J. Vis. Commun. Image R. 24 (2013) 752–759

Contents lists available at SciVerse ScienceDirect

J. Vis. Commun. Image R.

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

A fast algorithm of bitstream extraction using distortion prediction based on simulated annealing

Kaifang Yang *, Shuai Wan, Yanchao Gong, Yan Feng

School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China

ARTICLE INFO

Article history:
Received 13 September 2012
Accepted 17 April 2013
Available online 3 May 2013

Keywords:
Scalable Video Coding (SVC)
Bitstream extraction
Simulated annealing
Medium-grain Quality Scalability (MGS)
Interlayer dependency
Optimization
Decoding times
Fast algorithm

ABSTRACT

Scalable video streams can be extracted to meet the bandwidth limitation of different networks and end-users. Bitstream extraction is usually performed at the network proxy or gateway during transmission, where a low computational complexity is always preferred. How to quickly and accurately select the best resolution combination for a video to meet different bandwidth requirements by each user is crucial in bitstream extraction. In this paper a fast algorithm of bitstream extraction for scalable video is proposed. The interlayer dependency between the base quality layer and the first quality layer was used to predict the distortion of higher quality layers. When quality of every layer is available, the proposed method searches for the optimized combination of quality layers based on simulated annealing. Experimental results show that the proposed method provides an optimized performance, which is significantly higher than that can be achieved by the basic extraction method. Compared to the quality layer based extraction method in the reference software model of H.264/SVC (i.e., JSVM), the proposed algorithm can greatly decrease the decoding times from 2*NT* to only 2 without losing rate-distortion performance. Furthermore, the proposed method obtains a more smoothed video quality which is always favorable to the observer.

1. Introduction

Scalable video coding (SVC) was designed to satisfy the need of various multimedia services such as the IPTV, wireless networks, and video conferencing [1,2]. In SVC, the video sequence can be encoded into one base layer (BL) and several enhancement layers (ELs) to enable spatial, temporal, and quality scalabilities [3]. As a result, when transmitted over the heterogeneous networks, the SVC bitstream can be extracted with different combinations of resolutions to satisfy the different requirements and demands of the end-users. The key point of bitstream extraction is how to select a subset bitstream under a given target bit-rate while maximizing quality. Therefore the quality contribution of each packet in the bitstream is crucial to the extraction process.

The most accurate way to get the quality contribution of each packet is through actual decoding and comparison, which is a time-consuming process and cannot be allowed in practical applications. A basic algorithm for bitstream extraction was provided in the software implementation of H.264/SVC, i.e., the Joint Scalable Video Model (JSVM) [4]. This algorithm was computational efficient with no actual decoding needed. However, the resulting video quality was far from optimal due to lack of content adaptivity. A rate-distortion optimized extraction method was proposed by Amonou et al. [5] which utilized the concept of quality layers and sorted them based on rate-distortion optimization. This method can get optimal solutions, however, to obtain the impact of every quality increment on the overall video quality, 2*NT* times of decoding were needed (*N* is the number of quality layers and *T* is that of the time level). A more accurate method taking drift into account was proposed in [6], where the estimated distortion including distortion drift was used to assign quality layers to NAL (Network Abstraction Layer) units for a more efficient extraction. However, 2*NT* times decoding were still needed to calculate the drift and truncation distortion of the sequence to obtain the priority for extraction according to the rate-distortion slope [5,6]. To reduce the computational complexity, Lee et al. established three independent utility equations from the quality, spatial and temporal domain, respectively, and determined the extraction priority using the second derivative of these equations [7]. However, this method still requires *NT* times of decoding to establish the three functions. The steepest descent method was used in [8] to find a path with the smallest underneath area using the convex rate-distortion characteristics, whereas *NT* times of decoding were still needed to get the convex rate-distortion curves.

Since decoding is rather time-consuming and the relative computational complexity is high, we should minimize the times of decoding for speeding up bitstream extraction and satisfy the demand of practical applications. If we can predict the distortion introduced by the NAL unit, the extraction process can be much

* Corresponding author.
E-mail address: yangkaifang5@hotmail.com (K. Yang).

1047-3203/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jvcir.2013.04.013

accelerated in avoidance of decoding. Therefore an accurate distortion or quality prediction model is useful for a fast implementation of bitstream extraction. Over the past decades, many rate and distortion models for scalable video coding have been proposed through theoretical analysis or empirical approaches. Hassan Mansour et al. established the rate and distortion model for coarse-grain quality scalability (CGS), whereas the distortion model was related to mean absolute difference (MAD) of different quality layers which is not available when doing the bitstream extraction [9]. Jiaying Liu etc. analyzed the inter-layer dependency characteristics in both spatial and quality scalability and proposed a linear distortion model for bit allocation regarding the quality scalability. However, the slope used in this distortion model is not readily available and the model does not address the temporal scalability [10]. Yao Wang etc. proposed a two dimensional subjective video quality model. However, its application in bitstream extraction is not straightforward since several parameters need to be adjusted for each specific video [11].

In order to meet the increasing demand of scalable video transmissions, a fast bitstream extraction method is proposed with only two times of decoding needed to get the quality of all layers. Then the optimized resolution combination is searched using simulating annealing with a reward and punishment mechanism applied for different temporal layers. As a result, the optimized extraction point can be found at a very low computational complexity. Since in practical applications, the spatial resolution and the frame rate are often pre-determined according to the capability of the end user, this paper focuses on the scalability in the quality domain. However, the proposed bitstream extraction algorithm is not limited to quality scalability and it can be well extended to the spatial and temporal scalabilities.

The rest of the paper is organized as follows. Two bitstream extraction methods provided in the reference model of scalable video (i.e., JSVM) [4] are introduced in Section 2. The proposed low complexity bitstream extraction method based on distortion prediction and simulated annealing is described in Section 3. In Section 4, performance evaluation is presented where the proposed method is compared with the two bitstream extraction methods in JSVM in terms of R–D performance and complexity. This paper closes with conclusions given in Section 5.

2. Bitstream extraction in JSVM

Two bitstream extraction methods are provided in JSVM, namely the basic extraction and the quality layer based extraction.

2.1. Basic extraction (BE)

The basic extraction method determines the extraction resolution combination according to the bit-rate of different scalable levels only. A spatial–temporal resolution combination whose bit-rate is the closest to but not greater than the target bit rate will be firstly selected, and then for each lower spatial resolution, NAL units are gradually included with the quality level increased until the targeted bit-rate is reached [6]. Although BE is fast, the reduced complexity is at the expense of loss of rate-distortion performance since the independence of the video content is not considered [12]. As the result, its performance is far from optimal. Furthermore, using BE, frames from the same temporal level will have the same quality layer, which cannot make efficient use of the bandwidth.

2.2. Quality layer based extraction (QLE)

Quality layer based extraction is first proposed by [5], which employs rate-distortion (R–D) optimization to improve the performance of extraction and can find a resolution combination which has the optimal rate-distortion performance through evaluating the influence of different NAL units on the overall rate-distortion performance. Here the priority of each unit is sorted according to the actual R–D slope which needs decoding the bitstream by $2NT$ times to get the R–D points. When the appropriate order is obtained, one can get the resolution combination with the best R–D performance. During the computation of the distortion, QLE assumes that all quality layers of each frame at a lower temporal resolution are available. This assumption does not always hold in practice and drift errors are inevitably introduced. Although QLE can find a optimized resolution combination, its application in practice is rather limited due to its repeated decoding process. It is noticeable that the QLE is simplified when implemented in JSVM, whereas $\{(N-1)\cdot T+1\}$ times of decoding is still needed, which is still about NT times of complexity.

3. Optimized bitstream extraction with distortion prediction based on simulated annealing

3.1. Distortion prediction along quality layers

The quality scalability which varies the fidelity (signal-to-noise ratio) of the encoded video stream is provided with two mechanisms, namely the coarse grain scalability (CGS) and medium grain scalability (MGS) [13]. Since MGS is involved with much more complexity when QLE is performed, in this paper MGS is addressed. It is noticeable that the distortion prediction model proposed in this paper is well adapted to CGS since similar distortion relationship can be observed. To achieve a high coding efficiency, SVC has an adaptive inter-layer prediction mechanism which makes scalable layers dependent with its base layer [14].

In order to investigate the relationship between dependent quality layers, JSVM (version 9.19.11) was used to encode video sequences at the CIF (352×288) resolution. Here each sequence was encoded with three MGS quality ELs (i.e., MGS1, MGS2, MGS3) and one quality BL (i.e., BL). Each MGS quality layer was split into 3 MGS fragments following suggestions given by [16,17]. The dependent layer for inter-layer prediction was set to its adjacent lower layer. MGS Control was set to 2, which means that pictures of the highest EL can be used for motion estimation and motion compensation. Each layer was coded using the IPPP structure at 30 frames per second (fps). The QP at the base layer varied ranging from 30 to 42 with 2 as the step size. The QP difference (DQP) which means the QP for EL will be reduced from its base layer was set as 4 as suggested in [15]. As shown in Fig. 1, the Peak Signal to Noise Ratio (PSNR) of quality ELs changes linearly with the PSNR of its dependent layer, and the slope for different EL is almost the same.

For quality scalability combined with temporal scalability, SVC with the hierarchical B-frames structure was also tested.

Here the GOP length was set to 8 with the other settings the same as those under the coding structure of IPPP. The corresponding results regarding hierarchical B-frames are shown in Fig. 2. For the limit of paper length we only present the result of sequence "Foreman" and "Football" here, and the other sequences have similar results. It can be seen in Fig. 2 that the relationship between two interdependence layers with different time levels is still linear and the slope for different time level is almost the same as well.

For the same linear slop and almost the same constant parameter between different MGS layers, the linear relationship can be expressed as

$$\mathrm{PSNR}_{MGS_n}=k\cdot\mathrm{PSNR}_{MGS_{n-1}}+a\quad n>0 \tag{1}$$

where MGS_n means the nth MGS quality EL, MGS_{n-1} is the quality layer on which MGS_n is directly dependent. When the quality