TRANSCODING FROM H.264/AVC TO A WAVELET-BASED SCALABLE VIDEO CODEC

Eduardo Peixoto, Toni Zgaljic and Ebroul Izquierdo

Queen Mary, University of London, London, UK
Email: eduardopeixoto@ieee.org, toni.zgaljic@elec.qmul.ac.uk, ebroul.izquierdo@elec.qmul.ac.uk

ABSTRACT

Scalable Video Coding (SVC) enables low complexity adaptation according to transmission and display requirements, providing an efficient solution for video content delivery through heterogeneous networks. However, legacy video and most commercially available content capturing devices use conventional non-scalable coding, e.g., H.264/AVC, to compress and store video streams. As a consequence and in order to fully exploit the advantages of SVC technology, efficient transcoding from conventionally coded to scalable content is urgently needed. In this paper an efficient transcoder from H.264/AVC to a wavelet-based SVC is proposed. The complexity of the transcoder is kept very low by using information extracted directly from the decoded H.264/AVC bitstream, such as motion vectors and the presence of residual data. The proposed approach has been tested with well known benchmarking sequences, showing a good performance in terms of decoded video quality and system complexity.

Index Terms— Transcoding, Scalable Video Coding

1. INTRODUCTION

Scalable video coding (SVC) allows a real-time adaptation of video content since the extraction of a lower resolution, frame-rate and/or quality is possible by simple parsing of the compressed bitstream. However, video stored on the server is often encoded using conventional, non-scalable, codecs, such as H.264/AVC [1]. In this case, low-complexity, storage-efficient video adaptation cannot be achieved. To tackle this issue, a non-scalable bitstream can be converted into a scalable stream by transcoding. Therefore, transcoding is required only once and the transcoded video can then be adapted many times.

The most straightforward way to transcode from one format to another is to cascade the required decoder and encoder, known as the cascaded pixel-domain transcoder [2, 3]. Here, the trivial transcoder is defined as the transcoder in which no intermediate processing is performed, i.e., when the sequence is simply decoded and re-encoded. Such an approach results in high quality of the transcoded video, but its complexity is also very high. A possible way to reduce the complexity is to reuse motion information from the decoded video during the re-encoding process.

Transcoding between hybrid-based video coding structures has been extensively studied before [2, 3], even targeting hybrid-based scalable codecs [4]. However, hybrid-based to wavelet-based scalable video transcoding has not been fully investigated in the literature. Although a hybrid based technology was chosen for standardization of scalable video coding within MPEG [5], a great amount of research continued also on Wavelet-based Scalable Video Coding (W-SVC). Several recent W-SVC systems have shown a very good performance in different types of application scenario [6, 7, 8], while still being able to deliver some attractive features not supported by the standard, such as Fine Grain Scalability (FGS).

The employed codec, here denoted as W-SVC [6], supports quality (with FGS), spatial and temporal scalability and any of their combinations. Its main features are: hierarchical variable size block matching motion estimation, flexible selection of filters for both spatial and temporal wavelet transforms at each level of spatio-temporal decomposition, user-defined flexible decomposition path, support for conventional frame-based coding and object-based coding, bit-plane coding based on Embedded ZeroBlock Coding (EZBC), binary arithmetic coding and low-complexity post compression rate-distortion optimization for bit-stream allocation.

In this paper we investigate the impact of exploiting motion information from H.264/AVC bitstreams on transcoding to W-SVC. Since H.264/AVC supports choosing reference frames (RF) in a very flexible way, not all MVs can be directly reused in the fixed RF structure used in W-SVC. Here, these MVs are obtained by approximation and refinement. An additional issue is that the motion information from H.264/AVC is optimized for a single rate-distortion point, and therefore it has to be modified for the application in the scalable scenario. The key novelty of the proposed transcoder is efficient handling of the partitioning to tackle this optimization issue. Two new concepts to reduce the transcoder complexity are considered: using information on the number of DCT coefficients in the H.264/AVC stream and measuring the similarity of H.264/AVC MVs.

2. PROBLEM FORMULATION

Since the two codecs are fundamentally different, the transcoder architecture chosen for this work is the cascade pixel-domain approach [2, 3]. It consists of decoding the source H.264/AVC sequence, performing an intermediate processing on the decoded motion information and re-encoding the sequence using the W-SVC codec and processed motion information. In this way, complexity is largely reduced, since motion estimation (ME) is the most time consuming task in a video encoder.

In W-SVC the input sequence is subjected to Motion Compensated Temporal Filtering (MCTF) [9], which aims to reduce the correlation between consecutive frames and provide the basis for temporal scalability, without drift errors. Here, rigid limitations are imposed on the RF selection in the ME module for a particular step of filtering. This is depicted in Fig. 1, which shows a 3-level dyadic temporal decomposition of 9 frames, using MCTF with a bidirectional prediction. The markings $\mathbf{L_x}$ and $\mathbf{H_x}$ in the figure represent low-pass and high-pass tempo-

This research was partially supported by the European Commission under contract FP7-247688 3DLife and FP7-248474 SARACEN.

Table 1. Profile of the chosen partitions

	Soccer CIF		Crew CIF	
Partition Size	H.264	W-SVC	H.264	W-SVC
INTRA	5.28%	12.66%	21.16%	12.66%
16×16	39.33%	76.12%	23.86%	77.78%
16×8 or 8×16	15.43%	N/A	22.31%	N/A
8×8	19.53%	10.12%	16.42%	9.00%
8×4 or 4×8	17.18%	N/A	14.39%	N/A
4×4	3.25%	1.08%	1.89%	0.54%

ral subbands, respectively, at the x-th decomposition scale, and the arrows point from the frames being predicted to the RFs in the process of MCTF. The structure of RFs is fixed at each scale of the decomposition - frames at higher decomposition scales are never used as reference at lower decomposition scales. This strict condition straightforwardly enables the temporal scalability. Spatial and quality scalability are achieved by spatial wavelet transform and embedded coding.

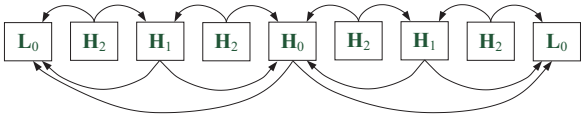

Fig. 1. Hierarchical structure for selecting the reference frame with 3 levels of temporal decomposition.

Contrasting the above rigid structure of RFs, H.264/AVC can choose reference in a more flexible way. The different RF structure in the two codecs restricts direct re-using of H.264/AVC motion information in transcoding. Only those MVs that point to the same RF in W-SVC and H.264/AVC can be directly reused. However, even these MVs may not be optimal for W-SVC and therefore they should be refined. In the H.264/AVC codec, the macroblock (MB) partitions and MVs are optimized to a single rate-distortion target, while the W-SVC framework should perform encoding tailored to a wide range of bitrates and spatio-temporal resolutions. To illustrate this difference, Table 1 shows the percentage of different block sizes chosen for two different sequences at CIF (352×288) resolution, both for H.264/AVC (using *IPPP* configuration and a quantization parameter $QP = 20$) and W-SVC codecs using 3 temporal decompositions (TD).

Refining MVs for all matching RF increases the complexity of transcoding. Therefore, a method for selecting which MVs need to be refined is required. On the other hand, when RFs do not match in W-SVC and H.264/AVC, neither MVs nor partitioning structure can be directly re-used for W-SVC. These issues are addressed in the following section.

3. THE TRANSCODER

For a given MB, which is set as 16×16 pixels in the W-SVC encoder, the following partition sizes can be tested: 16×16, 8×8 (four blocks in the given MB) and / or 4×4 partition (four blocks in a 8×8 partition). Here, we define the testing of a partition as the approximation (or reuse) and refinement of MVs, the mode selection (forward, backward or bidirectional prediction) and the computation of the cost for that partition, in rate-distortion sense. If more than one partition size is tested, the transcoder chooses the one that yields the lowest cost.

3.1. Handling the partitioning for each macroblock

When deciding which partition sizes will be tested for a given MB, the proposed transcoder takes into consideration the partition of the H.264/AVC MB. If the MB was encoded in inter mode, then only those partitions that are equal or larger than that of the H.264/AVC will be tested. However, since the employed W-SVC framework uses a strict quadtree partitioning of the MB, this procedure has to be slightly adapted. As an example, if the H.264/AVC MB was partitioned in two 16×8 blocks, then both 16×16 and 8×8 partitions will be tested in W-SVC. If the MB was encoded in intra mode in H.264/AVC, then the partitions tested follow the usual W-SVC scheme: first 16×16 and 8×8 partitions are tested and, if the latter has a lower cost, 4×4 partitions are tested.

3.2. Using the DCT coefficients to drive the transcoder

In the H.264/AVC codec, a decoded block is given as : $B_{DEC} = P + R + D_F(P + R)$, where B_{DEC} is the decoded block, P is the prediction for this block, R is the residual and $D_F(\cdot)$ is the effect of the deblocking filter [1], which is applied to $P + R$. Thus, if no coefficient is transmitted, then the residual is zero, and the decoded block is given as: $B_{DEC} = P + D_F(P)$. Since the W-SVC codec uses MCTF, ME is performed using the original frames, which, in the transcoder are the decoded H.264/AVC frames. To further reduce the complexity, the transcoder avoids refining the MVs when the residual for the corresponding block is zero. When this happens, H.264/AVC MVs are directly assigned to that partition, without further refinement. Other partitions may also be tested even when this happens, in which case the usual approach of approximation and refinement is used.

In the transcoder implementation, two different things were tested to check if the residual is zero: the actual number of non-zero DCT coefficients (which, if zero, guarantees that the residual is zero), and the syntax parameter coded block pattern (CBP). The CBP only tells the presence or absence of AC DCT coefficients in each 8×8 block inside the MB [5]. Thus, if the CBP for a given block is zero, it does not mean that the residual is zero, necessarily, since it can still have a DC coefficient. Also, the CBP does not have the information for each 4×4 block separately. However, the results using both methods are very close, and since the information on the CBP is more readily available, the CBP is used to drive the MV refinement for a partition.

3.3. Grouping Similar MVs

The proposed transcoder has an optional setting to further reduce the complexity that consists in measuring the similarity of the H.264/AVC MVs to decide whether or not smaller partitions will be tested. First, a mean MV is computed using the four 4×4 block within each 8×8 partition. Then, each of the four MVs is compared to the mean MV, using a similarity measure that considers each component, $MV.x$ and $MV.y$, separately. A single MV is considered similar to the mean MV if $\left|\overline{MV}.y - MV.y\right| < T$ and $\left|\overline{MV}.x - MV.x\right| < T$, where $\overline{MV}$ is the mean MV and T is a threshold, which can be set as a parameter. If all available H.264/AVC MVs inside an 8×8 block are considered similar, then 4×4 partitions will not be tested for this 8×8 block, regardless of the H.264/AVC partition. The similarity measure is applied to the four 8×8 blocks within a MB in a similar manner: if they are considered similar, 8×8 partitions will not be tested for this MB. This applies whether or not these MVs can be directly reused.

3.4. Framework for MV Approximation and Refinement

For each partition tested, the proposed transcoder produces two MV candidate lists, one for each direction. For each candidate in the two lists the cost is computed in a conventional way, considering the residual and the rate of the chosen MV. The best candidate is selected for each direction and then a further refinement step can be applied. In all cases, the refinement considered is a Hexagon search [10], starting at the best candidate for each direction. When some MVs within the partition being tested can be directly reused, all MVs corresponding to smaller partitions within the partition are added to the appropriate candidate lists. Otherwise, several strategies, enlisted in the remainder of this section, are used to populate the candidate list(s).

3.4.1. Spatial Approximation

Here, the approximated MV is formed as the weighted average of MVs of blocks above and on the left of the currently observed block. The weights for each MV are defined according to the sizes of its corresponding blocks. The advantage of this method is that it uses the MVs already refined by the W-SVC.

3.4.2. MV Composition

The MV composition method used here is similar in spirit to Forward Dominant Vector Selection [11] and Telescopic Vector Composition [12]. However, since these two methods have been developed to work with fixed-size motion blocks they were adapted to support variable-size motion blocks. If a block in a particular frame consists of several motion blocks that are of different sizes, weighted average of their MVs will be used, where the weight for each MV is proportional to its corresponding block size, relative to other observed blocks.

3.4.3. MV Scaling

When a H.264/AVC MV cannot be reused, it can be scaled to be used as a candidate for its direction. The scale factor is directly proportional to the distance between the H.264/AVC and the W-SVC RFs and the current frame.

3.4.4. MV Inversion

This method can be performed to create backward MV candidates from already found forward MV candidates or vice versa. This method can be useful for instance in *IPPP* coding configuration when only forward MVs are available.

4. EXPERIMENTAL RESULTS

In all tests, the full length (300 frames) of the original sequences was used, where the PSNR shown is the average among the frames of the luma component, while the rate considers also the chrominance components. The PSNR is always computed using the original sequence as the reference. In all cases, the W-SVC codec uses MCTF with bidirectional prediction.

The first point analyzed here is the loss caused by transcoding. Fig. 2 shows the rate-distortion point of the decoded H.264/AVC sequence (used in transcoding), the performance of the W-SVC codec using the original sequence and the results of the trivial transcoder using full ME. The H.264/AVC stream was encoded in *IPPP* configuration, with $QP = 28$ for Crew sequence and $QP = 20$ for Soccer and City sequences. The lower QP was chosen to enable a wide range of decoding points in W-SVC, and also because the transcoder is likely to operate on high-quality video. It can be seen that a certain loss of quality is present in transcoding, specially at medium and higher bitrates.

Table 2. SAD Calculations for Crew, Soccer and City sequences, normalized by the full ME.

Method	Crew	Soccer	City
Trivial - full ME	100%	100%	100%
Trivial - hexagon	0.84%	0.86%	0.63%
Proposed (i)	1.27%	1.31%	1.01%
Proposed (ii)	0.70%	0.78%	0.54%
Proposed (iii)	0.62%	0.71%	0.50%
Proposed (iv)	0.57%	0.54%	0.32%

This is expected, since the transcoder operates on a quantized sequence, being limited to the quality of this sequence.

In Fig. 3(a) to 3(c), two possible configurations of the proposed transcoder are compared to two versions of the trivial transcoder: (i) the proposed transcoder, always testing all partition sizes; and (ii) the proposed transcoder, using the proposed approach, as mentioned in Sec. 3.1. It can be seen that the proposed transcoder outperforms the trivial transcoder using hexagon search for all sequences and bitrates, with a gain of up to 1.2 dB for low bitrates in City sequence. Also, the loss of the proposed transcoder compared to the trivial transcoder with full ME is always lower than 0.2 dB for configuration (i) and 0.3 dB for configuration (ii). The performance difference between the two configurations is only noticeable for the City sequence, which uses 5 TD levels, while the other two sequences use only 3. In this case, the RF can be up to 15 frames apart from the current frame, which makes the H.264/AVC partition less reliable for choosing the W-SVC partition.

In Fig. 3(d) to 3(f) the results for other configurations are shown. The configurations tested are: (ii) the proposed transcoder; (iii) the proposed transcoder, using the CBP information, as mentioned in Sec. 3.2; and (iv) the proposed transcoder, using CBP information and the MV similarity, as mentioned in Sec. 3.3, using a threshold $T = 0.5$ at the integer-pixel scale. It can be seen that the difference between the configurations (ii) and (iii) is hardly noticeable for the sequences encoded with $QP = 20$, but it is more pronounced (but still small, up to 0.05 dB) when $QP = 28$ is used. This is because there are less residual encoded when a higher QP is used, and thus the H.264/AVC MV is directly reused more often. The difference between the configurations (iv) and (ii) is higher for City sequence, which uses 5 TD levels, (up to 0.3 dB), but still very small for Soccer and Crew sequences (up to 0.1 dB and 0.05 dB, respectively).

To analyze the transcoder complexity, the average number of SAD computations per pixel is computed. A SAD calculation represents calculating an absolute difference between two pixels or coefficients and adding this difference to a previously accumulated difference. The presented numbers show the number of SAD calculations normalized to the number of SAD calculations of the trivial transcoder using full ME. The results are shown in Table 2. It can be seen that the method (i) has a higher complexity than the trivial transcoder using hexagon search, and that the method (ii) has a complexity that is similar to this, but still slightly lower. The use of the CBP to avoid MV refinement reduces the transcoder complexity by up to 8% when $QP = 20$ is used, and by up to 12% when $QP = 28$ is used. Again, the refinement of MVs is skipped more often when a higher QP is used. The use of CBP together with grouping the MVs reduces the complexity by up to 40% when comparing to method (ii).

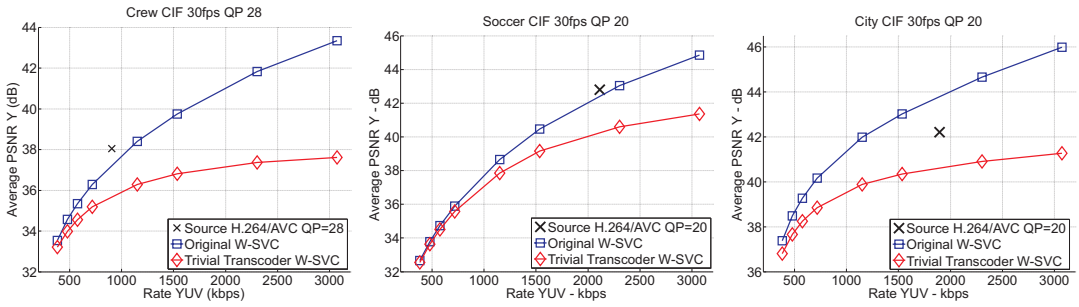

Fig. 2. Results of the W-SVC codec and the trivial transcoder for: (a) Crew (3 TD); (b) Soccer (3 TD); (c) City (5 TD).

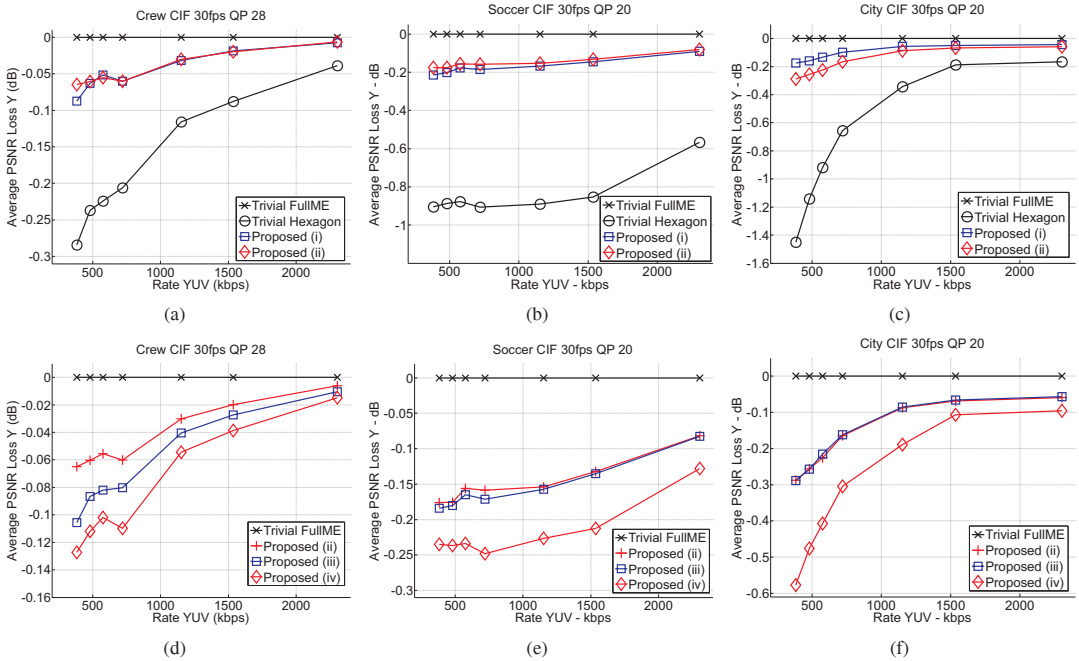

Fig. 3. Transcoder results: (a) Crew (3 TD); (b) Soccer (3 TD); (c) City (5 TD); (d) Crew (3 TD); (e) Soccer (3 TD); (f) City (5 TD).

It can be seen that the proposed transcoder has lower or similar complexity as a fast search algorithm, but it still yields better results, specially in the low to mid bitrates. All the transcoding strategies were outperformed by the full-search ME, but it is important to notice that, in the original H.264/AVC encoded sequence, there was no backward MV, so the backward MVs for the refined sequences had to be entirely calculated using the approximation-refinement framework.

5. CONCLUSION AND FUTURE WORK

In this paper, a transcoder from H.264/AVC to a Wavelet-Based SVC codec was presented. The proposed transcoder has a complexity similar or smaller than that of the trivial transcoder using a fast search algorithm (hexagonal search), but it consistently outperforms it at all bitrates, showing a performance close to the trivial transcoder using full motion estimation. Two new ways of reducing the transcoder complexity are discussed, yielding similar results, but further reducing the complexity by up to 40%.

For future work, the effect of performing more temporal decompositions on the W-SVC codec will be analyzed. Also, other H.264/AVC configurations and qualities will be studied.

6. REFERENCES

[1] T. Wiegand, G. Sullivan, G. Bjontegaard and A. Luthra, "Overview of the H.264/AVC Video Coding Standard", in *IEEE Trans. on Circuits and Systems for Video Technology*, vol.13, pp. 560-576, Jul. 2003.

[2] A. Vetro, C. Christopoulos and H. Sun, "Video Transcoding Architectures and Techniques: An Overview", in *IEEE Signal Proc. Mag.*, vol.20, pp. 18-29, Mar. 2003.

[3] J. Xin, C.-W. Lin and M.-T. Sun, "Digital Video Transcoding", in *Proceedings of the IEEE*, vol.93, pp. 84-97, Jan. 2005.

[4] J.D. Cock, S. Notebaert, P. Lambert and R.V. de Walle, "Architectures for Fast Transcoding of H.264/AVC to Quality-Scalable SVC Streams", in *IEEE Trans. on Multimedia*, vol. 11, n. 7, pp 1209-1224, Nov. 2009.

[5] , *Advanced Video Coding for Generic Audiovisual Services*, ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), ITU-T and ISO/IEC JTC 1, V.8, Nov. 2007.

[6] N. Sprljan, M. Mrak, T. Zgaljic and E. Izquierdo, "Software proposal for Wavelet Video Coding Exploration group", in *ISO/IEC JTC1/SC29/WG11/MPEG2005, M12941, 75-th MPEG Meeting*, Jan. 2006.

[7] Y. Andreopoulos, A. Munteanu, J. Barbarien, M. van der Schaar, J. Cornelis and P. Schelkens,"In-band motion compensated temporal filtering," Signal Processing: Image Communication, Vol. 19, No. 7, pp. 653 - 673, Aug. 2004.

[8] N. Adami, M. Brescianini, M. Dalai, R. Leonardi, and A. Signoroni, "A fully scalable video coder with inter-scale wavelet prediction and morphological coding", Proc. SPIE Visual Comm. and Image Proc. (VCIP), Vol. 5960, Jul. 2005.

[9] S.-J. Choi and J. W. Woods, "Motion-compensated 3-D subband coding of video", in *IEEE Trans. on Image Processing*, vol.8, pp.155-167, Feb. 1999.

[10] C. Zhu, X. Lin, L. P. Chau, K. P. Lim, H. A. Ang and C. Y. Ong "A novel hexagon-based search algorithm for fast block motion estimation", in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 1593-1596, Jun. 2001.

[11] J. Youn and M. Sun, "Motion vector refinement for high-performance transcoding", in *IEEE Trans. on Multimedia*, vol.1, pp. 30-40, Mar. 1999.

[12] T. Shanableh and M. Ghanbari, "Heterogeneous video transcoding to lower spatio-temporal resolutions and different encoding formats", in *IEEE Trans. on Multimedia*, vol.2, pp. 101-110, Jun. 2000.