Signal Processing: *Image Communication* 24 (2009) 89–100

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

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

View synthesis prediction for multiview video coding

Sehoon Yea *, Anthony Vetro

MERL – Mitsubishi Electric Research Laboratories, Cambridge, MA, USA

ARTICLE INFO

Article history:
Received 8 October 2008
Accepted 19 October 2008

Keywords:
Multiview video coding
View synthesis
Prediction
Depth
3DTV
Free-viewpoint video

ABSTRACT

We propose a rate-distortion-optimized framework that incorporates view synthesis for improved prediction in multiview video coding. In the proposed scheme, auxiliary information, including depth data, is encoded and used at the decoder to generate the view synthesis prediction data. The proposed method employs optimal mode decision including view synthesis prediction, and sub-pixel reference matching to improve prediction accuracy of the view synthesis prediction. Novel variants of the skip and direct modes are also presented, which infer the depth and correction vector information from neighboring blocks in a synthesized reference picture to reduce the bits needed for the view synthesis prediction mode. We demonstrate two multiview video coding scenarios in which view synthesis prediction is employed. In the first scenario, the goal is to improve the coding efficiency of multiview video where block-based depths and correction vectors are encoded by CABAC in a lossless manner on a macroblock basis. A variable block-size depth/motion search algorithm is described. Experimental results demonstrate that view synthesis prediction does provide some coding gains when combined with disparity-compensated prediction. In the second scenario, the goal is to use view synthesis prediction for reducing rate overhead incurred by transmitting depth maps for improved support of 3DTV and free-viewpoint video applications. It is assumed that the complete depth map for each view is encoded separately from the multiview video and used at the receiver to generate intermediate views. We utilize this information for view synthesis prediction to improve overall coding efficiency. Experimental results show that the rate overhead incurred by coding depth maps of varying quality could be offset by utilizing the proposed view synthesis prediction techniques to reduce the bitrate required for coding multiview video.

1. Introduction

Emerging applications in multiview video such as free-viewpoint TV (FTV) [13,5,6], 3D displays [2], and high-performance imaging [14] require dramatic increase in bandwidth for their dissemination and make compression ever more important. Consequently, there are growing interests in coding techniques that take advantage of the correlation among neighboring camera views. In response to such needs and interests, recent multiview coding standardization activities by MPEG/JVT have been focused on developing generic coding toolsets geared mainly toward compression efficiency improvement by capitalizing on the inter-view correlation existing among views [12].

Disparity-compensated prediction (DCP) is a well-known technique for exploiting the redundancy between different views. This prediction mode provides gains when the temporal correlation is lower than the spatial correlation, e.g., due to occlusions, objects entering or leaving the scene, or fast motion. Martinian et al. [8] first proposed the use of view synthesis prediction (VSP) as an additional method of prediction in which a synthesized picture is generated from neighboring views using depth

* Corresponding author. Tel.: +1 617 621 7509.
E-mail addresses: yea@merl.com (S. Yea), avetro@merl.com (A. Vetro).
URL: http://www.merl.com/people/?user=yea (S. Yea).

0923-5965/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2008.10.007

information and used as a reference for prediction. This prediction mode is expected to be complementary to disparity compensation due to the existence of non-translational motion between camera views and provide gains when the camera parameters and estimated depth of the scene are accurate enough to provide high-quality synthetic views. A related VSP scheme by Shimizu et al. [11] proposes to encode view-dependent geometry information that enables a VSP.

In this paper, we first incorporate view synthesis into the block-based rate-distortion (RD)-optimization framework and describe a joint search algorithm for depth and correction vectors that are needed for high-quality view synthesis [15]. We encode this side information using CABAC and the rate overhead for this prediction mode is included in the final experimental results. We also introduce a novel extension of the skip and direct coding modes with respect to synthetic reference pictures. With these methods, we are able to infer the side information, thereby saving bits to generate the view synthesis reference data for a block, while maintaining prediction efficiency. Experimental results demonstrate VSP brings about additional coding gain when combined with disparity-compensated view prediction in multiview video coding.

Next, we demonstrate that the proposed VSP can also be a useful coding tool when generation of intermediate views is required at the receiver. For instance, to support auto-stereoscopic displays that require a higher number of views than transmitted, it is necessary to generate additional intermediate views. Intermediate views must also be generated in free viewpoint navigation scenarios. Given a discrete number of views with sufficient overlap, one synthesizes arbitrary intermediate views of interest using camera geometry and depth information; this scenario is commonly referred to as FTV.

From the coding perspective, this added requirement of generating intermediate views poses a challenging problem of efficiently compressing not only the multiview video itself, but also the associated depth maps of the scene. Since the requirement on the fidelity of the encoded depth maps will be often dictated by the expected rendering quality at the receiver side, it could imply a huge rate overhead in terms of coding and transmission. Therefore, it is desirable to be able to capitalize on the similarity between the multiview video and its associated depth maps for improved overall coding efficiency. In this context, we propose the (re-)use of encoded depth maps available both at the encoder and the decoder to improve coding efficiency of multiview video. In other words, the VSP technique utilizes the depth information that is already being encoded and transmitted; the depth is being made available to the receiver primarily for rendering purposes, but we utilize it for more efficient encoding. It is shown that the rate overhead incurred by coding high-quality depth maps can be offset by reducing the rate for coding multiview (texture) video with the proposed VSP technique. The results, however, also indicate that the amount of such rate reduction is not necessarily proportional to high rate overheads, e.g., when smaller quantization parameters (QPs) or sub-sampling ratios are used.

The rest of this paper is organized as follows. An overview of VSP is given in Section 2. In Section 3, we describe the RD-optimization framework including VSP. We also introduce the synthetic skip and direct modes that are natural extensions of the H.264/AVC standard. In Section 4, we discuss various issues related to generating and using side information such as depth and correction vectors. We then discuss issues and techniques for encoding the generated side information in Section 5. Experimental results are provided in Section 6 followed by concluding remarks in Section 7.

2. Overview of VSP

DCP typically utilizes a block-based disparity vector that provides the best matching reference position between a block in the current (predicted) view and reference view. In contrast, VSP attempts to utilize knowledge of the scene characteristics, including scene depth and camera parameters, to generate block-based reference data used for prediction. Fig. 1 illustrates a case where, in contrast with DCP which always maps every pixel in a rectangular MB in the predicted view to the corresponding pixel in the reference view displaced by the same amount using a single disparity vector, VSP maps to the matching pixels displaced not necessarily by the same amount and hence potentially provides better prediction when the matching area in the reference view

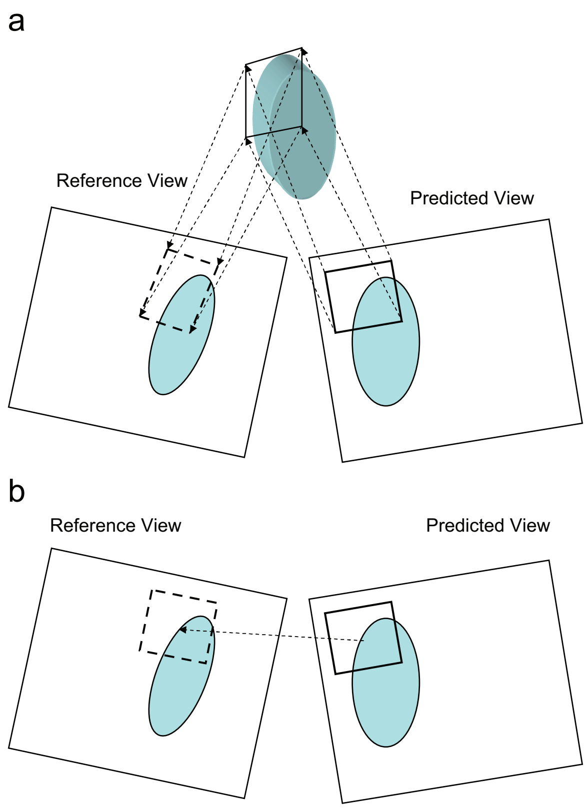

Fig. 1. Disparity compensated prediction (DCP, b) vs. view synthesis prediction (VSP, a).