Signal Processing: *Image Communication* 29 (2014) 332–344

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

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

Content-based group-of-picture size control in distributed video coding

Enrico Masala [a,1], Yanmei Yu [b,*], Xiaohai He [b,2]

[a] *Control and Computer Engineering Department, Politecnico di Torino, corso Duca degli Abruzzi 24, 10129 Torino, Italy*
[b] *College of Electronics and Information Engineering, Sichuan University, No. 24 South Section 1, Yihuan Road, 610065 Chengdu, China*

ARTICLE INFO

Article history:
Received 28 May 2013
Received in revised form
23 January 2014
Accepted 27 January 2014
Available online 6 February 2014

Keywords:
Distributed video coding
Adaptive GOP size
Content-based adaptation
Support vector machine

ABSTRACT

Controlling the group of picture (GOP) size in distributed video coding (DVC) is a difficult but important task since it has a direct impact on the coding performance. This paper presents a framework to adaptively control the size of GOPs in a Wyner–Ziv encoder by means of encoder-side decisions based on support vector machines (SVM) that uses simple features extracted from the original video content. To train the SVM, firstly this work proposes how to compute the sequence of GOP sizes with the best rate-distortion performance given the set of GOP sizes that can be used during the encoding process. Then, an algorithm based on the previously trained SVMs is presented to control the actual GOP size each time a new decision can be taken at the encoder. Results show that the proposed algorithm can achieve a rate distortion performance close to the ideal one. Moreover, comparisons with a reference adaptive GOP size selection algorithm in the literature shows gains up to 2 dB PSNR in the best conditions.

1. Introduction

Distributed video coding (DVC) is a video compression paradigm different from traditional video compression standards such as H.264/AVC. Traditional video compression standards, usually with a complex encoder and a low-complexity decoder, are very well suited for applications such as broadcasting. However, for other applications such as compression and transmission of video from wireless surveillance cameras to a central server – a scenario characterized by low power, scarce memory and limited computational resources – other approaches may be more beneficial. In particular, DVC has been proposed for these types of applications. DVC is based on the Slepian–Wolf (SW) [1] and Wyner–Ziv (WZ) [2] theorems, which show that data could be compressed efficiently by independent encoders in lossless coding and lossy coding modes, respectively. The DVC approach reverses the complexity of the encoder and the decoder in traditional video coding standards by shifting the complex tasks to the decoder. Two early DVC practical architectures, i.e. the Berkeley power-efficient, robust, high compression syndrome based multimedia coding (PRISM) Architecture [3,4] and the Stanford Architecture [5,6], were proposed independently in 2002. Among these two architectures, the Stanford architecture is the most popular one in the literature. This one firstly implemented Wyner–Ziv coding in the pixel domain [5], then it has been extended to the transform domain [7] to exploit the spatial statistical redundancy within a frame. In this paper, we focus on the Wyner–Ziv coding in the transform domain.

For the case of Wyner–Ziv coding, a number of technical issues need to be addressed to provide acceptable

* Corresponding author. Tel.: +86 28 854 62766; fax: +86 28 854 63466.
E-mail addresses: masala@polito.it (E. Masala), yuyanmei@scu.edu.cn (Y. Yu), hxh@scu.edu.cn (X. He).
[1] Tel.: +39 0110907036; fax: +39 0110907099.
[2] Tel.: +86 28 854 62766; fax: +86 28 854 63466.

http://dx.doi.org/10.1016/j.image.2014.01.013
0923-5965

performance. For instance, the decision about whether to encode a frame as a key or a Wyner–Ziv (WZ) one is particularly important since it can have a large impact on the efficiency of the encoding process. To clarify the point, consider the case when the sequence can be easily predicted, due to, e.g., the presence of a large amount of static content. In this case, it may be more efficient to encode frames using more than one WZ frame between two key frames, thus making larger group of pictures (GOP). The opposite holds when the content is more difficult to predict from the key frames. Therefore, the encoder needs to decide, for each new frame, if the frame has to be compressed using traditional intra-coding techniques (i.e., as a key frame) or using the WZ approach. This activity must be performed using a low-complexity algorithm not to lose one of the main advantages of DVC, i.e., low encoder complexity.

In the case of traditional hybrid video codec scheme, motion vectors are available at the encoder and they can give good hints about how to optimally control the GOP size [8]. The DVC case, instead, requires to measure the activity present in the video content with low complexity techniques to run the algorithm at the encoder, as in [9–11], or analyze the video characteristics at the decoder side, where more complex algorithms could be used, and send the results back at the encoder by means of a feedback channel. A review of the state-of-the-art in the field is presented in Section 2.2.

This work builds on the DVC architecture by proposing a framework comprising two main parts: (1) the sequence analysis part which shows how to compute, with linear complexity in the number of frames, the ideal GOP structure for a video sequence which maximizes the rate-distortion performance given a set of available GOP sizes; (2) the use of the previous results on a set of sequences to train three SVMs that are the basic building blocks of the adaptive GOP size selection algorithm proposed in this work. For this aspect, a new content-adaptive frame type decision scheme is proposed which allows to select the best encoding mode by analyzing some features of the current and a few past frames. The frame type is decided by linear support vector machines (SVMs), which are trained with the sequence of GOP sizes that provides the best rate-distortion performance, as computed in (1). Although the proposed method is not – strictly speaking – a distributed algorithm, since the SVM works jointly on frames belonging to the key-frame group and the WZ group, its characteristics make the method suitable for the typical contexts in which DVC is used.

The main contributions of this work can be summarized as follows: (1) An algorithm to determine, with linear complexity in the number of frames, the best rate-distortion (RD) performance and the corresponding sequences of GOP sizes for a video sequence, given the results of a limited set of offline DVC encodings of the video sequence itself. (2) The definition of sets of features useful to determine the local characteristics of the video sequence. (3) An algorithm to control the GOP size on the basis of the previous features. Simulation results based on many video sequences with different characteristics show the effectiveness of the proposed algorithm by means of comparisons with the ideal performance and with the technique proposed in [9].

The paper is organized as follows. Section 2 briefly reviews the DVC architecture used in this work, focusing on the GOP size issue. In Section 3 an algorithm is proposed to compute the best RD performance for any DVC sequence given a set of available GOP sizes and a number of offline DVC encodings of the video sequence itself. A proposal for a GOP size decision algorithm based on SVM is given in Section 4, followed by Section 5 which presents simulation results and comparisons with a reference technique. Conclusions are drawn in Section 6.

2. Background

2.1. DVC coding architecture

Several approaches have been proposed in the literature to implement DVC. In this work, we focus on the DISCOVER architecture [6,11,12]. In this framework, the encoder sends parity bits to the decoder in order to improve the quality of the side information (SI) available at the decoder. If the correlation between the side information and the current frame is high, less parity bits are needed to achieve a given quality, and vice versa. In order to generate the side information, the decoder uses interpolation mechanisms relying on the key frames which have been encoded using a traditional hybrid video coding approach, e.g., H.264/AVC.

As in the traditional hybrid video coding, frame interpolation is performed over groups of consecutive frames referred to as GOPs. However, the GOP size is typically allowed to vary during the encoding process to better exploit the temporal correlation, resulting in better performing temporal interpolation. In our method, for each new frame the encoder has to decide whether the frame has to be encoded as a key frame to start a new GOP, or as a WZ frame which increases the current GOP size.

WZ frames are typically more efficient in RD terms with respect to key frames if good side information estimation can be performed at the decoder. Static or easily predictable content is a typical situation where the WZ frames achieve high efficiency.

Determining the key and WZ frame position and frequency is very important to achieve a good RD performance. Consider a GOP size equal to two (Fig. 1): the intermediate WZ frame will be decoded using SI generated by means of the immediately previous and next key frames. If more WZ frames are present between key frames, the decoding process will need to perform several steps before reconstructing all WZ frames. For instance, for the case of GOP size 4 it may reconstruct an intermediate WZ frame, e.g., WZ_2, then it will proceed to decode the WZ_1 and WZ_3 frames.

Clearly, the characteristics of the video should be carefully considered in order to optimize the choice of the GOP size during the encoding process. Intuitively, when the amount of motion is high, the correlation among frames is low and smaller GOP sizes must be chosen. Conversely, if the amount of motion is low, correlation is higher and longer GOPs should be used, so to avoid the penalty of