Remote interactive browsing of video surveillance content based on JPEG 2000

F.-O. Devaux, J. Meessen, C. Parisot, J.-F. Delaigle, B. Macq and C. De Vleeschouwer

***Abstract*— In video surveillance applications, pre-stored images are likely to be accessed remotely and interactively upon user request. In such a context, the JPEG 2000 still image compression format is attractive because it supports flexible and progressive access to each individual image of the pre-stored content, in terms of spatial location, quality level, as well as resolution. However, when the client wants to play consecutive frames of the video sequence, the purely INTRA nature of JPEG 2000 dramatically penalizes the transmission efficiency. To mitigate this drawback, conditional replenishment mechanisms are envisioned. They convey arbitrary spatio-temporal segments of the initial video sequence directly through sporadic and rate-distortion optimized refresh of JPEG 2000 packets. Hence, they preserve JPEG 2000 compliance, while saving transmission resources. The replenishment algorithms proposed in this paper are original in two main aspects. First, they exploit the specificities of the JPEG 2000 codestream structure to balance the accuracy (in terms of bit-planes) of the replenishment across image subbands in a rate-distortion optimal way. Second, they take into account the still background nature of video surveillance content by maintaining two reference images at the receiver. One reference is the last reconstructed frame, as proposed in [2] and [3]. The other is a dynamically-computed estimate of the scene background, which helps to recover the background after a moving object has left the scene. As an additional contribution, we demonstrate that the embedded nature of the JPEG 2000 codestream easily supports prioritization of semantically relevant regions of interest while browsing video content. An interesting aspect of this JPEG 2000-based prioritization is that it can be regulated a posteriori, after the codestream generation, based on the interest expressed by the user at browsing time. Simulation results demonstrate the efficiency and flexibility of the approach compared to INTER-based solutions.**

***Index Terms*—Video server, JPEG 2000, Conditional replenishment, Adaptive and interactive media delivery**

I. INTRODUCTION

We consider application scenarios for which a client - typically a human controller behind a PC or a wireless PDA - remotely accesses pre-encoded content captured by possibly multiple (overlapping) surveillance cameras, to figure out what happened in the monitored scene at some earlier time. In such a context, a desired browsing interface should allow the end-user to randomly select any spatio-temporal segment of the video(s) at arbitrary resolution so that, in a typical interactive browsing scenario, the end-user can first survey the (multiple) video(s) at low temporal and spatial resolution, and then focus on higher resolution displays of short video segments of interest, or decide to zoom in on a specific spatial area or object of interest, either in a particular frame or video segment. Regarding deployment, we are interested in a browsing architecture that can scale to handle large volumes of content, captured by distinct cameras, on multiple sites, at distinct time instants. Therefore, the content has to be stored efficiently in a compressed format, and the computational load associated to content storage, access and distribution has to be limited[1].

To address the above requirements, we have decided to build our system on the JPEG 2000 compression standard [4]. JPEG 2000 indeed provides a natural solution to support the required access flexibility, through low complex manipulation of pre-encoded bitstreams [5] [6], without the need for computationally expensive transcoding -i.e. decompression followed by compression- operations. In the meantime, we have also renounced to exploit temporal prediction during compression, so as to preserve the capability of random temporal access to each individual frame of the sequence. To mitigate the penalty induced by a strict INTRA coding structure, we have adapted conditional replenishment principles to the specificities of JPEG 2000 and of video surveillance scenes. This has been done at two levels. First, next to the previously reconstructed frame, the pre-computed estimation of the scene background has been considered as a potential candidate to reconstruct the current frame in absence of replenishment information. Second, decisions about the replenishment of JPEG 2000 packets have been optimized in the rate-distortion (RD) sense by taking into account potential semantic information, e.g. defining some knowledge about the regions interesting the user in the scene. Interestingly that knowledge is exploited independently of the compression engine, which means that it can be provided a posteriori, at transmission time by each individual user.

In final, the integrated contributions of our paper result in a video server that:

- implements a multi-reference replenishment scheme for pre-encoded JPEG 2000 content, and demonstrates the relevance of the approach in scenarios capturing the video

Part of this work has been funded by the FP6 IST-2003-507204 project WCAM [1].

F.-O. Devaux and C. De Vleeschouwer are funded by the Belgian NSF. They are with the Communications and Remote Sensing Laboratory (TELE), Université catholique de Louvain (UCL), Belgium. E-mail: {devaux,devlees}@tele.ucl.ac.be.

J. Meessen, C. Parisot and J.-F. Delaigle are with Multitel A.S.B.L, Belgium. E-mail: {meessen,parisot,delaigle}@multitel.be

[1]Even in cases for which a given content is only accessed by a few clients, the server is expected to handle a large number of contents simultaneously, thereby making computational load on the server an important issue

sequence with still cameras, as often encountered in a video surveillance context.
- promotes adaptive and user-driven access to video content by defining a JPEG 2000 scheduler that adapts to heterogeneous channel conditions and user requirements (in terms of spatio-temporal interest) at low computational cost and in a post-compression way, based on a set of pre-calculated annotations.
- circumvents the drawbacks of closed-loop prediction systems by restricting transmissions to INTRA content. This is especially relevant when addressing heterogeneous clients dealing with different prediction references in lossy environments.
- does not aim at competing with state-of-the-art hybrid video compression algorithms [7] [8]. Instead of compression efficiency, our proposed solution rather emphasizes the capabilities for spatio-temporal random access required for interactive navigation through the (individual) frames or segments of the video sequence.

The novelty of our proposed server mainly lies in (1) the exploitation of multiple references in a replenishment framework, (2) the RD optimal and semantically weighted scheduling of pre-computed JPEG 2000 packets, and (3) the efficient implementation of the scheduling algorithm, to handle numerous heterogeneous clients simultaneously.

The outline of this paper is the following. Section II presents an overview of the interactive browsing system. Section III details the scheduler implemented on the server side to select the JPEG 2000 packets that provide rate-distortion optimal replenishment, given the reference(s) expected on the client side. Section IV further explains how this system can be deployed to adapt to client resources and interest in the scene in a cost-effective way. Section V presents the background estimation algorithm used in our replenishment system. The integrated approach is validated through Section VI, and conclusions are drawn in Section VII.

II. Application scenario and system overview

A. Remote interactive browsing in a surveillance context

To motivate the use of JPEG 2000 to store and disseminate surveillance video content, it is interesting to consider a typical interactive browsing surveillance scenario, and to compare the channel and computational resources required when accessing pre-recorded content remotely either based on hybrid (INTER) or JPEG 2000 (INTRA) compression formats.

The envisioned surveillance scenario significantly extends the common VCR functionalities [9]. Typically, a graphical user interface (GUI) allows the human controller to visualize the chronology of recorded - and possibly pre-analyzed - events through a timeline of low-resolution key frames (scenario 1). The user can then select some time segments of the video to display at higher resolution (scenario 2). (S)he can also interactively select and further zoom in on some areas of interest, in a particular video segment (scenario 3) or frame (scenario 4) of the displayed scene. For illustrative purposes, Table I reviews the four access scenarios involved in the above described typical browsing session, manipulating content captured at 15 fps, with a still 2 Mpixels camera. The scenarios differentiate themselves by the spatial resolution at which they access the content, and by the particular segment of the video they actually access. In particular, scenario 1 envisions the display of a chronological time-line of very low resolution frames. Scenario 2 considers the display of a video segment at low resolution. Scenario 3 considers a cropped and subsampled version of the video, while scenario 4 considers the access to a 384x288 window in a randomly selected frame of the original video sequence.

	Scenario	Encoded signal signal resolution	Displayed fraction of initial image
1	Time-line of very low-resolution frames	192×144	$1/1$
2	Low-resolution video segment	384×288	$1/1$
3	Zoom in (spatially) random video segment	768×566	$1/4$
4	Zoom$^+$ in (spatio-temporally) random frame segment	1536×1132	$1/16$

TABLE I
CONTENT ACCESS SCENARIOS DEFINITION. CONTENT HAS BEEN CAPTURED AT 15 FPS, WITH A 2 MPIXELS CAMERA.

Scenario	J2K	**Proposed CRB**	AVC I + 14 P	AVC All I	SVC	AVC FMO I + 14P
1 [kbit/ sample]	24	**24**	20	20	20	20
2 [kbit/ sec]	1020	**189**	78	840	93	78
3 [kbit/ sec]	702	**148**	215	2190	251	101
4 [kbit/ sample]	32	**32**	494	415	537	57

TABLE II
AVERAGE BANDWIDTH CONSUMPTION FOR EACH ACCESS SCENARIO AND FOR DISTINCT ENCODING SCHEMES, AT $35dB$. FOR THE J2K AND CRB METHODS, A SINGLE FINE-GRAINED CODESTREAM IS GENERATED FOR THE FOUR SCENARIOS AND COULD BE USED TO MEET OTHER RATE CONSTRAINTS. SVC AND AVC STREAMS ARE GENERATED TO TARGET THE FOUR PRE-DEFINED SCENARIOS, AND DIFFERENT VERSIONS OF THE AVC STREAM ARE GENERATED FOR EACH SCENARIO WHILE SVC ONLY REQUIRES A SINGLE STREAM.

For each scenario, Table II then compares the average bitrate required to access a typical surveillance content based on distinct codecs. For each coding scheme and each spatial resolution, the encoding parameters have been tuned to reach an approximate PSNR of 35 dB, so that the displayed signals are roughly comparable for a given scenario, but distinct encoding schemes.

In our analysis, we first consider the four codecs corresponding to the first four columns of Table II. J2K encodes and decodes the video images based on the JPEG 2000 algorithm. CRB refers to the original solution described and validated in the rest of the paper. It relies on JPEG 2000 packets, but implements multiple-reference and RD optimized conditional replenishment mechanisms to reduce the bandwidth consumption when accessing video segments characterized by still

backgrounds. The two next solutions build on the H.264/AVC standard, and encode one INTRA frame every second (column 3) or all frames in INTRA (column 4). For both AVC solutions, four distinct streams are generated, corresponding to the four spatial resolutions considered by the scenarios in Table I. The two last solutions respectively built on SVC and AVC FMO are detailed below.

In Table II, the bandwidth is defined in kbits/sample or kbits/sec depending on whether the scenario considers the access to an individual frame or to a (several seconds) video segment. As AVC is not supposed to provide spatio-temporal random access capabilities, we assume that entire frames have to be decoded to access the frame/video segment of interest in scenarios 3 and 4. Moreover, partial GOPs have to be decoded to access a single and randomly selected frame with AVC I+14P in scenario 4. Hence, depending on the position of the frame to access in the GOP, a number of P frames have to be decoded in addition to the first Intra frame of the GOP. This explains why the cost to access a sample in scenario 4 is higher for AVC I+14P than for AVC I.

A careful analysis of the first four columns of Table II reveals that the INTRA nature of JPEG 2000 strongly penalizes J2K compared to AVC (I+14P) when video segments have to be transmitted. It also reveals that J2K provides an attractive solution when random spatial and/or temporal access is desired (scenarios 1 and 3) or when a single frame has to be displayed (scenario 1 and 4). The lack of spatio(-temporal) random access capabilities significantly penalize AVC-based solutions compared to J2K and CRB solutions in scenarios 3 and 4. Interestingly, we observe that our proposed CRB solution preserves the advantages of J2K, while smoothing out its main drawback. Specifically, CRB appears to be the only solution that is able to deal with all scenarios with a bandwidth of 200 kbps and a latency smaller than one second for scenario 4. This definitely demonstrates the relevance of our study that, we should remind it, relies on the stationarity of the scene background, and is thus specially suited to surveillance contexts.

Before moving to the actual description of our CRB solution, it is worth making two comments about AVC-based video coding schemes.

First, the scalable extension of MPEG-4 AVC, namely SVC [10], enables the encoding of a high-quality video bitstream that contains one or more subset bitstreams that can themselves be decoded with a complexity and reconstruction quality similar to that achieved using MPEG-4 AVC with the same quantity of data as in the subset bitstream. Hence, SVC prevents the multiplication of streams, but does not fundamentally affect the conclusions drawn from Table II. This is illustrated by column 5 in Table II. There we present a SVC solution for which the first resolution has been encoded based on a I + 14 P GOP structure. For the second and third resolutions, frames are predicted based on the highest lower resolution and the previous frame. To improve random access capabilities, the last and finest resolution only exploits the lower resolution as a reference (and not the previous frame). We observe in Table II that SVC achieves about the same performance as the four versions envisioned for AVC I + 14P. This is not surprising since SVC encounters some (minor) penalty when embedding the four versions in a single bitstream. Beyond that example, it is also worth mentioning that the medium-grained scalable (MGS) version of SVC supports rather fine "on the fly" quality adaptation on entire frame, based on a smart design of temporal prediction loops, and on the frequency-based partitioning of enhancement coefficients. However, such MGS setting does not support neither region of interest based transmission, neither random temporal or multi-resolution access to video.

Second, it is possible to exploit the flexible macroblock ordering concept of MPEG-4 AVC to define a grid of block-shaped slices that can be accessed independently, thereby improving the spatially random access capabilities of AVC, at the expense of some coding efficiency[2]. Column 6 in Table II presents the bandwidth requirements corresponding to the four envisioned scenarios when the AVC I+14P codec considers independent slices of 64×64 pixels, thereby significantly improving the bandwidth requirement when random spatial access is required (for scenarios 3 and 4).

Bottom line, we conclude that, for a pre-defined set of access scenarios characterized by a given set of targeted resolutions or bit budgets, results equivalent or even slightly better than the one obtained with J2K could be obtained with MPEG-4 AVC or SVC standards for the fourth scenario, by encoding high resolution frames in INTRA (to allow for random temporal access) and based on a set of independent slices. However, despite this observation, JPEG 2000-based solutions still remain attractive due to their inherent fine grained embedded nature, which gives them the ability to deal with heterogeneous bandwidth constraints and RoI user requests. With JPEG 2000 or the proposed replenishment framework, there is no need to work with sophisticated decoder architectures, able to handle a discrete set of (embedded) versions of the same content, encoded with a discrete set of distinct quality and resolution levels. With replenishment-based solutions, the client simply handles and decodes conventional JPEG 2000 and parity packets to browse arbitrary portions of the content in a progressive and fine grained manner, both in quality and resolution. Such progressivity is especially desired when serving heterogeneous terminals, for which transmission resources and interest in the scene are defined by each individual user at transmission time.

A summary of this comparison is presented in Table III. There, we differentiate the "fixed scalability" and the "adaptive scalability". By "adaptive scalability", we refer to the dynamic adaptation of transmitted content according to the requirements defined by each individual client at transmission time. In contrast, the "fixed scalability" refers to the preparation and exploitation of (embedded) codestreams(s) dedicated to a discrete set of pre-defined transmission conditions. The ambiguous characterization of the SVC codec in Table III reflects the fact that the compression efficiency and adaptive

[2]For example, [11] considers a low resolution base layer encoded with motion compensation, and a high resolution enhancement layer encoded in a set of independent slices that are only predicted from the base layer.

scalability of SVC depends on the envisioned scenarios and parametrization of the codec. More interestingly, based on Table III, we conclude that J2K and proposed replenishment-based methods outperforms other approaches in terms of "adaptive scalability".

Codec	Compression efficiency	Fixed scalability	Adaptive scalability
AVC	High	Low	Low
SVC and/or FMO	Medium-High	High	Low-Medium
J2K	Low	High	High
Proposed CRB	**Medium**	**High**	**High**

TABLE III
SUMMARY OF CODECS COMPARISON. FIXED SCALABILITY REFERS TO THE FLEXIBILITY RESULTING FROM THE CREATION OF SEVERAL EMBEDDED VERSIONS OF A COMPRESSED CONTENT IN A SINGLE PRE-ENCODED CODESTREAM. ADAPTIVE SCALABILITY REFERS TO THE ADDITIONNAL FLEXIBILITY ARISING WHEN THE TRANSMITTED CONTENT CAN BE ADAPTED ON-LINE TO ROI USER REQUIREMENTS AND CHANNEL CONDITIONS.

Hence, the core of the paper mainly consists in explaining and demonstrating how dedicated conditional replenishment mechanisms efficiently preserve the fine grained flexible nature of JPEG 2000 to adapt streamed content to individual user needs while saving some bit budget when serving surveillance video segments, thereby reaching the performance presented in the CRB column of Table II.

B. Proposed video server overview

As explained in the previous section, the purpose of our paper is to explore how JPEG 2000 can support the efficient transmission of video sequences. As a still image compression standard, JPEG 2000 encodes the video frames independently, and does not exploit the potential temporal correlation existing between consecutive frames. The approach makes the access to each individual image direct and flexible, but penalizes the costs associated to the transmission of an entire video sequence. To mitigate this drawback, we propose to adopt a rate-distortion formalism so as to restrict the transmission of each image to the data units that bring a sufficient benefit per unit of transmission cost.

Our approach follows the conditional replenishment principle [2] [3] in that only the parts of the current image that significantly differ from a reference maintained at the receiver are transmitted. However, our work extends the initial replenishment scheme in three major aspects, which correspond to the three novel contributions of our paper:

- First, it exploits the specificities of the JPEG 2000 standard in that, for a given bit budget, it balances the size (in terms of code-blocks) and the accuracy (in terms of bit-planes) of the replenishment in a rate-distortion optimal way.
- Second, it proposes to maintain two reference images at the receiver. Next to the last reconstructed frame, as proposed in [3], our system maintains an estimate of the scene background as a second reference. Hence, for each frame, the system only transmits the JPEG 2000 data units that are not properly approximated at the receiver, neither based on the background estimate, nor based on the previous reconstructed frame. In our experiments, the background is estimated based on Gaussian mixtures that collect the statistics of past image samples in specific pixel locations, as described in Section V. When the current background estimate sufficiently differs from the reference background available at the client, the current background is transmitted to the receiver, and the reference background is updated. The simulations presented in Section VI demonstrate that this second reference significantly decreases the required transmission resources in case of stationary scene background, as encountered in video surveillance contexts.
- Third, as an additional original and crucial contribution, our study also demonstrates that most of the computation needed to take the replenishment decisions can be performed off-line, without preventing the server to adapt its replenishment (scheduling) decisions to the actual transmission resources and semantic interest defined on-line by a particular user during the browsing session. In practice, all these pre-computed informations are gathered in a file, named *Rate-Distortion* (RD) *Index file* in the following. The above statement has important and interesting practical consequences. In particular, it means that a single index file is pre-computed and exploited to cover the multiple scenarios considered in Table II, each scenario being considered as a particular interest expressed by the user. It also means that our proposed server naturally adapts to fluctuating and heterogeneous bandwidth conditions.

The proposed video server is depicted in Figure 1. Thanks to the information gathered in the pre-computed index file, the server is able to schedule pre-encoded JPEG 2000 packets to address the needs and resources of each individual client. On-line processing driving adaptive scheduling decisions only implies minor computational cost.

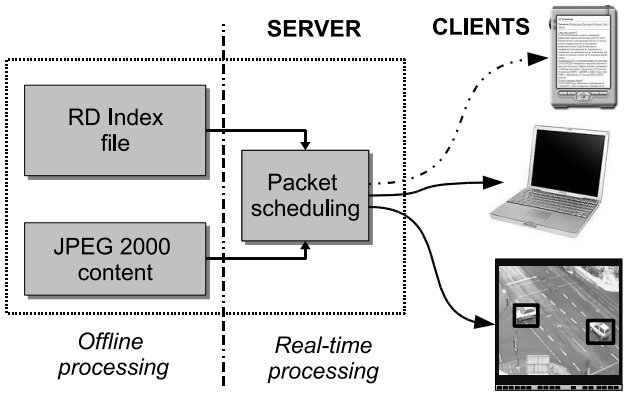

Fig. 1. Proposed video server architecture. Using a pre-calculated index, the server selects the optimal packets to transmit to each client, based on their individual needs and resources. Three types of clients are illustrated: a PDA client with a low resolution and low bandwidth, a laptop client with a high bandwidth and a third client focusing on two regions of interest for which he is expecting a high quality.

III. JPEG 2000 CONDITIONAL REPLENISHEMENT

In this section, we first review the JPEG 2000 standard. Then, we explain how conditional replenishment is implemented to transmit a JPEG 2000 frame in a rate-distortion optimal way, when a given reference is known to be available at the receiver. Finally, we define three methods for efficient streaming of a video segment based on JPEG 2000 packets replenishment. These three schemes differ by their ability to exploit the background estimate as a replenishment reference and to support the prioritized transmission of regions of images that are of particular interest to the user.

A. JPEG 2000 image representation and codestream abstraction

The JPEG 2000 standard describes images in terms of their discrete wavelet coefficients. An important question raised by conditional replenishment of JPEG 2000 coefficients is related to the granularity of refreshment of those coefficients. Specifically, one needs to understand to which extent it is possible to define the resolution, the subband, the position and the reconstruction accuracy of the coefficients that are refreshed. That issue is directly related to the JPEG 2000 format, which can be summarized as follows.

According to the JPEG 2000 standard, the subbands issued from the wavelet transform are partitioned into *code-blocks* that are coded independently [4] [5] [12]. Each code-block is coded into an embedded bitstream, i.e. into a stream that provides a representation that is (close-to-)optimal in the rate-distortion sense when truncated to any desired length. To achieve rate-distortion (RD) optimal scalability at the image level, the embedded bitstream of each code-block is partitioned into a sequence of increments based on a set of truncating points that correspond to the various rate-distortion trade-offs [13] defined by a set of Lagrange multipliers. A Lagrange multiplier λ translates a cost in bytes in terms of distortion. It defines the relative importance of rate and distortion. Given λ, the RD optimal truncation of a code-block bitstream is obtained by truncating the embedded bitstream so as to minimize the Lagrangian cost function $\mathcal{L}(\lambda) = D(R)+\lambda R$, where $D(R)$ denotes the distortion resulting from the truncation to R bytes. Different Lagrange multipliers define different rate-distortion trade-offs, which in turn result in different truncation points. For each code-block, a decreasing sequence of Lagrange multipliers $\{\lambda_q\}_{q>0}$ identifies an ordered set of truncation points that partition the code-block bitstream into a sequence of incremental contributions [13]. Incremental contributions from the set of image code-blocks are then collected into so-called quality layers, $\mathcal{Q}_q$. The targeted rate-distortion trade-offs during the truncation are the same for all the code-blocks. Consequently, for any quality layer index l, the contributions provided by layers $\mathcal{Q}_1$ through $\mathcal{Q}_l$ constitute a rate-distortion optimal representation of the entire image. It thus provides distortion scalability at the image level. Resolution scalability and spatial random access to the image result from the fact that each code-block is associated to a specific subband and to a limited spatial region.

Although they are coded independently, code-blocks are not identified explicitly within a JPEG 2000 codestream. Instead, the code-blocks associated to a given resolution are grouped into *precincts*, based on their spatial location [4], [14]. Hence, a precinct corresponds to the parts of the JPEG 2000 codestream that are specific to a given resolution and spatial location. As a consequence of the quality layering defined above, a precinct can also be viewed as a hierarchy of *packets*, each packet collecting the parts of the codestream that correspond to a given quality among all code-blocks matching the precinct resolution and position. Hence, packets are the basic access unit in the JPEG 2000 codestream.

B. Rate-distortion optimal replenishment of a JPEG 2000 frame

The conditional replenishment framework originally introduced in [2] and exploited more recently in multicast transmissions [3] has to be adapted to JPEG 2000 features.

Given a targeted transmission budget and a reference image available at the receiver, we now explain how to select the JPEG 2000 packets of the current image codestream so as to maximize the reconstructed image quality. As the JPEG 2000 codestream consists in a set of precincts organized in a hierarchy of layers (see Section III-A), the problem consists in selecting the indexes of the precincts to refresh and their quality of refreshment, so as to maximize the reconstructed quality (or minimize the distortion) under the bit budget constraint, knowing that non-refreshed precincts are approximated based on the wavelet coefficients of the reference image.

To simplify notations, and without loss of generality, the precincts are labeled by a single index i. To solve the problem efficiently, we assume an additive distortion metric, for which the contribution provided by multiple precincts to the entire image distortion is equal to the sum of the distortion computed for each individual precinct. We define $d^q(i)$ and $d^{ref}(i)$ to denote the distortion computed when the i^{th} precinct of the image to transmit is approximated based on its q first layers and based on the reference image, respectively. We also denote $s^q(i)$ to be the size in bytes of the q first packets of the i^{th} precinct and T is the available bit budget.

The problem of RD optimal allocation of a bit budget across a set of image blocks (precincts in our case) characterized by a discrete set of RD trade-offs has been extensively studied in the literature [15]–[17]. Under strict bit budget constraints, the problem is hard and its solution relies on heuristic methods or dynamic programming approaches [17] [18] [19]. In contrast, when some relaxation of the rate constraint is allowed[3], Lagrangian optimization and convex-hull approximation can be considered to split the global optimization problem in a set of simple block-based local decision problems [15], [16]. The convex-hull approximation consists in restricting the eligible transmission options for each block (or precinct) to the RD

[3]This is the case in a streaming context since buffers absorb momentary rate fluctuations. Hence, the bits that are saved (overspent) on the current frame just slightly increment (decrement) the budget allocated to subsequent frames, without really impairing the global performance of the communication.

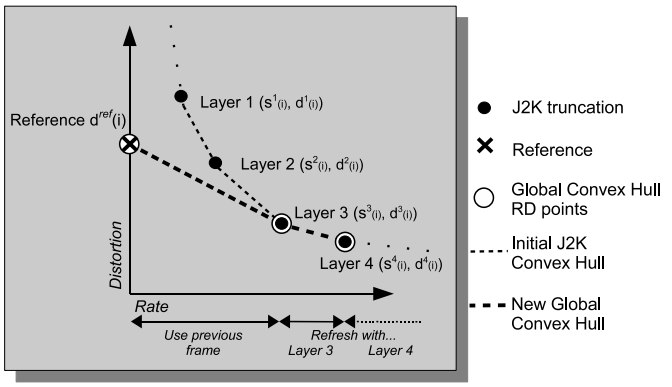

Fig. 2. Rate-Distortion points for a given precinct. The figure presents both the set of RD points corresponding to the JPEG 2000 codestream truncation points (*dots*), and the additional RD point corresponding to the approximation of the precinct by the reference image (*cross*). The set of RD trade-offs provided by the JPEG 2000 codestream lie on a convex-hul. The novel trade-off authorized by the reference image lie on the distortion axis ($R = 0$). Global convex-hull (circles) has to be considered for RD optimal allocation at image level.

points sustaining the lower convex hull of the available RD points of the block. In our case, this corresponds to the computation, for each precinct, of the convex-hull sustaining both the JPEG 2000 and the reference RD points, as depicted in Figure 2.

Hence, given a bit-budget and the set of accessible convex-hull RD points for each precinct, overall RD optimality is achieved at the image level by transmitting the packets corresponding to the convex-hull RD points selected in decreasing order of benefit per unit of rate, up to exhaustion of the transmission budget [15]. The approach is detailed in [20]. It is equivalent in principle to the one defined in [14], but accounts for the availability of a reference image by pre-computing for each precinct the convex-hull sustaining all accessible RD points.

The solution is RD optimal in the sense that, for the achieved bit-budget, it is not possible to attain a lower reconstructed image distortion based on different refreshment decisions. This is because, by construction, it is not possible to find a non-transmitted packet that provides a larger gain per unit of rate than the gain provided by a transmitted packet.

C. Video segment replenishment methods

In this section, we introduce three different replenishment mechanisms to stream a video segment. They all follow the algorithm defined in Section III-B, but differ in the reference they use for replenishment, or in the way they compute the distortion associated to a precinct.

In a video streaming context, we consider the transmission of frame t, and denote $d_t^{k,q}(i)$ to be the distortion measured when approximating the i^{th} precinct of frame t, based on the q first layers of the corresponding precinct in frame $(t-k)$. In particular, the replenishment of precinct i at time t with q layers is denoted by $d_t^q(i) = d_t^{0,q}(i)$. In absence of replenishment, the reference distortion for precinct i at time t is denoted $d_t^{ref}(i)$. The size in bytes of the first q JPEG 2000 packets of precinct i of frame t is noted $s_t^q(i)$, for each $q \in \mathcal{Q}$.

We now introduce the three replenishment methods considered in the simulation results presented in Section VI. They are denoted and defined as follows:

The **CR** – Conditional Replenishment – method follows the conventional replenishment mechanism originally introduced in [3] and adapted to the wavelet domain. The reference image is the previously reconstructed image, and the MSE is considered to measure the distortion. Formally, when the latest replenishment of precinct i occurred k_t^i frames earlier than t with q_t^i, the reference distortion $d_t^{ref}(i)$ for precinct i at time t is equal to $d_t^{k_t^i,q_t^i}(i)$. Here, all distortion metrics are computed based on the Square Error (SE) of wavelet coefficients, so as to approximate the reconstructed image square error [13]. Formally, let $\mathcal{B}_i$ denote the set of code-blocks associated to precinct i, and let $c_b[n]$ and $\hat{c}_b[n]$ respectively denote the two-dimensional sequences of original and approximated subband samples in code-block $b \in \mathcal{B}_i$. The distortion $d(i)$ associated to the approximation of the i^{th} precinct by $\hat{c}_b[n]$ coefficients is then defined by

$$d(i) = \sum_{b \in \mathcal{B}_i} w_\sigma^2 \sum_{n \in b} (\hat{c}_b[n] - c_b[n])^2 \qquad (1)$$

where w_σ denotes the L2-norm of the wavelet basis functions for the subband σ to which code-block b belongs [13]. Equation (1) is adopted to compute $d_t^{k,q}(i)$ and $d_t^{ref}(i)$ for all i, k, q and t.

The **CRB** – Conditional Replenishment with Background – method is novel and proposes to consider both the previous image and the estimated background as possible references for each precinct. In practice, for a given precinct, the image that best approximates the precinct is selected as the reference for that specific precinct. Hence, letting $b_t(i)$ to denote the distortion obtained when approximating the i^{th} precinct of frame t based on the latest version of the background, the reference distortion is now defined by $d_t^{ref}(i) = min[d_t^{k_t^i,q_t^i}(i), b_t(i)]$. As for the CR method, the distortion measures the SE of wavelet coefficients. Our simulations demonstrate that CRB significantly outperforms CR in the surveillance scenario.

The **CROI** – Conditional Replenishment with Regions of Interest – follows the mechanism introduced by CRB, but defines the distortion based on a weighted SE of wavelet coefficients, so as to take into account the knowledge the server may have about the semantic significance of approximation errors. We assume that the information about the semantic relevance of approximation errors is provided at the precinct level based on user feedback or on some kind of automatic pre-analysis of the scene. We define the semantically weighted distortion to be $d_{RoI,t}(i) = w_t(i) d_t(i)$, where $w_t(i)$ denotes the semantic weight assigned to the i^{th} precinct at time t. From a functional point of view, the CROI approach provides a mechanism to take the user needs and interest into account to define replenishment decisions. It is worth noting that the convex-hull analysis performed on non-weighted distortions (see Section III-B) remains valid, as long as the weighting affects in a similar way all the packets of a precinct, which is the case if weights are defined at the precinct level. Hence, the complexity of CROI is equivalent to the one of CRB,

independently of the interest (weights) defined by the user. This is a key difference with most earlier contributions that have considered semantically meaningful weighted distortion metrics in the past, e.g. in [21]. Most earlier contributions exploit those metrics either before or during the encoding step. In contrast, our work supports the posterior definition of semantics weights, at transmission time for each client, thereby allowing to serve multiple clients, with different semantic interests, based on a single JPEG 2000 codestream, and without any significant complexity increase.

IV. Serving multiple heterogeneous clients: Index file definition

In this section, we consider the practical deployment of the replenishment system described in Section III, to serve a large number of heterogeneous clients while preserving an acceptable computational complexity. The objective is thus to support low cost adaptation to user requests and resources. When the server has to cope with a large number of clients, possibly accessing distinct streams, the real-time calculation of the $d_t^q(i)$ and $d_t^{ref}(i)$ values of interest becomes computationally intractable. In order to decrease this complexity, we propose to separate the process in two phases. During an off-line phase, the server performs once and for all most of the computationally expensive operations, and stores the results in an index. This index is then exploited for on-line adaptive scheduling of packets, based on the actual resources and interest of a particular client.

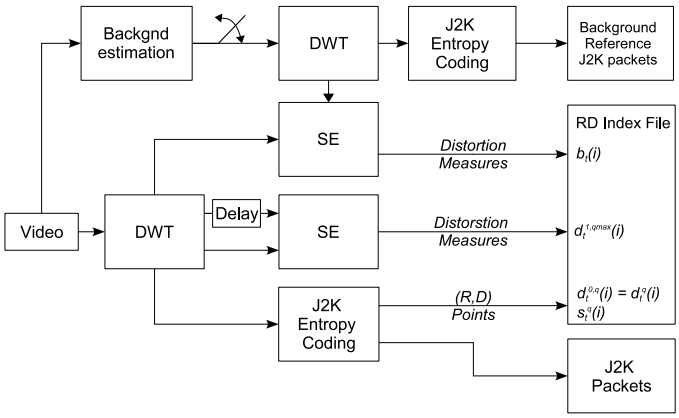

Fig. 3. Off-line operations leading to the creation of the RD index file at the server.

Based on the conditional replenishment concepts presented in Section III, the information that is needed by the server to take optimal scheduling decisions is strictly limited to the measures of JPEG 2000 packet sizes and both types of distortions $d_t^{k,q}(i)$ and $b_t(i)$. This information only depends on the input signal, and not on earlier and user-dependent replenishment decisions. Hence, they can be computed off-line[4], and stored in an index file that will then be used during the streaming session to adapt the replenishment decisions to each individual client. The process is illustrated in Figure 3, which can be described as follows.

The scene background is estimated based on the original video content, typically based on Gaussian mixtures, as detailed in Section V. Each background estimate that is transmitted to the client is then JPEG 2000 encoded, and used to compute the distortion values $b_t(i)$. Distortion values $d_t^q(i)$, also denoted $d_t^{0,q}(i)$, and $s_t^q(i)$ values are directly computed based on JPEG 2000 encoding of each individual frame of the original video. In contrast, the computation of the $d_t^{k,q}(i)$ values, for all $k > 0$, implies a significantly larger effort, both in terms of computation and memory resources. To mitigate this effort, we propose to use the following approximation:

$$d_t^{k,q}(i) \cong d_{t-k}^{0,q}(i) + \sum_{l=0}^{k-1} d_{t-l}^{1,q_{max}}(i) \qquad (2)$$

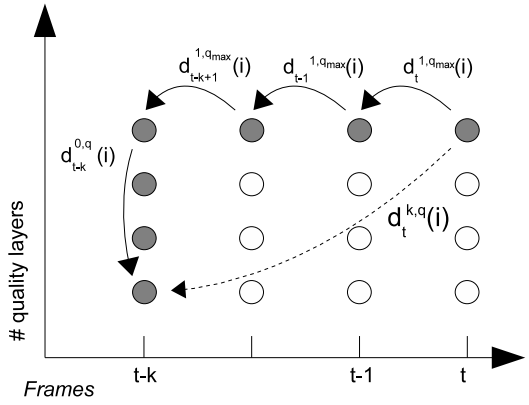

Fig. 4. Path used to approximate the distortion of the previous references, compared to the optimal path (dashed arrow). This approximation significantly decreases the pre-processing complexity and storage requirements, without significantly impairing the streaming performance.

In this equation, $d_{t-k}^{0,q}(i)$ denotes the distortion resulting from approximating the i^{th} precinct of frame $(t-k)$ based on q layers, with $0 \leq q < q_{max}$. The second term accounts for the fact that we are interested in the distortion measured when approximating the i^{th} precinct of frame t based on the first q layers in frame $(t-k)$. Therefore, we have to estimate how well the i^{th} precinct of frame $(t-k)$ approximates the corresponding precinct in frame t. By assuming that the errors on precinct coefficients are zero mean, we can interpret the SE distortion as a measure of the variance of the random variable associated to errors. In addition, if we admit that errors between consecutive pairs of frames are independent, then we can simply estimate the SE between the t^{th} and $(t-k)^{th}$ frames based on the sum of the SE measured between all pairs of consecutive frames between $(t-k)$ and k, ending in the second term of Equation 2. The approximation process is illustrated by Figure 4, which depicts the hierarchy of layers associated to frames indexed from t to $t-k$. We observe that any $d_t^{k,q}(i)$ can be approximated based on a distortion computation path that only relies on $d_X^{0,q}(i)$ and

[4]Off-line means here that computations are performed independently of the actual semantic weights $w_t(i)$ or transmission resources experienced by a particular user.

$d_Y^{1,q_{max}}(i)$ values, which significantly reduces the amount of values to compute and store in the index file, compared to $d_X^{Y,Q}(i)$, where X and Y variables take all possible values, and $0 < Q < q_{max}$.

We will see in the next section that this approximation does not have a significant impact on the system performances. To estimate the complexity, we define I_d to be the depth of the index file. In other words, I_d defines the number of previous frames considered for the calculation of the previous reference distortion. Hence, $d_t^{k,q}(i)$ is computed for all $k < I_d$, which shows that the complexity increases linearly with the index depth I_d for the optimal algorithm. By using the approximation, we make the number of calculations independent of the value of I_d, and turn the $O(I_d * q_{max})$ complexity into $O(I_d + q_{max})$.

The gain is further illustrated in Figure 5, which plots the memory and computational complexity requirements for three different implementations of our proposed replenishment framework within a video server. The sequence considered to plot this figure is the *Speedway* CIF sequence. Details regarding this sequence and the compression parameters considered are provided in Section VI.

The *Optimal Online* strategy computes the reference distortion required for the rate allocation of each individual user at transmission time, knowing the exact scheduling history of the user. The *Optimal Offline* strategy computes and stores all possible reference distortions off-line, without any approximation, by anticipating all the precinct references resulting from possible earlier transmission strategies. The *Proposed approximations* strategy only pre-computes the distortion resulting from the approximation of a precinct based on its previous correspondence, encoded at the highest quality level, i.e. it only computes $d_t^{1,q_{max}}(i)$ for all i and t. It then uses Equation 2 to estimate the missing reference distortions $d_t^{k,q}(i)$, with $k \neq 1$ and $q \neq q_{max}$.

In Figure 5(a), the memory requirements are determined by summing the number of bytes required to store each JPEG 2000 packet rate and distortion. For the *Optimal Offline* solution, the distortion of all possible references must also be stored, while this information is calculated on the fly for the *Optimal Online* and approximated for the third method. The number of JPEG 2000 packets in a frame, which directly influences the memory requirements, is determined by the sequence compression parameters[5]. As expected, we observe that the memory required by the *Optimal Offline* strategy increases linearly with I_d, while it remains constant for the other strategies.

In Figure 5(b) the server complexity is measured in terms of the numbers of arithmetic operations required for the computation of the distortion. I_d is set to 10. We observe that the complexity of the *Optimal Online* strategy increases significantly with the number of users, since the computation are performed independently for each client. For the other strategies, most operations are performed offline, and the additional required online operations appear to be insignificant compared to the offline operations. This is reflected by nearly horizontal curves for these strategies in Figure 5(b).

[5]The number of JPEG 2000 packets in a single tiled frame corresponds to the product of the number of resolutions, layers, precincts and components.

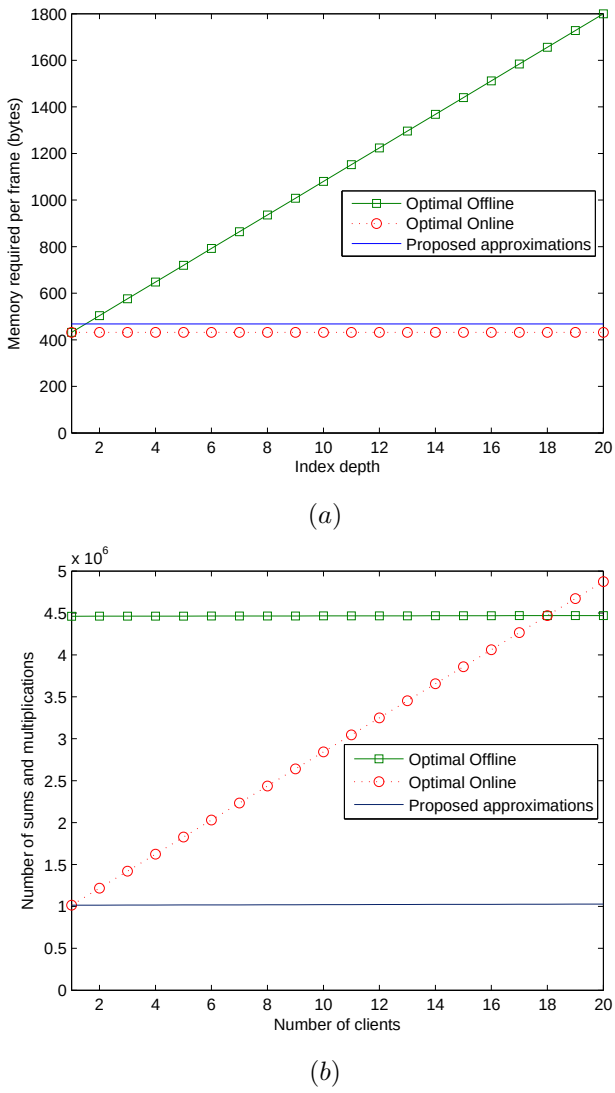

Fig. 5. Comparison of (a) memory and (b) computational resources for three different implementations of the proposed replenishment framework. Memory is depicted as a function of the index depth I_d, while computational resources are presented as a function of the number of server clients.

V. Background Estimation

The goal of the background estimation process is to create the additional reference frames for the replenishment module. Each background pixel is estimated on a sliding window, based on a mixture of Gaussians model [22] [23] [24]. This approach automatically supports backgrounds characterized by multiple states, like blinking lights, grass and trees moving in the wind, acquisition noise, etc. Furthermore, it naturally updates the background model -in an unsupervised manner- when the scene conditions are changing.

Figure 6 shows the mixture of Gaussians model for a pixel at some given time. Computation of that model is based on the aggregation of all luminance values observed for that specific pixel in the previous frames belonging to the sliding

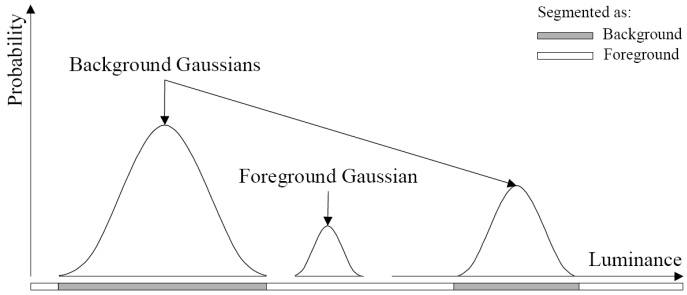

Fig. 6. Statistical modeling of a background pixel using three Gaussians. Multiple Gaussians aggregate the pixel luminance values observed in a sliding window.

window. The model is updated as follows, each time a novel luminance value is observed for the pixel. The novel luminance observation is compared to the current mixture. The pixel is assigned to one of the Gaussians if the distance between the its luminance and the Gaussian mean is lower than a given threshold, chosen to be proportional to the Gaussian standard deviation -typically 1.6 times the standard deviation. If the pixel belongs to one of the most probable Gaussians, the pixel is classified as background and the relevant Gaussian parameters, i.e. mean, variance, and frequency, are updated. Otherwise, the pixel is classified as foreground and the parameters of the associated Gaussian are updated according to this additional luminance value. At the beginning of the process, a new Gaussian is initialized each time a pixel is classified as foreground until the pre-defined maximum number of Gaussians is reached. The maximum number of Gaussians is a parameter that should theoretically be adapted to the number of different states a pixel of the background can have according to the different noises (acquisition, vibrations, etc.). In practice three Gaussians per mixture perform well in most indoor and outdoor conditions.

At any time, the background can thus be estimated based on the mean of the most probable Gaussian for each pixel. In our replenishment system, those background frames are used to update the reference background at the client when major background changes are detected. Note that at the very beginning of the sequence, typically during the first 2 seconds, the background estimate is unstable since the number of samples to model each Gaussian is very small. In order to avoid prohibitive background updates during this period and because our system deals with pre-encoded content, the initial background reference available from the beginning of the sequences is the stable estimate obtained after a few seconds of Gaussian mixture processing. In a real-time transmission context, the first frame would be considered as being the best reference until the Gaussian mixtures can be considered as stable.

VI. Experimental validation

In this section, we first analyze the performances of the proposed replenishment method when serving a single pre-encoded content at multiple rates. For comparison purposes, we provide the compression performance achieved by MPEG-4 AVC and SVC at similar bitrates. We then illustrate how the transmission of a video segment can be adapted to favor semantically relevant areas of the content. In practice, those regions of interest can be defined either automatically based on scene analysis mechanisms, or interactively by each individual user. The remarkable feature of our system lies in the fact that the adaptation of forwarded content to user needs and resources is performed at low computational cost by the scheduler, without the need to generate and manipulate multiple encoded versions of the same content. Finally, we briefly consider the behavior of our proposed scheme in presence of transmission errors.

Our approach has been tested exhaustively, but we present the results on *Speedway* and *CAVIAR* video sequences. Those sequences correspond to CIF video-surveillance sequence, captured with a fixed camera at 25 fps, and are available on the WCAM and CAVIAR projects websites [1], [25]. Regarding the JPEG 2000 compression parameters, each sequence has been encoded with four quality layers (corresponding to compression ratios of 2.7, 13.5, 37 and 76) and with three code-blocks per precinct (one in each subband). Precinct sizes have been set to 64x64, 32x32, and 16x16 for the four remaining lowest resolutions. In all simulations, the reference background ($\sim$ 50 kbytes) is sent only once at the beginning of the transmission because it remains sufficiently constant during the 8 seconds corresponding to the whole sequence duration.

A. Bandwidth usage efficiency

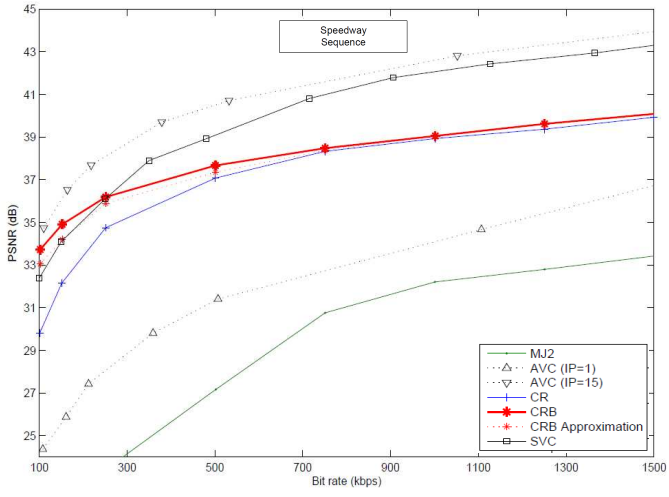

Fig. 7. Rate distortion curves of the proposed algorithms compared with MJ2, AVC and SVC MGS for the *Speedway* sequence. Frame rates and encoding parameters are defined in the text.

Figures 7 and 8 present the rate distortion curves of the proposed CRB system for the *Speedway* and *CAVIAR* sequence respectively. We will focus on the latter one for the following analysis. Two curves correspond to CRB: the optimal algorithm and the suboptimal algorithm using the approximation described in Section IV. The system is compared to MJ2, CR, and SVC which respectively correspond to independent

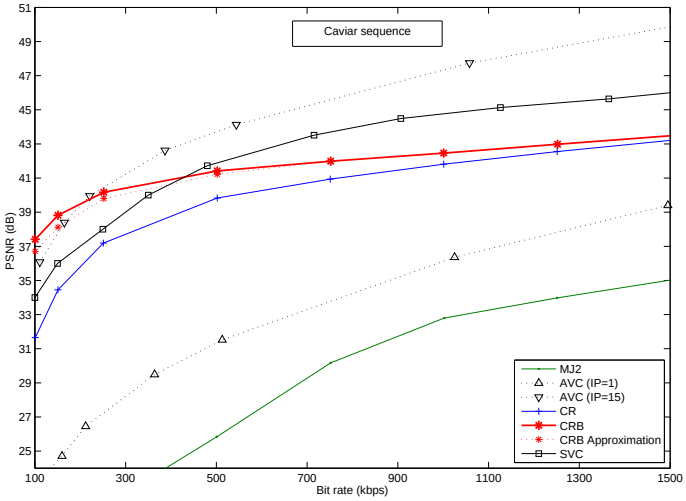

Fig. 8. Rate distortion curves of the proposed algorithms compared with MJ2, AVC and SVC MGS for the *Caviar* sequence. Frame rates and encoding parameters are defined in the text.

transmission of JPEG 2000 frames, to conventional conditional replenishment (i.e. without background reference) and to the scalable extension of MPEG-4 AVC. For completeness, the graph also plots MPEG-4 AVC with two different Intra Periods (IP). It is worth noting however that AVC relies on distinct encoded versions to address each target bitrate, while JPEG 2000 and SVC based solutions address multiple rate constraints based on a single pre-encoded codestream. Regarding the rate control, the bit-rate has been uniformly distributed on all frames for JPEG 2000-based methods. With AVC, we have adapted the quantization parameters to reach the expected bit-rates. The SVC solution is characterized by a GOP of 16 frames with medium-grain SNR scalable layers (MGS).

Unsurprisingly, MJ2 appears to be the worst scheme from a compression efficiency point of view. At very low bitrates, the CR method improves MJ2 by 15 dBs and CRB further improves CR by more than 5 dBs. The difference between CR and CRB tends to decrease with the bitrate. This is because at high bitrates, INTRA refresh can be performed with high quality, i.e. based on a large number of quality layers. As a result, the quality provided by the background approximation is not good enough compared to the accurate reconstruction quality offered by the INTRA refresh option, so that the background replenishment option is not selected anymore. Hence, the relative gain brought by the background approximation decreases as available bitbudget increases.

In Figures 7 and 8, we also observe that the suboptimal CRB approximation behaves very similarly to the optimal algorithm. At 200 kbps, the difference is smaller than 0.5 dB, and decreases as the target bitrate increases. This illustrates that the approximations described in Section IV to implement a computationally efficient video server do not alter significantly the proposed system.

Compared to MPEG-4 AVC and SVC, which do not offer an independent access to each frame nor the possibility to define RoI at transmission time, CRB results are very convincing, given the increased flexibility offered by a JPEG 2000-based low complexity server (see Section II-A). At 250 kbps, CRB PSNR is 13 dB above AVC IP-1, and comparable to SVC and AVC IP-15.

Surprisingly, we observe that the CRB curve is flatter than the AVC, SVC and MJ2 curves. This relative slower increases of quality with the bitrate does not reflect any sub-optimality regarding the way CRB uses the available bit budget. Rather, it can be understood by considering the relatively small increment of quality provided by the precinct that switch from a reference-based approximation to an INTRA refresh replenishment option when the available bitbudget increases. At low bitrates, those precincts were approximated based on the reference, at zero transmission cost. As a consequence, when they switch to an INTRA refresh replenishment mode, the increment in rate has to be compared to zero, while the increment in quality is computed with respect to the reference approximation, and not to the signal reconstructed in absence of any reference. In contrast, for MJ2, AVC and SVC schemes, an increment of quality always results from the refinement of already partially transmitted coefficients (finer quantization with AVC and SVC or additional layer with MJ2), and not to the complete transmission of the information needed to switch from a reference-based approximation to an actual transmission of JPEG 2000 coefficients.

Another way to apprehend the relative flatness of the CRB curve compared to MJ2, SVC or AVC consists in considering the evolution of quality when going from high to low bitrates. At extremely high bitrates, MJ2 and CRB provide the same quality (not depicted on the graphs). When the available bitbudget decreases, our CRB method manages to better preserve quality than J2K, by relying on the background approximation to save some bit budget (that can then be allocated to INTRA refreshed areas). Hence, the CRB curve drops more slowly than the MJ2 curve when going from high to low bit rates, thereby explaining the flatter trend of the CRB curve compared to MJ2, and in turns to AVC and SVC.

To figure out what the RD curves plotted in the two previous figures mean from a perceptual point of view, Figure 9 presents snapshots of the *Speedway* sequence compressed with the MJ2, CR, and CRB methods at 250 kbps. We observe that at very low bitrates, CR considerably improves the MJ2 method, but still remains blurry compared to CRB.

B. Semantically weighted adaptive streaming

We now consider two scenarios for which the server adapts its packet scheduling decisions to the specific interest expressed by the client about the scene content. In both scenarios, moving objects are considered to be more important than the scene background. In the first scenario, this knowledge is used to prioritize the refresh of moving objects. In the second scenario, the same knowledge is exploited to mitigate the impact of a noisy content acquisition process on the replenishment decisions. Bottom line, both scenarios illustrate the flexibility of the proposed CRB method, and its ability to integrate individual user needs at transmission time, based on a single JPEG 2000 pre-encoded codestream.

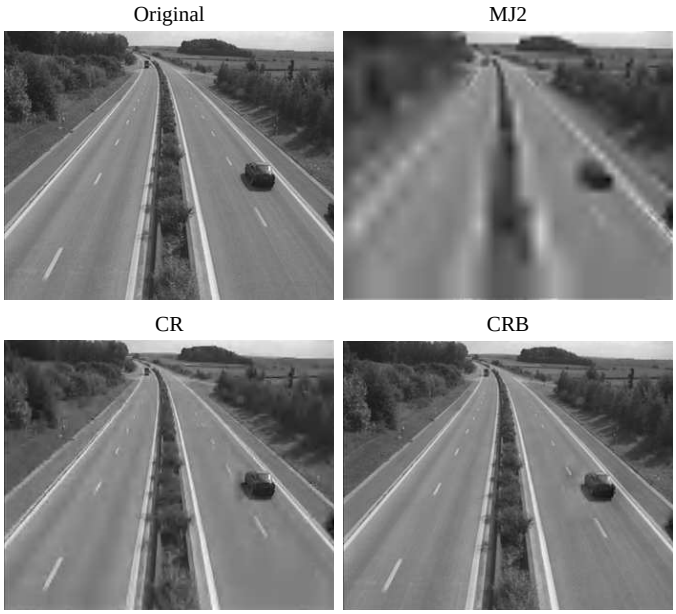

Fig. 9. MJ2, CR, and CRB methods for the 10th frame of the *Speedway* sequence transmitted at 250 kbps, 25 fps and in CIF format.

1) RoI-based streaming: In a video surveillance context, Regions of Interest are generally defined to be mobile objects, possibly matching pre-defined features (e.g. size, position, texture, etc.) or behaviors (e.g. people entering restricted areas). In our simulation, without loss of generality, we have considered that the user is interested in the moving vehicles of the *Speedway* sequence, so that the RoI segmentation mask is simply derived from the background estimation module, which inherently partitions the current image into moving foreground and still background regions (see Section V). Note that any other RoI definition could be envisioned, including a RoI defined interactively during the streaming process. This is because, as explained in Section III-C, the encoding process is totally independent of the RoI definition.

To maximize the impact of RoI prioritarization, the semantic weights $w(i)$ are set to one (zero) for precincts that belong to the RoI (non-RoI) areas[6]. The strategy is aggressive but defines a limit case that allows to get a clear idea about the potential benefit to draw from a semantic weighting of distortion.

Figure 10 presents the PSNR of RoI and non-RoI regions of *Speedway* for several JPEG 2000-based streaming methods. We observe that, for the MJ2 method, the non-RoI quality is always higher than the RoI because most of these background regions, like the road and the sky, are very efficiently compressed. Indeed, since these regions are quite predictable, the JPEG 2000 entropy coder easily reduces the number of bits used to code them compared to regions with a lower predictability. The RoI contains the cars that are characterized by a large amount of details, which are less efficiently compressed. Compared to MJ2, the CR method offers a higher quality for the RoIs, which correspond to the zones that are more often refreshed. This trend gets reinforced by the CROI method, which rapidly maximizes the RoI quality but maintains constant background quality. This is explained by the fact that the non-ROI areas are never refreshed by CROI, and are only defined using the background reference transmitted at the same high quality for all bitrates[7]. The CRB method behaves like CR at high bit rates, but offers a higher non-RoI quality at low bit rates, since the background reference can be used to increase non-RoI quality.

2) Noisy sequence: In this paragraph, we consider a noisy version of the *Speedway* sequence to further illustrate the flexibility of our proposed streaming server. Specifically, we show that our proposed method naturally support the exploitation of a priori knowledge about the relevance of approximation errors in the scene. In the scenario considered here, we have added white Gaussian noise with a standard deviation of 10 to the *Speedway* sequence. The noise simulates the effect of adverse surveillance conditions: noisy camera acquisition, bad weather, presence of traffic lights or moving objects (trees, ...). The noise causes luminance changes in the background regions, but these changes are not relevant with respect to the surveillance purpose of the application and should not trigger replenishment mechanisms. Hence, the approximation error observed on background areas should be neglected compared to errors measured in the foreground moving areas. In our simulation, this is simply done by using the CROI method, with distinct weights assigned to foreground and background precincts. Indeed, one characteristic of the segmentation algorithm presented in Section V is that the background Gaussians widths are automatically adapted to the sequence noise, i.e. the Gaussians have a higher standard deviation in noisy sequences than sequences with a lower noise. This feature prevents the pixels of the background to be considered as foreground pixels, even in case of strong noise, which in turns guarantees that the RoI replenishment prioritization allocates transmission resources to the objects moving in the scene, and not to the non-relevant variations of background caused by the noise.

Moreover, the background estimation process filters the sequence temporally and provides a denoised version of the background. Thus, we expect the CROI method to offer a denoised, and perceptually more pleasant version of the sequence at the client side. This is confirmed visually, and illustrated in Figure 11 where the original sequence is taken as reference to compute the PSNR values obtained when transmitting the original and noisy sequences based on the CROI, CRB and AVC methods, respectively. The left part of the figure focuses on the *RoI*. In normal conditions, all transmitted bits of the CROI method are dedicated to the RoI, which explains the higher performances of this method compared to CRB, and even to AVC for sufficiently large rates. In noisy conditions, the RoI quality of all methods sharply decreases since it is computed with respect to the original sequence, while all codecs attempt to describe the noise. The right part of the figure represents the *non-RoI* quality. In normal conditions, AVC outperforms CRB and CROI. In more

[6]Here, we consider that a precinct belongs to the RoI if at least 5% of its supporting pixels are labeled as RoI pixels. The supporting pixels of a precinct are obtained by dyadic upsampling of the precinct subband support.

[7]Note that in our simulation, once the RoI reaches its maximal quality, CROI does not transmit additional data to improve the non-RoI region, even if some bit-budget is available.

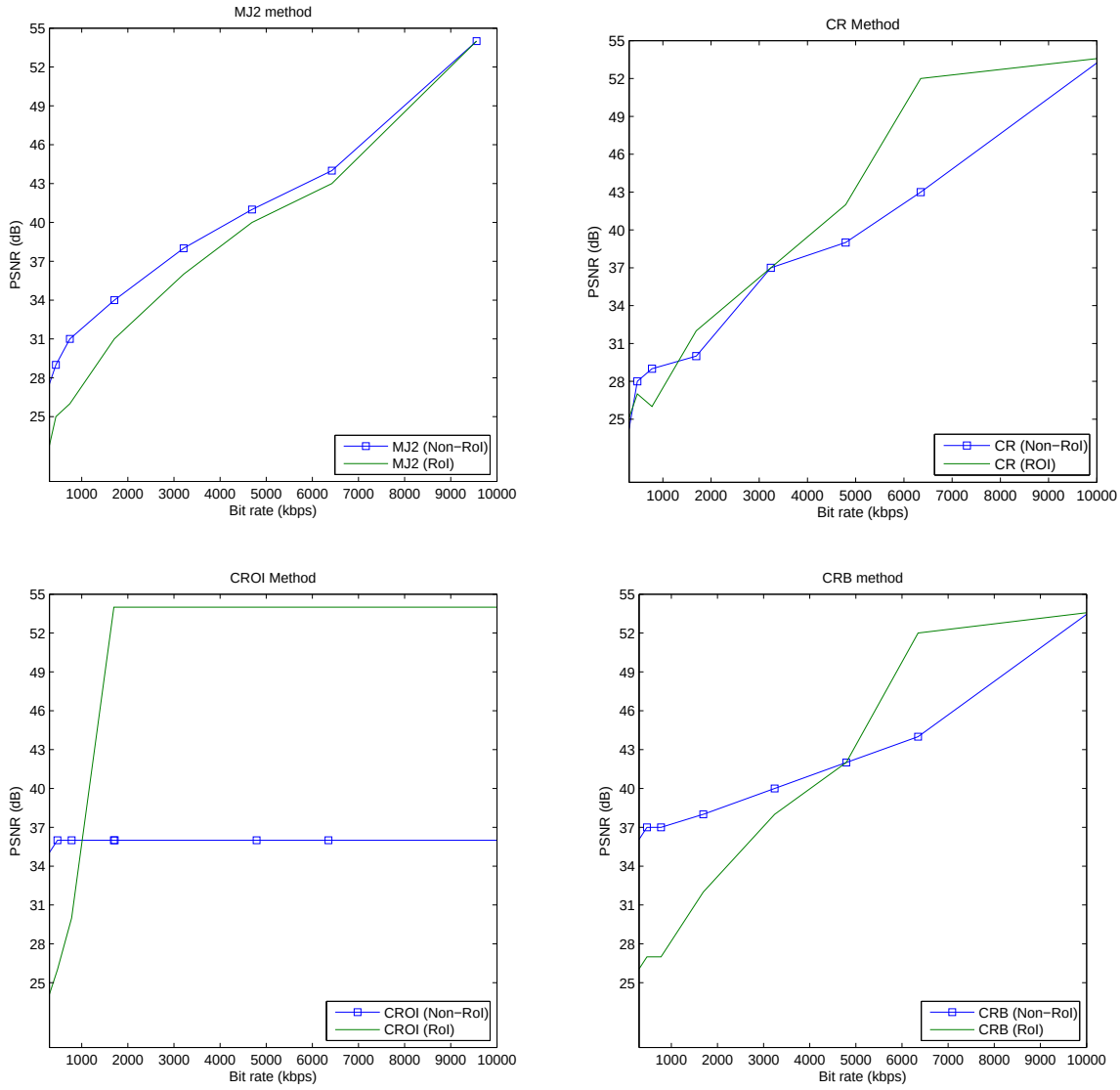

Fig. 10. RoI and non-RoI quality as a function of the total transmission rate for the CR, CROI, CRB and MJ2 methods (*Speedway* sequence).

details, the CROI method minimizes the rate allocated to non-RoI regions, thereby preventing the non-RoI quality to increase with the bit budget. In contrast, CRB progressively refreshes the non-RoI regions as the global (RoI + non-RoI) available rate increases, providing a higher overall non-RoI quality. In noisy conditions, we observe that CROI outperforms both CRB and AVC. Since the non-RoI regions are modified by the noise at each frame, the CRB (AVC) method regularly refreshes (corrects prediction errors for) those regions, mainly to render noise effects, which ends up in decreasing the quality compared to the original signal. On the contrary, since the CROI method knows a priori that most of the changes affecting the background are due to noise, it concentrates the refresh on RoI regions and almost never refreshes the non-RoI regions, thereby providing a higher background quality compared to the original (without noise) sequence. The same argument also explains why the non-RoI quality -measured with respect to the original sequence- is higher than the RoI quality when considering the CROI encoding scheme. In short, RoI noise is coded accurately while a denoised filtered background is used as the reference for the non-RoI, resulting in a non-RoI signal which is closer to the original.

C. Error resilience capabilities

The conditional replenishment transmission framework is characterized by the fact that the refreshed information is transmitted in INTRA, without any reference to the past. Hence, it naturally provides some resilience to transmission errors, since an error only remains perceptible until the next successful refresh. Unfortunately, this assertion also means that areas that are rarely refreshed become more sensitive to transmission errors than other regions. In order to prevent persistent errors when transmitting video in noisy environments, a particular attention should thus be devoted to those regions that become temporally stable after a period during which they were significantly changing. Indeed, for those regions,

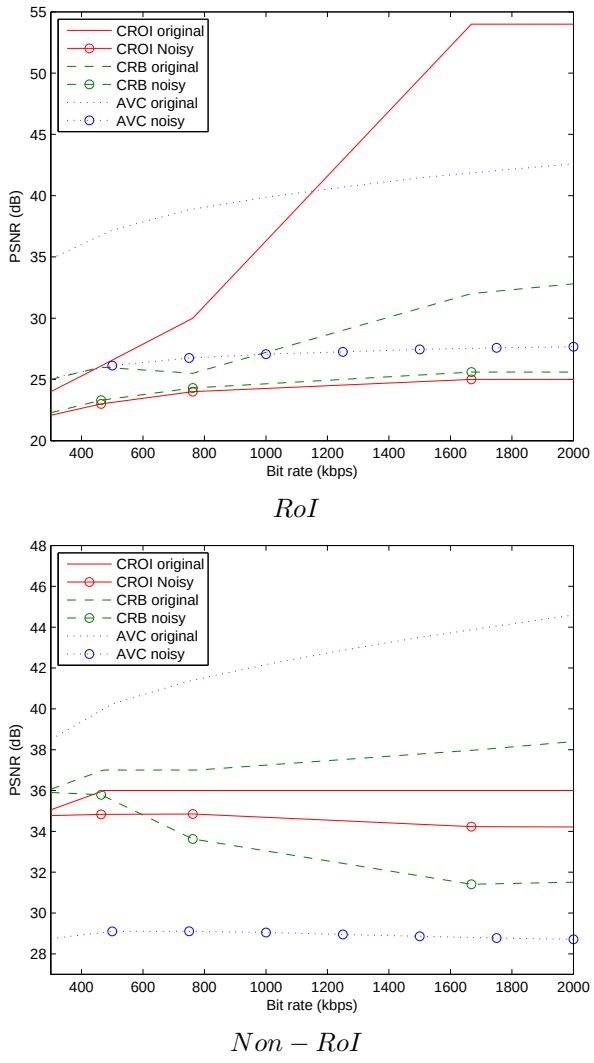

Fig. 11. RoI and non-RoI quality for the CROI, CRB, and AVC methods in normal and noisy conditions (*Speedway* sequence). In all cases, the PSNR is calculated using the original (non noisy) sequence as reference.

if the last refreshment before a stable period is lost, then the resulting reconstruction error affects the whole stable period, with dramatic perceptual impact.

In this section, we propose a preliminary analysis of the replenishment system performances in a noisy environment and propose simple methods to validate our intuition. The problem formalization should be done in a rate distortion framework, similarly to what we have done in [26] for the resilient transmission of JPEG 2000 codestreams, and incorporate temporal considerations related to the variability of the temporal impact of errors in our replenishment framework. The optimization of the replenishment scheduling taking into account this variability is beyond the scope of this paper. However, we provide an illustrative example based on an adaptive scheduling and a heuristic protection which retransmits important refresh.

To validate our intuition, Figure 12 considers an error-prone channel characterized by independent and identically distributed (iid) bit errors, and assumes that a packet is lost as soon as one of its bits becomes erroneous. The figure compares three scheduling methods. The first one is the conventional replenishment method (*CR*). The second one, denoted *CR Robust I*, knows the channel BER and takes it into account to schedule JPEG 2000 packets. Specifically, it uses a first order approximation to compute the reference distortion, and accounts for the packet loss probability to compute the benefit expected from JPEG 2000 packets transmissions. The third one, denoted *CR Robust II*, extends the previous method by adding a simple heuristic to improve the robustness of critical refresh.

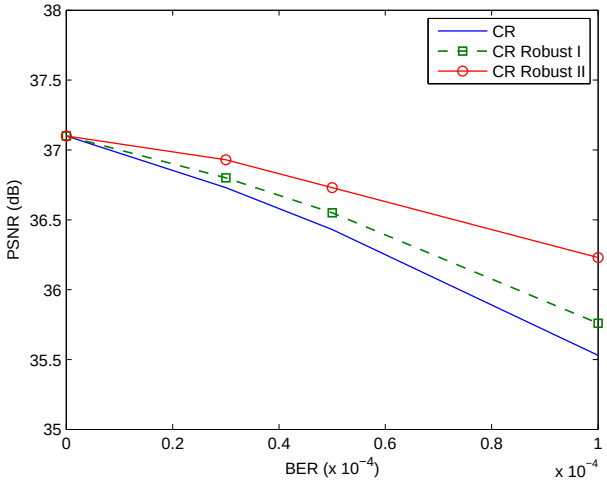

Fig. 12. Comparison of three replenishment methods as a function of the channel bit error rate for the *Speedway* sequence at 500 kbps. Transmitted JPEG 2000 packet are considered as lost as soon as one of their bits becomes erroneous. The three methods are described in the text.

Formally, according to the notations introduced in Section III-C, (k_t^i, q_t^i) and $(k_{t-k_t^i}^i, q_{t-k_t^i}^i)$ denote the index and quality level associated to the two latest refreshment of the i^{th} precinct. Those refreshments occurred respectively at time $(t - k_t^i)$ and $(t - k_t^i - k_{t-k_t^i}^i)$. Here, to simplify notations, we omit the precinct and time indexes for k and q, and just use (k_1, q_1) and (k_2, q_2), with $k_1 = k_t^i, k_2 = k_t^i + k_{t-k_t^i}^i$, to denote the instants and quality levels associated to the two latest refreshments of precinct i. Hence, $d_t^{k_1,q_1}(i)$ and $d_t^{k_2,q_2}(i)$ denote the distortion measured when approximating the i^{th} precinct at time t either based on the last or last but one refreshment. If we denote p_1 the probability that the last refreshment of precinct i at time t has been lost, the first order approximation of the reference distortion can be computed as $d_t^{ref}(i) \cong (1-p_1)\, d_t^{k_1,q_1}(i) + p_1\, d_t^{k_2,q_2}(i)$. It corresponds to a first order approximation since it ignores the fact that the last but one refreshment might also be lost. Similarly, the benefit expected from the refreshment of the i^{th} precinct at time t can be estimated based on the knowledge of the channel BER. Those refinements of the expected distortions are implemented by the *CR Robust I* method. However, they are unable to prevent the appearance of persistent errors in regions that become stable, e.g. after a moving object discloses a still background. This is because p_1 is typically small, making $p_1\ d_t^{k_2,q_2}(i)$ insignificant compared to changing areas in the

frame. To circumvent this drawback, we propose a simple heuristic to identify the regions that are expected to remain stable for some time after significant changes, and force an additional refreshment for them, so as to ensure with a high probability that they will be correctly received at the client. In practice, this is done by setting $d_t^{ref}(i)$ to $d_t^{k_2,q_2}(i)$ for the regions for which $d_t^{k_2,q_2}(i) \gg d_t^{k_1,q_1}(i)$, which are regions that appear to be significantly changing before $(t-k_2)$ and stable at time $(t-k_1)$. The curve *CR Robust II* in Figure 12 implements that heuristic. Unsurprisingly, we observe that it significantly improves the resilience of the conditional framework to losses. We conclude after this preliminary analysis that an adaptation of the scheduling algorithm should enable the replenishment framework to efficiently support error-prone channels.

VII. Conclusions

This paper considers remote interactive browsing of JPEG 2000 content captured by still cameras. Rather than transmitting each frame independently to the clients as it is generally done in the literature for JPEG 2000 based systems, our proposed streaming server adopts a conditional replenishment scheme to exploit the temporal correlation of the video sequence. As a first contribution, we propose a rate-distortion optimal strategy to select the most profitable packets to transmit. As a second contribution, we provide the client with two references, the previous reconstructed frame and an estimation of the current scene background calculated at the server side, which significantly improves the transmission system rate-distortion performances. As a third and significant outcome, we describe a post-compression rate allocation mechanism, which enables the server to adapt in real-time the content forwarded to heterogeneous -both in terms of resources and interest- clients using a single pre-compressed version of the sequence. An index is pre-calculated offline to reduce the computational load at the server while scheduling the packets according to the needs and resources of each individual client. Extensive simulations have revealed that the proposed system significantly outperforms both naive independent transmission of consecutive frames, and conventional replenishment mechanisms. At 500 kbps, the distortion of the proposed method for the *CAVIAR sequence* is 16 dB above MJ2, 10 dB above INTRA AVC and 2.5 dB below AVC with an Intra Period of 15. These results encourage the development of integrated and entirely JPEG 2000-based storage and transmission video surveillance systems, without the need to transcode the content to an MPEG-like format before its transmission.

References

[1] FP6 IST-2003-507204 WCAM, Wireless Cameras and Audio-Visual Seamless Networking, http://www.ist-wcam.org, 2004.

[2] F. W. Mounts. A video encoding system with conditional picture-element replenishment. *Bell Systems Technical Journal*, 48, no. 7:2545–2554, September 1969.

[3] S. McCanne, M. Vetterli and V. Jacobson. Low-complexity video coding for receiver-driven layered multicast. *IEEE Journal of Selected Areas in Communications*, 15(6):982–1001, 1997.

[4] ISO/IEC 15444-1. JPEG 2000 image coding system, 2000.

[5] M. Rabbani and R. Joshi. An overview of the JPEG 2000 image compression standard. *Signal Processing: Image processing*, 17:3–48, 2002.

[6] D. Santa-Cruz and T. Ebrahimi. An analytical study of JPEG 2000 functionalities. In *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Vancouver, September 2000.

[7] Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG. Joint Final Commitee Draft (JFCD) of Joint Video Specification (ITU-T Rec. H.264 – ISO/IEC 14496-10 AVC). Doc. JVT-D157, July 2002.

[8] T. Wiegand, G.J. Sullivan, G. Bjntegaard, A. Luthra. Overview of the H.264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):560–576, July 2003.

[9] C. Lin, J. Zhou, J. Youn, and M. Sun. MPEG video streaming with VCR-functionality. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(3):415–425, March 2001.

[10] MPEG and ITU-T. Scalable Video Coding Standard ISO/IEC 14496-10. August 2007.

[11] A Mavlankar, D. Varodayan, and B. Girod. Region-of-interest prediction for interactively streaming regions of high resolution video. In *Proceedings of 16th IEEE International Packet Video Workshop (PV)*, Lausanne, Switzerland, November 2007.

[12] D. Taubman D. and M. Marcellin. *JPEG 2000: Image compression fundamentals, standards and practice*. Kluwer Academic Publishers, 2001.

[13] D. Taubman. High performance scalable image compression with EBCOT. *IEEE Transactions on Image Processing*, 9(7):1158–1170, July 2000.

[14] D. Taubman and R. Rosenbaum. Rate-distortion optimized interactive browsing of JPEG 2000 images. In *IEEE International Conference on Image Processing (ICIP)*, September 2003.

[15] Y. Shoham and A. Gersho. Efficient bit allocation for an arbitrary set of quantizers. *IEEE Transactions on Signal Processing*, 36(9):1445–1453, September 1988.

[16] Antonio Ortega. Optimal bit allocation under multiple rate constraints. In *Data Compression Conference*, pages 349–358, Snowbird, UT, April 1996.

[17] A. Ortega, K. Ramchandran, and M. Vetterli. Optimal trellis-based buffered compression and fast approximation. *IEEE Transactions on Image Processing*, 3(1):26–40, January 1994.

[18] H. Kellerer, U. Pferschy, and D. Pisinger. *Knapsack problems*. Springer Verlag, 2004. ISBN 3-540-40286-1.

[19] L. Wolsey. *Integer Programming*. Wiley, 1998.

[20] F.O. Devaux, J. Meessen, C. Parisot, J.F. Delaigle, B. Macq and C. De Vleeschouwer. A flexible video transmission system based on JPEG 2000 conditional replenishment with multiple references. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 07)*, Hawaii, USA, April 2007.

[21] A. Cavallaro, O. Steiger and T. Ebrahimi. Semantic video analysis for adaptive content delivery and automatic description. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(10):1200–1209, October 2005.

[22] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 246–252, June 1999.

[23] D.S. Lee. Effective Gaussian Mixture Learning for Video Background Subtraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):827–832, 2005.

[24] X. Desurmont, C. Chaudy, A. Bastide, C. Parisot, J.F. Delaigle and B. Macq. Image analysis architectures and techniques for intelligent systems. In *IEE proceedings on Vision, Image and Signal Processing, Special issue on Intelligent Distributed Surveillance Systems*, volume 152, pages 224–231, 2005.

[25] ThreePastShop1front sequence from the CAVIAR Project (Context Aware Vision using Image-based Active Recognition). http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1, 2001.

[26] M. Agueh, F.O. Devaux, M. Diop, J.-F. Diouris, C. De Vleeschouwer, and B. Macq. Optimal Wireless JPEG 2000 compliant Forward Error Correction rate allocation for robust JPEG 2000 images and video streaming over Mobile Ad-hoc Networks. *EURASIP, Journal on advances in Signal Processing*, 2008.