A FLEXIBLE VIDEO TRANSMISSION SYSTEM BASED ON JPEG 2000 CONDITIONAL REPLENISHMENT WITH MULTIPLE REFERENCES

François-Olivier Devaux[1], *Jérôme Meessen*[2], *Christophe Parisot*[2], *Jean-François Delaigle*[2], *Benoit Macq*[1] *and Christophe De Vleeschouwer*[1]

[1] Communications and Remote Sensing Laboratory, Université catholique de Louvain (UCL), Belgium

[2] Multitel, Belgium

{devaux, macq, devlees}@tele.ucl.ac.be

{meessen, parisot, delaigle}@multitel.be

ABSTRACT

The image compression standard JPEG 2000 offers a high compression efficiency as well as a great flexibility in the way it accesses the content in terms of spatial location, quality level, and resolution. This paper explores how transmission systems conveying video surveillance sequences can benefit from this flexibility. Rather than transmitting each frame independently as it is generally done in the literature for JPEG 2000 based systems, we adopt 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, which improves the transmission system performances.

Index Terms— JPEG 2000, Intra Coding, Replenishment, Adaptive Delivery, Semantic Based Coding

1. INTRODUCTION

Nowadays, an increasing number of video surveillance systems use digital video coding standards and IP networks to compress and transmit a huge amount of video data from cameras and storage servers to a wide variety of terminals, from control rooms to wireless PDAs. While Motion JPEG and MPEG-2 codecs have been largely deployed, MPEG-4, AVC and JPEG 2000 codecs are now emerging in video surveillance devices and systems.

Motion JPEG 2000 (MJ2) [1], the video file format encapsulating JPEG 2000 frames [2], presents several important and attractive features for video surveillance systems. Compared to MPEG-based systems, this intra compression standard offers a high robustness to transmission errors and provides fine-grained temporal, spatial, resolution and quality scalability [3]. The coded bitstream can easily be parsed and adapted in real-time following each of these scalabilities without the need of expensive transcoding operations. This enables the server to optimize the transmitted video quality according to the client needs and decoding capabilities, and according to the varying network resources, with a minimum impact on its processing requirements.

In this paper, we focus on JPEG 2000 video surveillance systems with fixed cameras. Rather than transmitting each frame independently to the clients as it is generally done in the literature for JPEG 2000 based systems, we adopt 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. The use of a second reference appears to significantly improve the transmission system rate-distortion performances.

The goal of our work is not to compete with other existing video coding systems like AVC, but to propose a rate-distortion optimized transmission system adapted to a JPEG 2000 video surveillance environment. Our simulations encourage the deployment of such video surveillance systems taking advantage of the JPEG 2000 features throughout the acquisition, analysis and transmission chain.

This paper is structured as follows. In Section 2, we present an overview of the proposed replenishment system. In Section 3, we remind the JPEG 2000 concepts useful for this work, and define two replenishment methods. Section 4 presents the simulation results. Conclusions are provided in Section 5.

2. SYSTEM OVERVIEW

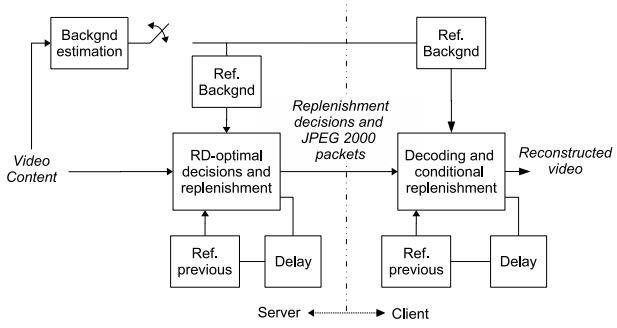

Fig. 1. Overview of the proposed JPEG 2000 video transmission architecture.

Figure 1 depicts the proposed transmission architecture. For each frame, the system only transmits the JPEG 2000 data units that are not properly approximated at the decoder, neither based on the background estimate, nor based on the previous reconstructed frame. As a consequence, the main concern of the sender is related

to the selection of (i) the parts of the JPEG 2000 image that have to be refreshed, and (ii) the level of quality associated with the corresponding refreshments. The second issue addressed by the sender is related to the background estimation. In the proposed system, an average background is computed based on Gaussian mixtures that collect the statistics of past image samples in specific pixel locations, as presented in [4]. At regular time intervals, or 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.

3. CONDITIONAL REPLENISHMENT

The section is organized as follows. First, we review the specificities of the JPEG 2000 standard that are relevant to the design of our replenishment decision engine. Then, we explain how rate-distortion optimal replenishment decisions are taken in agreement with the JPEG 2000 structure. Finally, we define two replenishment schemes that differ by their ability to exploit the background estimate as a replenishment reference.

3.1. JPEG 2000 image representation and code stream abstraction

The JPEG 2000 standard describes images in terms of their discrete wavelet coefficients. Hence, a replenishment scheme dedicated to JPEG 2000 contents decides to refresh or approximate the current image wavelet transform, based on the knowledge of the wavelet coefficients describing the reference background and previous images. An important question raised by conditional replenishment is related to the granularity of access to the current JPEG 2000 image 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 [2] [3] [5]. 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 [6] 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 [6]. 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 [2, 7]. 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.

3.2. RD optimal replenishment

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 is composed of sets of precincts organized in a hierarchy of layers, the problem consists of 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. Non-refreshed precincts are approximated based on the wavelet coefficients of the reference image. The use of multiple reference images is described in Section 3.3.

To simplify notations, and without loss of generality, the precincts, originally defined by their (r, p) indexes, are now 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 $d^0(i)$ to denote the distortion computed when the i^{th} precinct is approximated based on its q first packets, i.e. 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 the bit budget. Based on the additivity assumption and because a packet is only useful upon reception of all its ancestors, the problem can be formulated as a Knapsack problem with precedence constraints [8]. Let $q(i)$ denote the number of quality layers transmitted for the i^{th} precinct. Then, the RD optimal refreshment decisions are defined by the set $\{q(i)\}_{i \leq N}$ that maximizes $\sum_{i<N}(d^0(i) - d^{q(i)}(i))$, subject to $\sum_{i<N} s^{q(i)}(i) \leq T$. Formally, this Knapsack problem can be solved based on dynamic programming [8, 9]. However, two specificities of our problem simplify it, and make an iterative greedy solution RD optimal.

First, the lower RD convex-hull of a precinct originates in the RD point defined by the reference image ($R = 0$) and goes through all the refreshment solutions that involve a sufficient number of quality layers. This is because, in absence of a reference frame, the benefit per transmission cost of a precinct packet decreases as the layer index increases [6]. Hence, the succession of RD points corresponding to an increasing number of layers sustains the lower RD convex-hull in absence of reference. In the replenishment case, the lower RD convex-hull is affected by the existence of a reference frame, and the refreshment of a precinct only becomes worthwhile in the convex-hull sense beyond a quality level for which the benefit (compared to the quality achieved based on the reference frame) per unit of rate

becomes larger than the relative gain offered by subsequent layers of the precinct. Hence, for the i^{th} precinct, the set of convex-hull RD optimal solutions contains the reference precinct (R=0) and the refreshment solutions involving more than $q_r(i)$ quality layers, with $q_r(i)$ being the smallest value q such that

$$\frac{d^0(i) - d^q(i)}{s^q(i)} \geq \frac{d^0(i) - d^{q+1}(i)}{s^{q+1}(i)} \quad (1)$$

Second, the bit budget constraint can be somewhat relaxed, without impairing the overall performance of the communication system. This is because all video communication applications rely on buffers to absorb momentary rate fluctuations. As a consequence, the few bits that are saved (or overspent) compared to the bit budget allocated to a frame just slightly increments (or decrements) the budget allocated to the next frame.

As a consequence of the above observations, overall RD optimality can be achieved at the image level by selecting the packets to transmit so as to refresh the image precincts in decreasing order of benefit per unit of rate, up to exhaustion of the transmission budget. This approach is equivalent in principle to the one defined in [7], but is adapted to account for the availability of a reference image. Formally, the iterative process can be defined as follows.

Let $q_t(i,m)$ denote the number of layers already transmitted for the i^{th} precinct at step m, and let $q_t^+(i,m)$ denote the next convex-hull optimal refreshment level for the i^{th} precinct at step m. Based on the above discussion, $q_t^+(i,m) = q_r(i)$ when $q_t(i,m) = 0$, and $q_t^+(i,m) = q_t(i,m) + 1$ in other cases. Based on these definitions, at the initial step, we have $q_t(i,1) = 0\ \forall i$. Then, at each step m, the greedy process decides to improve the quality of the precinct i_m^* that provides the largest decrement in distortion per unit of transmission, i.e.

$$i_m* = \underset{1 \leq i \leq N}{\operatorname{argmax}} \frac{\left(d^{q_t(i,m)}(i) - d^{q_t^+(i,m)}(i)\right)}{\left(s^{q_t^+(i,m)}(i) - s^{q_t(i,m)}(i)\right)} \quad (2)$$

To prepare the next iteration, $q_t(i,m+1)$ is set to $q_t(i,m)\ \forall i \neq i_m^*$, and to $q_t^+(i_m^*,m)$ when $i = i_m^*$. The process goes on iterating on m as long as the bit budget is not exhausted.

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.

In practice, in our work, the distortion metric is computed based on the Square Error (SE) of wavelet coefficients, and approximates the reconstructed image square error [6]. Formally, let $\mathcal{B}_i$ denote the set of code-blocks associated to precinct i, and let $c_b[k]$ and $\hat{c}_b[k]$ 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 is then defined by

$$d(i) = \sum_{b \in \mathcal{B}_i} w_{sb}^2 \sum_{k \in b} (\hat{c}_b[k] - c_b[k])^2 \quad (3)$$

where w_{sb} denotes the L2-norm of the wavelet basis functions for the subband sb to which code-block b belongs [6]. As an alternative to the conventional SE metric, one can also consider a distortion defined based on semantically meaningful weighting of the SE, so as to take into account the a priori knowledge one may get about the semantic significance of approximation errors. Assuming that the information about the semantic relevance of approximation errors is provided at the precinct level, we define the semantically weighted distortion to be $d'(i) = w(i)d(i)$, where $w(i)$ denotes the semantic weight assigned to the i^{th} precinct. Semantically meaningful weighted distortion metrics have already been considered in the past. However, most earlier contributions exploit these metrics either before or during the encoding step. In contrast, our work supports the posterior definition of semantics weights, at transmission time, given the pre-encoded stream.

3.3. Replenishment methods definition

We now introduce the two replenishment methods that are considered in the simulations presented in Section 4. They are all based on the greedy approach described in Section 3.2, but differ in the way they define the reference image.

The **CR** (Conditional Replenishment) method follows the conventional replenishment mechanism originally introduced in [10] and adapted to the wavelet domain.

The **CRB** (CR 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. Our simulations demonstrate that CRB significantly outperforms CR in the surveillance scenario of interest in our study.

4. RESULTS

The transmission methods have been tested exhaustively, but we present the results on *Speedway*, a CIF video-surveillance sequence captured with a fixed camera at 25 fps. The original sequence, its estimated background and the segmentation masks are available on the WCAM project website [11]. Regarding the JPEG 2000 compression parameters, the 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). In order to have a spatial coherence between the precincts at different resolutions, we have chosen decreasing precinct sizes of 32x32, 16x16, 8x8, and 4x4 for the three remaining lowest resolutions. Regarding the rate control, the bit-rate has been uniformly distributed on all frames in the three intra methods. With AVC, we have adapted the quantization parameters to reach the expected bit-rates.

In these simulations, the background is sent only once at the beginning of the transmission because it remains sufficiently constant during the whole sequence. The transmission overhead is negligible, as the compressed estimated background of *Speedway* has a size of 55 Kbytes.

Figure 2 compares the PSNR at different bit-rates of the CR, CRB, MJ2 and MPEG-4 AVC (with three different Intra Periods, IP) methods.

The CR method improves the MJ2 compression, which is the less efficient compression scheme, by 2 dBs at low bit rates because only the most relevant blocks are refreshed. The CRB method outperforms these methods thanks to the estimated background that concentrates the refreshment in the most changing areas.

At very low bit rates, the CRB method results are close to MPEG-4 AVC. At 300 kbps, its PSNR is 1.5 dB below IP-10, 1.5 dB above IP-5 and 7 dB above IP-2. Its performances are comparable to AVC IP-2 at 1300 kbps. As mentioned in the introduction, the goal of this paper is not to propose a new compression scheme competing with existing ones like AVC, but rather to increase the perfor-

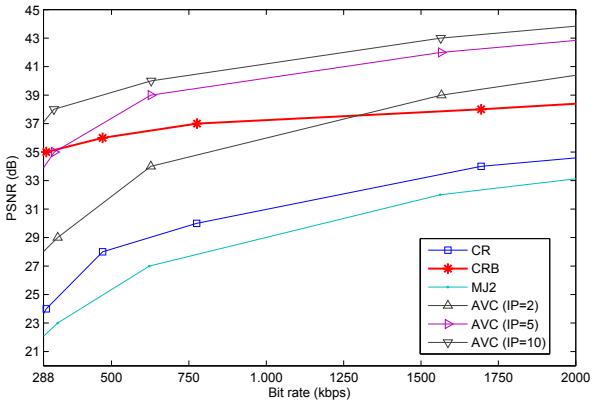

Fig. 2. Rate distortion curves of the proposed algorithms compared with MJ2 and AVC. Frame rates and encoding parameters are defined in the text.

mances of flexible video surveillance transmission systems based on JPEG 2000.

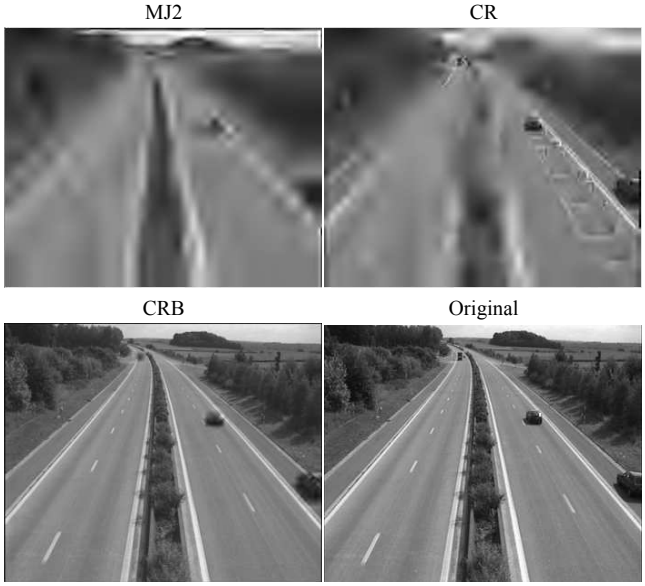

Fig. 3. MJ2, CR, CRB and AVC methods for the 25th frame of the *Speedway* sequence transmitted at 235 kbps.

Snapshots of the *Speedway* sequence compressed with the CR, CRB, MJ2 and AVC methods at 235 kbps are shown in Figure 3. As we can observe, CR improves slightly the MJ2 method, increasing mostly the precision on the vehicles. The CRB method improves the global sequence quality by offering a high quality for the background and a fair quality for the vehicles.

5. CONCLUSION

In this work, we have investigated the use of conditional replenishment mechanisms to transmit JPEG 2000 video surveillance content. We have explained how to take the refreshment decisions in a RD optimal way. We have also demonstrated the benefit of using multiple reference images for non-refreshed areas. In particular, we have proposed to compute an estimate of the background of the scene captured by a still camera, and have shown that such estimate significantly improves rate-distortion performances in video surveillance scenarios. In addition, we have highlighted the flexibility offered by a JPEG 2000 transmission of video content by prioritizing the refresh of scene areas that are a priori known to be semantically significant.

Interestingly, as a consequence of the JPEG 2000 intrinsic scalability, the prioritization allows to dynamically allocate transmission resources to the video content, but is independent of the JPEG 2000 codestream creation. Hence, it allows to allocate the rate to the content according to the user needs *a posteriori*, once the images have been compressed and stored. For the same reason, our system can be extended to a transmission to several clients, each client being characterized by its own resources. Eventually, simulations have revealed that the proposed system achieves close to AVC performance at low rates, and significantly outperforms both naive independent transmission of consecutive frames, and conventional replenishment mechanisms. These results encourage the deployment of integrated solutions able to store and transmit video surveillance content in JPEG 2000 format.

6. REFERENCES

[1] "Motion JPEG 2000 Final Committee Draft, 1.0, ISO/IEC JTC 1/SC 29/WG1 N2117," March 2001.

[2] ISO/IEC 15444-1, "JPEG2000 image coding system," 2000.

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

[4] J. Meessen, C. Parisot, X. Desurmont and J.F. Delaigle, "Scene Analysis for Reducing Motion JPEG 2000 video Surveillance Delivery Bandwidth and Complexity," in *IEEE International Conference on Image Processing (ICIP 05)*, Genova, Italy, September 2005, vol. 1, pp. 577–580.

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

[6] D. Taubman, "High performance scalable image compression with EBCOT," *IEEE Trans. on Image Processing*, vol. 9, no. 7, pp. 1158–1170, July 2000.

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

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

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

[10] S. McCanne, M. Vetterli and V. Jacobson, "Low-complexity video coding for receiver-driven layered multicast," *IEEE Journal of Selected Areas in Communications*, vol. 15, no. 6, pp. 982–1001, 1997.

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