Complexity Scalable $2:1$ Resolution Downscaling MPEG-2 to WMV Transcoder with Adaptive Error Compensation

Guobin Shen*, Yuwen He†, Wanyong Cao*, Shipeng Li*
* Microsoft Research Asia, Beijing, 100080, P.R.China
† Panasonic Research Lab, Singapore
E-mail: {jackysh, wanycao, spli}@microsoft.com; Yuwen.He@sg.panasonic.com

***Abstract*— In this paper, we focus on $2:1$ spatial resolution downscaling transcoding from MPEG-2 to WMV. We propose two architectures (for sequences with or without B-frames respectively) that are unique in their complexity scalability and efficient control over the drifting error, which in return provide a flexible mechanism to achieve desired tradeoff between the complexity and the quality. We achieve resolution downscaling completely in the DCT domain and show that the standard IDCT (as in all the MPEG series standards) can be merged with other DCT-like transform (e.g., the integer transform in WMV) with proper one-time per-element scaling. Extensive experimental results verified the effectiveness of proposed structures against several design objectives such as complexity scalability and performance tradeoffs.**

I. Introduction

Transcoding refers to the general process of converting one compressed bit stream into another compressed one subjected to specific requirement on the format conversion, and/or changes of coding parameters such as bit rate, frame rate, spatial resolution and/or their combinations. We are particularly interested in transcoding from MPEG-2 format to Windows Media Video (WMV) format [1],[1] given the dominant position of MPEG-2 in the content space and that of WMV in the streaming world. More specifically, we study efficient *spatial resolution downscaling transcoding* solutions besides format conversion and bit rate reduction to fulfill the ever increasing universal access requirement to multimedia content.

Typical spatial resolution downscaling transcoding scenarios can be classified into two classes, namely regular $2:1$ downscaling and arbitrarily downscaling. The former class includes applications that transcode, for example, from PAL Standard Definition (704×576) to CIF (352×288) and from CIF to QCIF (176×144). The latter class includes transcoding High Definition (e.g., 720p) video to SD video such as 480p.

In this paper, we focus on $2:1$ spatial resolution downscaling transcoding from MPEG-2 to WMV. Having identified that specific shortcuts exist which can significantly improve the transcoding speed, we propose an architecture that is unique in its complexity scalability and efficient control over the drifting error, which in return provides a flexible mechanism to achieve desired tradeoff between the complexity and the quality.

The rest of the paper is organized as follows: in Section II, we perform a short literature survey with emphasis on works that are closely related to this work. We propose an efficient MPEG-2 to WMV transcoding architecture with complexity scalability (using dynamic switches) and adaptive error compensation in Section III. Experimental results are presented in Section IV. Finally, Section V concludes the paper.

[1]WMV is often referred to by its SMPTE codename VC-1. We will use them interchangeably in this paper.

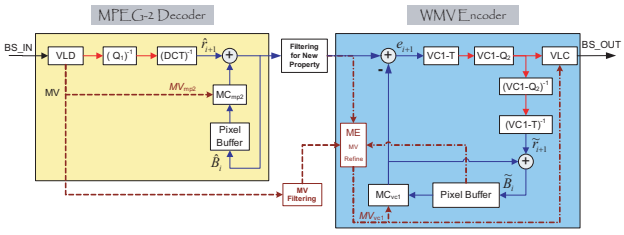

Fig. 1. Block diagram of the reference cascaded pixel-domain transcoder for MPEG-2 to WMV transcoding.

II. Related Works

A. Reference Cascaded Pixel-Domain Transcoder

We first present in Fig. 1 the cascaded pixel-domain MPEG-2 to WMV transcoder (CPDT) which will serve as the reference for the derivation of the proposed transcoders. Some symbols are embedded in the figure. Throughout the paper, subscripts $mp2$ and $vc1$ indicate operations or parameters in the MPEG-2 decoding stage and the WMV encoding stage, respectively. For example, $MC_{vc1}(X, \overrightarrow{mv})$ stands for motion compensation with motion vector $\overrightarrow{mv}$ on reference X, using WMV interpolation filtering. $D(\cdot)$ stands for downscaling operation. B and b represent reconstructed frames at original resolution and reduced resolution, respectively.

B. Complexity Scalability And Adaptive Error Compensation

A complexity scalable MPEG-2 transcoder (CST) for bit rate reduction with graceful quality degradation was proposed in [2] by introducing two switches. However, there is no control over the error introduced by the switches and leads to quality degradation over frames. In our recent work [3], we developed an efficient MPEG-2 to WMV transcoder (hereafter referred to as AEC-DST) that features adaptive error compensation and complexity scalability, as shown in Fig. 2. The complexity scalability is achieved via various switches (see Table I) and adaptive error compensation is done by introducing an error accumulation buffer and monitoring the accumulated error. Once that error exceeds a certain threshold, an error update is performed to compensate it back and stop the propagation. A switch is dedicated to control the computation overhead of error accumulation.

C. Spatial Resolution Downscaling Transcoder

1) DCT-domain downscaling: Conventional transcoder achieves downscaling via spatial domain low-pass filtering and decimation processes. However, as shown in [4], the low-pass filtering and decimation processes can be combined and performed directly in the DCT domain which leads to high computation savings because the downscaled signal needs not to go through another DCT process again. For example, directly performing a 4×4 inverse DCT on the

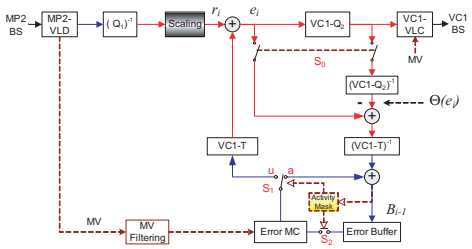

Fig. 2. MPEG-2 to WMV transcoder with adaptive error compensation and dynamic switches, without spatial resolution downscaling [3].

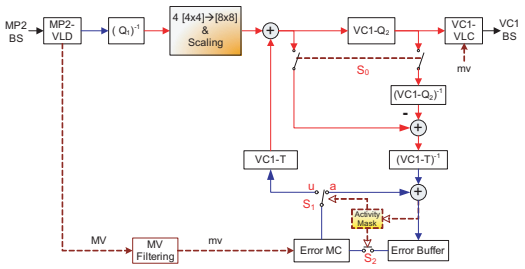

Fig. 3. Simplified DCT-domain 2:1 resolution downscaling transcoder.

4×4 low-frequency DCT coefficients of an 8×8 block gives a low-pass filtered and half-decimated version of that block. It was also advocated that using DCT-domain downscaling, both the PSNR and the visual quality are better besides the complexity reduction [5].

2) MV and mode composition: One key factor in fast transcoding is the reuse of the information (mainly motion information) carried in the incoming bit stream. The common idea is to derive the new candidate MV and mode from the MVs and modes associated with the corresponding macroblocks at the original resolution.

For MV composition, different weighting methods were proposed in [6]–[8]. Note but MV filtering is not always necessary if the target format supports finer MV coding modes. The derived MVs can further serve as the seeds for refining MV search coding efficiency optimization.

For mode composition, it is challenging to decide a good coding mode if the four original MBs have different coding modes. This process is called *mixed-block processing*. Typical mode decision strategy is a majority-based decision whereas other weighted decision logics are possible. Basically, there are two possibilities for mixed-block processing, namely Intra-to-Inter and Inter-to-Intra, i.e., to convert between Inter and Intra modes [9]. Note that these mode modification mechanisms implies a decoding loop to reconstruct the full resolution picture.

III. Proposed MPEG-2 to WMV Transcoder with 2:1 Resolution Downscaling

Generally, there are three sources of errors for transcoding with spatial resolution downscaling:

1) Downscaling: since we intend to obtain a downscaled video, this kind of errors are inevitable.
2) Requantization error: it is considered most unnoticeable compared with other two sources of errors for spatial resolution downscaling transcoding. With proper residue error compensation and higher bit rate, this error can be eliminated.
3) Motion error: Incorrect motion vector will lead to wrong motion compensated prediction. Worse even, this error can only be compensated by re-doing motion compensation based on the new MVs and modes.

We will address the last two sources of errors one by one below and propose corresponding architectures to cope with them.

A. Requantization Error Compensation

From the reference CPDT transcoder shown in Fig. 1, we can derive the input to the VC-1 encoder for frame (i+1) as follows:

$$e_{i+1} = D(\hat{r}_{i+1}) + D\left(MC_{mp2}(\hat{B}_i, \overrightarrow{MV}_{mp2})\right) - MC_{vc1}(\tilde{b}_i, \overrightarrow{mv}_{vc1})$$

With assumptions that downscaling and motion compensation processes are commutable, motion compensation is linear and $MC_{vc1} = MC_{mp2}$, and reuse the motion vectors, we can derive

$$e_{i+1} = D(\hat{r}_{i+1}) + MC_{mp2}(\hat{b}_i - \tilde{b}_i, \overrightarrow{mv}_{mp2}) \quad (1)$$

The first term in (1), $D(\hat{r}_{i+1})$, refers to the downscaling process to the decoded MPEG-2 residue signal. We adopt the *DCT-domain downscaling*. In WMV, for P-frames and B-frames, the 4×4 transform size is allowed. Therefore, we can directly map the an input MB at the original resolution to a constituent 8×8 block of the new output MB at the reduced resolution, all in DCT domain. In other words, the DCT-domain downscaling can be achieved extremely simple by retaining only the top-left 4×4 low-frequency DCT coefficients of an incoming 8×8 block. Note that we need to scale each remained 4×4 sub-block with S_{44}, which is given below because WMV adopts a DCT-akin integer transform (called VC1-T hereafter).

$$S_{44} = diag(T_4 C_4') \cdot diag(C_4 T_4')' \circ N_{44}$$

where C_4 and T_4 are the 4×4 transform matrices for standard DCT and VC1-T, respectively. N_{44} is the normalization matrix of VC1-T, given by $N_{44} = c_4' \cdot c_4$, with $c_4 = \left[\frac{1}{1156} \ \frac{1}{1168} \ \frac{1}{1156} \ \frac{1}{1168}\right]$.

However, for I-frames, only the 8×8 transform type is allowed. Consequently, we need to merge the four 4×4 low-frequency DCT sub-blocks into an 8×8 VC1-T block. This is a well studied topic [10]. The difference is still the replacement of standard DCT with VC1-T and the normalization to the final results with N_{88} [3].

The second term in (1), $MC_{mp2}(\hat{b}_i - \tilde{b}_i, \overrightarrow{mv}_{mp2})$, implies requantization error compensation on a downscaled resolution. Clearly, the MC in MPEG-2 decoder and that in WMV encoder are merged to a single MC process that operates on accumulated requantization errors at the reduced resolution.

1) Complexity scalability with dynamic switches: Comparing (1) against the second equation in [3], one immediately finds that they are almost the same except that now it is operating on the reduced resolution while the one in [3] is on the original resolution. Therefore, the proposed complexity scalable scheme can be applied here as well. The resulting simplified DCT-domain 2:1 resolution downscaling transcoder is shown in Fig. 3. The only structural difference between this one and that in Fig. 2 is the *scaling* module (which is denoted differently on purpose). All the switches have the same function. Note that in Fig. 3, the first two modules (MPEG-2 VLD and inverse quantization) can be more efficiently implemented since only the top-left 4×4 portion out of the 8×8 block needs to be processed.

2) Mixed-block processing: An interesting observation is that the mixed block processing module is avoided in Fig. 3, thanks to the fact that WMV supports *mixed mode* by allowing up to three consisting 8×8 blocks of an Inter coded MB to be coded with Intra mode. In other words, we allow an Intra MB at the original resolution to be mapped into an Intra 8×8 block of an Inter MB at the reduced resolution. Recall that mixed block processing requires a decoding

loop to reconstruct the full resolution picture. Therefore, the removal of mixed block processing module implies significant computation savings. The final MB mode mapping rule is simply:

$$\text{Mode}_{vc1} = \begin{cases} Intra & \text{if all Mode}_{mp2} = Intra \\ SKIP & \text{if all Mode}_{mp2} = SKIP \\ Inter & \text{otherwise} \end{cases}$$

B. Motion Error Compensation

Although WMV supports four MV coding mode, it is intended for P-frames only. As a result, the architecture shown in Fig. 3 is recommended to use only when there are *no* B-frames in the input MPEG-2 stream or the B-frames are to be discarded during the transcoding towards a lower temporal resolution.

Due to the constraint that only one MV is allowed for B-frame MBs in WMV, we have to compose a new motion vector from the four MVs associated with the MBs at the original resolution. All the aforementioned MV composition methods can be applied here. In our implementation, we adopted median filtering. As mentioned earlier, incorrect MV will lead to wrong motion compensated prediction. Worse even, one can never get it back if not redoing motion compensation based on the new MVs. Therefore, we have to come up with an architecture that allows such motion errors to be compensated.

Under this circumstance, the assumption that $\overrightarrow{mv}_{mp2} = \overrightarrow{mv}_{vc1}$ does not hold any more. However, we can manipulate (1) and obtain:

$$e_{i+1} = D(\hat{r}_{i+1}) + \left[MC_{mp2}(\hat{b}_i, \overrightarrow{mv}_{mp2}) - MC_{vc1}(\hat{b}_i, \overrightarrow{mv}_{vc1})\right] + MC_{vc1}(\hat{b}_i - \tilde{b}_i, \overrightarrow{mv}_{vc1}) \quad (2)$$

Clearly, the two terms in the square brackets in (2) implies the compensation of the motion errors caused by inconsistent MVs or caused by different MC filtering methods between MPEG-2 and WMV. Note that in (2), $MC_{mp2}(\hat{b}_i, \overrightarrow{mv}_{mp2})$ is performed for all the 8×8 blocks that correspond to original Inter MBs, and $\overrightarrow{mv}_{mp2} = \overrightarrow{MV}_{mp2}/2$ with quarter pixel precision. The $\overrightarrow{mv}_{vc1}$ is a single MV, which is the median of the MVs of the four corresponding MBs at the original resolution and can go to quarter-pixel precision. The last term in (2) is to compensate the requantization error of reference frames. Since B-frames are not used as reference frames, this error compensation can be safely turned off for B-frames to achieve higher speed.

As to the mode composition, we can either apply Intra-to-Inter or Inter-to-Intra conversion easily since we have reconstructed the B-frame and the reference frames at the MPEG-2 decoder part, both at already reduced resolution. This conversion is done in the mixed block processing module in Fig. 4. Two mode composition methods are possible: one method is to select the dominant mode as the new mode. The other method is to select the mode as the one will lead to largest error. Experimental results shows that the latter method offers slightly better quality because it provides an opportunity to compensate the large error. Similar thinking is revealed in [8] where the align-to-worst offers better quality than other schemes.

The final architecture according to (2) is shown in Fig. 4, where the modules highlighted and grouped into a shaded block perform compensation to motion errors. The resulting architecture is seemingly just as complex as the reference cascaded pixel-domain transcoder. It is *not* actually. The explicit pixel-domain downscaling process is avoided. Instead, it is implicitly achieved in the DCT domain by simply discarding the high DCT coefficients. More importantly, the resulting architecture has excellent complexity scalability which can be achieved by various switches.

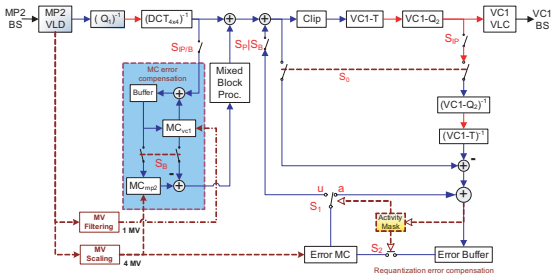

Fig. 4. Simplified 2:1 downscaling transcoder with full drifting error compensation.

Since both architectures in Fig. 3 and Fig. 4 are intended for the same purpose but to handle different cases where B-frames exist or not, we refer to both of them as AEC-DST hereafter.

There are three block-level switches and four frame-level switches in in Fig. 4. Their functions are listed in Table I.

TABLE I
FUNCTIONALITIES OF VARIOUS SWITCHES IN AEC-DST. REFER TO FIG. 2, FIG. 3 AND FIG. 4 FOR THEIR POSITIONS.

Symbol	Description
S_0	Block level error accumulation switch
S_1	Block level error update switch
S_2	Block level early skip block decision switch
S_{IP}	Frame level switch, closed for I- and P-frames
$S_{IP/B}$	Same as S_{IP}, applicable only if there are B-frames
S_P	Frame level switch, closed for P-frames
S_B	Frame level switch, closed for B-frames (= $\sim S_{IP}$)

The three block-level switches are discussed in detail in [3]. The four frame-level switches ensure different coding paths for different frame types to achieve complexity scalability and performance tradeoffs between coding efficiency and speed. Specifically, no residue-error accumulation is performed for B-frames (S_{IP}), no MV error compensation is performed for I- and P-frames (S_B), and no reconstruction of reference frames if there is no B-frames to be generated ($S_{IP/B}$).

In short summary, thanks to the support of four-MV and mixed coding mode for P-frames in WMV, both the requantization error and motion error compensation can be efficiently achieved and controlled by various switches towards complexity scalability. However, for B-frames, there is constraint of one-MV coding mode. As a result, motion error compensation has to be performed by full reconstruction of the input signal but at a reduced resolution, i.e., through partial MPEG-2 decoding. Various frame-level switches are introduced for complexity reduction.

IV. PERFORMANCE ANALYSIS

We have performed extensive experiments to verify the effectiveness of the proposed transcoding architectures. The experimental platform is Windows XP PC with Pentium-IV 3-GHz CPU and 512 MB memory. Two test sequences were used. One is BestCap whose resolution is 640×480. The bit rate of the input MPEG-2 bitstream is 5.7 Mbps and the PSNR is 44.42 dB. The other is SmallTrap which is of standard definition (720x480). The bit rate of the MPEG-2 input bitstream is 5.2 Mbps and the PSNR is 43.22 dB. Both sequences consist of quite a few typical video scenes such as slow motion, high motion, fading, low texture and high texture, etc.

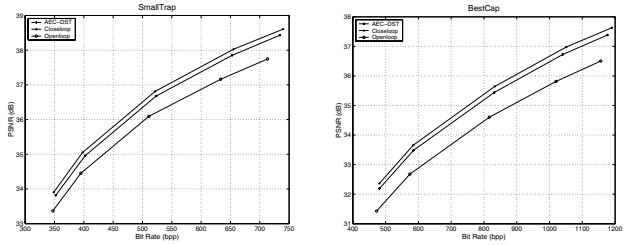

Fig. 5. Coding efficiency comparison of different transcoding schemes.

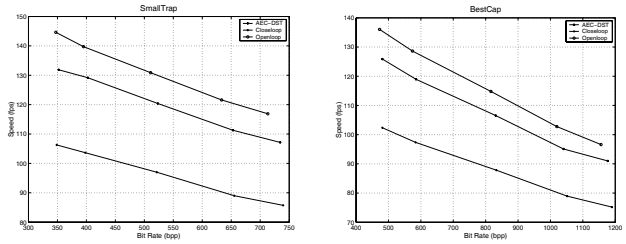

Fig. 6. Speed comparison of different transcoding schemes.

A. Coding Efficiency Comparisons

Fig. 5 depicts the coding efficiency performances for the open-loop (i.e., transcoder with no feedback loop at all), close-loop (i.e., transcoder with full connected feedback loop) and proposed AEC-DST for SmallTrap and BestCap sequences. The ground truth for PSNR calculation is obtained by applying the same DCT-domain down-sampler (as the one used in the MPEG-2 decoding part in the transcoder) on the original video. Doing this way, we minimize the impact of different downscaling filters and focus on the transcoding efficiency.

From the figures, we can see that the coding efficiency of AEC-DST is very close to that of the close-loop transcoder and is significantly better than that of the open-loop transcoder.

B. Speed Comparisons

Fig. 6 shows the speed comparison. From these figures, we can clearly seen that the speed of AEC-DST is very close to that of the open-loop transcoder and is significantly faster than that of the close-loop transcoder. Another observation, which echoes the observation in [11], is that the transcoding complexity is closely related to the output bit rate.

C. Complexity Scalability And Performance Tradeoffs

As stated before, with the proposed schemes, the application can find a desired tradeoff between quality and speed. The tradeoff is controlled by the switches which is application adjustable using thresholds. The speed changes against the thresholding levels for SmallTrap sequence is shown in Fig. 7. In the figure, the bit rate change and PSNR change are depicted by the dash-dotted short line and dashed short line on each anchor point. Clearly, a faster speed generally comes at a larger PSNR penalty. However, the loss may not be as significant as the numbers indicate since the corresponding rate is slightly reduced as well.

V. Concluding Remarks

In this paper, we studied the problem of efficient transcoding from MPEG-2 to WMV format with $2:1$ spatial resolution downscaling.

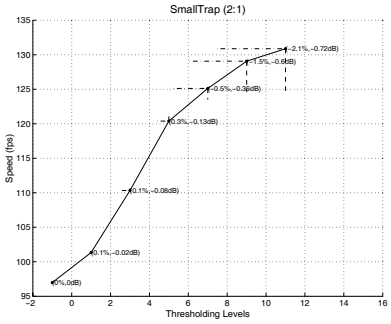

Fig. 7. Complexity scalability of AEC-DST scheme. Thresholding level -1 means Close-loop and 11 stands for Open-loop. Coding efficiency (PSNR, bit rate) benchmark is Close-loop.

We proposed two architectures (for sequences with or without B-frames) with adaptive error compensation and dynamic switches which in return offer excellent complexity scalability and adaptive drifting error control. We achieved resolution downscaling completely in the DCT domain and showed that the standard IDCT (as in all the MPEG series standards) can be merged with other DCT-like transform (e.g., the integer transform in WMV) with proper one-time per-element scaling. Extensive experiments demonstrated that the proposed architectures indeed provide excellent complexity scalability and performance tradeoff.

As a final remark, due to the significant overlap between the WMV syntax and that of MPEG-4, this work can also be applied to MPEG-2 to MPEG-4 transcoding applications.

References

[1] Sridhar Srinivasan, et al, "Windows Media Video 9: Overview and Applications," *Signal Processing: Image Communication*, vol. 19, no. 9, pp. 851–875, Oct. 2004.

[2] E. Barrau, "A Scalable MPEG-2 Bit-Rate Transcoder with Graceful Degradation," *IEEE Transactions on Consumer Electronics*, vol. 47, no. 3, pp. 378–384, August 2001.

[3] G. Shen, Y. He, W. Cao, and S. Li, "Complexity Scalable MPEG-2 to WMV Transcoder with Adaptive Error Compensation," accepted for publication by ISCAS 2006.

[4] K. N. Ngan, "Experiments on Two-Dimensional Decimation in Time and Orthogonal Transform Domains," *Signal Processing*, vol. 11, pp. 249–263, 1986.

[5] R. Dugad and N. Ahuja, "A Fast Scheme for Image Size Change in the Compressed Domain," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 4, pp. 461–474, April 2001.

[6] G. Shen, B. Zeng, Y.-Q. Zhang, and M. L. Liou, "Transcoder with Arbitrarily Resizing Capability," in *IEEE International Symposium on Cirsuits and Systems*, vol. 5, 2001, pp. 25–28.

[7] M.-J. Chen, M.-C. Chu, and S.-Y. Lo, "Motion Vector Composition Algorithm for Spatial Scalability in Compressed Video," *IEEE Transactions on Consumer Electronics*, vol. 47, no. 3, pp. 319–325, August 2001.

[8] B. Shen, I. K. Sethi, and B. Vasudev, "Adaptive Motion-Vector Resampling for Compressed Video Downscaling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 6, pp. 929–936, Sept. 1999.

[9] P. Yin, A. Vectro, B. Liu, and H. Sun, "Drifting Compensation for Reduced Spatial Resolution Transcoding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 11, pp. 1009–1020, Nov. 2002.

[10] Y.-R. Lee, C.-W. Lin, and C.-C. Kao, "A DCT-domain Video Transcoder for Spatial Resolution Downconversion," in *Proc. Visual Information System*, Hsinchu, Taiwan, Mar. 2003.

[11] J. Xin, M.-T. Sun, B.-S. Choi, and K.-W. Chun, "An HDTV-to-SDTV Spatial Transcoder," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 11, pp. 998–1008, Nov. 2002.