Neurocomputing 175 (2016) 544–554

Contents lists available at ScienceDirect

Neurocomputing

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

Efficient multi-view video coding using 3D motion estimation and virtual frame

Manoranjan Paul

Centre for Research in Complex Systems, School of Computing & Mathematics, Charles Sturt University, Bathurst, NSW 2795, Australia

ARTICLE INFO

Article history:
Received 23 October 2014
Received in revised form
19 June 2015
Accepted 27 October 2015
Communicated by Dr. Y. Yuan
Available online 4 November 2015

Keywords:
3D motion estimation
3D Coding
Uncovered background
Multiple reference frames
Hierarchical B-picture

ABSTRACT

Three dimensional (3D) i.e., multi-view extension of *High Efficiency Video Coding* (HEVC) standard provides a better compression compared to the simulcast coding (i.e., HEVC) technique by exploiting inter- and intra-view redundancy for multi-view video coding. However, this technique imposes a *random access frame delay* (RAFD) problem as well as requires huge computational time. In this paper three novel techniques are proposed to overcome the problems mentioned above. Firstly, a simulcast video coding technique is proposed where each view is encoded individually using hierarchical bi-predictive structure with extra virtual reference frame generated by dynamic background modeling (popularly known as McFIS – *the most common frame in a scene*) of the corresponding view. Secondly a novel technique is proposed using 3D *motion estimation* (3DME) where a 3D frame is formed using the same temporal frames of all views and motion estimation is carried out for the current 3D frame using dual 3D reference frames. Finally, a modification of 3DME dual reference frame technique (3DME-McFIS) is proposed where 3D McFIS is used as the second reference frame. Experimental results confirm that the proposed three techniques reduce the overall computational time and reduce the RAFD problem with comparable or better rate-distortion performance compared to the HEVC-3D extension. Specifically the proposed 3DME-McFIS technique outperforms the HEVC-3D coding standard by improving 0.90 dB PSNR on average, by reducing computational time by 50%, and by reducing RAFD problem compared to the existing HEVC-3D coding standard.

1. Introduction

Capturing a scene using multiple cameras from different angles is expected to provide the necessary interactivity in the *three-dimensional* (3D) space to satisfy end-users' demands for observing objects and actions from different angles and depths. Multi-view broadcast is becoming increasingly popular in commercial television networks for the added user-level interactivity. Multi-view also improves effectiveness of video surveillance systems. In addition to providing different perspectives, multiple views can offer a natural solution to the occlusion/de-occlusion problem [24,34,41], which often leads to incorrect object recognition and tracking. An occluded object is one that is obscured from view as there is another object in front of it. To encode multi-view video captured by multiple cameras, requires an efficient *multi-view video coding* (MVC) technique. MVC covers a wide range of active viewing experience, including stereoscopic (two-view) video (popularly known as 3DTV), free viewpoint television (2D view from any viewing angle can be generated from 3D scene modeling) and multi-view 3DTV. FutureSource predicts that 60% of US households will have a 3DTV by 2015 and 75% of households will have 3D Blu-ray players in a similar time frame [29]. More importantly, 3DTV adoption is outpacing HDTV adoption by around 50%. However, 3DTV viewer satisfaction does not meet customer expectations due to lack of content and interactivity. Otherwise, the 3DTV adoption rate would have been much higher. Considering the significant overlapping of the views and, more importantly, the availability of a rich set of relations on the geometric properties of a pair of views from camera properties, known as the epipolar geometry [1], joint encoding/decoding of views can achieve significant compression by exploiting inter-view correlation, in addition to the traditional intra-view correlation.

Usually multi-view video can be defined as the simultaneous multiple video streams from multi-view cameras. Multi-view cameras can be set up in different ways to capture a distinct scene simultaneously e.g., cameras may be trained on (i) a spherical shape to cover all 360° angles in *x*-, *y*-, *z*-axes, (ii) inwards semi-circular arrangement at the same height position (i.e., convergent), or (iii) straight line parallel to the scene. The convergent position (Fig. 1(a)), is the most popular for its wide application in movies, advertising, educational video (e.g., surgical instructions), sports and general event broadcasting. However, the multi-view content captured from convergent viewpoints is more difficult for an encoder than that captured from a parallel camera setup, which

http://dx.doi.org/10.1016/j.neucom.2015.10.094
0925-2312/

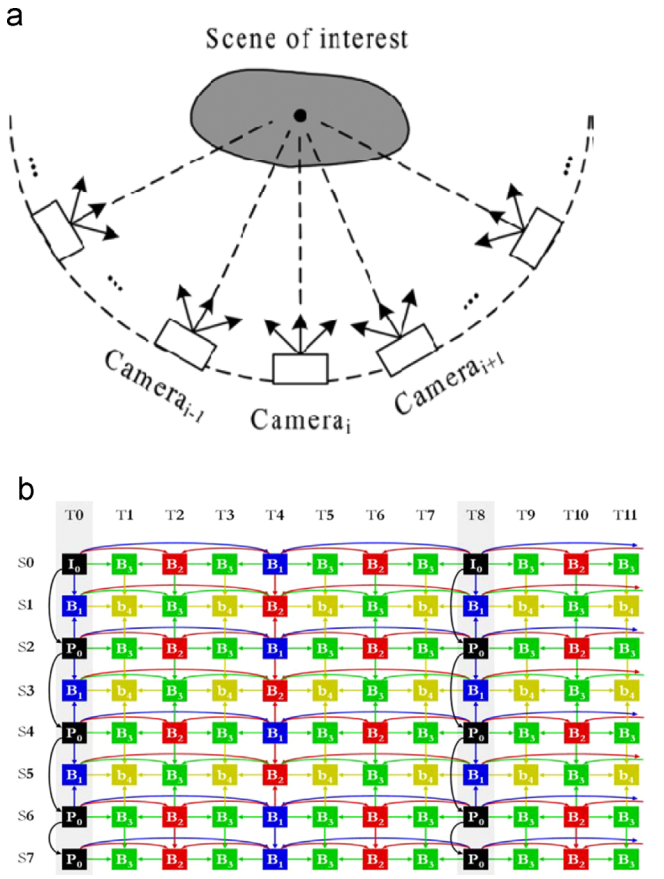

Fig. 1. (a) Convergent camera setup [1] and (b) images from views and temporal positions with prediction structure recommended by the HEVC-3D extension standard for referencing different views (S) and different temporal (T) images for MVC [2–4].

can be regarded as a simplified form of the convergent setup once the angular difference between two adjacent cameras' viewing directions is decreased to zero. The reason is that the disparity [1,5] (i.e., the translational displacement of a frame in one view into the frames of the same temporal instance in other views) intensity and disparity directions for the scenes are often much more intensive and heterogeneous across the multi-view frames captured using a convergent camera setup [1]. Fig. 1(b) depicts the concepts of multi-view video with eight views (S), nine temporal (T) images of each view.

Obviously, transmission and storing of multi-view video requires a huge amount of computation and data manipulation compared to single view video, although there is a significant amount of data redundancy among views. Recently, H.264/MVC and HEVC-3D [2–4] proposed a reference structure i.e., a mechanism for encoding an image using other already encoded images to exploit intra- and inter-view redundancy for improving *rate-distortion* (RD) gain in MVC. Fig. 1(b) depicts the *hierarchical bi-predictive* (HBP) reference structure [6,7], for example, T_4-th frame of S_1 view (i.e., B_2) uses two B_1 frames of the same view and two B_1 frames from S_0 and S_2 views respectively as reference frames. This structure clearly exploits intra- and inter-view redundancy from neighboring frames for maximum compression gain which provides 20% more bitstream reduction compared to the simulcast technique where no-inter-view redundancy is exploited i.e., each view is encoded separately [2]. However, it introduces huge computational complexity and more importantly *random access frame delay* (RAFD) problem due to the dependency on other inter/intra-frames. The enormous requirement of computational time limits the scope of MVC applications especially for electronic devices with limited processing and battery power such as smart phones. Moreover, RAFD problem limits the interactivity capacity of the coding scheme results in long time intra-frame and inter-view video switching in immersive interactivity system, 3D video on demand, etc. [8]. The RAFD for the highest hierarchical order is given by: $F_{\max} = 3 \times l_{\max} + 2 \times (N-1)/2$ where $l_{\max}$ is the highest hierarchical order and N is the total number of views [4]. For instance, in order to access a B-frame in the 4th hierarchical order (B4-frames in Fig. 1(b)), 18 frames must be decoded. By sacrificing compression gain, simulcast coding can remove inter-view switching and uni-directional intra-view referencing can remove intra-view image switching. Huayi et al. [8] modified the exhaustive referencing scheme (Fig. 1(b)) by encoding third view and sixth view independently and other views using those views. By doing this, they reduced inter-view and intra-view image switching time 30% and 0% respectively. Liu et al. [42] proposed three approaches using SP/SI frame coding, interleaved view coding, and secondary representation coding to provide low-delay RAFD. The method [42] should not outperform the MVC as it requires extra frames to encode and could not exploit nearest frame correlations. Abreu et al. [43] analyzed two types of frame reference prediction structure using one key view and two key views. Obviously this procedure could not provide better rate-distortion performance compared to the MVC structure (see Fig. 1(b)) as it exploits less number of references compared to the conventional MVC prediction structure. In the proposed scheme, our aim is to reduce inter-view latency and achieve the lowest possible intra-view random access latency using a 3D video compression technique.

HEVC video coding standard improves the coding performance by reducing up to 50% bitstreams compared to its predecessor H.264/AVC by increasing computational complexity around 4 times [9–11] for a single view video. In addition, when the HEVC encodes multi-view videos, it requires multiple amounts of computational time compared to the HEVC. The enormous requirement of computational time limits the scope of 3D video coding applications especially for electronics devices with limited processing and battery power such as mobile, iPhone etc.

Although simulcast coding technique (where each view is encoded individually) using HEVC for multi-view videos is inferior compared to HEVC-3D in terms of RD performance, it does not have a RAFD problem. Recently, a video coding technique was proposed where a dynamic background frame (i.e., McFIS – *the most common frame in a scene*) was used as an extra reference frame [12,13,41] for encoding the current frame assuming that the motion part of the current frame would be referenced using the immediate previous frame and the static background part would be referenced using McFIS. McFIS is generated using *Gaussian mixture model* [14–16]. Other background frame generation techniques [30,31] are also available for video coding purpose; however, Gaussian-based McFIS is better due to its capability to capture uncovered background area. *In this paper the first proposed scheme is a simulcast video coding approach (named HEVC-McFIS) based on HEVC using HBP prediction structure is proposed where McFIS of each view is used as an extra reference frame to encode corresponding view.* This technique improves RD performance compared to HEVC-3D and simulcast HEVC schemes on multi-view videos which have significant amount of background areas.

The first proposed scheme HEVC-McFIS has no RAFD problem and provides better RD performance compared to HEVC-3D scheme, however, it may suffer RD performance for motion-active video sequences as it does not exploit any inter-view redundancy. Moreover, it takes relatively more computations. To improve RAFD problem of HEVC-3D and better compression, Li et al. [34] propose a MVC scheme where they form a 3D cube by