MULTIPLE DESCRIPTION FOR ROBUST SCALABLE VIDEO CODING

D. Alfonso1, R. Bernardini2, L. Celetto1, R. Rinaldo2, P. Zontone2

1 STMicroelectronics - AST - Agrate Brianza (MI) - Italy
2 University of Udine, DIEGM, Udine, Italy

ABSTRACT

Scalable Video Coding (SVC) was recently standardized as an extension of the H.264/AVC standard. A scalable coder allows to combine different layers of spatial, temporal and quality scalability, and is a viable solution for adaptation to user characteristics and network conditions. On the other hand, Multiple description coding (MDC) can provide error resilience and graceful quality degradation for transmission over error-prone channels. In this paper, we propose a video codec that combines the SVC and MDC coding paradigms. In particular, the enhancement data of the SVC coder is transmitted using MDC. The resulting fully compatible coder takes advantage of the efficiency of SVC and of the robustness of MDC techniques. The effectiveness of the proposed solution is confirmed by experiments.

1. INTRODUCTION

Scalable Video Coding (SVC) [1] was recently standardized as an extension of the H.264-AVC video coding standard [2]. In scalable coding, the video sequence is coded as a base layer (BL) and a certain number of enhancement layers (ELs), which carry additional information to provide different levels of spatial, temporal and quality (SNR) resolution. In particular, SVC offers two different possibilities for SNR scalability, called Coarse Grain Scalability (CGS) and Medium Grain Scalability (MGS) [1]. CGS can be considered a particular case of spatial scalability, obtained when the layers have the same spatial resolution, but different quality. A scalable scheme can be useful to adapt to user characteristics and to varying network conditions. A scalable scheme, however, is typically vulnerable to packet errors, since different layers depend on each other and errors in one layer can prevent decoding of the other layers. To achieve robustness to packet losses, one can adopt an Unequal Error Protection (UEP) strategy [3, 4], typically using Forward Error Correction (FEC) codes with different error recovery capabilities at different layers.

Multiple Description Coding (MDC) paradigm has been proposed for robust transmission [5]. In many practical schemes, the input video signal is downsampled temporally and/or spatially to generate a set of different “descriptions,” which are then separately encoded. At the decoder side, the received descriptions are progressively merged to improve the quality of the reconstructed sequence. In the presence of transmission errors, the inherent correlation among descriptions can be exploited for error concealment. In general, MDC allows a very high level of resiliency to errors especially with error-prone channels with high error rates [5]. When compared with FEC coding, the major benefit of MDC techniques is their capability of showing higher quality when the error rate can greatly vary over time. When the target FEC error correction capacity is exceeded by the actual error percentage in the channel, FEC can not recover the missing data. MDC techniques instead may offer a nice degradation of the quality with the increase of the bit error rate on the channels [5, 6].

The proposed video codec scheme is shown in Fig. 1, while the structure of the preprocessing stage is described as follows. At the encoding side, the full resolution (4CIF) input video signal is spatially downsampled to obtain a set of four descriptions, corresponding to the polyphase decomposition of each input frame. These four descriptions are then averaged to obtain one additional description. All the descriptions have the same spatial resolution (CIF for a 4CIF full resolution video sequence). The additional description is then encoded as the BL of an SVC compliant bitstream, whereas the other representations are coded as CGS ELs, each one of them exploiting inter-layer prediction with respect to the same Base Layer. At the decoder (see Fig. 1), errors in the enhancement layers are concealed by taking advantage of the inter-layer spatial correlation. Then, the five descriptions are optimally combined, as described in Section 2, to reconstruct the full resolution video sequence.

The obtained coding scheme is thus a combination of SVC and MDC coding paradigms. In fact it exploits the SVC framework, since the representations of the enhancement layer are coded independently to effectively exploit the MDC paradigm. Thanks to this new coding scheme, it is possible to take advantage of the benefits of the SVC and MDC coding paradigms and to get rid of the respective weaknesses. In fact, the decreased coding efficiency of the MDC scheme with respect to the SVC scheme, which is due to the inherent redundancy of the data encoded in the separate description streams, is greatly reduced by the fact that the descriptions are actually encoded as enhancement layers in the scalable hierarchy. In this way, a relevant part of the redundancy is eliminated by the SVC inter-layer prediction process, effectively encoding in each CGS layer only the difference between the description and the BL representation. On the other hand, the decreased robustness of the SVC scheme with respect to the MDC scheme, which is due to the fact that an error occurring in the data relative to any layer in the

---

**Fig. 1.** The proposed Multiple Description - Scalable Video codec scheme.
hierarchy affects all the uppermost layers, is reduced by the fact that the enhancement layers are independently encoded. Since CGS enhancement layers only depend on the base layer, any error occurring in a CGS layer will not affect the other ones, allowing very effective error correction at the decoder side. Of course, an error in the BL can affect any of the CGS layers, so that the proposed scheme can be better applied in case of Unequal Error Protection (UEP), giving higher robustness to the BL than the CGS layers.

2. RECEIVER DESIGN

In this section we describe two possible approaches to the problem of reconstructing a block of $2 \times 2$ pixel from the corresponding pixels in the five descriptions. In the following, we will denote by $x = [x_1, x_2, x_3, x_4] \in \mathbb{R}^4$ the vector whose components are the samples of a $2 \times 2$ block, by $y = [y_1, y_2, y_3, y_4] \in \mathbb{R}^4$ the vectors whose components are the samples of the five descriptions and with $\hat{x} = [\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4]$ and $\hat{y} = [\hat{y}_1, \hat{y}_2, \hat{y}_3, \hat{y}_4]$ the corresponding coded versions.

The procedure used to compute the five descriptions can be written as a blockwise processing $y = Fx$, where $x$ and $y$ are as defined above and $F = [u_1, I_2^t]$, where $u_1 = [1/4, 1/4, 1/4, 1/4]$ and $I_2$ is the $4 \times 4$ identity matrix. Let $\hat{y} = \hat{y} + \epsilon$ be the decoded version of $y$, where vector $\epsilon$ represents the coding error with $\mathbb{E}[\epsilon] = 0$ and co-variance matrix $\mathbf{R} = \mathbb{E}[\epsilon\epsilon^t]$. The decoder problem is to reconstruct $x$ from $\hat{y}$. The typical solution, which does not require knowledge of the statistics of $x$, is to do a least-square reconstruction by means of the pseudo-inverse $F^\dagger$ of $F$, that is, $\hat{x} = F^\dagger \cdot \hat{y}$. Such a solution is the least-square optimal one in the sense that $\hat{x}$ is the vector which minimizes $\|y - F\hat{x}\|_2$. Note that since $F^\dagger F = I$, one can recover $x$ exactly when $\epsilon = 0$. Moreover, it is possible to show that if all the quantization noises are uncorrelated and with the same variance (i.e., $\mathbf{R} = \sigma^2 I$), then $F^\dagger$ is the left-inverse of $F$ that minimizes the expected distortion $\mathbb{E}[\|x - \hat{x}\|_2^2]$.

In the scalable video coding setup we are considering, however, the enhancement layer subsequences are coded using the base-layer as a reference. It is therefore possible to code the video subsequences with different distortions, in order to optimize the overall performance or meet a desired rate or quality target for the BL. An interesting problem, therefore, is to design the optimal left inverse $F_{\text{opt}}^\dagger$ of $F$ that minimizes $\mathbb{E}[\|F_{\text{opt}}^\dagger \hat{y} - x\|_2^2]$ for a given $\mathbf{R}$. Matrix $F_{\text{opt}}^\dagger$ does not depend on the statistics of $x$, as claimed in the following property, whose proof is omitted for the sake of space.

**Property 1.** Let $\hat{y} = Fx + \epsilon$ where $\epsilon$ is a random vector with zero mean and covariance matrix $\mathbf{R} = \mathbb{E}[\epsilon\epsilon^t]$. The left inverse $F_{\text{opt}}^\dagger$ of $F$ that minimizes $\mathbb{E}[\|F_{\text{opt}}^\dagger \hat{y} - x\|_2^2]$ is

$$F_{\text{opt}}^\dagger = (F^t \mathbf{R}^{-1} F)^{-1} F^t \mathbf{R}^{-1}.$$  

(1)

Note that for $\mathbf{R} = \sigma^2 I$, one obtains $F_{\text{opt}}^\dagger = F^\dagger$, as already known. Eq. (1) allows to compute the optimal left inverse on the basis of the coding error correlation matrix $\mathbf{R}$. In the experiments we will consider in the following, we suppose that the coding errors are uncorrelated, with a variance that can be determined on the basis of the QP used for each video subsequence. As a matter of fact, the QP value for subsequence $i$ determines the quantization step used for encoding, according to the approximate relation $\Delta_i = \Delta_0 2^{\text{QP}_i}/6$. $\Delta_0 = 0.625$, and the corresponding distortion is approximated as $\Delta_i^2/12$.

3. EXPERIMENTAL RESULTS

For the simulations presented in this section, we used version JSVM 9.7 of the SVC codec, with the following parameters: GOP size: 8, IntraPeriod: 32, BaseLayerMode: 1, FrameRate: 30; QP range: 24–45. Each slice covers 22 macroblocks and has a size of $16 \times 352$ pixels. Coding results are relative to the 4CIF sequences City and Ice. In particular, City is a highly spatially detailed sequence, while Ice has very little spatial detail.

For a given bit budget $R$, the optimal allocation requires detailed knowledge of the overall rate-distortion function, which in turns depends on the reconstruction matrix. Moreover, the problem is further complicated by the fact that the four descriptions in the EL of the proposed SVC scheme are actually coded taking the BL as a reference, and therefore their rate-distortion function depends on the rate allocated to the BL. In order to determine a reasonable coding strategy, in this paper, we performed experiments by using different values of the H.264 quality parameter QP for the base layer BL and the enhancement layers EL. In particular, once a given QP(BL) is chosen for the base layer, we coded the enhancement layers with a different QP(BL) = QP(EL) + $\Delta_Q$, $\Delta_Q \in \{-10, \ldots, 10\}$. Experiments with different video sequences and reconstruction matrices consistently showed that choosing QP(BL) = QP(EL) – 6 (which corresponds to code the BL with a quantization step that is about half of the one used for the EL) gives some advantage for most bitrates. As an example, Table 1 shows the PSNR results obtained with the optimal left-inverse for QP(BL)=31. In the following, we will use QP(BL) = QP(EL) – 6 to code the base and enhancement layers. Other choices are possible to decrease the percentage of the bitrate allocated to the BL. The bitrate of the BL, in Mbit/s, is reported within parenthesis in Table 1, while the last column reports the total bitrate. The difference between the total bitrate and the BL bitrate is used to code the 4 ELs, which require approximately the same bitrate.

<table>
<thead>
<tr>
<th></th>
<th>PSNR (dB) as a function of $\Delta_Q$ (QP(BL)=31)</th>
</tr>
</thead>
<tbody>
<tr>
<td>City</td>
<td>$\Delta_Q$=6</td>
</tr>
<tr>
<td>City</td>
<td>32.56 (1.206)</td>
</tr>
<tr>
<td>Ice</td>
<td>38.33 (.574)</td>
</tr>
</tbody>
</table>

**Comparison of the reconstruction techniques with no-losses:** we compare the performance of the proposed Multiple Description Scalable Video Coder (MD-SVC), using the different reconstruction techniques considered in this paper, with the performance of the standard CIF-4 CIF SVC scalable coder (SVC). As mentioned before, we expect that, when there is no loss, the MD-SVC has lower performance than SVC, due to the inefficiency of separate coding of the enhancement layer descriptions. We also compare the proposed scheme with a robust Multiple Description scheme (MD), where the four polyphase components of the input sequence are coded by 4 independent H.264/AVC coders [8] and transmitted in multicast as four descriptions. Figures 2–3 compare the performance of MD and SVC with the proposed schemes, when different reconstruction strategies are employed at the receiver. The plots report the luminance PSNR vs. the total bitrate. Note that the proposed scheme actually transmits 5 descriptions, i.e., the BL and four EL descriptions. Besides considering the pseudo-inverse (PI) and the optimal left-inverse (OLI), we also plot the results obtained by using the Mean Squared error reconstruction (MS), where the optimal matrix $M$ that minimizes $\sum_i \|x_i - M\hat{x}_i\|_2^2$ (where $x_i$ is the $i$-th $2 \times 2$ block and $\hat{y}_i$ is the corresponding $2 \times 2$ coefficient vector) is computed at the encoder at the beginning of each GOP and sent to the receiver as side information. We also plot the results obtained when the BL is used at the receiver.
to decode the four EL descriptions, but is not used for the reconstruction. In other words, this solution reconstructs the 4CIF sequence via an inverse polyphase transform, by simply joining together the four descriptions (Join). It can be seen from the plots that the proposed scheme has better performance than the MD scheme. Even if five description are transmitted, coding of the EL descriptions can efficiently make use of the correlation with the BL. Moreover, we will show below that, in the presence of transmission errors, the proposed scheme has an error resilience comparable to that of the MD scheme, and much superior to that of the SVC solution. The plots show that using the BL for the reconstruction of the 4CIF sequence gives some advantage with respect to the Join solution (more than 1 dB at lower bitrates). The pseudoinverse, the optimal left inverse and the mean squared error reconstruction give similar results, with MS providing a gain of about 0.2 dB. Note that, since the BL is coded with QP(BL)=QP(EL) - 6, independently of QP(EL), the distortion in the BL is 1/4 of that of the EL and, as a consequence, the optimal pseudo-inverse does not depend on QP(BL).

Performance of the Optimal Left-Inverse solution in the presence of losses: in the following, we compare the proposed MD-SVC scheme with MD and SVC in the presence of losses. In all schemes, we suppose that one CIF description, i.e., the BL in MD-SVC and SVC, and one polyphase component in MD, are received without errors. This requires to adequately protect one of the substreams, and this is a reasonable assumption in a practical scheme where one wants to assure that the user receives at least one low spatial resolution description. We do not take into account the required additional bitrate. Errors are due to packet losses, each packet containing one slice. We assume that packets are lost independently with probability $P_{\text{loss}}$. For the MD and MD-SVC, we conceal errors in the descriptions using the low-complexity algorithm described in [9]. Then, in the proposed solution, the actual EL reconstruction is performed by using the appropriate matrix on the BL and the concealed EL descriptions. For conciseness, for MD-SVC we report the results relative to the use of the optimal left-inverse for reconstruction at the receiver. As for the no-loss case, the use of the MS reconstruction gives a slight advantage at the expense of an increased coder complexity. For the SVC scheme, EL error concealment is performed at the receiver via interpolation of the BL using the upsampling method 0 of the JSVM software. Concealed frames are copied to the decoder frame buffers in order to prevent error propagation. Figure 4 shows the luminance PSNR vs. the bitrate for different values of $P_{\text{loss}}$ for the video sequence City. Similar figures are obtained for other sequences. The plot also report the performance of the MD scheme (dash-dotted lines). It is evident from the figure that both MD and MD-SVC allow for a graceful quality degradation in the presence of losses. Moreover, MD-SVC has clear advantages for all sequences and almost all bitrates (the MD scheme has a slight advantage with respect to MD-SVC at high bitrates for City). Note that the visual quality of SVC in the presence of errors may not be acceptable, even if the PSNR decrease is not significant for moderate $P_{\text{loss}}$ (see Fig. 5). We further investigate this matter by considering the case when entire descriptions are lost in the proposed MD-SVC scheme, and compare the performance with SVC at a corresponding $P_{\text{loss}}$. Note that the loss of one or more entire descriptions can be typical of a transmission scheme that uses multiple routing paths such as, for instance, a P2P transmission system with multiple trees. Figure 6 shows the frame-by-frame luminance PSNR results for the sequence City, when one entire description is lost in MD-SVC, and at corresponding $P_{\text{loss}} = 0.25$ for SVC. As a comparison, we report in the figures the PSNR obtained with SVC and no error in the EL. It is evident from the figure that the PSNR degradation for the SVC scheme is very significant and that MD-SVC provides better results and consistent quality.

4. CONCLUSIONS

In this paper, we proposed an MD technique for the transmission of the EL in a scalable CIF-4CIF video coder. The proposed coder is fully compatible with SVC, and requires limited post-processing at the decoder. The problem of subsequence distortion allocation and optimal reconstruction have been considered. The proposed solution has several advantages. It shares with other multiple description techniques the characteristic of being very robust to packet losses and allowing a graceful quality degradation. This is very important in emerging applications like P2P video streaming. On the other hand, the proposed MD-SVC scheme mitigates the efficiency loss of some multiple description techniques, since the descriptions are coded by taking the BL as a reference in the framework of the SVC standard.
Fig. 4. PSNR vs. rate using the optimal left inverse for different values of \( P_{\text{loss}} \) (City sequence - continuous lines). Comparison with MD multicast (dash-dotted lines).

Fig. 5. Sequence City. Packet loss probability \( p = 0.1 \) (the rectangle evidences the position of the lost slices in the SVC scheme). (a) Reconstruction from the SVC bitstream using the BL for concealment. (b) Reconstruction using the proposed robust scheme (one entire description lost).

Fig. 6. Sequence City: PSNR comparison between the proposed MD-SVC scheme and a standard CIF-4 CIF scalable bitstream (SVC 1 lost) when one description is lost (\( P_{\text{loss}} = 0.25 \)).

5. REFERENCES


