A flexible video server based on a low complex post-compression rate allocation

François-Olivier Devaux and Christophe De Vleeschouwer
Communications and Remote Sensing Laboratory,
Université catholique de Louvain (UCL), Belgium
{devaux, devlees}@tele.ucl.ac.be

Laurent Schumacher
Computer Science Institute,
The University of Namur (FUNDP), Belgium
lsc@info.fundp.ac.be

***Abstract*— This paper proposes a highly flexible streaming video server that is particularly well suited for video-surveillance content. To achieve the required flexibility, the sequences are encoded with the JPEG 2000 coding scheme. To preserve compression efficiency, the temporal redundancy is exploited by a conditional replenishment technique. The main novelty of this paper consists in a post-compression rate allocation, which enables the server to adapt in real-time the content forwarded to heterogeneous 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 in real-time, according to the needs and resources of the clients.**

***Index Terms*— JPEG 2000, Rate Allocation, Conditional Replenishment, Adaptive Delivery, Semantic Based Coding**

I. Introduction

An increasing number of video surveillance applications integrate digital video streaming over IP networks. Data acquired from cameras have to be transmitted to a large variety of clients in terms of display resolutions and network resources. Moreover, these users also differ in their interest for the content, and require a flexible access to the data in several ways: they wish to be able to obtain a higher quality for a given region of interest, zoom on this zone and have a fast and precise temporal access to the sequence.

To address this issue, we propose a highly flexible video server able to stream content to a large number of users with heterogeneous needs and resources. First, to answer the flexibility constraint, our server stores the video sequences in the JPEG 2000 format [1], an image coding standard that offers a high flexibility in terms of spatial, resolution and quality access to the compressed data. Secondly, in order to increase the transmission efficiency of these intra coded sequences, a conditional replenishment scheme is used, which transmits prioritarily the most beneficial parts of the JPEG 2000 data based on a rate-distortion optimal strategy. In the mean time, the conditional replenishment only sends intra information, circumventing the drawbacks of closed-loop prediction systems when addressing heterogeneous clients dealing with different prediction references. Hence, this JPEG 2000 replenishment enables the adaptation of a single compressed version of the content to clients with different resources, answering the heterogeneity constraint. Finally, to serve a large number of users while keeping a low computational complexity, the server uses a pre-calculated index that guides him in adapting the JPEG 2000 packet scheduling decisions as a function of the client needs and resources.

François-Olivier Devaux and Christophe De Vleeschouwer are funded by the Belgian NSF.

The goal of our work is not to compete with other existing video coding systems like AVC in terms of compression efficiency. Instead, we aim at providing a solution for applications that require a flexible access to the content, meaning random access to an individual frame or interactive navigation through the video. Our system is thus particularly well adapted to a video surveillance environment. We extend the framework developed in [2] to serve multiple heterogeneous clients based on rate-distortion optimal scheduling of a single pre-computed bitstream.

The outline of this paper is the following. Section II presents an overview of the streaming system and Section III details the replenishment technique. Section IV presents how this system can serve multiple heterogeneous clients. Results are presented in Section V and the paper is concluded in Section VI.

II. System overview

The proposed system is depicted in Figure 1. The server adapts in real-time the content to each user by using information on the sequence generated prior to transmission.

Conditional replenishment enables the system to reach a high transmission efficiency. In this technique, the video server schedules for each client the JPEG 2000 packets that contribute the most efficiently to the quality of the sequence rendering. The other packets are not transmitted and are approximated at the client side using two references: the previously decoded image and an estimation of the background of the image.

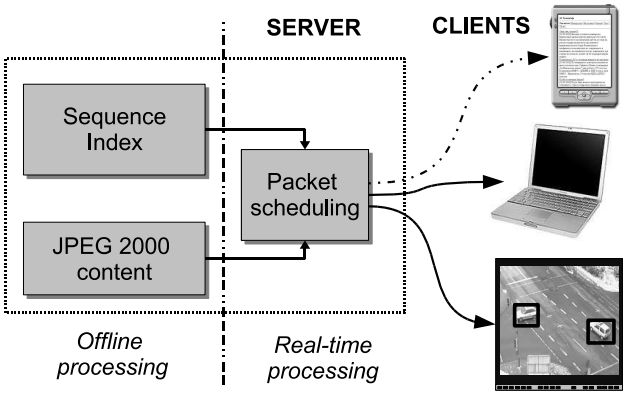

Fig. 1. System Overview. Using a pre-calculated index, the server selects the optimal packets to transmit to each client, based on their individual needs and resources. Three types of clients are illustrated: a PDA client with a low resolution and low bandwidth, a laptop client with a high bandwidth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

During the streaming session, the server requires two types of data: the JPEG 2000 content to transmit and a pre-calculated index. This index is used by the server to schedule the packets to transmit, based on each user requirements and resources. It is processed once when the server acquires the sequence and lists the distortion gain offered by the different replenishment solutions.

The flexibility of JPEG 2000 is exploited by offering a real-time spatial, in terms of resolution and localization, and qualitative adaptation. Temporal granularity is provided by the fact that JPEG 2000 is an intra video coding scheme, enabling an simple access to each independent frame.

III. JPEG 2000 Conditional Replenishment

The section is organized as follows. First, we review the JPEG 2000 standard. Then, we explain how conditional replenishment supports efficient video streaming based on JPEG 2000.

A. JPEG 2000 structure

The JPEG 2000 standard describes images in terms of their discrete wavelet coefficients. The subbands issued from the wavelet transform are partitioned into *code-blocks* that are coded independently [3] [4]. 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 [5]. 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 [1], [6]. 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. Replenishment and Packet scheduling

We now explore how the optimal JPEG 2000 replenishment decisions are taken.

The conditional replenishment framework originally introduced in [7] and exploited more recently in multicast transmissions [8] has been adapted here to the wavelet domain. In this framework, the server only transmits the JPEG 2000 packets that correspond to changing regions. The other regions, which are not refreshed, are approximated at the client side by the most efficient of the two available references: the previously reconstructed frame, and a reference background that is pre-computed at the server and is based on Gaussian mixtures estimations as presented in [9], and transmitted at regular time intervals to the client. In the simulations of Section V, a high quality version (lossless compression) of the reference background has been transmitted once every 8 second.

We consider the transmission of frame t, knowing the replenishment decisions taken for the previous frames. We 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 $(t-k)$. By extension, the replenishment of precinct i with q layers results in a distortion $d_t^{0,q}(i)$. Besides, we denote by $b_t(i)$ the distortion obtained when approximating the i^{th} precinct of frame t based on the latest update of the background. When the latest replenishment of precinct i occurred k frames earlier than t, the distortion using

the best reference for this precinct is noted $d_t^{ref}(i) = min[d_t^{k,q}(i), b_t(i)]$.

In addition to the distortion information, we are also interested in the size in bytes of the q first JPEG 2000 packets of precinct i of frame t, noted $s_t^q(i)$, for each $q \in \mathcal{Q}$.

Figure 2 gives a RD representation of these possible conditional replenishment decisions for a given precinct.

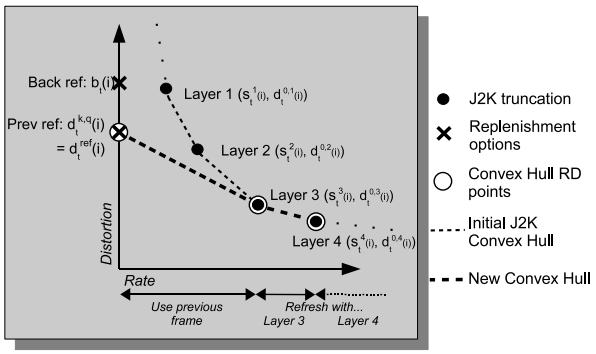

Fig. 2. Rate-Distortion representation of the distortion introduced in a given precinct by the four possible replenishments and the use of the background and previous references.

As explained in Section III-A, the truncation points of the embedded bitstream correspond to different rate-distortion trade-offs that lie on a RD convex-hull (*dots* in Figure 2). In addition, the possibility to approximate the precinct based on one of its two replenishment reference introduces two new points located on the distortion axis of the RD diagram (*crosses* in Figure 2).

Given a bit-budget and the set of accessible RD points for each precinct, the RD optimal allocation of the bit budget over the image is derived by considering the RD points sustaining the lower convex hull of each precinct, as depicted in Figure 2. Specifically, the increments of bitstream corresponding to these points are transmitted in decreasing order of distortion reduction per transmitted bytes, until the frame bit-budget is reached [6]. The approach is further detailed in [2].

IV. Serving multiple heterogeneous clients

In this section, we first describe how the replenishment decisions can take into account the definition of Regions Of Interest (ROI) by the user. Then, we study the practical implementation of the replenishment system to address a large number of clients while preserving an acceptable computational complexity.

A. Heterogeneous needs: Interactive ROI selection

In order to take the user needs into account, we integrate in the replenishment system the a priori knowledge one may have about the semantic significance of the approximation errors. To do so, we consider semantically weighted versions of the distortions: $d'(i) = w(i) * d(i)$, where $w(i)$ is the semantical weight associated to the precinct.

The transmission of the packets in decreasing order of semantical gain per unit of cost is RD optimal as long as the only packets that correspond to the convex-hull optimal RD points are considered. Please note that the convex-hull analysis performed on non weighted distortions remains valid, as long as the weighting affects in a similar way all the packets of a precinct. Indeed, in this case, the relative positions of the possible replenishment decisions of a precinct remain exactly the same.

Semantically meaningful weighted distortion metrics have already been considered in the past [10]. 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 for each client, and without any significant complexity increase.

B. Complexity issues: Index generation

When the server has to cope with a large number of clients, the real-time calculation of the $d_t^{k,q}(i)$ 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 the most complicated operations, and writes the results in an index. This index is then used in the second phase, which consists in taking the replenishment decisions for a particular streaming session.

Figure 3 depicts the first phase.

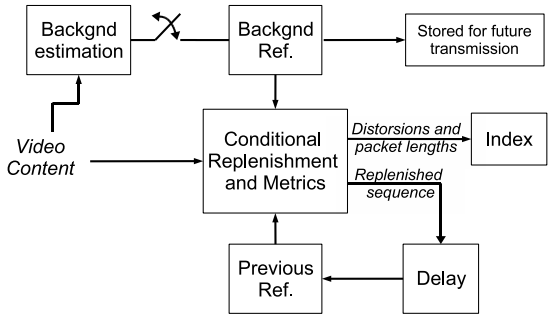

Fig. 3. Server off-line operations leading to the creation of the index.

The information needed by the server to take optimal decisions, based on the conditional replenishment concepts presented in section III-B, are the sizes $s_t^q(i)$ of the q first packets of precinct i and the three types of distortions $d_t^{k,q}(i)$, $b_t(i)$, and $d_t^{0,q}(i)$. These

informations are computed off-line and written to an index file that will be used during the streaming session to take the optimal decisions for each client.

The reference backgrounds and associated distortion $b_t(i)$ are generated based on Gaussian mixtures estimations and MSE computations. The $d_t^{0,q}(i)$ and $s_t^q(i)$ values are available from the JPEG 2000 compression of individual images. In contrast, the computation of the $d_t^{k,q}(i)$ values, which have to be calculated for each k and q, implies a significantly larger effort, both in terms of computation and memory resources. To reduce this effort, we use the following approximation:

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

The approximation is justified in Figure 4, where the different representations of frames t to $t-k$ are depicted. We observe that $d_t^{k,q}(i)$ is approximated based on a distortion computation path that only relies on $d_X^{0,q}(i)$ and $d_Y^{1,q_{max}}(i)$ 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, Y and Q variables take all possible values.

We will see in the next section that this approximation does not have a significant impact on the system performances. In terms of complexity, if we denote I_d to be the number of previous frames considered for the calculation of the previous reference distortion, $d_t^{k,q}(i)$ is computed for all $k < I_d$, which shows that the complexity increases linearly with the index depth for the optimal algorithm. By using the approximation, we limit the number of calculations to a constant value, independently of the index depth.

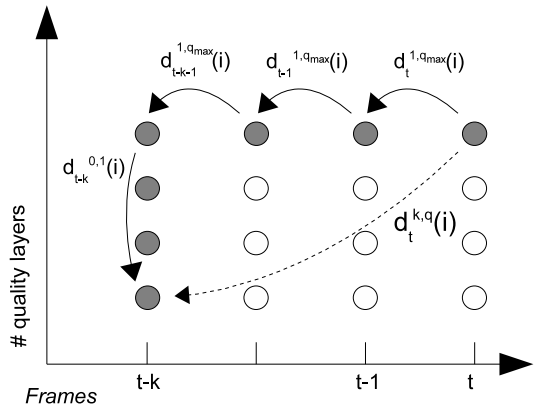

Fig. 4. Path used to approximate the distortion of the previous references, compared to the optimal path (dashed arrow). This approximation significantly decreases the pre-processing complexity and does not have a significant impact on the system performances.

V. Performances

In this section, we present the performances of the proposed streaming system. First, we compare our solution with two widely used coding scheme, MJ2 (Motion JPEG 2000) and MPEG4-AVC. Then, we illustrate how semantical weights can be used to focus on a given region.

The system has been tested on various video-surveillance sequences, but we present here the results on a sequence of the *CAVIAR* project [11]. The resolution is half PAL standard (384 x 288 pixels) and it has been encoded at 25 frames per second. 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 bitrate has been uniformly distributed on all frames in the three intra methods. With AVC, we have adapted the quantization parameters to reach the expected bitrates.

Figure 5 represents the rate distortion curves of the proposed system called CRB (Conditional Replenishment with Background). Two curves correspond to CRB: the optimal algorithm and the suboptimal algorithm using the approximation describes in Section IV-B. The system is compared to MJ2 and MPEG-4 AVC with three different Intra Periods (IP). As mentioned in the introduction, the goal of this paper is not to propose a solution competing with AVC in terms of compression efficiency, but rather to increase the performances of flexible video servers based on JPEG 2000.

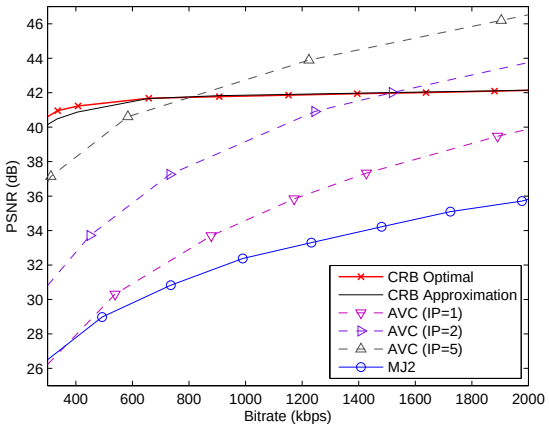

Fig. 5. Comparison of the system performance with MJ2 and MPEG-4 AVC in terms of PSNR at various bitrates for the *CAVIAR* sequence.

At very low bitrates, we observe that the proposed

method is more efficient than the other coding methods. The suboptimal CRB approximation behaves very similarly to the optimal algorithm, except under 600 kbps (at 300 kbps, the difference is 0.5 dB). This illustrates that the approximations described in Section IV-B do not alter significantly the proposed system efficiency. At higher bitrates, the AVC compression gives better performances. It is interesting to mention that with this sequence, intra MPEG4-AVC (IP=1) is more efficient than MJ2, thanks to its very efficient intra-coding algorithm. CRB outperforms MJ2 at all bitrates (+13.6 dB at 300 kbps and +6.4 dB at 2000 kbps). We realize that AVC with larger GOP would outperform the proposed scheme at the expense of reduced video access flexibility. Similar results have been observed with other video-surveillance sequences [2].

At high bitrates, the flatness of the CRB curves compared with the other coding schemes can be explained by the following. In order to improve the quality of a precinct that was estimated by the reference at lower bitrates, the whole compressed data of the precinct must be transmitted. For the other schemes, this increment in quality only requires a refinement of quantization (e.g. an increase in the number of coding passes to transmit for MJ2). Hence the CRB curve slope, corresponding to the ratio *quality increment* vs *rate increment*, is lower than for other methods.

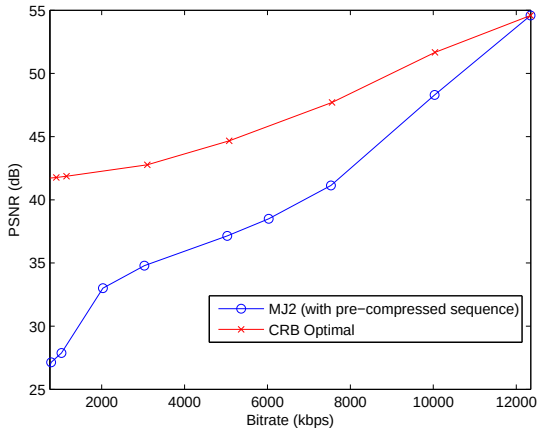

Fig. 6. Comparison of the system performance with MJ2 considering the adaptation of a single pre-encoded sequence to various bitrates for the *CAVIAR* sequence.

In Figure 6 we consider that the server adapts a single pre-compressed JPEG 2000 bitstream to the several targeted bitrates of the graph. This context corresponds more to the focus of this paper than the one considered in the previous figure where optimal MJ2 and AVC bitstreams were generated for each targeted bitrates. Figure 6 illustrates the benefit of using the references in the CRB method. As expected, this benefit decreases with the bitrate, as an increasing number of packets can be replenished. At the highest bitrate, where the bandwidth is sufficient to transmit all the available pre-encoded packets, both MJ2 and CRB methods give the same performances as the references.

To evaluate the interactive ROI selection, we have considered that a user viewing a speedway sequence wishes to focus on the moving vehicles. Segmentation masks have been calculated, and the areas defined by these masks have been considered as ROI. The semantical weights have been set to 1 for precincts belonging to the ROI and to 0 for the other precincts, which is of course an extreme choice. The *Speedway* sequence we have worked with is available on the WCAM project website [12] with its estimated background and the segmentation masks.

The background is sent only once with a high quality 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.

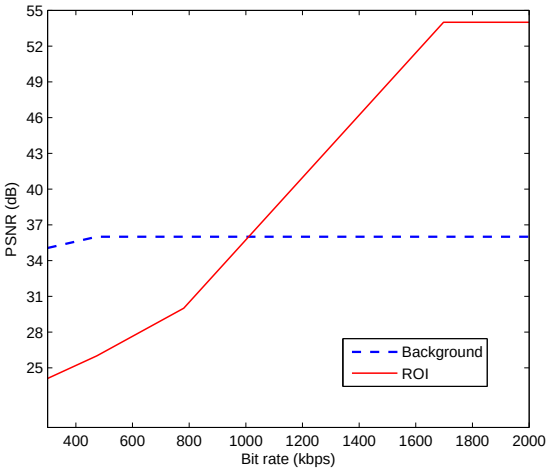

Fig. 7. Background and Region of Interest quality evolution for user focusing on the ROI of the *Speedway* sequence.

As we can observe on Figure 7, the background quality remains constant. This is explained by the fact that these non-ROI areas are never refreshed, and are only defined using the background reference transmitted at the same high quality for all bitrates. The ROI quality increases until a given threshold where all the code-blocks defining this ROI are refreshed. After this threshold (at 1700 kbps), neither the non-ROI nor the ROI quality is increased, as no additional data are transmitted.

VI. Conclusion

In this paper, we have presented a video transmission system designed to serve multiple clients with heterogeneous needs and resources in the context of video-surveillance applications. The sequences are stored on the server side in the JPEG 2000 format which offers a high flexibility. The temporal redundancy of the video content is exploited by the conditional replenishment of the JPEG 2000 precincts. The RD optimization framework within JPEG 2000 is extended to address this case of replenishment with two references. This allows a post compression rate allocation based on individual semantic needs that can be expressed by the clients during the streaming session. The server is able to handle these multiple heterogeneous clients in real-time by using a pre-processed index containing useful information for the replenishment decisions.

The system has been compared to MJ2 and MPEG4-AVC, and has been proved to keep the same flexibility as MJ2 while increasing the compression efficiency, and to be efficient in trading off the compression efficiency for flexibility compared to AVC.

References

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

[2] 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.

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

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

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

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

[8] 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.

[9] 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.

[10] A. Cavallaro, O. Steiger and T. Ebrahimi, "Semantic video analysis for adaptive content delivery and automatic description," *IEEE trans. on CSVT*, vol. 15, no. 10, pp. 1200–1209, October 2005.

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

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