R-D OPTIMIZED AUXILIARY INFORMATION FOR INPAINTING-BASED VIEW SYNTHESIS

Ismael Daribo, Gene Cheung

National Institute of Informatics (NII)
Tokyo, Japan

Thomas Maugey, Pascal Frossard

Signal Processing Laboratory (LTS4)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland

ABSTRACT

Texture and depth maps of two neighboring camera viewpoints are typically required for synthesis of intermediate virtual views via depth-image-based rendering (DIBR). However, the bitrate overhead required for reconstruction of multiple texture and depth maps at decoder can be large. The performance of multiview video encoders such as MVC is limited by the simple fact that the chosen representation is inherently redundant: a texture or depth pixel visible from both camera viewpoints is represented twice. In this paper, we propose an alternative 3D scene representation without such redundancy, yet at decoder, one can still reconstruct texture and depth maps of two camera viewpoints sufficient for DIBR-based synthesis of virtual intermediate views. In particular, we propose to first encode texture and depth maps of a single viewpoint, which are used to synthesize the uncoded viewpoint via DIBR at decoder. Then, we encode additional rate-distortion (RD) optimal auxiliary information (AI) to guide an inpainting-based hole-filling algorithm at decoder and complete the missing information due to disocclusion. For a missing pixel patch, the AI can: i) be skipped and let the decoder retrieve the missing information by itself, ii) identify a suitable spatial region in the reconstructed view for patch-matching, or iii) explicitly encode missing pixel patch if no satisfactory patch can be found in the reconstructed view. Experimental results show that our alternative representation can achieve up to 41% bit-savings compared to H.264/MVC implementation.

Index Terms — Texture-plus-depth format, depth-image-based rendering, compact representation

1. INTRODUCTION

Recently, the interest in end-to-end 3D video communication services is increasing rapidly and has led to interactivity and 3D perception improvements in related applications, including Three-Dimensional Television (3D-TV) and Free Viewpoint Television (FTV). This breakthrough has been aided by the recent development of auto-stereoscopic displays, multi-camera-captured systems and depth acquisition technologies. In particular, Multiview Video (MVV) communication systems can provide user navigation with a look-around sensation by view synthesis via depth-image-based rendering (DIBR) at decoder [1]. Views that are not captured from a real camera can be synthesized using texture and depth maps of two neighboring camera-captured views. With DIBR-based view synthesis, only texture and depth videos of a subset of *reference* views are needed at decoder for reconstruction of all intermediate virtual views used for smooth view transition.

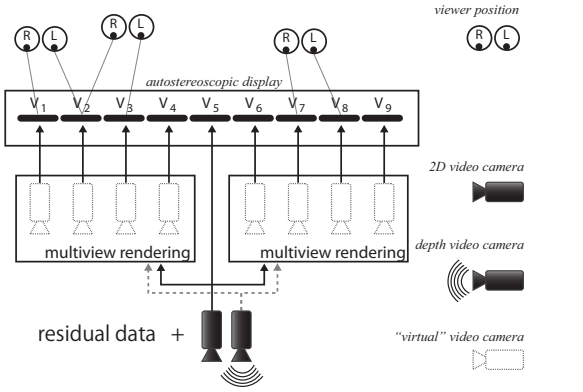


Figure 1. Proposed interactive multiview imaging system at decoder. *Virtual* views are synthesized around the camera-captured views by DIBR assisted with residual data.

To reduce the bitrate required for reconstruction of multiple texture maps from multiple viewpoints at decoder, multiview video coding (MVC) schemes [2] encode them using disparity compensation to exploit cross-view correlation. While MVC has shown coding gain over more naïve independent view coding approaches, its performance is limited by the simple fact that the chosen representation is inherently redundant[1]: namely, a texture or depth pixel visible from two reference views is represented twice.

In this paper, we propose an alternative 3D scene representation, one without pixel redundancy, that encodes reference frames for DIBR-based synthesis of intermediate views at decoder. In particular, we propose to first encode texture and depth maps of a single viewpoint, which are used to synthesize a second uncoded viewpoint via DIBR at decoder. Then, we design *auxiliary information* (AI) that is used to guide an inpainting-based hole-filling[2] algorithm [5] at decoder. The idea here is that hole-filling algorithm can maximally exploit *non-local* but correlated pixel patches in the reconstructed image to complete missing pixels due to disocclusion. We selectively encode AI in a rate-distortion (RD) optimal way. Specifically, for a given missing pixel patch in the synthesized view, the AI can: i) be skipped and let the decoder by itself retrieve the missing information, ii) identify a suitable spatial region in the reconstructed view for patch-matching, or iii) ex-

[1]For reference viewpoints that are close to each other, the intensity difference of the same pixel viewed from different viewpoints for most objects is likely small. Further, it is not always true that encoding the pixel difference contributes to view synthesis quality during pixel blending.

[2]Unlike typical 2D image inpainting scenarios, partial 3D geometric information (depth map) can be exploited during pixel-filling [3, 4].

plicitly encode missing pixel patch if no satisfactory patch can be found in the reconstructed view. Experimental results show that our alternative representation can reduced bitrate by up to 41% compared to MVC for the same synthesized view quality.

The outline of the paper is as follows. We first discuss related work in Section 2. We then overview our interactive multiview image system in Section 3. We discuss how AI are designed and selected in an RD-optimal manner in Section 4. Finally, experimental results and conclusions are presented in Section 5 and 6, respectively.

2. RELATED WORK

From a representation perspective, the most related work in the literature is the *layered depth video* (LDV) representation [6], where texture and depth maps of a single viewpoint is first encoded as the main layer, then occluded spatial regions in other camera viewpoints are added as enhancement layers. We first note that LDV, like our proposed representation, also avoids the pixel representation redundancy problem in MVC. However, we differ from LDV in the following aspects. First, we use a hole-filling algorithm[3] to complete missing pixels in the projected anchor view, while LDV typically used traditional coding tools based on transform plus entropy coding to explicitly encode disoccluded regions. Second, we design and employ RD-optimal AI to guide the hole-filling algorithm to further improve quality of the synthsized reference view. In the experimental section, we will show the performance gain of our scheme against LDV.

From a methodology perspective, the most similar work is an image compression algorithm in [8], where *assistant information* (edges in a code block) was encoded to aid a decoder edge-based inpainting scheme to reconstruct missing blocks. Though similar in spirit to our proposed AI, our proposal differs in the following aspect. First, our AI can provide location information to guide a *non-local* exemplar-based hole-filling algorithm to a spatial region with similar textural patches. In contrast, assistant information in [8] provides only edge information, which is used only for a local structural inpainting method that uses prior assumptions about the smoothness of the structures in the missing regions to propagate boundary data. It has been shown that non-local textural exemplar-based inpainting methods [5] often outperform local structural methods when the smoothness assumption is not necessarily valid. Second, unlike block-based image coding, a disoccluded patch can be of arbitrary shape, so in the case when it is not possible for a hole-filling algorithm to locate a satisfactory similar patch, we efficiently encode the arbitrarily shaped pixel patch using the Graph-Based Transform (GBT) [9].

3. PROPOSED SYSTEM

3.1. Encoder/Decoder Communication

In the proposed interactive multiview communication system, a user freely navigates from currently observed views to selected neighboring views. If the requested view is itself a reference view, the full view is explicitly encoded and transmitted. Otherwise, the user is first given the closest reference view (color and depth maps). Then, RD-optimal AI are additionally transmitted by encoder, so that another reference view—one where the requested virtual view becomes an intermediate view between the two references—can be constructed via DIBR using texture and depth maps of the first reference, plus hole-filling guided by the transmitted AI. The desired virtual view is finally synthesized via DIBR using the two constructed reference views. It is important to note that the proposed RD optimization of the AI at encoder ought to result in good quality for the continuum of virtual views between the two references, and not just for a particular virtual view synthesized at decoder.

In the next section, let us review the hole-filling method based on the well-known Criminisi's algorithm [5]. Though we chose this specific implementation of exemplar-based techniques for concreteness, it is important to note that our proposed optimization framework extends beyond this specific scheme.

[3] In our earlier work [7], an inpainting algorithm was used in a straightforward manner for hole-filing in the projected view, but no RD-optimal AI was designed and deployed.

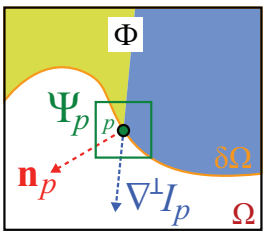


Figure 2. Notation diagram of an exemplar-based inpainting technique (from [5]). The current patch Ψ_p to be filled in, centered at the point p on the boundary $\delta\Omega$, is overlaying over the missing region Ω and the source region Φ.

3.2. Inpainting-based Hole-Filling

Because there often exist recurring patterns in pixel patches across a typical image, one solution is to search and identify exemplar-matching patches in order to fill in missing pixels. Criminisi *et al.* [5] first reported that exemplar-based texture synthesis contains the process necessary to replicate both texture and structure.

With input image I and missing region Ω, the source region Φ is defined as $\Phi = I - \Omega$, and the boundary of the missing region is indicated by $\delta\Omega$ as illustrated in Fig. 2. For every patch Ψ_p centered at the point p, where $p \in \delta\Omega$, the patch Ψ_p can be decomposed into two disjoint sub-regions such that

$$\Psi_p = \big(\Psi_p \cap \Phi\big) \cup \big(\Psi_p \cap \Omega\big) \quad \text{and} \quad \emptyset = \big(\Psi_p \cap \Phi\big) \cap \big(\Psi_p \cap \Omega\big) \quad (1)$$

where both $\Psi_p \cap \Phi$ and $\Psi_p \cap \Omega$ are known at the encoder, while the decoder only has knowledge of $\Psi_p \cap \Phi$.

3.2.1. Priority computation

It has been shown that the quality of the output image synthesis is greatly influenced by the order in which the inpainting is processed [5]. In addition, in the context of DIBR system, disocclusions are the result of displaced foreground object that reveals some background areas. Filling in the disoccluded regions using background pixels therefore makes more sense than foreground ones [4]. More priority is then given to patches that overlay regions where the depth variance is low, excluding regions at the foreground/background boundaries. The selection of the current patch to be filled in can be formulated as

$$\Psi_{p^*} = \arg\max_{p \in \delta\Omega} \Big\{ C(\Psi_p) \cdot D(\Psi_p) \cdot L(\Psi_p) \Big\}, \quad (2)$$

where C is the *confidence* term that indicates the reliability of the current patch, D is the *data* term that gives special priority to the isophote direction, and L is the *level regularity* term as the inverse square variance of the depth patch. For the sake of brevity, we

will not describe the different terms: for more details, the reader is referred to [4].

3.2.2. Patch matching

As originally defined by Criminisi *et al.* [5], once the highest priority patch Ψ_{p^*} is selected, a block matching algorithm derives the best exemplar patch Ψ_{q^*} to fill in the missing pixels under the patch Ψ_{p^*} such that

$$\Psi_{q^*} = \arg\min_{q\in\Phi} \left\{ d(\Psi_{p^*} \cap \Phi, \Psi_q \cap \Phi) \right\} \quad (3)$$

where the distance $d(.,.)$ is defined as the Sum of Squared Differences (SSD).

Having found the source exemplar Ψ_{q^*}, the value of each pixel-to-be-filled $p' \in \Psi_{p^*} \cap \Omega$ is copied from its corresponding pixel in Ψ_{q^*}. After the patch Ψ_{p^*} has been filled, the confidence term C(p) is updated as follows

$$C(p) = 1 \quad \forall p \in \Psi_{p^*} \cap \Omega. \quad (4)$$

4. DESIGN OF AUXILIARY INFORMATION

4.1. Types of AI

The solution of Eq. (3) can diverge, however. This is due to the fact that the minimization is done only on the sub-region $\Psi_{p^*} \cap \Phi$. To tackle this issue, we propose to assist the inpainting process with AI that prevents the aforementioned solution divergence. The proposed framework supports four different AI φ_p, where $\varphi_p \in \left\{\varphi_{\text{skip}}, \varphi_{\text{intra}}, \varphi_{\text{pred}}, \varphi_{\text{mv}}\right\}$ such that

- $\varphi_{\text{skip}} \equiv$ no information is sent. As a result, at the decoder side, the patch is classically inpainted by minimization of the distance function over the source sub-region $\Psi_{p^*} \cap \Phi$ as expressed in Eq. (3).
- $\varphi_{\text{intra}} \equiv$ the quantized transformed coefficients of the decoder-side-missing-regions $\Psi_p \cap \Omega$ are explicitly delivered directly to the decoder such that

  $$\varphi_{\text{intra}} := Q\big(\zeta\left(\Psi_p \cap \Omega\right)\big)$$

  where the transform domain function ζ represents the Graph-Based Transform (GBT) [9], which fits well the arbitrarily shaped region $\Psi_p \cap \Omega$. Q is a uniform quantization function.
- $\varphi_{\text{pred}} \equiv$ after inpainted prediction, such that the inpainting process at the decoder side is reproduced at the encoder, the quantized transformed coefficients of the remaining residual is sent as follows

  $$\varphi_{\text{pred}} := Q\big(\zeta\left(\Psi_{\text{res}} \cap \Omega\right)\big), \quad \text{with} \quad \Psi_{\text{res}} = \Psi_p - \Psi_{q^*}$$

  where

  $$\Psi_{q^*} = \arg\min_{q\in\Phi} d(\Psi_p \cap \Phi, \Psi_q \cap \Phi)$$

  where the distance $d(.,.)$ is defined as the Sum of Squared Differences (SSD).
- $\varphi_{\text{mv}} \equiv$ in a more traditional way, the ground truth is fully utilized to compute the motion vector $\mathbf{mv}$ that minimizes the Lagrangian function cost such that $\varphi_{\text{mv}} := \mathbf{mv}^*$ with

  $$\mathbf{mv}^* = \arg\min_{p+\mathbf{mv}\in\Phi} \Big\{ d(\Psi_p, \Psi_{p+\mathbf{mv}}) + \lambda \cdot \mathrm{R}\left(\mathbf{mv}\right) \Big\},$$

  where all possible motion vectors are restrained within a search window.

At decoder side, we then propose to modify Eq. (3) to support the proposed AI as follows

$$\Psi_{q^*} = \begin{cases} \Psi^0_{q^*} & \text{if } \varphi_{p^*} = \varphi_{\text{skip}} \\ \zeta^{-1}\left(Q^{-1}\left(\varphi_{\text{intra}}\right)\right) & \text{if } \varphi_{p^*} = \varphi_{\text{intra}} \\ \Psi^0_{q^*} + \zeta^{-1}\left(Q^{-1}\left(\varphi_{\text{pred}}\right)\right) & \text{if } \varphi_{p^*} = \varphi_{\text{pred}} \\ \Psi_{p^*+\varphi_{\text{mv}}} & \text{if } \varphi_{p^*} = \varphi_{\text{mv}} \end{cases} \quad (5)$$

where the functions ζ^{-1} and Q^{-1} are the inverse GBT and quantization function, respectively. $\Psi^0_{q^*}$ being defined in Eq.(3), represents the selected patch in a traditional inpainting algorithm, *i.e.*, no AI is utilized.

4.2. RD Optimized Coding of AI

Given a delivered AI represented by $\boldsymbol{\varphi} = \left\{\varphi_p\right\}$, we propose to re-formulate the hole-filling problem in an RD manner as follows

$$\arg\min_{\Psi_p} \int_{\delta\Omega} \Big(\mathrm{SSD}\left(\Psi_p \cap \Omega \mid \varphi_p\right) + \lambda \cdot \mathrm{R}\left(\varphi_p\right) \Big) \mathrm{d}p \quad (6)$$

where at the location p the SSD measurement quantifies an estimate of the inpainted reconstructed quality of the missing regions, while R measures the bits needed to encode the AI φ_p that assists the inpainting process. Here, $\lambda \geq 0$ is the Lagrangian multiplier.

Under the assumption that both encoder and decoder are using the same inpainting algorithm, it is possible to RD-optimize AI $\boldsymbol{\varphi}$ being transmitted to the decoder, which will improve the overall reconstruction quality as described in Eq. (6). For a given quantization parameter qp, finding the optimal RD-driven AI $\boldsymbol{\varphi}$ can be formulated through the minimization of the following Lagrangian criterion:

$$\arg\min_{\boldsymbol{\varphi}=\{\varphi_p\}} \int_{\delta\Omega} \Big(\mathrm{SSD}\left(\Psi_p, \varphi_p \mid qp\right) + \lambda \cdot \mathrm{R}\left(\varphi_p \mid qp\right) \Big) \mathrm{d}p \quad (7)$$

with λ as defined in H.264 standard

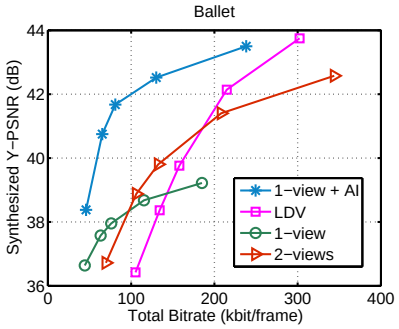
$$\lambda = 0.85 \cdot 2^{qp-12} \cdot 4$$

In addition, it is important to note that the ground truth of the missing-regions Ω is known at the encoder side.

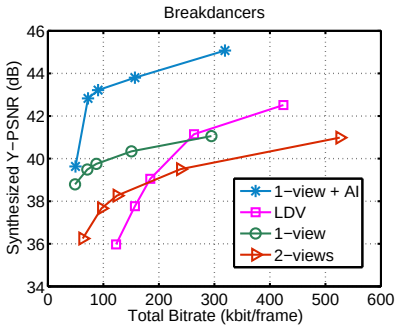
5. EXPERIMENTAL RESULTS

The performance of the proposed framework was evaluated using the multiview video dataset `Ballet` and `Breakdancers` (1024×768 @15 Hz) provided by Microsoft In the experiments, the camera 4 is used as anchor view, and the view 5 as synthesized one.

The comparison of objective compression performance is illustrated in the rate-distortion (RD) curves plotted in Fig. 3, where the peak signal-to-noise ratio (PSNR) of the synthesized texture video is plotted against bitrate (kbits/frame) over 100 frames. The RD results correspond to five qp quantization parameters: 24, 28, 32, 34, and 38. The bitrate consists of the sum of the anchor view rate plus the residual data rate. As shown in Fig. 3 we compare our proposed "1-view+AI" scheme against three others schemes: LDV [6], "1-view", "2-views". The proposed "1-view+AI" scheme consists in encoding one anchor view and AI to assist the hole-filling at decoder, as described previously. LDV corresponds to the specific case of sending only INTRA AI. The "1-view" scheme consists in sending only the anchor view, which is equivalent to delivering no AI (*i.e.*, SKIP mode). The "2-views"



(a)

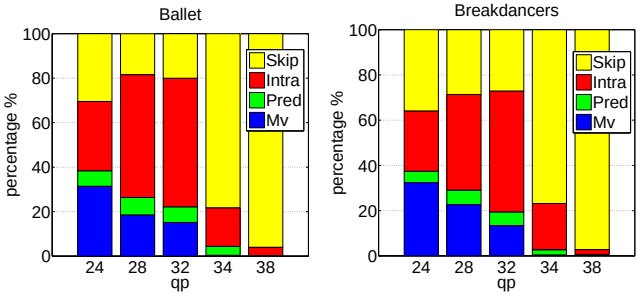


(b)

Figure 3. RD comparison of our proposed scheme "1-view+AI" against: LDV [6], "1-view" scheme where no AI is sent, and "2-views" scheme where two anchor views are sent.

scheme consists of explicitly sending the two closest anchor views. We used the implementation of H.264/MVC standard JMVC 7.0, to exploit the cross-layer and inter-view correlation in the LDV and "2-views" representation, respectively.

We see that our new compact representation "1-view+AI" results in significant compression gain. Specifically, an average bitrate reduction up to 41% and 35% for the multiview dataset `Ballet` and `Breakdancers`, respectively, are observed. It can be also observed in Fig. 4 that the average distribution of the different AI at different quantization parameter qp. As expected, at low bitrate (*i.e.*, high quantization parameter qp) the bitrate saving comes from the over selection of SKIP AI, while at high bitrate the motion vector AI gradually replaces the INTRA AI.



(a) `Ballet` (b) `Breakdancers`

Figure 4. AI mode distribution for different values of the quantization parameter qp.

6. CONCLUSION

In this paper we proposed an alternative 3D scene representation without pixel redundancy. We first encode texture and depth videos of a single view, which are used to synthesize a second reference view at decoder. Then, we encode additional RD-optimal auxiliary information (AI) to guide an inpainting-based hole-filling algorithm at decoder to complete missing information due to disocclusion. Experimental results show an overall bitrate reduction up to 41% over a classical H.264/MVC implementation.

7. REFERENCES

[1] Liu Zhan-wei, An Ping, Liu Su-xing, and Zhang Zhao-yang, "Arbitrary view generation based on DIBR," in *Proc. of the International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, 2007, pp. 168–171.

[2] P. Merkle, K. Müller, A. Smolic, and T. Wiegand, "Efficient compression of multi-view video exploiting inter-view dependencies based on H.264/MPEG4-AVC," in *Proc. of the IEEE International Conference on Multimedia and Expo (ICME)*, 9–12 July 2006, pp. 1717–1720.

[3] Kwan-Jung Oh, Sehoon Yea, and Yo-Sung Ho, "Hole filling method using depth based in-painting for view synthesis in free viewpoint television and 3-D video," in *Proc. of the Picture Coding Symposium (PCS)*, Chicago, IL, USA, May 2009, pp. 1–4.

[4] I. Daribo and B. Pesquet-Popescu, "Depth-aided image inpainting for novel view synthesis," in *Proc. of the IEEE Workshop on Multimedia Signal Processing (MMSP)*, oct. 2010, pp. 167 –170.

[5] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1200–1212, 2004.

[6] K. Muller, A. Smolic, K. Dix, P. Kauff, and T. Wiegand, "Reliability-based generation and view synthesis in layered depth video," in *Proc. of the IEEE Workshop on Multimedia Signal Processing (MMSP)*, Cairns, Queensland, Australia, Oct. 2008, pp. 34–39.

[7] I. Daribo and H. Saito, "A novel inpainting-based layered depth video for 3DTV," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 533–541, june 2011.

[8] Dong Liu, Xiaoyan Sun, Feng Wu, Shipeng Li, and Ya-Qin Zhang, "Image compression with edge-based inpainting," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 10, pp. 1273–1287, 2007.

[9] G. Shen, W.-S. Kim, S.K. Narang, A. Ortega, Jaejoon Lee, and Hocheon Wey, "Edge-adaptive transforms for efficient depth map coding," in *Proc. of the Picture Coding Symposium (PCS)*, Nagoya, Japan, Dec. 2010, pp. 566–569.