J. Vis. Commun. Image R. 40 (2016) 149–158

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Spatially misaligned HEVC transcoding with computational-complexity scalability ☆

Johan De Praeter [a,*], Glenn Van Wallendael [a], Thijs Vermeir [a,b], Jürgen Slowack [b], Peter Lambert [a]

[a] Data Science Lab (Ghent University – iMinds), Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium
[b] Barco N.V., President Kennedypark 35, 8500 Kortrijk, Belgium

ARTICLE INFO

Article history:
Received 15 March 2016
Revised 22 June 2016
Accepted 24 June 2016
Available online 24 June 2016

Keywords:
High Efficiency Video Coding (HEVC)
Machine learning
Transcoding
Picture composition

ABSTRACT

In control rooms, video walls display footage from multiple sources. Often, a composition of these sources is sent to other devices in a single video stream. To minimize the computational complexity of this composition process, information from the original bitstreams can be reused. However, in High Efficiency Video Coding (HEVC), simply copying the original encoding decisions is not compression efficient if the individual videos are spatially misaligned with the grid of coded blocks of the composition. Our proposed HEVC-based transcoder reduces the computational complexity by predicting encoding decisions of misaligned sequences by using a trivial method or a more adaptive, computational-complexity scalable machine learning method. Higher compression efficiency is observed when more alignment is preserved with the original block grid. Overall, the machine learning method achieves a higher compression efficiency than the trivial method. Both methods attain a complexity reduction of up to 82% compared to the reference software.

1. Introduction

Visual information such as surveillance footage plays an important role in many industries. This information is gathered in a control room where it is displayed as a composition of multiple input streams on a video wall. These individual video sequences are either tiled next to each other, or partially overlap. However, (parts of) this information is sent to other devices such as desktop computers, laptops, and other handheld devices. All of these devices receive a personalized composition of input bitstreams. This composition must be freely arrangeable, meaning that each separate sequence can be dragged across the screen. Since decoding all these input bitstreams requires multiple decoders, which are not available on the client device, the composition is created in the network by specialized hardware. This new bitstream is then decoded by a single decoder at the receiver.

The same approach is used for sending personalized advertisements embedded in the picture during live video broadcasts. Depending on the location of the viewers, broadcasters insert different advertisements into the encoded video stream of the content providers without requiring the client device to decode both the video and advertisement.

Another similar scenario occurs during video conferencing between more than two locations. If a participant wants to display all video streams on his handheld device, the power and memory of the decoder might be insufficient for decoding the multiple input bitstreams.

In the above use cases, each different composition needs to be encoded. This encoding step requires much computational power. Moreover, current displays support resolutions up to 3840 × 2160 pixels and will further evolve towards 7680 × 4320 pixels. With such high resolutions becoming common, both the input video sequences and the composition will be compressed using High Efficiency Video Coding (HEVC) [1], since this compression standard offers a bit rate reduction of 50% for the same perceptual quality as its predecessor, Advanced Video Coding (H.264/AVC) [2]. However, the computational complexity of the encoder is also higher compared to the predecessors of HEVC. Consequently, reducing this increased complexity is crucial in order to limit the resources necessary to generate compositions.

Previous work on compositions focuses on picture-in-picture insertion of video content in H.264/AVC by re-using encoding decisions from the original video sequences [3–5]. A similar approach was also used for HEVC in order to achieve high complexity reductions. In this approach, the block structures of the input bitstreams are merged and passed to the encoder, reducing the number of encoding decisions that need to be evaluated [6]. However, this solution is only applicable if the inserted content is constrained

☆ This paper has been recommended for acceptance by M.T. Sun.
* Corresponding author.
E-mail address: johan.depraeter@ugent.be (J. De Praeter).

http://dx.doi.org/10.1016/j.jvcir.2016.06.025
1047-3203/© 2016 Elsevier Inc. All rights reserved.

to a fixed grid based on the size of the Coding Tree Units (CTUs). In compositions that do not constrain sequences to this grid, misalignment of the CTU-grids is introduced as seen in Fig. 1. As such, copying the block structures is not feasible in applications that require more flexible compositions. Therefore, a solution is necessary in order to re-encode such misaligned content in HEVC.

The problem of handling misaligned video sequences is classified as transcoding, which is defined as converting one compressed input bitstream to an outgoing bitstream. A naive cascaded transcoding approach consists of decoding the input bitstream, applying the desired operation, and then re-encoding the video. However, this re-encoding step is computationally complex, whereas all of the coding information of the input bitstream is discarded. One way to accelerate this computationally intensive task is by determining the coding information of the re-encode in parallel [7,8]. Another way would be to not completely discard the coding information of the original input bitstream. Therefore, transcoding techniques focus on exploiting the information of the input bitstream in order to skip encoding decisions in the re-encoding step [9–11]. Examples of transcoding operations include spatial downscaling [12,13], frame rate reduction [14,15], requantization [16,17], and changing the compression format [18,19].

In this paper, we propose an efficient HEVC-based transcoder for spatially misaligned sequences that can be applied to any misalignment of an input sequence. To achieve this, both a trivial and a machine learning method are presented. The latter method also introduces a probability threshold, which allows the model to only apply decisions with a higher confidence than the threshold. This results in a transcoder that allows a trade-off between complexity and quality depending on the available resources.

The rest of the paper starts with a description of related work on transcoding. Section 3 then describes the proposed methods followed in Section 4 by an in-depth analysis of the parameters for the machine learning method. The results are presented in Section 5. Finally, Section 6 ends with the conclusion.

2. Related work

In other works on transcoding, motion re-estimation is often used in order to reduce complexity during re-encoding [12,18]. However, in HEVC, CTUs of 64×64 pixels can be recursively split into Coding Units (CUs) which can be as small as 8×8 pixels. These CUs can then be split further into 8 possible Partition Unit (PU) modes which are used for inter prediction. These modes are $2N \times 2N, 2N \times N, N \times 2N, N \times N$, and four asymmetrical partitioning modes, for which a CU is defined as having a size of $2N \times 2N$ pixels. With the first mode, the entire CU has one PU of size $2N \times 2N$. For modes $2N \times N, N \times 2N$, and the asymmetrical modes, the CU is split into two PUs of a size as described by the mode. For mode $N \times N$, the CU is split into four PUs. Consequently, even when motion re-estimation is applied, the most optimal block structure still needs to be determined.

Another way to reduce complexity is by using mode mapping in order to limit the amount of partitioning modes that the encoder needs to test. In related transcoding works, machine learning techniques have often been applied to find a correlation between coding information of the input bitstream and the partitioning modes of the output bitstream [20–28]. The applied techniques include support vector machines (SVM) [20,21], decision trees [22–26], and linear discriminant functions (LDF) [27,28]. Some of these algorithms have also been compared to each other, showing that a random forest performed the best in the case of spatial transcoding [29]. Several algorithms have been trained offline [20–26], i.e., using a subset of sequences for training and a separate set for evaluation. However, online-trained models, which are trained on the first k frames of a sequence and are then applied to the rest of the sequence, are better adapted to the content of the current sequence [27–29].

Many of the above papers focus on transcoding using older standards than HEVC. Applying these techniques on HEVC is impossible due to the more complex coding tools of HEVC compared to its predecessors. Despite this difference in coding tools, some research tries to bridge the gap between the different compression technologies by translating an older compression standard to HEVC [19,27,28]. However, because of the big difference in compression standards, techniques that try to learn these conversions are suboptimal to be used as transcoding techniques within the same standard.

Transcoding in HEVC itself has been considered for transrating [23] and spatial transcoding [29], but in these cases the encoder has been accelerated only by limiting CU partitioning decisions. Complexity in HEVC can be further reduced by also limiting PU partitioning modes, which is part of the proposed method in this paper.

None of the above papers have considered transcoding of spatially misaligned sequences. As far as to the authors' knowledge, this type of transcoding in HEVC has only been considered in [26]. In this work, a mode mapping model was created based on an offline-trained decision tree. This tree was trained for pixel-shifts of 32, 16 and 8 pixels. However, this means that the model is restricted to these shifts, which limits the applicability and does not allow the end user to freely arrange the videos in the composition. Moreover, only CU decisions were skipped, meaning that all possible PU partitioning modes are still evaluated for each predicted CU.

As its main novelty, this paper overcomes the drawbacks of being restricted to certain shifts and of only accelerating CU decisions. This is achieved by proposing two transcoding methods that can be applied to all possible pixel-shifts and further improve complexity reduction by predicting PU information as well. Moreover, the second method achieves an improved trade-off between encoding complexity and compression efficiency.

3. Proposed transcoding methods

An efficient transcoder for misaligned sequences is a transcoder that reduces the complexity of re-encoding while minimizing the

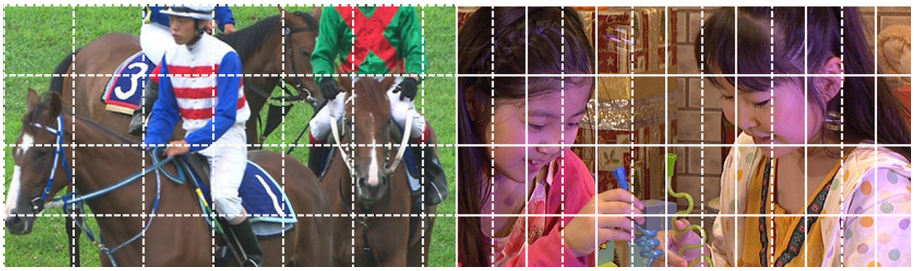

Fig. 1. The CTU-grid of the sequence on the right (full lines) is misaligned with the grid of the composition (dashed lines).