J. Vis. Commun. Image R. 33 (2015) 255–264

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Optimal layered representation for adaptive interactive multiview video streaming ☆

Ana De Abreu [a,b,*], Laura Toni [a], Nikolaos Thomos [c], Thomas Maugey [d], Fernando Pereira [b], Pascal Frossard [a]

[a] Signal Processing Laboratory (LTS4), Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
[b] Instituto Superior Técnico, Universidade de Lisboa – Instituto de Telecomunicações (IST/UL-IT), 1049-001 Lisbon, Portugal
[c] University of Essex, Colchester, United Kingdom
[d] Inria Rennes Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes Cedex, France

ARTICLE INFO

Article history:
Received 15 April 2015
Accepted 21 September 2015
Available online 28 September 2015

Keywords:
Multiview video
Layered representation
Navigation window
Depth image based rendering (DIBR)
Depth maps
Texture
View synthesis
Dynamic programming

ABSTRACT

We consider an interactive multiview video streaming (IMVS) system where clients select their preferred viewpoint in a given navigation window. To provide high quality IMVS, many high quality views should be transmitted to the clients. However, this is not always possible due to the limited and heterogeneous capabilities of the clients. In this paper, we propose a novel adaptive IMVS solution based on a *layered multiview representation* where camera views are organized into layered subsets to match the different clients constraints. We formulate an optimization problem for the joint selection of the views subsets and their encoding rates. Then, we propose an optimal and a reduced computational complexity greedy algorithms, both based on dynamic-programming. Simulation results show the good performance of our novel algorithms compared to a baseline algorithm, proving that an effective IMVS adaptive solution should consider the scene content and the client capabilities and their preferences in navigation.

1. Introduction

In emerging multiview video applications an array of cameras captures the same 3D scene from different viewpoints in order to provide the clients with the capability of choosing among different views of the scene. Intermediate virtual views, not available from the set of captured views, can also be rendered at the decoder by depth-image-based rendering (DIBR) techniques [1], if texture information and depth information of neighboring views are available. As a result, *interactive multiview video* clients have the freedom of selecting a viewpoint from a set of captured and virtual views that define a navigation window. The quality of the rendered views in the navigation window depends on the quality of the captured views and on their relative distance, as the distortion of a virtual view tend to increase with the distance to the views used as references in the view synthesis process [2]. This means that in the ideal case, all the captured views, encoded at the highest possible rate, would be transmitted to all the clients. However, in practice, resource constraints prevent the transmission of all the views. In particular, clients may have different access link bandwidth capabilities, and some of them may not be able to receive all the captured views. In this context, it becomes important to devise adaptive transmission strategies for interactive multiview video streaming (IMVS) systems that adapt to the capabilities of the clients.

In this work, we consider the problem of jointly determining which views to transmit and at what encoding rate, such that the expected rendering quality in the navigation window is maximized under relevant resource constraints. In particular, we consider the scenario illustrated in Fig. 1, where a set of views are captured from an array of time-synchronized cameras. For each captured view, both a texture and a depth map are available, so that intermediate virtual viewpoints can eventually be synthesized. The set of captured and virtual views defines the navigation window available for client viewpoint request. Clients are clustered according to their bandwidth capabilities; for instance, in Fig. 1 only one client per cluster is illustrated for three groups with 1Mbps, 5Mbps and 10Mbps bandwidth constraints. Then, the set of captured views are organized in layers or subsets of views to be transmitted to the different groups of clients in order to maximize the overall navigation quality. With a layered organization of the captured

☆ This paper has been recommended for acceptance by M.T. Sun.

* Corresponding author at: Signal Processing Laboratory (LTS4), Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland.
E-mail addresses: ana.deabreu@epfl.ch (A. De Abreu), laura.toni@epfl.ch (L. Toni), nthomos@essex.ac.uk (N. Thomos), thomas.maugey@inria.fr (T. Maugey), fp@lx.it.pt (F. Pereira), pascal.frossard@epfl.ch (P. Frossard).

http://dx.doi.org/10.1016/j.jvcir.2015.09.010
1047-3203/

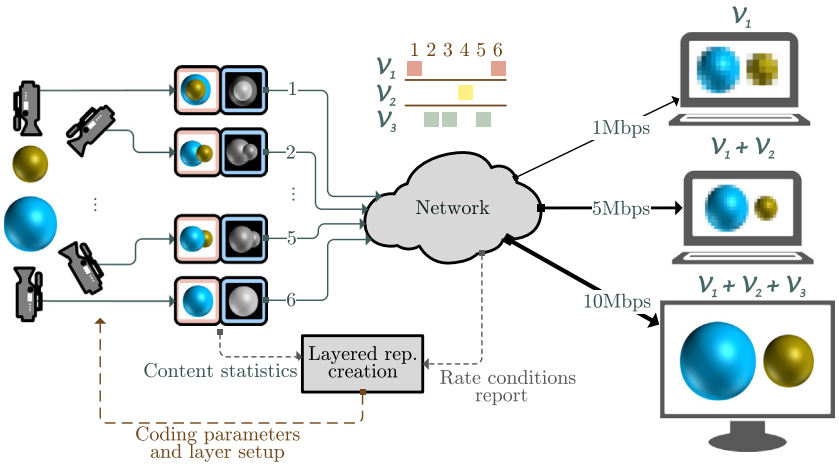

Fig. 1. Illustration of an IMVS system with 6 camera views and 3 heterogeneous clients. The optimization is done by the *layered representation creation* module considering three layers defined by the set of views $\{\mathcal{V}_1, \mathcal{V}_2, \mathcal{V}_3\}$.

views in the navigation window, we aim at offering a progressive increase of the rendering quality, as the quality of the navigation improves with the number of layers (subset of views) that clients are able to receive. In the example of Fig. 1, three layers or subsets of views are formed as: $\mathcal{V}_1 = \{1, 6\}, \mathcal{V}_2 = \{4\}$ and $\mathcal{V}_3 = \{2, 3, 5\}$. Depending on the clients' bandwidth capabilities, they receive the views in $\mathcal{V}_1$, or in $\mathcal{V}_1$ and $\mathcal{V}_2$, or in $\mathcal{V}_1, \mathcal{V}_2$ and $\mathcal{V}_3$. In particular, the client with the lowest bandwidth capability (i.e., the client with a mobile phone) is able to receive only the subset of views $\mathcal{V}_1$ in the first layer, and needs to synthesize the rest of the views. On the other hand, the client with the highest bandwidth capability (i.e., the client with a TV), is able to receive all the views, and therefore reaches the highest navigation quality.

We formulate an optimization problem to jointly determine the optimal arrangement of views in layers along with the coding rate of the views, such that the expected rendering quality is maximized in the navigation window, while the rate of each layer is constrained by network and clients capabilities. We show that this combinatorial optimization problem is NP-hard, meaning that it is computationally difficult and there are not known algorithm that optimally solves the problem in polynomial time. We then propose a globally optimal solution based on the dynamic-programing (DP) algorithm. As the computational complexity of this algorithm grows with the number of layers, a greedy and lower complexity algorithm is proposed, where the optimal subset of views and their coding rates are computed successively for each layer by a DP-based approach. The results show that our greedy algorithm achieves a close-to-optimal performance in terms of total expected distortion, and outperforms a distance-based view and rate selection strategy used as a baseline algorithm for layer construction.

This paper is organized as follows. First, Section 2 discusses the related work. Then, the main characteristics of the layered interactive multiview video representation are outlined in Section 3 where also our optimization problem is formulated. Section 4 describes the optimal and greedy views selection and rates allocation algorithms for our layered multiview representation. Section 5 presents the experimental results that show the benefits of the proposed solution and the conclusions are outlined in Section 6.

2. Related work

In this section, we review the work related to the design of IMVS systems by focusing on the problem of data representation and transmission in constrained resources environments.

In general, the limited bandwidth problem in IMVS has been approached by proposing some coding/prediction structure optimization mechanisms for the compression of multiview video data. In [3–5], the authors have studied the prediction structures based on redundant P- and DSC-frames (distributed source coding) that facilitate a continuous view-switching by trading off the transmission rate and the storage capacity. To save transmission bandwidth, different interview prediction structures are proposed in [6] to code in different ways each multiview video dataset, in order to satisfy different rate-distortion requirements. In [7–9], a prediction structure selection mechanism has been proposed for minimal distortion view switching while trading off transmission rate and storage cost in the IMVS system.

A different coding solution to the limited bandwidth problem has been proposed in [10], where a *user dependent multiview video streaming for Multi-users* (UMSM) system is presented to reduce the transmission rate due to redundant transmission of overlapping frames in multi-user systems. In UMSM, the overlapping frames (potentially requested by two or more users) are encoded together and transmitted by multicast, while the non-overlapping frames are transmitted to each user by unicast. Differently, the authors in [11,12] tackle the problem of scarce transmission bandwidth by determining the best set of camera views for encoding and by efficiently distributing the available bits among texture and depth maps of the selected views, such that the visual distortion of reconstructed views is minimized given some rate constraints.

Although these works propose solutions to the constrained bandwidth problem in IMVS, they do not consider the bandwidth heterogeneity of the clients, and rather describe solutions that do not adapt to the different capabilities of the clients.

The adaptive content concept in multiview video has been mostly used in the coding context, where the problem of heterogeneous clients has been tackled via scalable multiview video coding. For instance, some extensions of the H.264/SVC standard [13] for traditional 2D video have been proposed in the literature for multiview video [14] [15]. In [16–18], the authors propose a joint view and rate adaptation solution for heterogeneous clients. Their solution is based on a wavelet multiview image codec that produces a scalable bitstream from which different subsets can be extracted and decoded at various bitrates in order to satisfy different clients bandwidth capabilities.

In addition, multiview video permits the introduction of a new type of adaptive content compared to classical video. For instance, instead of transmitting the complete set of views of the multiview