J. Vis. Commun. Image R. 21 (2010) 474–486

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Multi-view video coding with view interpolation prediction for 2D camera arrays

Tae-Young Chung, Il-Lyong Jung, Kwanwoong Song, Chang-Su Kim *

School of Electrical Engineering, Korea University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 30 July 2009
Accepted 6 October 2009
Available online 20 October 2009

Keywords:
Multi-view video coding
Hierarchical B prediction
View interpolation
Bilateral criterion
2D camera array
3D-TV
H.264/AVC

ABSTRACT

An efficient compression algorithm for multi-view video sequences, which are captured by two-dimensional (2D) camera arrays, is proposed in this work. First, we propose a novel prediction structure, called three-dimensional hierarchical B prediction (3DHBP), which can efficiently reduce horizontal inter-view redundancies, vertical inter-view redundancies, and temporal redundancies in multi-view videos. Second, we develop a view interpolation scheme based on the bilateral disparity estimation. The interpolation scheme yields high quality view frames by adapting disparity estimation and compensation procedures using the information in neighboring frames. Simulation results demonstrate that the proposed multi-view video coding algorithm provides significantly better rate–distortion (R–D) performance than the conventional algorithm, by employing the 3DHBP structure and using interpolated view frames as additional reference frames.

1. Introduction

Recently, three-dimensional (3D) video technologies have been widely researched because of their potential for offering vivid reality and richer experience in various applications such as 3D-TV broadcasting, entertainment, education, and games [1]. For example, the European ATTEST project developed a 3D-TV system, which can support the backward compatibility with the standard 2D-TV using a layered coding scheme [2]. The base layer is encoded by the MPEG-2 standard for traditional 2D video systems, and the enhancement layer transmits additional depth and occlusion information. The ATTEST project also presented the notion of 3D processing chain, consisting of 3D content acquisition, coding, transmission, and display. A European consortium, 3D-TV Network of Excellence (NoE) also researched a 3D-TV system and expanded the 3D processing chain by identifying scene capture, representation, coding, transmission, and display as major components [3]. Vetro et al. [4] proposed another 3D-TV system using 16 high resolution multi-cameras and a multi-projector display.

To provide realistic 3D experience, a multi-view sequence supports the rendering of a scene from various viewpoints. However, a multi-view video sequence requires a large storage space or wide transmission bandwidth, since its data amount is proportional to the number of viewpoints. Many attempts have been made to encode a multi-view sequence efficiently to reduce the data amount. MPEG called for the evidence and requirements for achieving high coding gains for multi-view videos, and several multi-view video coding (MVC) algorithms have been proposed [5,6]. The Joint Video Team (JVT), which was organized by ISO/IEC MPEG and ITU-T VCEG, has standardized an efficient MVC technique [7]. Several prediction schemes such as the view synthesis prediction [8], the disparity vector prediction [9], and the hierarchical B prediction [10] have been proposed to exploit inter-view correlations in multi-view videos. Among them, JVT has adopted the hierarchical B prediction structure, which employs inter-view and intra-view predictions in a hierarchical manner, as the reference prediction structure in the Joint Multi-view Video Model (JMVM) [7].

View interpolation prediction is an efficient approach to exploiting inter-view correlations in MVC. A view frame is interpolated from neighboring view frames, and the interpolation error is encoded instead of the original frame, achieving a coding gain. In general, a view is interpolated based on the depth or disparity information. The depth information represents the distances of scene points from the imaging plane, and the disparity information represents the difference between a pixel location in a view frame and the corresponding pixel location in another view frame. The depth information can be reconstructed from the disparity information, and vice versa [11].

Martinian et al. [12,13] proposed a view interpolation method based on a pinhole camera model. For each candidate depth value, their method computes the sum of absolute differences (SAD) between a block in a view frame and the corresponding block in the other view frame, whose location is specified by the candidate depth. Then, the depth value giving the smallest SAD is used for the view interpolation. Based on the Martinian et al.'s interpolation method, Yea et al. proposed a depth estimation algorithm using

* Corresponding author. Fax: +82 2 921 0544.
E-mail addresses: lovelool17@korea.ac.kr (T.-Y. Chung), illyong@korea.ac.kr (I.-L. Jung), kwsong71@korea.ac.kr (K. Song), changsukim@korea.ac.kr (C.-S. Kim).

1047-3203/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2009.10.001

variable block sizes in [14], and developed an R–D optimized encoding algorithm including the view interpolation prediction mode in [15]. However, in [12–15], camera matrices, which represent the intrinsic and extrinsic properties of cameras, are required to estimate depth values, and thus the quality of interpolated view frames is sensitive to the fidelity of camera matrices. Özkalaycl et al. [16,17] proposed a dense depth-based interpolation method using Markov random field (MRF) modeling and 3D warping. Their 3D warping algorithm is similar to [12,13], but utilizes a regular mesh to reduce interpolation artifacts. Their algorithm, however, uses the belief propagation technique [18] to solve the MRF problem, which is computationally too complex to be used in most video coding applications. Droese et al. [19] proposed the ray-space interpolation method using pixel level disparities. Their method computes disparities by minimizing an energy function, which consists of a similarity term and a regularization term. Kitahara et al. [20] proposed an MVC algorithm based on the Droese et al.'s interpolation method. Also, Yamamoto et al. [21] employed the chrominance information as well as the luminance information in the disparity estimation to improve the estimation accuracy. However, the main disadvantage of [19–21] is that they require high computational complexity to estimate pixel level disparities.

According to the arrangement of cameras, multi-view videos can be categorized into 1D parallel, 1D arc, 1D convergent, 2D parallel, etc. Most previous work on MVC has been focused on 1D camera arrangements. However, 2D multi-view videos can provide more realistic experience than 1D multi-view videos. Several capturing systems for 2D multi-view videos have been developed. Yang et al. [22] developed an 8×8 light field camera system. Zhang et al. [23] developed a large multi-camera system, whose camera positions can be reconfigured. Wilburn et al. [24] constructed camera arrays in variable arrangements using 100 cameras to achieve high resolution imaging and hybrid aperture imaging. Tanimoto et al. [25] set up a 100 camera system, composed of 20 cameras in the horizontal direction and 5 cameras in the vertical direction. Recently, Cao et al. [26] also developed an 8×8 camera system. In contrast to the development of these 2D multi-view video capturing systems, there are few researches on 2D MVC. In general, the amount of data for a 2D multi-view sequence is even larger than that for a 1D multi-view sequence. It is therefore essential to develop an efficient coding technique for 2D multi-view sequences.

In this work, we propose a novel prediction structure, called 3DHBP, that can exploit horizontal and vertical inter-view correlations as well as temporal correlations in 2D multi-view videos. In addition, we develop an efficient view interpolation scheme, which estimates disparity vectors based on the bilateral criterion with an adaptive smoothness constraint and constructs high quality disparity-compensated frames. Simulation results demonstrate that, by incorporating the view interpolation scheme into the 3DHBP structure, the proposed MVC algorithm provides much better R–D performance than the conventional algorithm.

The rest of this paper is organized as follows. Section 2 reviews conventional prediction structures for MVC and proposes the 3DHBP structure for 2D multi-view videos. Section 3 describes the proposed view interpolation scheme, and Section 4 explains how to incorporate the interpolation scheme into the MVC algorithm. Section 5 discusses experimental results. Finally, we conclude the paper in Section 6.

2. Multi-view prediction structure

To compress multi-view video sequences compactly, it is necessary to develop an appropriate prediction structure. Several approaches, such as the group of GOP (GoGOP) structure, the sequential view prediction (SVP) structure, and the hierarchical B prediction structure, have been proposed [10,12,19,27,28]. Kimata et al. [27] proposed the GoGOP structure to exploit temporal and inter-view redundancies and to support low delay decoding. In the SVP structure [28], the first view in a multi-view sequence is encoded using the temporal prediction only. Then, the nth view is encoded using the prediction from the $(n-1)$th view as well as using the temporal prediction $(n = 2, 3, 4, \ldots)$. The hierarchical B prediction structure [10] provides better compression efficiency than the other structures in general, by employing inter-view and intra-view predictions hierarchically. In this section, we propose a novel prediction structure for 2D MVC based on the hierarchical B prediction.

2.1. Conventional hierarchical B prediction structure

The hierarchical B picture mode can compress video frames efficiently by applying bi-directional prediction hierarchically. In MVC with the hierarchical B prediction structure [10], a frame can be predicted from temporally adjacent frames in the same view and/or from spatially adjacent frames in neighboring views. Fig. 1 illustrates the hierarchical B prediction structure for a 1D 8-view video sequence, when the group of pictures (GOP) contains eight temporal frames in each view.

A view in the hierarchical B prediction structure can be classified into I-view, P-view, or B-view according to the type of its first frame, called the key frame, in the GOP. A key frame in an I-view is intra-coded without prediction from the other frames. A key frame in a P-view is predicted from a single neighboring view frame. A key frame in a B-view is bi-directionally predicted from two neighboring view frames. Note that non-key frames are encoded using the temporal prediction as well as the inter-view prediction.

Several extended structures of the hierarchical B prediction were proposed to encode 2D multi-view videos in [29]. These prediction structures use horizontal inter-view relationships efficiently, but do not fully exploit vertical inter-view relationships. Fig. 2(a) shows the one of the structures in [29], when cameras form a 5×9 array. In Fig. 2(a), X and Y denote the axes for horizontal and vertical indices of multi-view cameras, respectively, and T is the temporal axis. Note that the temporal hierarchical B prediction is maintained within each view, although it is not illustrated for the simplicity of the figure. Let $F_{x,y,t}$ denote the frame at time t, which is taken from the (x, y)th camera. $F_{0,0,t}$ is the only I-view, which is predicted in the temporal direction only. Except for $F_{0,0,t}$, in Fig. 2(a), the views in every

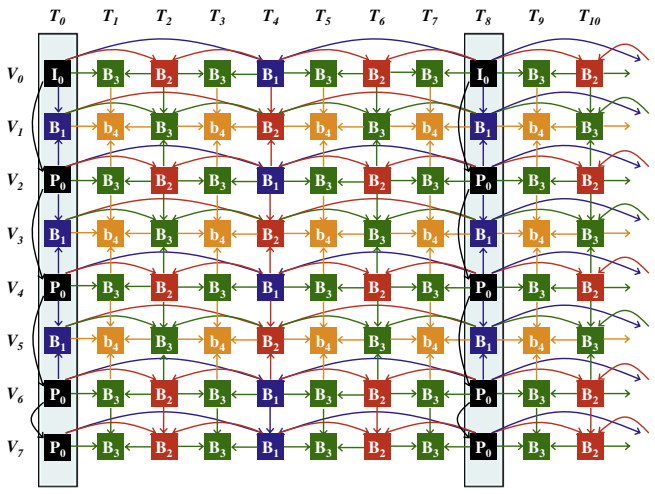

Fig. 1. The hierarchical B prediction structure [10] for a 1D 8-view sequence, when the GOP length is 8.