Packet Scheduling in Multicamera Capture Systems

Laura Toni, Thomas Maugey, and Pascal Frossard

Signal Processing Laboratory (LTS4), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
{laura.toni, thomas.maugey, pascal.frossard}@epfl.ch

***Abstract*—In multiview video services, multiple cameras acquire the same scene from different perspectives, which results in correlated video streams. This generates large amounts of highly redundant data, which need to be properly handled 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 sent to a central server through a bottleneck channel, and eventually delivered to interactive clients. We propose a dynamic correlation-aware packet scheduling optimization under delay, bandwidth, and interactivity constraints. A novel trellis-based solution permits to formally decompose the multivariate optimization problem, thereby significantly reducing the computation complexity. Simulation results show the gain of the proposed algorithm compared to baseline 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, the multiview system should be able to deliver the requested viewpoints to each client. This might result in large storage/bandwidth requirements that are not always sustainable by the network. In order to provide effective video quality in resources constrained environments, opportunistic resource allocation strategies are essential.

Only a few works have studied the packet scheduling problem in multi-camera systems [1]–[3]. In [1], a spatial correlation model is proposed for static camera selection in sensor networks, while the work in [3] focuses on an adaptive correlation-aware packet scheduling algorithm, for a simplistic independent view coding framework. In this work, 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 consider an acquisition scenario in which multiple cameras acquire successive frames of the same scene but from different perspectives. Each camera stores the acquired frame in its short buffer in three different encoded versions: intra-coded (or key frame), temporally predicted frame (P frame) or distributively coded frame (i.e., Wyner-Ziv (WZ) frame). Theoretically, each acquired frame should then be sent from the cameras to a central server through a bottleneck channel before its deadline, imposed either by camera buffer limitations or decoding deadlines. However, network limitations might impose to send only a portion of the frames. The server gathers the received camera frames and possibly reconstructs the missing ones. The latter can be reconstructed from correlated neighboring views if available at the decoder. The contribution that each view can offer to missing ones is estimated by each view knowing the neighboring camera position. This is the only minimal information exchanged among cameras. No visual information is exchanged among cameras due to the system settings or resource limitations, which prevents the use of multiview video coding [4] to encode the video information. Finally, the server eventually serves the clients requests, see Fig. 1. Our objective is to maximize the amount of information at the server, such that clients can be served with high quality images. This is possible by optimizing the scheduling policy such that the quality in the reconstruction of the multi-camera data is maximized while both bandwidth and time constraints are met.

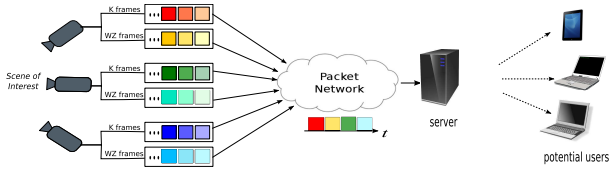

Figure 1. Multi-camera system with bottleneck network.

We propose, a *dynamic correlation-aware packet scheduling algorithm* for encoded packets for multi-camera system in bandwidth-limited networks. Differently than other packet scheduling algorithms, our framework considers the correlation between cameras in a novel *correlation-based rate distortion (RD)* model. Then, the coding structure is dynamically adapted as a result of the packet scheduling optimization that depends on the channel conditions, the content information as well as the user preferences. We propose a *novel solving algorithm* that is able to reduce the computational complexity by decomposing the multivariate scheduling optimization problem, while preserving the optimality of the solution. Simulation results demonstrate that the proposed dynamic scheduling algorithm outperforms scheduling policies with a static coding strategy and agnostic transmission schemes. We show that including information about interactive clients in the problem formulation leads to an improvement in terms of perceived quality w.r.t. baseline algorithms. Simulation results also show the limitation of static encoding strategies,

IEEE VCIP'14, Dec. 7 - Dec. 10, 2014, Valletta, Malta.
978-1-4799-6139-9/14/$31.00 2014 IEEE.

mainly in highly constrained scenarios. These results open the path to novel navigation-based transmission strategies optimizations that take into account not only the source and channel information but also the behavior of interactive clients.

II. Multi-view Video Framework

We consider a multicamera system with *correlated sources*. Acquired frames can be correlated among each other both in time (one camera acquiring temporally consecutive frames) and in space (neighboring cameras might acquire overlapping portions of the same scene). Both the temporal and the spatial correlations might help in estimating one frame from neighboring ones. In particular, let denote by $F_{t,m}$ the frame acquired from the m-th camera at time t. Frame $F_{t,m}$ can be estimated from frame $F_{\tau,l}$ via depth-image based rendering (DIBR). Typically, DIBR algorithms use depth information in order to estimate by projection the position of pixels from view k in the missing view n. The projected pixels are generally of good precision but do not cover the whole estimated image, due to visual occlusions. We denote by $\rho(F_{t,m}|F_{\tau,l})$ the portion of the image $F_{t,m}$ that can be estimated (i.e., not occluded) by $F_{\tau,l}$, which can be neighboring in either the temporal or the spatial dimension. This corresponds exactly to the level of correlation between $F_{t,m}$ and $F_{\tau,l}$, denoted by $\rho(F_{t,m}|F_{t,l})$ and ranging from 0 to 1.

On the *camera side*, each frame is encoded as a key or predictive (i.e., WZ and P) frame. The key frame is encoded at rate R^K. Predictive frames are encoded with no exact information on which side information (SI) will be available at the decoder.[1] Let $\mathcal{N}_S(F_{t,m})$ and $\mathcal{N}_T(F_{t,m})$ be the set of candidate SIs for WZ and P frames, respectively, defined as the set of cameras such that the correlation level with $F_{t,m}$ is greater than a pre-imposed threshold. We then encode predictive frames at a rate which guarantees the decoding even in the worst-case scenario, in which the SI is the least correlated view among the ones in $\mathcal{N}_S(F_{t,m})$ and $\mathcal{N}_T(F_{t,m})$. This means that WZ and P versions, respectively, are encoded at a rate of

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

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

On the *receiver side*, each received key frame is decoded independently, while received predictive frame is decodable if there is at least a key frame available as SI. Key frames and decodable predictive frames can be decoded at a distortion of $d(R^K)$. Not transmitted or not decodable images are recovered from the neighboring key frames available at the receiver by DIBR techniques.[2] The distortion at which missing frames are reconstructed depends on the correlation level as follows

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

where $d_{\max}$ is the maximum distortion at which occluded areas are reconstructed (e.g., impainting distortion), and $\boldsymbol{\chi}$ is the set of key frames available at the decoder.

III. Packet Scheduling Optimization

We now propose a novel problem formulation for rate-distortion optimal packet scheduling.

Each image acquired at a given time instant from a particular camera is packetized into multiple data units (DUs) (one per encoded version), and stored in the camera buffer.[3] Lossless transmissions are considered, such that scheduled packets are available at the decoder. At each transmission opportunity τ, the scheduler decides the best set of DUs to schedule. Let F_l be a generic view, acquired at $T_{\mathrm{A},l}$ and expiring at $T_{\mathrm{TS},l}$.[4] 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 views for being sent at τ as $\mathcal{L} = \{F_l \text{ s.t. } T_{\mathrm{A},l} \le \tau \le T_{\mathrm{TS},l}\}$. Theoretically, the encoded versions of any view in $\mathcal{L}$ are candidate DUs for being scheduled. In practice, we impose the following scheduling policies: i) only one version among WZ, P, and key frames of the same view can be scheduled; ii) a predictive frame can be scheduled only if at least one SI frame is already available at the decoder. Note that both channel conditions and content models may vary over time, leading to different scheduling policies at different transmission opportunities. Thus, in order to have a dynamic transmission strategy that constantly adapts to both the content and the network, 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

$$\min_{\boldsymbol{\pi}} \overline{D}_{\boldsymbol{\pi}} = \sum_{l:T_{\mathrm{A},l}\le\tau\le T_{\mathrm{D}}} P_l D_l(\boldsymbol{\pi}|\boldsymbol{\chi}) \quad (3a)$$

$$\text{s.t. } \sum_i \pi_{l,i} \le 1, \quad \forall l \quad (3b)$$

$$\sum_l \pi_{l,1} R^{(K)}_l + \pi_{l,2} R^{(WZ)}_l + \pi_{l,r} R^{(P)}_l \le C \quad (3c)$$

$$\pi^T_{l,2} \le \sum_{F_l\in\mathcal{N}_S(F_l)} \pi_{l,1} \quad (3d)$$

$$\pi^T_{l,3} \le \sum_{F_l\in\mathcal{N}_T(F_l)} \pi_{l,1} \quad (3e)$$

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

[1] Knowing the scheduling policy, we could know a priori which SI will be sent to the server. However, the scheduling policy is optimized in real-time so it is not known a priori at the encoder side.

[2] To avoid error propagation, we assume that missing frames can be reconstructed from key frames only.

[3] 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.

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

Algorithm 1 OPT-p optimization

Init: Let $\mathcal{A}^p$ be the set of actions of 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 available BW.

Solve:

$$V_{opt} : \max_{\mathcal{T} \subseteq \mathcal{A}^p} \sum_{l \in \mathcal{T}} \delta(a_l) \tag{4}$$

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

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 it is equal to $d(R^K)$ otherwise.

What makes the scheduling optimization above challenging, in terms of solving method, is both the inter-dependency and the redundancy that subsist among candidate DUs. The *coding-dependency* is imposed by the coding structure and it is such that a predictive frame can be decoded only if at least one SI frame can also be decoded. The *reward-dependency* is rather coming from the correlation among neighboring key frames. Since a scheduled key frame can reconstruct missing frames, the reward of scheduling a key DU is not limited to the distortion gain of the key frame, but it extends to the distortion gain of reconstructed missing frames. However, the reward of scheduling a key DU is not known a priori, but it depends on the scheduling policy of the correlated DUs.

Because of coding- and reward-dependency, the optimization in Eq. (3) cannot be solved by conventional optimization frameworks. Solutions proposed in [5], [6] could be adopted in the case of coding-dependency, but they do not address the reward-dependency. 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 proposed in [3], where branches in the trellis are pruned to reduce the complexity. The pruning is performed in such a way that only the most innovative DUs (i.e., least correlated to the previously scheduled frames) are left as candidate for transmission. However, this pruning applies only among key candidate DUs and not among key and predictive candidate frames, as in this work.

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 construction. 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 now describe the trellis-based solution proposed to solve the optimization problem in Eq. (3). We start from an initial

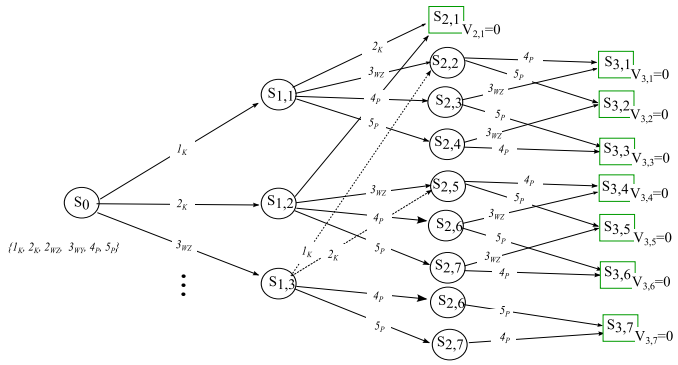

Figure 2. Trellis-Based Solution.

state (or node) S_0, characterized by the initial set of DUs candidate for being transmitted at time τ. We then construct a trellis, as depicted in Fig. 2, where each branch is a possible action (i.e., a scheduled DU). 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 node among the ones in the i-th column (corresponding to i DUs scheduled). Each action a_i has a cost (size of the scheduled DU) and a reward in terms of distortion gain $\delta(a_i)$, derived as the difference of the distortion $\overline{D_{\boldsymbol{\pi}}}$ before and after scheduling the action a_i. An action $a_i \in \mathcal{A}(S_{i,k})$ taken from the state $S_{i,k}$ leads to a successor state $S_{i+1,j}$; we denote this transition 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 $C(S_{i+1,j})$. The latter is the bandwidth still available at $S_{i+1,j}$ and it is evaluated as C minus the transmission cost of the decisions taken along the path from S_0 to $S_{i+1,j}$. If the remaining channel bandwidth is zero, the state is a final state and the value function is set to 0. Moreover, we define $\mathcal{A}^p(S_{i,k})$ and $\mathcal{A}^k(S_{i,k})$ as the set of predictive and key candidate DUs, respectively, such that $\mathcal{A}(S_{i,k}) = \mathcal{A}^p(S_{i,k}) \cup \mathcal{A}^k(S_{i,k})$.

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 [8]

$$V(S_{i,k}) = \max_{a_i \in \mathcal{A}(S_{i,k})} \{\delta(a_l) + V(S_{i+1,j}|S_{i,k}, a_i)\}. \tag{5}$$

The problem is NP-hard and suffers of large computational complexity, exponentially growing with C and the cardinality of set of candidate DUs. We then impose the following rules

Rule 1: If a_i is the action of scheduling 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 actions along a full paths represent DUs that will be scheduled in the same transmission opportunity, thus the order of the actions does not matter. For example, in Fig. 2, scheduling 1_K and then 3_{WZ} leads to the state $S_{2,2}$, which is

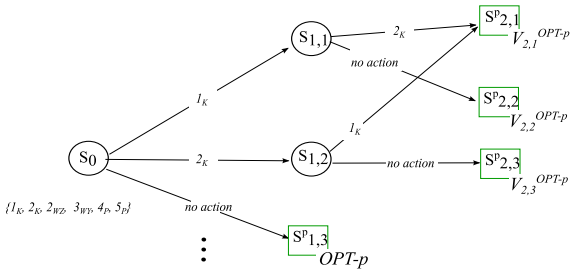

Figure 3. Trellis-Based Solution.

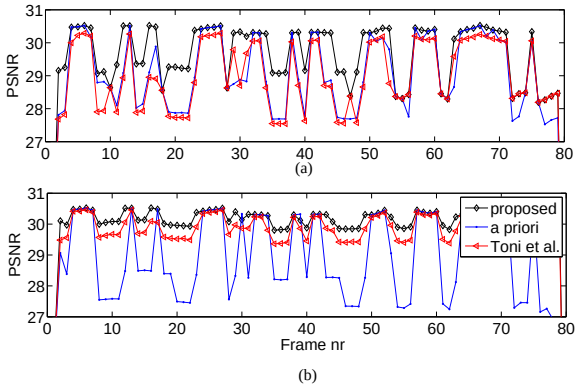

Figure 4. Mean PSNR results for different scenarios. (a) $\rho_S = 2$, (b) $\rho_S = 4$.

the same that can be reached by scheduling 3_{WZ} first and 1_K after. Rule 1 limits the redundancy among branches, but more importantly it allows us to *separate* branches with reward-dependent DUs from branches with reward-independent ones. Then we have the following rule:

Rule 2: If a_i schedules a predictive frame, then a_i and all successor states/actions are equivalently 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 solved by DP programming (e.g., knapsack problem [8]), thereby reducing the computational complexity to $\mathcal{O}(C^p|\mathcal{A}^p|)$. Trellis solution in Fig. 2 is then equivalent to the one in Fig. 3, where branches are constructed only for key actions in $\mathcal{A}^k$. Final states S^P have no more a value function of 0, they rather have an *equivalent value function* derived from the $OPT-p$ algorithm, which optimizes the scheduling of predictive candidate frames.

V. Results

Results are provided for "Ballet" video sequence [9]. A dynamic channel is considered such that in good and bad conditions, respectively, 3 and 2 key frames (or the equivalent in predictive frames) can be scheduled per transmission opportunity. The channel is modeled as a Markov model with transition probability from good to bad (and from bad to good) of 0.8. Every two transmission opportunities, cameras acquire a new frame, which is stored in the buffer, while previously acquired frames are discarded from the buffer. 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) "a priori": a priori selection of the coding scheme with no information neither about the channel nor about the correlation; ii) "Toni et al.", scheduling optimization of only key frames introduced in [3].

In Fig. 4, the mean PSNR (average over views) vs frame index is depicted for the case of maximum spatial correlation $\rho_S = 2$ (Fig. 4 (a)) and $\rho_S = 4$ (Fig. 4 (b)) with no temporal correlation. In both cases, the proposed algorithm outperforms baseline ones, although the gap between the proposed solution and the one in [3] reduces in Fig. 4 (b). This is justified by the fact that, for large ρ_S, missing frames are well reconstructed from received key frames.

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 the reward and coding dependency that subsists among frames, 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. This allows to reduce the computational complexity but still finding the optimal scheduling policy. Simulation results have demonstrated the gain of the proposed method compared to classical resource allocation techniques.

References

[1] 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.

[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] 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.

[4] 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.

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

[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.

[8] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein *et al.*, *Introduction to algorithms*. MIT press Cambridge, 2001, vol. 2.

[9] [Online]. Available: http://research.microsoft.com/en-us/um/people/sbkang/3dvideodownload/