaminsky, E. & Hadar, O., (2010). ROI adaptive scalable video coding for limited bandwidth wireless networks, *Wireless Days (WD), 2010 IFIP*, pp.1-5, 20-22 Oct. 2010.

Grois, D.; Kaminsky, E. & Hadar, O. (2010). Adaptive bit-rate control for Region-of-Interest Scalable Video Coding, *Electrical and Electronics Engineers in Israel (IEEEI), 2010 IEEE 26th Convention of*, pp.761-765, 17-20 Nov. 2010.

Grois, D.; Kaminsky, E. & Hadar, O. (2010). Optimization Methods for H.264/AVC Video Coding, The Handbook of MPEG Applications: Standards in Practice, (eds M. C. Angelides and H. Agius), John Wiley & Sons, Ltd, Chichester, UK.

Hanfeng, C.; Yiqiang, Z. & Feihu, Q. (2001). Rapid object tracking on compressed video, in *Proc. 2nd IEEE Pacific Rim Conference on Multimedia*, Oct. 2001, pp.1066-1071.

Haritaoglu, I.; Harwood, D. & Davis, L. S. (2000). W4: real-time surveillance of people and their activities, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol.22, no.8, pp.809-830, Aug 2000.

Hu, Y.; Rajan, D.; Chia, L. (2008). Detection of visual attention regions in images using robust subspace analysis, *Journal of Visual Communication and Image Pepresentation,* 19(3): 199-216, 2008.

Jamrozik, M. L. & Hayes, M. H. (2002). A compressed domain video object segmentation system, *in Proc. IEEE Int. Conf. Image Processing*, 2002, vol. 1, pp.113-116.

Jeong, C. Y.; Han, S. W.; Choi, S. G. & Nam, T. Y., An Objectionable Image Detection System Based on Region of Interest, *Image Processing, 2006 IEEE International Conference on*, vol., no., pp.1477-1480, 8-11 Oct. 2006.

Ji, S. & Park, H. W. (2000). Moving object segmentation in DCT-based compressed video, *Electronic Letters*, Vol. 36, No. 21, October 2000.

JSVM (2009). JSVM Software Manual, Ver. JSVM 9.19 (CVS tag: JSVM_9_19), Nov. 2009.

Kaminsky, E.; Grois, D. & Hadar, O. (2008). Dynamic Computational Complexity and Bit Allocation for Optimizing H.264/AVC Video Compression, *J. Vis. Commun. Image R.*, Elsevier, vol. 19, iss. 1, pp. 56-74, Jan. 2008.

Kas, C. & Nicolas, H. (2009). Compressed domain indexing of scalable H.264/SVC streams," Signal Processing Image Communication (2009), Special Issue on scalable coded media beyond compression, pp. 484-498, 2009.

Kim, D.-K. & Wang, Y.-F. (2009). Smoke Detection in Video, *Computer Science and Information Engineering, 2009 WRI World Congress on*, vol.5, no., pp.759-763, March 31, 2009-April 2, 2009.

Kodikara Arachchi, H.; Fernando, W.A.C.; Panchadcharam, S. & Weerakkody, W.A.R.J. (2006). Unequal Error Protection Technique for ROI Based H.264 Video Coding, *Canadian Conference on Electrical and Computer Engineering,* pp. 2033-2036, Ottawa, 2006.

Kwon, H.; Han, H.; Lee, S.; Choi, W. & Kang, B. (2010). New video enhancement preprocessor using the region-of-interest for the videoconferencing, *Consumer Electronics, IEEE Transactions on*, vol.56, no.4, pp.2644-2651, Nov. 2010.

Lambert, P.; Schrijver, D. D.; Van Deursen, D.; De Neve, W.; Dhondt, Y. & Van de Walle, R. (2006). A Real-Time Content Adaptation Framework for Exploiting ROI Scalability in H.264/AVC, *Advanced Concepts for Intelligent Vision Systems*, pp. 442-453, 2006.

Li, Z.; Pan, F.; Lim, K. P.; Feng, G.; Lin, X. & Rahardja, S. (2003). Adaptive basic unit layer rate control for JVT, in 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, Mar. 2003.

Li, Z. G.; Yao, W.; Rahardja, S. & Xie, S. (2007). New Framework for Encoder Optimization of Scalable Video Coding, *2007 IEEE Workshop on Signal Processing Systems*, pp.527-532, 17-19 Oct. 2007.

Li, Z.; Zhang, X.; Zou, F. & Hu, D. (2010). Study of target detection based on top-down visual attention, *Image and Signal Processing (CISP), 2010 3rd International Congress on*, vol.1, no., pp.377-380, 16-18 Oct. 2010.

Lim, K-P.; Sullivan, G. & Wiegand, T. (2005). 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*, Busan, Korea, Apr. 2005, Doc. JVT-O079.

Liu, L.; Zhang, S.; Ye, X. & Zhang, Y. (2005). Error resilience schemes of H.264/AVC for 3G conversational video services, *The Fifth International Conference on Computer and Information Technology*, pp. 657- 661, Binghamton, 2005.

Liu, Y.; Li, Z. G. & Soh, Y. C. (2008). Rate Control of H.264/AVC Scalable Extension, *Circuits and Systems for Video Technology, IEEE Transactions on*, vol.18, no.1, pp.116-121, Jan. 2008.

Lu, Z.; Peng, W.-H.; Choi, H.; Thang T. C. & Shengmei, S. (2005). CE8: ROI-based scalable video coding, JVT-O308, Busan, KR, 16-22 April, 2005.

Lu, Z.; Lin, W.; Li, Z.; Pang Lim, K.; Lin, X.; Rahardja, S.; Ping Ong, E. & Yao, S. (2005). Perceptual Region-of-Interest (ROI) based scalable video coding, JVT-O056, Busan, KR, 16-22 April, 2005.

Manerba, F.; Benois-Pineau, J.; Leonardi, R. & Mansencal, B. (2008). Multiple object extraction from compressed video, JASP – *EURASIP Journal on Advances in Signal Processing*, Vol. 2008 (2008), Article ID 231930, 15 pages, doi:10.1155/2008/231930.

Mehmood, M. O. (2009). Study and implementation of color-based object tracking in monocular image sequences, *Research and Development (SCOReD), 2009 IEEE Student Conference on*, vol., no., pp.109-111, 16-18 Nov. 2009.

Micheloni, C.; Salvador, E.; Bigaran, F. & Foresti, G.L. (2005). An integrated surveillance system for outdoor security, *Advanced Video and Signal Based Surveillance, 2005. AVSS 2005. IEEE Conference on*, vol., no., pp. 480- 485, 15-16 Sept. 2005.

Mustafah, Y.M.; Bigdeli, A.; Azman, A.W. & Lovell, B.C. (2009). Face detection system design for real time high resolution smart camera, *Distributed Smart Cameras, 2009. ICDSC 2009. Third ACM/IEEE International Conference on*, vol., no., pp.1-6, Aug. 30 2009-Sept. 2 2009.

Ndili, O. & Ogunfunmi, T. (2006). On the performance of a 3D flexible macroblock ordering for H.264/AVC, *Digest of Technical Papers International Conference on Consumer Electronics*, 2006, pp. 37-38.

Qayyum, U. & Javed, M.Y. (2006). Real time notch based face detection, tracking and facial feature localization, *Emerging Technologies, 2006. ICET '06. International Conference on*, vol., no., pp.70-75, 13-14 Nov. 2006.

Ribas-Corbera, J.; Chou, P. A. & Regunathan, S. L. (2003). A generalized hypothetical reference decoder for H.264/AVC, *IEEE Trans. Circuit Syst. Video Technol.*, vol. 13, pp. 674-686, Jul. 2003.

Roodaki, H.; Rabiee, H. R. & Ghanbari, M. (2010). Rate-distortion optimization of scalable video codecs, *Signal Processing: Image Communication*, vol. 25, iss. 4, Apr. 2010, pp. 276-286.

Sadykhov, R. Kh. & Lamovsky, D. V. (2008). Algorithm for real time faces detection in 3D space, *Computer Science and Information Technology, 2008. IMCSIT 2008. International Multiconference on*, vol., no., pp.727-732, 20-22 Oct. 2008.

Schwarz, H.; Marpe, D. & Wiegand, T. (2007). Overview of the scalable video coding extension of the H.264/AVC standard, *IEEE Trans. Circ. Syst. for Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sept. 2007.

Schoepflin, T.; Chalana, V.; Haynor, D. R. & Kim, Y. (2001). Video object tracking with a sequential hierarchy of template deformations, *IEEE Trans. Circuits Syst. Video Technol.* 11, pp.1171-1182, 2001.

Shoaib, M. & Anni C. (2010). Efficient residual prediction with error concealment in extended spatial scalability, *Wireless Telecommunications Symposium (WTS)*, 2010, vol., no., pp.1-6, 21-23 Apr. 2010.

Shokurov, A.; Khropov, A. & Ivanov, D. (2003). Feature tracking in images and video," in *International Conference on Computer Graphics between Europe and Asia* (*GraphiCon-2003*), pp.177-179, Sept. 2003.

Sun, Z. & Sun, J. (2008). Tracking of Dynamic Image Sequence Based on Intensive Restraint Topology Adaptive Snake, *Computer Science and Software Engineering, 2008 International Conference on*, vol.6, no., pp.217-220, 12-14 Dec. 2008.

Sun, Q; Lu, Y. & Sun, S. (2010). A Visual Attention Based Approach to Text Extraction, *Pattern Recognition (ICPR), 2010 20th International Conference on*, vol., no., pp.3991-3995, 23-26 Aug. 2010.

Wang, T. & Zhu, Z. (2008). Intelligent multimodal and hyperspectral sensing for real-time moving target tracking, *Applied Imagery Pattern Recognition Workshop, 2008. AIPR '08. 37th IEEE*, vol., no., pp.1-8, 15-17 Oct. 2008.

Thang, T. C.; Bae, T. M.; Jung, Y. J.; Ro, Y. M.; Kim, J.-G.; Choi, H. & Hong, J.-W. (2005). Spatial scalability of multiple ROIs in surveillance video, JVT-O037, Busan, KR, 16-22 April, 2005.

Vezhnevets, M. (2002). Face and facial feature tracking for natural Human-Computer Interface," in *International Conference on Computer Graphics between Europe and Asia* (*GraphiCon-2002*), pp.86-90, September 2002.

Wang, J.-M.; Cherng, S.; Fuh, C.-S. & Chen, S.-W. (2008). Foreground Object Detection Using Two Successive Images, *Advanced Video and Signal Based Surveillance, 2008. AVSS '08. IEEE Fifth International Conference on*, vol., no., pp.301-306, 1-3 Sept. 2008.

Wang, H.; Leng, J. & Guo, Z. M. (2002). "Adaptive dynamic contour for real-time object tracking," in Image and Vision Computing New Zealand (IVCNZ2002), December 2002.

Wei, Z. & Zhou, Z. (2010). An adaptive statistical features modeling tracking algorithm based on locally statistical ROI, *Educational and Information Technology (ICEIT), 2010 International Conference on*, vol.1, no., pp.V1-433-V1-437, 17-19 Sept. 2010.

Wiegand, T. & Sullivan, G. (2003). Final draft ITU-T recommendation and final draft international standard of joint video specification (ITU-T Rec. H.264 ISO/IEC 14 496-10 AVC), in Joint Video Team (JVT) of ITU-T SG16/Q15 (VCEG) and ISO/IEC JTC1/SC29/WG1, Annex C, Pattaya, Thailand, Mar. 2003, Doc. JVT-G050.

Wiegand, T.; Schwarz, H.; Joch, A.; Kossentini, F. & Sullivan, G. J. (2003). Rate-constrained coder control and comparison of video coding standards, *IEEE Trans. Circuit Syst. Video Technol.*, vol. 13, iss. 7, pp. 688- 703, Jul. 2003.

Wiegand, T.; Sullivan, G.; Reichel, J.; Schwarz, H. & Wien, M. (2006). Joint draft 8 of SVC amendment, ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6 9 (JVT-U201), 21st Meeting, Hangzhou, China, Oct. 2006.

Xiang, G. (2009). Real-Time Follow-Up Tracking Fast Moving Object with an Active Camera, *Image and Signal Processing, 2009. CISP '09. 2nd International Congress on*, vol., no., pp.1-4, 17-19 Oct. 2009.

Xu, L.; Ma, S.; Zhao, D. & Gao, W. (2005). Rate control for scalable video model, *Proc. SPIE, Visual Commun. Image Process.*, vol. 5960, pp. 525, 2005.

Yang, M.-H.; Kriegman, D.J. & Ahuja, N. (2002). Detecting faces in images: a survey, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol.24, no.1, pp.34-58, Jan 2002.

You, W.; Sabirin, M. S. H. & Kim, M. (2007). Moving object tracking in H.264/AVC bitstream, *MCAM* 2007, LNCS, vol. 4577, Springer, Heldelberg, 2007, pp.483-492.

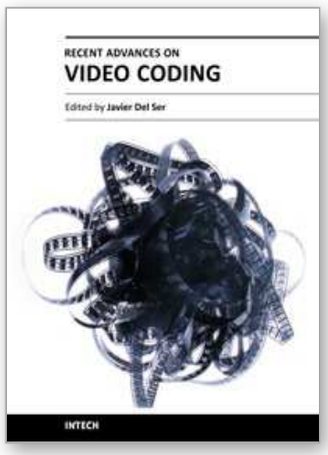
You, W.; Houari Sabirin, M. S. & Kim M. (2006). Real-time detection and tracking of multiple objects with partial decoding in H.264/AVC bitstream domain, *Proceedings of SPIE*, N. Kehtarnavaz and M.F. Carlsohn, San Jose, CA, USA: SPIE, 2009, pp. 72440D-72440D-12.

You, W. (2010). Object Detection and Tracking in Compresses Domain. Available from http://knol.google.com/k/wonsang-you/object-detection-and-tracking-in/3e2si9juvje7y/7#.

Youlu, W.; Casares, M. & Velipasalar, S. (2009). Cooperative Object Tracking and Event Detection with Wireless Smart Cameras, *Advanced Video and Signal Based Surveillance, 2009. AVSS '09. Sixth IEEE International Conference on*, vol., no., pp.394-399, 2-4 Sept. 2009.

Yuan, L. & Mu, Z.-C. (2007). Ear Detection Based on Skin-Color and Contour Information, *Machine Learning and Cybernetics, 2007 International Conference on*, vol.4, no., pp.2213-2217, 19-22 Aug. 2007.

IntechOpen



Recent Advances on Video Coding

Edited by Dr. Javier Del Ser Lorente

ISBN 978-953-307-181-7
Hard cover, 398 pages
Publisher InTech
Published online 24, June, 2011
Published in print edition June, 2011

This book is intended to attract the attention of practitioners and researchers from industry and academia interested in challenging paradigms of multimedia video coding, with an emphasis on recent technical developments, cross-disciplinary tools and implementations. Given its instructional purpose, the book also overviews recently published video coding standards such as H.264/AVC and SVC from a simulational standpoint. Novel rate control schemes and cross-disciplinary tools for the optimization of diverse aspects related to video coding are also addressed in detail, along with implementation architectures specially tailored for video processing and encoding. The book concludes by exposing new advances in semantic video coding. In summary: this book serves as a technically sounding start point for early-stage researchers and developers willing to join leading-edge research on video coding, processing and multimedia transmission.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Dan Grois and Ofer Hadar (2011). Recent Advances in Region-of-interest Video Coding, Recent Advances on Video Coding, Dr. Javier Del Ser Lorente (Ed.), ISBN: 978-953-307-181-7, InTech, Available from: http://www.intechopen.com/books/recent-advances-on-video-coding/recent-advances-in-region-of-interest-video-coding

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2011 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License, which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.

IntechOpen

IntechOpen