RATE-DISTORTION BASED RECONSTRUCTION OPTIMIZATION IN DISTRIBUTED SOURCE CODING FOR INTERACTIVE MULTIVIEW VIDEO STREAMING

Ngai-Man Cheung [#], *Antonio Ortega* [*], *Gene Cheung* [+]

[#] Stanford University [*] University of Southern California [+] National Institute of Informatics

ABSTRACT

Interactive multiview video streaming (IMVS) is an application where, as the streaming multiview video is played back in time, an observer iteratively requests one of many available views at the server. In response, the server sends the appropriate pre-encoded data to the observer, with data chosen for transmission depending on the specific transmitted data available in the observer's cache. The primary challenge in IMVS is to design a structure for the pre-encoded multiview data, so that during an IMVS streaming session, the transmission rate is appropriately traded off with the pre-encoded data storage size. Previously, we have developed novel distributed source coding (DSC) based frame configurations to optimize the said tradeoff, outperforming periodical insertions of I-frames in both transmission and storage costs. In this paper, we show that by exploiting the freedom to choose a target decoded frame at a view switching point for DSC, further performance gains can be achieved: by up to 0.7 dB in our experiments.

***Index Terms*—** Interactive multiview video streaming, distributed source coding.

1. INTRODUCTION

In this paper, we focus on *interactive multiview video streaming* (IMVS) of stored content [1, 2, 3], where, as the streaming multiview video is played back in time, an observer iteratively requests one of many available views from the server, and in response the server sends the appropriate pre-encoded data to the observer, based on already transmitted data in the observer's cache, for decoding and display. The primary challenge in IMVS is to design a structure for the pre-encoded multiview data so that the *transmission rate* is appropriately traded off with the *storage size*.

One simple structure for IMVS [1] is to insert I-frames at desired view-switching points in the pre-encoded video stream to facilitate view switching (Figure 1(b))[1]. A drawback is that high bit-rate I-frames need to be sent, incurring high transmission cost. Another approach is to use P-frames for different decoding paths at the view-switching points (Figure 1(c)). However, since the P-frame reconstructions are not identical, this would lead to coding drift, or, the subsequent picture $F^o_{i+1,j}$ of time instant $i+1$ and view j would need to be encoded into multiple P-frames using each of the different reconstructions $F_{i,j}$'s as the predictor, resulting in considerable storage requirement [1].

A more transmission-storage efficient solution is to use H.264 SP-frames [4] for different decoding paths (Figure 1(d)). This works because SP-frames allow identical reconstructions from different traversals, and there is no need to encode the subsequent picture $F^o_{i+1,j}$ into multiple P-frames for different SP reconstructions $F_{i,j}$'s. Note that view switching arises when an observer wishes to switch to one of a small set of neighboring views during continuous playback of the video, and one of only a handful of possible frames of previous time instant (albeit from a different view) must be available at the decoder. Therefore, inter-frame correlation can be exploited as in a SP-frame to achieve transmission rate far below that of an equivalent I-frame. However, the SP-frame solution would require storing, for each picture at the view-switching point, multiple SP-frames, with each SP corresponding to one of the decoding paths.

In our previous work, we have proposed application of distributed source coding (DSC) for IMVS [2]. In this setting, DSC frames can be seen to play a role similar to that of SP frames, in that they allow merging of multiple decoding paths with drift (Figure 1(e)). In particular, DSC is used to generate a single reconstructed frame at time instant i and view j, using as side information multiple decoded versions obtained from differentially coded frames $F_{i,j}$, each corresponding to a different decoding path along the set of views (e.g., a different reconstructed frame $F_{i,j}$ would be generated if the decoded frame at the previous time instant was in view $j-1$ instead of in view j).

In this paper, we further improve upon our previously proposed DSC-based frame configuration for IMVS by exploiting a new degree of freedom previously unexplored. We observe that in this problem, the target for the DSC frame can in fact be chosen, i.e., merging the decoding paths requires having a common reconstruction target for all paths, but it is up to the designer to choose what this reconstruction should be. In particular we note that, for each block in the reconstructed frame, values for individual transform coefficients can be chosen independently. This choice can be made to optimize a rate-distortion (RD) metric. Thus, we seek to minimize distortion for the reconstructed frame, for a given rate spent transmitting the DSC frame (which depends on the reconstruction and the side information). For example, a target coefficient can be selected to be the mid-point of the range of all corresponding coefficients of side information, which minimizes rate (since it minimizes the largest distance between target and any side information). Alternatively, a coefficient can be selected to be the coefficient of an intra-coded version of $F^o_{i,j}$, which minimizes distortion. An optimization that exploits this degree of freedom can find the best RD tradeoff given side information statistics. Experimental results suggest that up to 0.7 dB improvement can be achieved.

Application of DSC to enable random access in lightfield compression was studied in [5] and [6]. Our previous work has focused on the design of frame structures of pre-encoded data for IMVS application [1, 2, 3]. In particular, in [2] we have proposed DSC-based constructions to improve the transmission-storage tradeoff over structures that use I- and P-frames only. Our current work further improves the transmission-storage tradeoff of one previous DSC-based construction by exploiting the freedom to select an appropriate target for a DSC frame. Note that some DSC applications

[1] We denote an original picture at time instant i and view j by $F^o_{i,j}$, and its (I-, P-, SP- or DSC-frame) reconstruction by $F_{i,j}$.

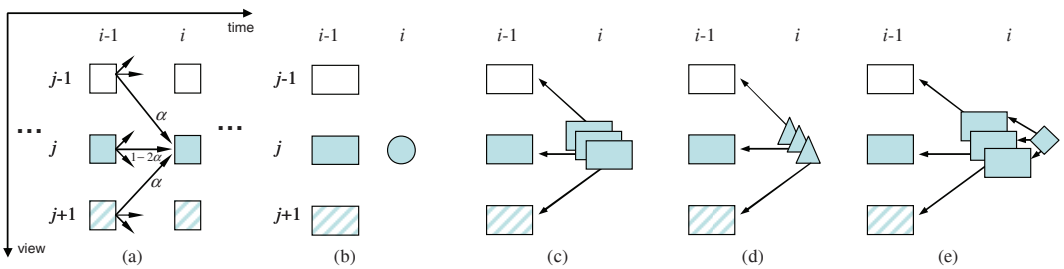


Fig. 1. (a) Interactive multiview streaming example. A client may switch to one of the adjacent views at time instant i with probability α. Each square represents an original picture $F^o_{i,j}$. Pictures belonging to the same view are shown with the same pattern. (b) Using an I-frame at the switching point: $F^o_{i,j}$ is encoded as an I-frame (represented by a circle) to facilitate view switching from adjacent views. Pictures at time instant $i-1$ are encoded as P-frames (rectangles) in this example. (c) $F^o_{i,j}$ is encoded as multiple P-frames for different decoding paths. Note that different P-frame reconstructions $F_{i,j}$'s are not identical. This may result in drifting. (d) Drift-free solution with $F^o_{i,j}$ encoded in SP-frames (triangles), which reconstructions are identical. (e) Using a DSC construction at the switching point: DSC frame $W_{i,j}$ (diamond) is used to remove the mismatch in different intermediate P-frames $P^{(k)}_{i,j}$'s.

may apply some simple reconstruction refinement at the decoder, e.g., [7]. The proposed RD optimized reconstruction in the present work, however, had not been studied in previous DSC work.

The outline of the paper is as follows. We provide an overview of IMVS optimization framework and our previously proposed DSC-based coding tool in Section 2. The key to successfully exploit the aforementioned degree of freedom is to find an appropriate target frame for DSC that optimizes the RD tradeoff. We discuss our formulation for this in Section 3. Results and conclusions are presented in Sections 4 and 5, respectively.

2. INTERACTIVE MULTIVIEW VIDEO STREAMING

2.1. IMVS Optimization Overview

While the main focus of this work is to investigate frame coding algorithms that can achieve good transmission-storage tradeoffs, we first briefly discuss in this section how some of these frames can be optimally combined to achieve good overall performance when encoding a sequence. Specifically, our IMVS optimization framework [1] designs a frame structure optimally trading off transmission rate with storage, given a known set of coding "tools" as building blocks. In this setting, we assume there are K spatially correlated views captured periodically and synchronously in time by a 1D array of cameras. During an IMVS session, an observer watches one view only, and requests a view change to a neighboring view k from current view j at time instant i with *view transition probability* $\alpha_{i,j}(k)$. We assume also that view switches only take place at multiples of M frames, i.e., $i = nM, n \in \mathcal{I}$. For the sake of simplicity, we will assume $M = 1$ in this discussion[2].

One simple IMVS structure is to insert an I-frame at each view-switching point for each view. See Figure 2(a) for an example when $K = 2$. This, however, leads to high transmission cost. Another structure is to differentially encode a P-frame for each possible view traversal for next instant $i+1$ given frames $F_{i,j}$'s in instant i. In Figure 2(b), assuming observers start from view 1 at instant $i = 0$, at instant $i = 1$, two P-frames of two different views are differentially coded using an I-frame at instant $i = 0$ as predictor. At instant $i = 2$, four P-frames are differentially coded and so on. While this *redundant P-frames* approach leads to low transmission cost, using this approach alone will lead to exponential storage with respect to switching instants. In [1], we proposed optimal frame structure designs using combinations of I- and redundant P-frames to optimize transmission-storage tradeoff. Figure 2(b) shows an example of a frame structure using combination of I- and P-frames. Clearly, some of these I- and P-frames can be replaced by more efficient frame constructions such as DSC frames to achieve better performance.

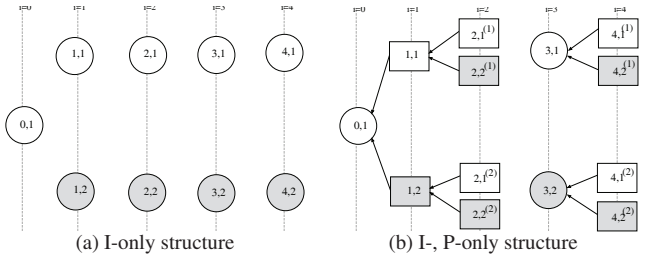


Fig. 2. Examples of frame structures for IMVS for two views (white and grey for view 1 and 2) and $M = 1$. Circles, rectangles are I- and P- frames, respectively.

2.2. DSC-based Frame Configuration

In this section we discuss the DSC construction proposed in [2] for IMVS. We first encode multiple *intermediate* P-frames $P^{(k)}_{i,j}$'s for $F^o_{i,j}$, each differentially coded from the set of possible previous frames $F_{i-1,k}$'s that can be in the decoder's buffer. In addition, we generate a *common* DSC frame $W_{i,j}$ *of the same time instant and view*, using $P^{(k)}_{i,j}$'s as side information, to remove the mismatch among $P^{(k)}_{i,j}$'s. Figure 1(e) depicts an example where three intermediate P-frames $P^{(k)}_{i,j}$'s are first encoded, followed by a DSC frame $W_{i,j}$.

In the DSC framework, $W_{i,j}$ is the parity information to "correct" each of the possible $P^{(k)}_{i,j}$'s into the same target $F_{i,j}$ [8]. The target reconstruction is carefully selected based on the discussion in Section 3 in the present work. Note that in order to achieve identical reconstruction from any decoding path, the encoder needs to: i) compute the statistics of the correlated noise $F_{i,j} - P^{(k)}_{i,j}$, and ii) provide a sufficient amount of parity bits to ensure recovery of $F_{i,j}$ from the *worst case* correlation noise [9]. In the case of IMVS, because the set $\mathcal{S}$ of possible frames $F_{i-1,k}$'s of previous instant $i-1$ that can be in the decoder's buffer is known *a priori*, the worst case noise statistics can be computed exactly given a reconstruction target. Note also that since $P^{(k)}_{i,j}$'s are reconstructed frames of the same time instant and view as the target, the statistics of $F_{i,j} - P^{(k)}_{i,j}$ are very similar for different k. Moreover, since we choose to select

[2]For the general case when $M \geq 1$, at a view-switching point, a carefully chosen I-, P- or DSC-frame can be selected by an optimizer, followed by $M-1$ P-frames of the same view to minimize bitrate. See [3] for details.

$P_{i,j}^{(k)}$ to have similar reconstruction quality as the target, the energy of the possible correlated noise $F_{i,j} - P_{i,j}^{(k)}$ is very small. Therefore, a small number of parity bits will be required to overcome the worst case correlated noise to reconstruct target $F_{i,j}$.

Note that the notion of pre-encoding multiple versions $F_{i,j}$'s of an original picture $F_{i,j}^o$ so that the exact same reconstruction can be achieved is similar to SP-frames. The difference, however, is that for SP-frames, $|\mathcal{S}| - 1$ secondary SP-frames, each employing lossless coding to remove mismatches, must be pre-encoded, while only one DSC frame is required for such lossless coding. Hence the DSC construction is much more storage-efficient compared to SP-frames. See Section 4 for comparison with SP-frames.

3. RD BASED RECONSTRUCTION OPTIMIZATION

3.1. Encoding Overview

Our proposed encoding of a DSC frame $W_{i,j}$ is illustrated in Figure 3. We apply DCT to the current picture $F_{i,j}^o$ to obtain transform coefficient X. We also apply DCT to the $|\mathcal{S}|$ intermediate P-frames $P_{i,j}^{(k)}$'s to obtain side information coefficients $\{Y_k \mid k = 0, ..., |\mathcal{S}| - 1\}$. We compute A, the mid-point of the range of side information $\{Y_k\}$ (This will be justified in Section 3.2). We select either X or A to apply quantization to generate the reconstruction target that optimizes the desired RD tradeoff, using algorithm to-be-discussed in Section 3.2. We losslessly encode the quantization index Q by converting it into a bit-plane representation and compressing it using an LDPC-based Slepian-Wolf encoder. The worst-case correlation information is used to compress the bit-planes to ensure exact recovery of Q under any decoder side information.

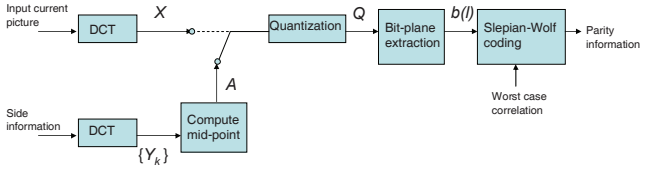


Fig. 3. DSC frame encoding.

3.2. Reconstruction optimization

Our proposed RD based reconstruction optimization algorithm adaptively selects, for each input transform coefficient X of the original picture, a target reconstruction that achieves a good tradeoff between the coding rate and distortion. Specifically, we propose to reconstruct each input coefficient into either a *minimum-distortion target*, $\hat{X}_0$, or a *minimum-rate target*, $\hat{X}_1$ (Figure 4). Denote the difference between $\hat{X}_i$ and X by $e_i, i = 0, 1$. Denote the side information that is farthest from $\hat{X}_i$ by V_i, $V_i \in \{Y_k\}$, and the distance between $\hat{X}_i$ and V_i by d_i; i.e.,

$$d_i = \max_{Y_k} |\hat{X}_i - Y_k|, \quad (1)$$

$$V_i = \arg\max_{Y_k} |\hat{X}_i - Y_k|. \quad (2)$$

Minimum-distortion target $\hat{X}_0$ is obtained by scalar quantization of X. $\hat{X}_0$ has a reconstruction distortion e_0^2. Also, $\hat{X}_0$ would require a coding rate that increases with d_0, as the correlation between $\hat{X}_0$ and V_0 would decrease when the distance between them, i.e., d_0, increases. We will discuss the rate approximation in more detail. Note that $\hat{X}_0$ could, in general, achieve the minimum distortion (given the quantizer configuration), as it is the closest reproduction point of the input X.

The minimum-rate target $\hat{X}_1$ is obtained as follow. As the coding rate of a target would increase with the maximum distance between this target and any of the side information, we could lower the rate by choosing one that minimizes this maximum distance. In particular, this can be achieved by selecting the mid-point of the range of side information values:

$$A = \frac{\max Y_k + \min Y_k}{2}. \quad (3)$$

In principle, A can lead to the minimum coding rate. We therefore obtain $\hat{X}_1$ by scalar quantization of A. Note that $\hat{X}_1$ may result in larger distortion compared to $\hat{X}_0$, as $\hat{X}_1$ may not be the closest reproduction level for X (when X and A are in different quantization bins).

To approximate the coding rate, we use the differential entropy $h(\hat{X}_i|V_i)$ [10][3]. Assume $\hat{X}_i = V_i + Z_i$, where Z_i is the correlation noise independent of V_i. Assume also that Z_i is Laplacian distributed with mean zero and standard derivation σ_i. It can be shown that

$$h(\hat{X}_i|V_i) = h(Z_i) = \log \sigma_i + \log \sqrt{2}e. \quad (4)$$

We assume each individual coefficient may have a different σ_i, and we estimate σ_i by d_i. We approximate the rate to encode $\hat{X}_i$ by $R(d_i) = \log d_i + \log \sqrt{2}e$, and select for each coefficient the target that achieves a smaller cost $e_i^2 + \lambda R(d_i)$, where λ is the Lagrange multiplier. We would like to remark that more accurate rate estimation can be achieved by using models that are specific to bit-plane based Slepian-Wolf coding [11]. Note that since $\{Y_k\}$ are available at the encoder in our problem, we can compute d_i precisely. For other DSC applications, similar reconstruction optimization may be applied if the coding rate or the correlation can be estimated. Note also that some DSC applications may use the correlation information to refine the reconstruction at the decoder, e.g., [7]. This can be readily accommodated in our proposed algorithm by modifying $\hat{X}_i$ accordingly (Our experiments employ such refinement as well).

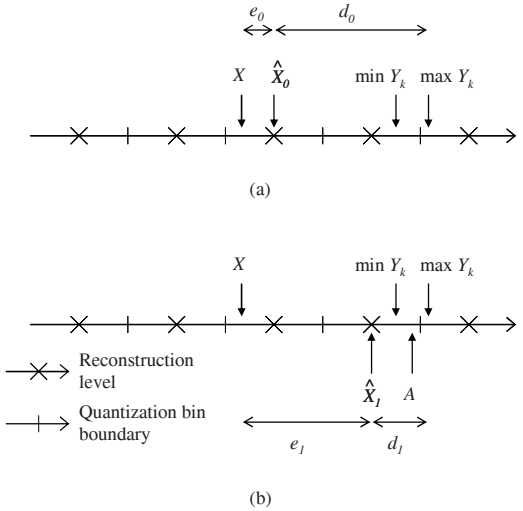


Fig. 4. RD based reconstruction optimization: (a) Minimum-distortion target. (b) Minimum-rate target.

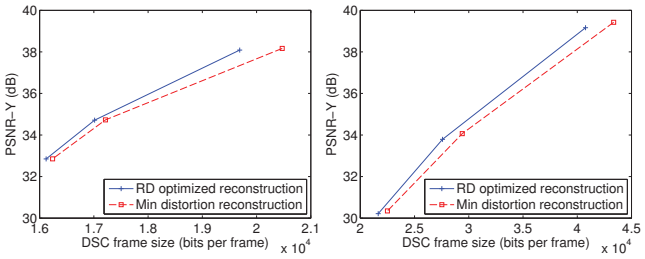
4. EXPERIMENTS

In this section we evaluate our proposed RD optimized reconstruction. We assume that, at each switching point, with probability $1 - 2\alpha$ clients remain in the same view, and can switch to one of the adjacent views with probability α (Figure 1(a)). We consider streaming solutions where into the P-frame-only structure, DSC frames or SP-frames are inserted periodically. In the case of SP-frames, the

[3] Note that since we estimate the rate for comparison only, there is no issue with that differential entropy could be negative.

(smaller) primary SP is used for the most probable request. We compare the bitrates at the switching points. Therefore, the results are largely independent of the length of the switching interval. The algorithm proposed in [9] is used to generate parity information $W_{i,j}$ in the DSC solution. The DSC system is developed from a H.263 codec to ease implementation, and uses H.263 coding tools (e.g., half-pixel motion estimation (ME)). For a more fair comparison, we use only 16×16 motion compensation in H.264 SP-frames. Note that the SP-frame solution still uses some advanced coding tools (e.g. quarter-pixel ME, CAVLC) compared to the DSC system. Thus, we have a slight disadvantage w.r.t. SP-frames. We use multiview video sequences Akko&Kayo and Ballroom in the experiments, which are in 320×240 and are encoded at 30 fps and 25 fps respectively. Figure 5 compares the proposed RD optimized reconstruction with previous work, which uses only the minimum-distortion target [2, 3]. The results suggest that by selecting target representations adaptively up to 0.7 dB improvement can be achieved.

Figure 6 depicts the comparison of storage cost. In this setting, the SP-frame solution requires storing of one primary SP-frame and two expensive secondary SP-frames (Figure 1(d)), while the DSC solution requires storing of three small P-frames and one set of parity information for mismatch elimination (Figure 1(e)). The results suggest that the DSC solution can achieve a better storage requirement as discussed in Section 2. Note that with each additional path, the storage requirement of DSC solution would increase only by the size of a small P-frame. For the SP-frame solution, however, the storage would increase by that of an expensive secondary SP-frame. Figure 7 depicts the comparison of the expected transmission cost, with switching probability α equal to 0.2. For SP-frame, the small primary SP-frame is transmitted with probability $1-2\alpha$ when clients stay in the same view (Figure 1(d)), and one of the expensive secondary SP-frames is sent with probability α when view switching is requested. For DSC solution, the same-view predicted intermediate P-frame and the parity bits $W_{i,j}$ are sent with probability $1-2\alpha$, and one of the cross-view predicted intermediate P-frames and the parity bits are sent with probability α. The results suggest that DSC solution could be comparable to or more bandwidth-efficient than SP-frames. Note that the sum bitrate of an intermediate P-frame and parity bits in DSC is larger than that of a primary SP-frame, but less than that of a secondary SP. Therefore, DSC tends to perform better in transmission cost when view switching is more likely (i.e., when α is large). In summary, the results suggest DSC solution can achieve competitive tradeoffs between storage requirement and transmission cost, in interactive streaming of multiview videos.

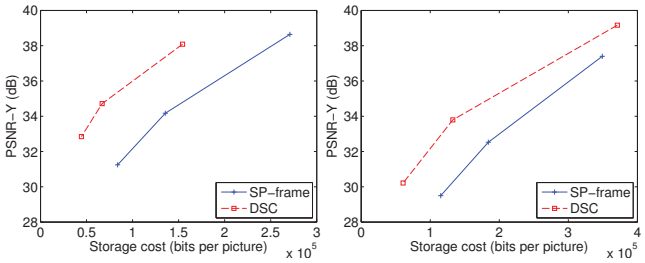


(a) Akko & Kayo (b) Ballroom

Fig. 5. Comparing the proposed RD optimized reconstruction with the minimum-distortion reconstruction. The horizontal axis represents the average size of the parity information, i.e., $W_{i,j}$.

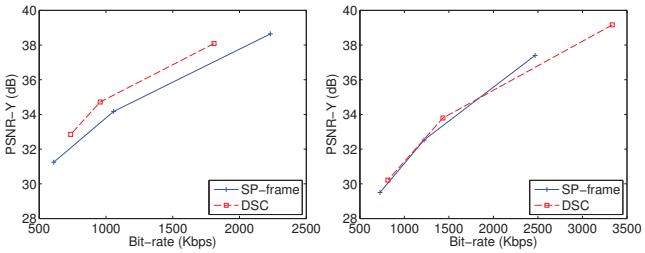
5. CONCLUSIONS

We have discussed RD based reconstruction optimization in DSC for IMVS. Our algorithm exploits the freedom to select a reconstruction target adaptively for each coefficient at a view switching point. Experimental results suggest that up to 0.7 dB improvement can be achieved with the proposed optimization. In addition, the DSC solution compares favorably to SP-frames in both the storage and transmission requirements. Future work includes investigation of sophisticated rate models to improve the RD decisions.



(a) Akko & Kayo (b) Ballroom

Fig. 6. Average storage cost per picture, at different reconstruction quality. In DSC, different reconstruction quality can be achieved by varying the quantization parameters for the intermediate P-frames and the targets.



(a) Akko & Kayo (b) Ballroom

Fig. 7. Expected transmission cost, with switching probability α equal to 0.2. DSC tends to perform better when view switching is more likely. See [2] for comparison with other settings of α.

6. REFERENCES

[1] G. Cheung, A. Ortega, and N.-M. Cheung, "Generation of redundant coding structure for interactive multiview streaming," in *Seventeenth International Packet Video Workshop*, Seattle, WA, May 2009.

[2] N.-M. Cheung, A. Ortega, and G. Cheung, "Distributed source coding techniques for interactive multiview video streaming," in *27th Picture Coding Symposium*, Chicago, IL, May 2009.

[3] G. Cheung, N.-M. Cheung, and A. Ortega, "Optimized frame structure using distributed source coding for interactive multiview streaming," in *IEEE International Conference on Image Processing*, Cairo, Egypt, November 2009.

[4] M. Karczewicz and R. Kurceren, "The SP- and SI-frames design for H.264/AVC," in *IEEE Transactions on Circuits and Systems for Video Technology*, July 2003, vol. 13, no.7, pp. 637–644.

[5] A. Aaron, P. Ramanathan, and Bernd Girod, "Wyner-Ziv coding of light fields for random access," in *IEEE International Workshop on Multimedia Signal Processing*, Siena, Italy, September 2004.

[6] A. Jagmohan, A. Sehgal, and N. Ahuja, "Compression of lightfield rendered images using coset codes," in *Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, November 2003, vol. 1, pp. 830–834.

[7] A. Aaron, S. Rane, and B. Girod, "Wyner-Ziv video coding with hash-based motion compensation at the receiver," in *Proc. IEEE Int'l Conf. Image Processing (ICIP)*, 2004.

[8] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proceedings of the IEEE, Special Issue on Advances in Video Coding and Delivery*, vol. 93, no. 1, pp. 71–83, Jan. 2005.

[9] N. Cheung and A. Ortega, "Distributed source coding application to low-delay free viewpoint switching in multiview video compression," in *Proc. of Picture Coding Symposium, PCS'07*, Lisbon, Portugal, Nov. 2007.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: Wiley, 1991.

[11] N.-M. Cheung, H. Wang, and A. Ortega, "Sampling-based correlation estimation for distributed source coding under rate and complexity constraints," *IEEE Trans. Image Processing*, vol. 17, no. 11, pp. 2122 – 2137, Nov. 2008.