Packet Scheduling in Multicamera Capture Systems

Anonymous VCIP Submission
Paper ID: 244

***Abstract*—In interactive multi-view streaming, neighboring cameras acquire frames which are generally correlated. This results in large amounts of highly redundant data, that makes essential to handle properly the correlation during encoding and transmission of the multi-view data. In this work, we study coding and transmission strategies in multicamera sets, where correlated sources need to be set to central server, to be then delivered to interactive clients. We propose a dynamic correlation-aware scheduling optimization of encoded packets from correlated sources under delay and bandwidth constraints, in order to enable effective navigation of the scene. A novel trellis-based solution is proposed, providing us with a formal decoupling of dependent from independent frames, thereby significantly reducing the computation complexity. Simulation results show the gain of the proposed algorithm when coding and scheduling policy are dynamically optimized based on knowledge of network and correlation model, compared to agnostic scheduling policies.**

I. Introduction

Advances in interactive services and 3D television have paved the road to the development of multi-camera capture systems, in which multiple sources acquire, encode and transmit correlated video information. To provide high quality navigation to interactive clients, however, large storage/bandwidth requirements are needed. Opportunistic resource allocation strategies are essential in order to provide effective video quality in resources constrained environments.

Resource allocation has obviously been addressed in single view systems [1] or multi-view systems with joint encoders, *e.g.,* [2], but only a few works have studied the packet scheduling problem in multi-camera systems. In [3], a spatial correlation model has been proposed for static camera selection in wireless sensor networks, while in [4], an adaptive correlation-aware packet scheduling algorithm has been proposed, for a simplistic independent view coding framework. In this work, we overcome the main limitations of previous works and we dynamically optimize the selection and scheduling of encoded packets from *correlated sources* under delay and bandwidth constraints, to enable effective reconstruction of scene from any view that can be potentially requested by interactive users.

We are interested in a live-acquisition scenario in which multiple cameras acquire frames from the same scene but from different perspectives. Each camera acquires successive frames of the scene and stores them in its short buffer in three different encoded versions: intra-coded (or key frame), P frame (*i.e.*, predictively encoded in time), and Wyner-Ziv (WZ) frame. We assume that no content information is exchanged among cameras due to the system settings or resource limitations, which prevents the use of multiview video coding [5] to encode the video information. The frames are then sent from cameras to a central server through a bottleneck channel. Each view can be sent to the server only before its deadline, which is imposed either by camera buffer limitations or decoding deadlines in the streaming application. A server gathers the camera frames and eventually serves the clients requests. The objective is obviously to maximize the amount of information at the server. When the channel constraints do not permit to send all captured views, it becomes important to optimize the scheduling policy in such a way that the quality in the reconstruction of the multi-camera data is maximized and both bandwidth and time constraints are met.

We propose, a *correlation-aware* packet scheduling algorithm for encoded packets for multi-camera streaming in bandwidth-limited networks. We consider a correlation-based rate distortion (RD) model that is specific to multi-camera systems and we formulate a packet scheduling optimization problem, aimed at minimizing the quality of the data available at the server. Rather than being known a priori, the coding structure is dynamically selected as result of the packet scheduling optimization. In this way, we constantly adapt the set of coded packets stored in the server to the channel conditions as well as to the content information. To solve the optimization problem, we propose a *novel solving method* able to reduce the computational complexity by decoupling dependent from independent encoded frames. This allows to reduce the computational complexity, still reaching optimality of the solving method. Simulation results demonstrate that the proposed dynamic scheduling algorithm outperforms scheduling policies with static coding strategy and agnostic transmission schemesd.

II. Framework

We consider M cameras that acquire images and depth information of a 3D scene from different viewpoints. At the *encoder side*, each frame can be encoded as a key-frame (i.e., as intra-coded frame) or *predictive* frames: P-frames are predictively encoded from the same view but previously acquired in time, while Wyner-Ziv (WZ) frames are coded with distributed source coding (DSC) techniques using key-frames of the same view as side information (SI).

At the *receiver side*, we target to have an almost constant quality of the scene across space and time, in such a way that a smooth interactive system can be offered to the user in ideal conditions. Since we assume that the characteristics of each frame are similar, we set an equal encoding rate R^K for every key frame which leads to a constant decoded quality $d(R^K)$. Targeting the same distortion $d(R^K)$, the encoding rate of the

P and WZ frames depends on the *level of correlation*, ρ, they have with the reference frames.

From the decoded key frame, the other frames might be estimated at the receiver using depth-image based rendering (DIBR) or motion compensation techniques. Typically, these algorithms use depth or motion information in order to estimate by projection the position of pixels of the reference frame in the estimated one. The projected pixels are generally of good precision (depending on the accuracy of the depth map and motion vector field) but do not cover the whole estimated image, due to visual occlusions. We thus denote by $\rho(F_{t,m}|F_{\tau,l})$ the portion of the image $F_{t,m}$ (m-th camera at time t acquires the frame) that can be recovered (i.e., not occluded) by the key version of $F_{\tau,l}$, which can be neighboring in either the temporal or the spatial dimension. For any acquired frame $F_{t,m}$, we define the set of possible SI in spatial and temporal domain respectively as

$$\mathcal{N}_S(F_{t,m}) = \{F_{t,l} \text{ s.t. } \rho(F_{t,m}|F_{t,l}) \geq \beta_S, \text{ with } l \in [1, M]\}$$
$$\mathcal{N}_T(F_{t,m}) = \{F_{\tau,m} \text{ s.t. } \rho(F_{t,m}|F_{\tau,m}) \geq \beta_T, \text{ with } \tau \leq t\}$$

where $\rho(F_{t,m}|F_{\tau,l})$ is the level of correlation between $F_{t,m}$ and $F_{\tau,l}$. In short, we consider as possible SI any frame which have a level of correlation with $F_{t,m}$ greater than a threshold value β. We assume that WZ and P versions of $F_{t,m}$ predictively encoded from $F_{\tau,l}$ would have an encoding rate of $\mathcal{R}(F_{t,m}|F_{t,l}) = [1 - \rho(F_{t,m}|F_{t,l})]\, R^K$, where R^K is the encoding rate of every key frame. Thus, in our work in which the worst case SI is considered, we have that each WZ and P frame is encoded at a rate of

$$R^{WZ}_{t,m} = \max_{F_{t,l} \in \mathcal{N}_S(F_{t,m})} \left\{[1 - \rho(F_{t,m}|F_{t,l})]\, R^K_{t,l}\right\} \quad (1)$$
$$R^{P}_{t,m} = \max_{F_{\tau,m} \in \mathcal{N}_T(F_{t,m})} \left\{[1 - \rho(F_{t,m}|F_{\tau,m})]\, R^K_{\tau,m}\right\}.$$

These rates allows every predictive frame to be decoded at a distortion of $d(R_K)$ when both predictive frame and at least one SI is available at the decoder.

Theoretically all views should be sent to a central server, but network limitations might impose to send only a portion of them. Missing images can be however reconstructed from decoded key frame available at the decoder, denoted by $\boldsymbol{\chi}$, using DIBR warping of neighboring views or motion compensation of past frames. [1] In that case, the distortion is equal to

$$D_{t,m} = \rho(F_{t,m}|\boldsymbol{\chi}) \cdot d\left(R^K\right) + (1 - \rho(F_{t,m}|\boldsymbol{\chi})) \cdot d_{\max} \quad (2)$$

where d_{max} is the maximum distortion at which occluded areas are reconstructed (e.g., impainting distortion). The distortion model proposed above is used in our packet scheduling optimization able to select the best set of DUs to be sent to a central server, such that the expected distortion is minimized, and the network constraints are met. We consider that each encoded image at a given time instant from a particular camera is packetized into multiple data units (DUs) (one per encoded version and encoding rate), and stored in the camera buffer. We also assume lossless channels, such that scheduled packets are available at the decoder, while missing frames are the not scheduled ones. DUs representing the key versions contain texture and depth information about the 3D scene, while WZ or P versions DUs only send the encoded texture information, since they will not be used to reconstruct missing views.

[1]The decoding process can be physically performed either at the central server or at the clients. Our problem formulation can consider both cases.

III. Packet Scheduling Optimization

At each transmission opportunity τ, the scheduler decides the best set of DUs to schedule. Let F_l be a generic view, acquired at $T_{\text{A},l}$ and expiring at $T_{\text{TS},l}$.[2] The interactivity offered to clients is captured by the camera popularity P_l, the portion of clients that can request the frame F_l. We then define the set of candidates for being sent at τ as $\mathcal{L} = \{F_l \text{ s.t. } T_{\text{A},l} \leq \tau \leq T_{\text{TS},l}\}$. The encoded versions of views in $\mathcal{L}$ are candidate for being scheduled, however we impose the following scheduling constraints: i) only one version among WZ, P, and key of the same view can be scheduled; ii) a predictive frame is scheduled only if some SI frame is already available at the decoder. Note that both channel conditions and content models may vary over time, leading to different scheduling policy at different transmission opportunities. Thus, the scheduling policy is refined at each transmission opportunity.

For the sake of clarity, we now provide the problem formulation for the case of three encoded frames per view. However, the optimization holds also for more than three coded versions. We define the scheduling policy as $\boldsymbol{\pi} = [\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_{|\mathcal{L}|}]^T$ where $\boldsymbol{\pi}_l = [\pi_{l,1}, \pi_{l,2}, \pi_{l,3}]$, with $\pi_{l,1}, \pi_{l,2}, \pi_{l,3}$ being the scheduling policy of respectively the key, WZ, and the P DU of F_l. We can then express our optimization problem as follows

$$\min_{\boldsymbol{\pi}} \overline{D}_{\boldsymbol{\pi}} = \sum_{l: T_{\text{A},l} \leq \tau \leq T_{\text{D}}} P_l D_l(F_l|\boldsymbol{\pi}) \quad (3a)$$
$$\text{s.t. } \sum_i \pi_{l,i} \leq 1, \quad \forall l \quad (3b)$$
$$\sum_l \pi_{l,1} R^{(K)}_l + \pi_{l,2} R^{(WZ)}_l + \pi_{l,r} R^{(P)}_l \leq C \quad (3c)$$
$$\pi^T_{l,2} \leq \sum_{F_l \in \mathcal{N}_S(F_l)} \pi_{l,1} \quad (3d)$$
$$\pi^T_{l,3} \leq \sum_{F_l \in \mathcal{N}_T(F_l)} \pi_{l,1} \quad (3e)$$

where Eq. (3b) imposes that only one encoded version of the same view is scheduled; Eq. (3c) imposes the bandwidth constraint, and Eq. (3d) and Eq. (3e) force a predictive frame to be scheduled only if at least one SI is available at the decoder. Note that D_l is derived from Eq. (2) if F_l is not decodable, and $d(R^K)$ otherwise.

What makes the scheduling optimization above challenging, in terms of solving method, is the inter-dependency and the redundancy that subsist among candidate DUs. The *coding-dependence* is imposed by the coding structure and it is such

[2]We have dropped the subscript (t, m) in favor of a general subscript l.

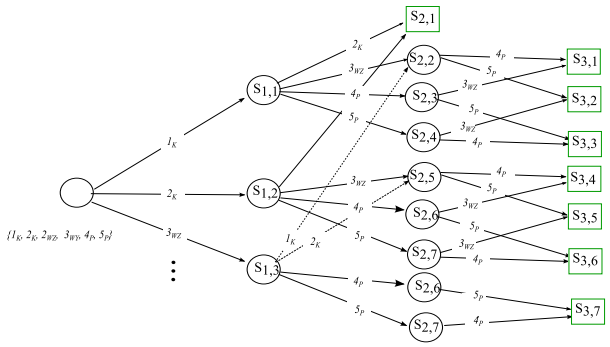


Figure 1. Trellis-Based Solution.

that a predictive frame can be decoded only if all key frames on which it depends can also be decoded. The *reward-dependence* is rather coming from the correlation among neighboring key frames. Since a key frame can help in the reconstruction of missing ones, if correlated, the reward of scheduling a DU is not known a priori, but it depends on the scheduling policy of the correlated DUs.

Because of coding- and reward-dependence, the greedy optimization in (3) cannot be solved by conventional optimization frameworks. Solutions proposed in [1], [6] could be adopted to optimize scheduling problem in the case of coding-dependence, but they do not address the reward-dependence. Although a formal scheduling optimization has been posed for redundant DUs in [7], computational complexity remains an open issue. A viable solution for *reward-dependent* DUs is the trellis-based algorithm in [4], where a pruning-branches technique for reducing the complexity is proposed. The pruning is performed in such a way that only the most innovative DUs (i.e., least correlated to the already scheduled frames) are left as survivors. However, when DUs are not homogeneous (i.e., not all key frames) the level of innovation can be not comparable between key and predictive frames.

Thus, the solving method to optimize the scheduling policy in multi-view systems is still a challenging problem. Here, we propose a trellis-based solution which allows to reach *optimality* reducing at the same time the computational complexity. The heterogeneity of the DUs enables us to express scheduling rules in the trellis constructions. These rules provide us with an elegant structure to decouple reward-dependent DUs (key frames) from the reward-independent ones (predictive frames), thereby significantly reducing the computation complexity.

IV. Proposed Solving Method

We start from an initial state S_0, characterized by a set of candidate DUs. We then construct a trellis, as depicted in Fig. 1, where each branch is an action (i.e., a DU scheduled). Let $\mathcal{A}(S_{i,k})$ be the set of feasible actions that can be taken from the node $S_{i,k}$ (i.e., set of possible DUs to schedule at $S_{i,k}$), which is the k-th among the nodes in the i-th column (corresponding to i DUs scheduled). Each action has a cost (size of the scheduled DU) and a reward in terms of distortion gain $\delta(a_i)$,

Algorithm 1 OPT-p optimization

Init: Let $\mathcal{A}^p$ be the set of candidate predictive DUs. Let c_l and $\delta(a_l)$ be the transmission cost and reward, respectively, of DU $l \in \mathcal{A}^p$. Let C^p be the the available BW.

Solve:

$$V_{opt} : \max_{\mathcal{T} \subseteq \mathcal{A}^p} \sum_{l \in \mathcal{T}} \delta(a_l) \tag{5}$$

$$\text{s.t.} \sum_{l \in \mathcal{T}} c_l \leq C^p \tag{6}$$

derived as difference of the distortion $\overline{D}_{\boldsymbol{\pi}}$ with and without the scheduling action a_i. An action $a_i \in \mathcal{A}(S_{i,k})$ taken from state $S_{i,k}$ leads to a successor state $S_{i+1,j}$ and we denote this by $(S_{i+1,j}|S_{i,k}, a_i)$. Each node $S_{i,k}$ is characterized by $\mathcal{A}(S_{i,k})$, the value function $V_{i,k}$, and the remaining channel bandwidth $C(S_{i+1,j})$, evaluated as C minus the transmission cost of the decisions taken along the path from S_0 to $S_{i,k}$. If the remaining channel bandwidth is zero, the state is a final state. Moreover $\mathcal{A}(S_{i,k}) = \mathcal{A}^p(S_{i,k}) \cup \mathcal{A}^k(S_{i,k})$, where $\mathcal{A}^p(S_{i,k})$ and $\mathcal{A}^k(S_{i,k})$ are the set of predictive and key candidate DUs, respectively.

The full-path (going from S_0 to a final state) which leads to the maximum distortion gain is the best set of DUs to be scheduled. From the Bellman's equations, the optimal solution can be found by backward induction as follows

$$V_{i,k} = \max_{a_i \in \mathcal{A}(S_{i,k})} \left\{ \delta(a_l) + V\left(S_{i+1,j}(S_{i,k}, a_i)\right) \right\}. \tag{4}$$

The problem is NP-hard and suffer of large computational complexity. We then imposes the following rules

Rule 1: If a_i is the scheduling of a predictive frame, then key frames cannot be scheduled in any successor state.

This rule avoids to construct redundant branches that would be pruned anyway, so optimality is still guaranteed. This is true since the order of the actions does not matter since selected DUs will be scheduled in the same transmission opportunity. For example, in Fig. 1, scheduling 1_K and then 3_{WZ} leads to the state $S_{2,2}$, which is the same that can be reached by scheduling 3_{WZ} first and 1_K after.

The ordered scheduling imposed by Rule 1 reduces redundancy among branches, but more importantly it allows us to *separate* reward-dependent DUs from reward-independent ones. Then,

Rule 2: If a_i schedules a predictive frame, then a_i and all successor states/actions are replaced by a single null action branch, leading to a final state with state value function $V_{opt}(S_{i+1})$. The latter is the results of the OPT-p optimization depicted in Algorithm 1.

Rule 2 allows to separate paths of predictive frames from the key ones. Since all DUs in $\mathcal{A}^p$ are reward-independent, the problem OPT-p can be easily solved by DP programming (e.g. knapsack problem), reducing the computational complexity.

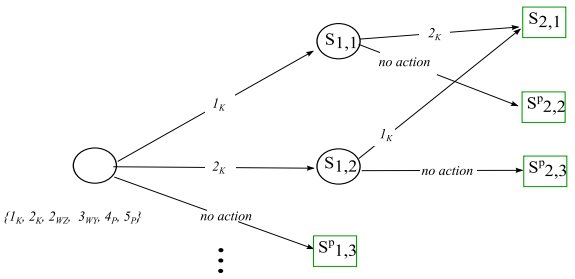
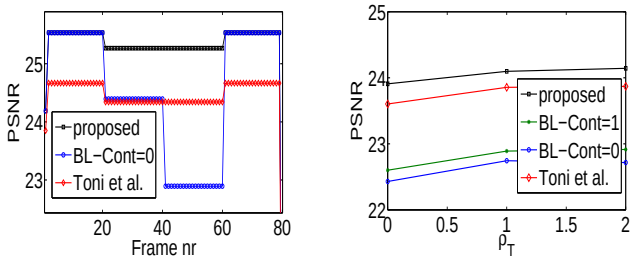


Figure 2. Trellis-Based Solution.



(a) $\rho_S = 4$ and $\rho_T = 0$. Static channel

(b) $\rho_S = 4$ for dynamic channel.

Figure 3. Mean PSNR results for different scenarios.

Hint on the proof of optimality: If a_i schedules a predictive frame, all future actions will schedule predictive frames only (Rule 1), which are all reward-dependent. From state $S_{i,k}$, the action a_i leads to the $S_{i+1,j}$ state, with a reward

$$V_{opt}(S_{i,k}) = \max_{a_i^p \in \mathcal{A}^p(S_{i,k})} \{\delta(a_i) + V\left(S_{i+1,j}(S_{i,k}, a_i^p)\right)\}$$
$$= \max_{a_i, a_{i+1}} \{\delta(a_i) + \delta(a_{i+1}) + V\left(S_{i+2,m}(S_{i,j}, a_i, a_{i+1})\right)\}$$
$$= \max_{\boldsymbol{a}} \left\{ \sum_{s=0}^{I} \delta(a_{i+s}) + \underbrace{V\left(S_{i+I,x}(S_{i,j}, \boldsymbol{a})\right)}_{0} \right\}$$

where $\boldsymbol{a} = [a_i, a_{i+1}, \ldots a_{i+I}]$ is the action vector, $S_{i+I,x}$ is the final state, and $V\left(S_{i+I,x}(S_{i,j}, \boldsymbol{a})\right) = 0$ since all final states in the original tree are set to 0. This leads to the optimization problem in (5), proving the optimality of Rule 2. □

We can then conclude that the trellis solution in Fig. 1 is equivalent to the one in Fig. 2, where only key frames can be considered as actions and any final state has a equivalent value function derived from the $OPT - WZ, P$ algorithm.

V. Results

Results are provided for a synthetic video sequence built with realistic correlation models both in temporal and spatial domain. We consider 16 views with a correlation model substantially varying every 20 frames. This creates a dynamic scene with a correlation model varying over time. At each time opportunity, 3 key frames (or the equivalent in predictive frames) can be scheduled in good channel conditions, while only 2 in bad conditions. Also 1 frame is acquired and expired every two transmission opportunities. Results are provided in terms of mean PSNR, weighted by camera popularity. The PSNR of the reconstructed scene is evaluated from the rate-distortion model described in Sec. II. We consider $d(R) = \mu\sigma^2\, 2^{-2R}$ where R is the number of bits per pixels, σ^2 is the spatial variance of the frame and μ is a constant depending on the source distribution.

Our optimization algorithm is compared with the following baseline algorithms: i) "BL - Cont=0": a priori selection of the coding scheme with no information neither about the channel nor about the correlation; ii) "BL - Cont=1": a priori selection of the coding scheme with no information about the channel but with information about the correlation; iii) "Toni et al.", scheduling optimization of only key frames introduced in [4].

In Fig. 3(b), the mean PSNR (average over views) vs frame index is depicted for the case of static channel (always good channel conditions are experienced) and $\rho_S = 4$ and $\rho_T = 0$ in the video sequence. Introduction of obstacles in the synthetic sequence is experienced every 20 frames. It can be observed that the proposed method is always able to outperform baseline algorithms.

In Fig. 3(b), we rather show the mean PSNR (averaged also over time) for a case of dynamic channel, in which at each transmission opportunity the channel has a 0.8 probability of changing state (from good to bad and viceversa). Also in this case we can see that the proposed solution is the best one.

VI. Conclusions

We have studied coding and scheduling strategies of redundant correlated sources in a multi-camera system. We have proposed a dynamic packet scheduling algorithm, which opportunistically optimizes the transmission policy based on the channel capacity and source correlation. Because of reward and coding dependence, conventional solving methods cannot be adopted in our work. We have then proposed a novel trellis-based solving method, able to decouple dependent and independent DUs in the trellis construction, reduces then the computational complexity but still reaching optimality. Simulation results have demonstrated the gain of the proposed method compared to classical resource allocation techniques.

References

[1] P. Chou and Z. Miao, "Rate-distortion optimized streaming of packetized media," *IEEE Trans. Multimedia*, vol. 8, no. 2, pp. 390 – 404, April 2006.

[2] J. Chakareski, "Transmission policy selection for multi-view content delivery over bandwidth constrained channels," *IEEE Trans. Image Processing*, vol. 23, no. 2, pp. 931–942, Feb 2014.

[3] P. Wang, R. Dai, and I. Akyildiz, "A spatial correlation-based image compression framework for wireless multimedia sensor networks," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 388 –401, Apr. 2011.

[4] L. Toni, T. Maugey, and P. Frossard, "Correlation-aware packet scheduling in multi-camera networks," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 496–509, Feb 2014.

[5] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *Proc. IEEE Int. Conf. on Image Processing*, Sept 2007.

[6] F. Fu and M. van der Schaar, "Structural solutions for dynamic scheduling in wireless multimedia transmission," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 5, pp. 727–739, May 2012.

[7] H. Wang and A. Ortega, "Rate-distortion optimized scheduling for redundant video representations," *IEEE Trans. Image Processing*, vol. 18, no. 2, pp. 225–240, Feb 2009.