EXTENDING SVC BY CONTENT-ADAPTIVE SPATIAL SCALABILITY

Yongzhe Wang[1], Nikolce Stefanoski[1], Manuel Lang[1,2], Alexander Hornung[1], Aljoscha Smolic[1] and Markus Gross[1,2]

[1]Disney Research Zurich [2]ETH Zurich

ABSTRACT

This paper provides details on a complete integration of *Content-adaptive Spatial Scalability* (CASS) into the scalable video coding extension of H.264/AVC (SVC). CASS enables the efficient encoding of a high-quality bit stream that contains several versions of an original image sequence. Thereby, each such image sequence has been created by content-adaptive and art directed retargeting to different display aspect-ratios and/or resolutions. Non-linear dependencies between spatial layers, which have been introduced through content-adaptive retargeting, are exploited by a generalization of the three inter-layer prediction tools of SVC, i.e. by content-adaptive inter-layer texture, motion and residual prediction. The CASS extended SVC enables the transmission of video content which has been specifically adapted in an *art-directed* way to multiple display configurations (e.g. to SD and HD displays with 4:3 and 16:9 aspect-ratios, respectively) using a single compressed bit stream. With our extension, video content of higher semantic quality can be transmitted in a scalable way by introducing an average overhead in bit rate of 9.3%.

Index Terms— H.264/AVC, scalable video coding, spatial scalability, content-adaptation, non-linear image warping

1. INTRODUCTION

Today, TV and video services are consumed using various types of display devices, like traditional TV sets, home theater projectors, or smart phones. Hence, we have a heterogeneous environment of display devices available, where different display aspect-ratios (e.g. 4:3, 5:3, 5:4, 16:9) and resolutions (e.g. SDTV, HDTV, VGA, XGA) are natively supported. Content providers usually employ single resolution video coding standards like H.264/AVC for transmission or storage. As a consequence, content producers face the problem that they cannot control the way how the decoded video content is retargeted at the consumer side to the available displays. However, SVC [1][2], the scalable extension of H.264/AVC, allows the joint transmission of several videos with different aspect-ratios and resolutions, which have been created through cropping and/or linear scaling (Fig. 1). Consequently, if SVC is used for coding and transmission, an appropriate video for the available display at the consumer side could be decoded, while inherent dependencies between the transmitted videos are exploited for an efficient overall compression. This can be achieved with SVC with help of the concept of Extended Spatial Scalability [3]. However, usually cropping and linear scaling do not provide enough flexibility for high quality video retargeting. These retargeting methods can change the original vision and intention of a director or cinematographer and distort salient image regions in an unacceptable way.

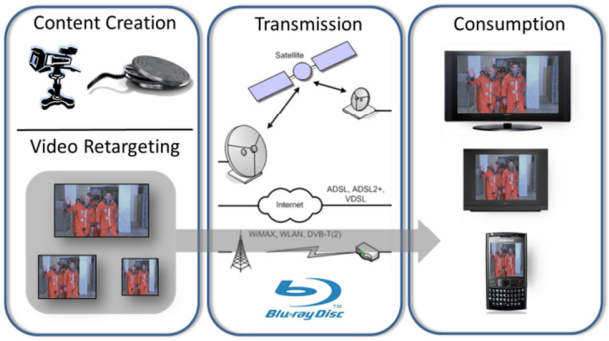

Fig. 1. General transmission scenario for retargeted video content.

Recently, a lot of research is conducted in the area of automatic and semi-automatic image and video retargeting [4]. Retargeting is performed in a content-adaptive way, e.g. by applying non-linear scaling operations through image warping [5], where visually salient regions are preserved and distortions are hidden in less important areas. A recent large-scale subjective study [4] showed that [5] is one of the currently best performing retargeting methods. A method for content-adaptive coding of single-resolution video was also recently presented in [6].

In our previous work, we presented a framework for content-adaptive and art-directable scalable video coding [7]. In particular, we showed that warps can be efficiently encoded and integrated in this framework. Furthermore, we introduced the concept of CASS for SVC, integrated in SVC a tool for content-adaptive inter-layer texture prediction and showed its high efficiency [8]. In this paper, we present generalizations of all three inter-layer prediction tools of SVC and provide an overall evaluation of our integration of CASS in SVC. First, in Section 2, a short overview of CASS extended SVC is given. Section 3 describes the three generalized inter-layer prediction tools in detail. Then, evaluation results are presented in Section 4, and finally, Section 5 concludes the paper.

2. OVERVIEW OF CASS EXTENDED SVC

Fig. 2 shows the block diagram of a typical CASS extended scalable video encoder. Although the encoding of multiple spatial layers is fully supported by SVC, for the sake of simplicity, we present the approach by encoding two successive spatial layers, namely a base layer (BL), and an enhancement layer (EL). First, the EL video sequence I_{EL} is content-adaptively retargeted to a BL video sequence I_{BL} with a lower resolution and a different aspect ratio. The corresponding per-pixel spatial position between EL and BL is defined by a sequence of non-linear warping functions W, calculated by the video retargeting algorithm presented in [5]. I_{BL} and I_{EL} are jointly encoded by the CASS extended scalable video coder with W as the side information. I_{BL} is encoded by a H.264/AVC compatible encoder so that it can be decoded

independently; I_{EL} is coded relative to the decoded I_{BL} by using inter-layer prediction mechanisms. There are three kinds of information that can be exploited from BL: texture, motion and residual information, and thus can serve as inter-layer prediction signals which provides the EL encoder with an additional prediction source to the single-layer prediction tools [1]. To enable CASS, the three standard inter-layer prediction tools that support only linear scaling relationships between spatial layers are extended to support also non-linear relationships as presented in Section 3. To transmit the warps efficiently, a new function block *Warp Coder* is added to the SVC framework. We have proposed a novel algorithm to encode the warps by exploiting spatial and temporal dependencies existing within and between warps [7].

The CASS extension for SVC supports non-linearly retargeted lower resolution layers which provides content-adaptivity and art-directability to the content providers, and subsequently provides a better subjective quality to the end-users.

3. CONTENT ADAPTIVE INTER-LAYER PREDICTION TOOLS

Inter-layer prediction tools are the main means to exploit BL information in order to improve the coding efficiency of EL. In the standard SVC, the inter-layer prediction tools support only linear scaling and cropping between spatial layers. In order to enable CASS, each of the inter-layer prediction tools is extended as described in the following subsections.

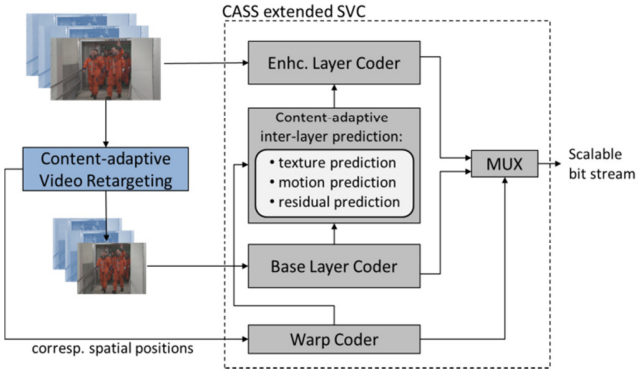

Fig. 2. Block diagram of a CASS extended SVC supporting two spatial layers.

3.1. Inter-layer motion prediction

In SVC, a coarse-to-fine projection approach is adopted [2]. For each 4x4 luma block in the current EL macroblock being encoded, the corresponding block in BL is first identified by using a single pixel. Then the motion data (prediction type, reference picture indices, and motion vectors) are inherited from this corresponding block in BL with proper scaling and offset. Last, the merging process takes place to determine the final mode and motion segmentation in the EL macroblock [2].

Our content-adaptive inter-layer motion prediction follows this approach with two modifications: the corresponding position identification between spatial layers and the scaling of the motion vectors. In the standard SVC, the corresponding position of an EL sample in BL can be directly calculated due to the linear spatial relationships, and is constant temporally. In our extension of CASS, this identifying process of this sample (the black dot in Fig. 3) is replaced by referring to the warp W_f which defines for each pixel in EL the corresponding pixel position in BL for the current frame f. This identifying process is also used in the content-adaptive inter-layer texture and residual prediction with higher precision.

The scaling factor of the motion vectors derived from BL is global in the standard SVC, whereas in our CASS extension, it depends on the local characteristics of the warp. The horizontal scaling factor is approximated by the ratio of the EL 4x4 block side length to the average between the top and the bottom edges of the deformed quad; and the vertical scaling factor is the ratio to the average of the left and the right edges of the deformed quad.

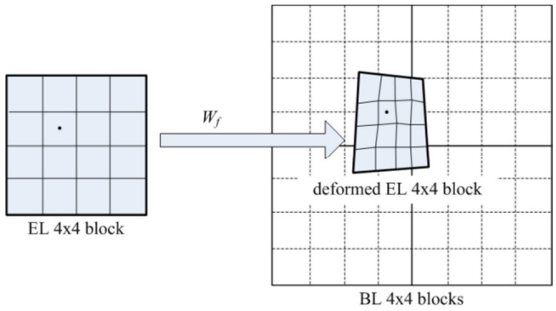

Fig. 3. Mapping of an EL 4x4 block in BL.

3.2. Inter-layer texture prediction

If all 4x4 luma blocks of the current EL macroblock are intra-coded, this macroblock is in "Intra_BL" mode and being predicted by the inter-layer texture (or intra) prediction. The process involves first decoding of the intra-coded macroblocks in BL, applying the deblocking filter, boarder extension and up-sampling [2]. In the standard SVC, a one-dimensional 4-tap FIR filter based on cubic splines is applied on luma component while on chroma components a simple bi-linear filter is applied. We have proposed a non-separable implementation of the original filter for the content-adaptive inter-layer texture prediction in [8]. A loop over each pixel in the predicted image $\hat{I}_{EL}$ is performed. Then the corresponding position in BL is identified with 1/16th sample precision based on the warp. Finally, the predicted sample is interpolated using the intensities of the neighboring 16 samples for luma component. For chroma components, similar non-separable implementation of the bi-linear filter applies. Test results show good prediction quality as well as overall coding efficiency in the intra-only coding.

3.3. Inter-layer residual prediction

Similar to the inter-layer texture prediction, the CASS extension of the inter-layer residual prediction applies a non-separable bi-linear filter to both luma and chroma residual components in order to up-sample them. Again the warp defines inter-layer positional correspondences. These up-sampled blocks are then used as a prediction for corresponding residual blocks of the EL. However, in contrast to the texture up-sampling, the up-sampling of the residual signals doesn't occur across the transform block boundaries.

Fig. 4. Comparison between enhancement layers (EL) and base layers created by content-adaptive retargeting (RT BL) and linear scaling (LN BL) for the sequences Kimono1, ParkScene, and Crew. Note that EL has HD resolution while RT BL and LN BL have SD resolution.

4. EVLUATION RESULTS

In this section, we show results on i) the impact of the content-adaptive inter-layer prediction tools on the coding efficiency and ii) the overall performance of CASS extended SVC in comparison to the SVC standard (JSVM version 9.18 is used). For coding experiments, we retargeted original test sequences from 1280x720 (HD) to 720x540 (SD) resolution using two different methods: i) content-adaptive retargeting [5] and ii) linear scaling. In Fig. 4, retargeted pictures are shown for each test sequences. Due to the change of aspect-ratio from 16:9 to 4:3, with linear scaling the complete image is uniformly distorted irrespective of the content, i.e. the complete image is more strongly scaled in horizontal than in vertical direction. However, with content-adaptive retargeting salient regions are still almost uniformly scaled, which obviously leads to an overall higher semantic quality.

4.1 The impact of inter-layer prediction tools

The effectiveness of the content-adaptive inter-layer prediction tools are evaluated in this section. A BL created by content-adaptive retargeting is encoded at a fixed QP while the EL QP varies. Simulations have been carried out using a GOP of 16 with hierarchical B frames, and an intra period of 32. Warps are encoded at a relative low bitrate point (56 kbps for the ParkScene sequence). Enabling each inter-layer prediction (ILP) tool successively, i.e. texture (I), motion (M) and residual (R), saves in average 3.1%, 17.3% and 21.7% of total bitrate compared to simulcast for the ParkScene sequence as shown in Fig. 5. In the domain around 37.5 dB, the overall gain in bit rate is about 15% compared to simulcast. This demonstrates the efficiency of our content-adaptive extension of the ILP tools.

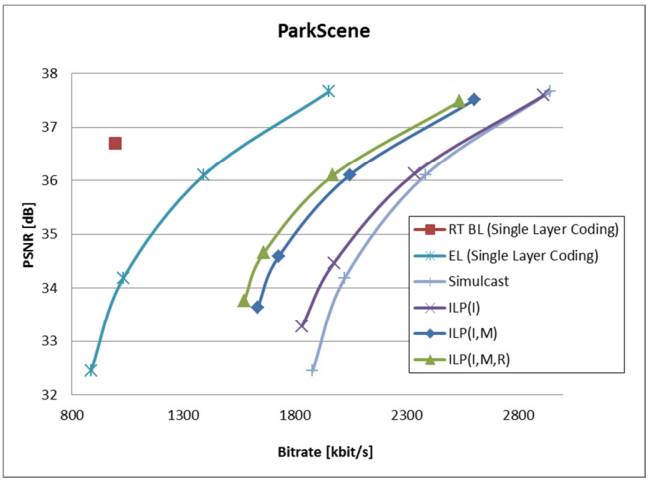

Fig. 5. RD performance improved by content-adaptive inter-layer prediction tools.

Table 1. Operational points of SVC-CASS at quality levels around 37 dB and 38 dB for base and enhancement layer, respectively. In brackets, the corresponding bit rate overhead expressed in % for SVC-CASS in comparison to SVC is shown as well as the deviation in objective quality between the reconstructed image sequences. Note that SVC-CASS and SVC encode different base layers.

Sequence Name	RT BL		SVC-CASS		
	PSNR of RT BL in dB	Bit rate of RT BL in kbit/s	PSNR of EL in dB	Aggregate bit rate w/o warp in kbit/s	Aggregate bit rate in kbit/s
Crew	37.4 (-0.05)	987.7 (0.3%)	38.2 (0.0)	3982.1 (1.7%)	4210.3 (7.5%)
Kimono1	37.8 (-0.1)	964.9 (-1.2%)	38.7 (0.0)	2123.1 (9.3%)	2198.3 (13.2%)
ParkScene	36.7 (-0.6)	992.9 (2.8%)	37.5 (0.0)	2480.4 (5.2%)	2536.6 (7.2%)

4.2 Comparative evaluation

We compared the coding efficiency of CASS extended SVC (SVC-CASS) with the standard SVC by encoding the same HD res. enhancement layer but different SD res. base layers. With SVC-CASS we encoded the base layer created by content-adaptive retargeting (RT BL), while with SVC we encoded the base layer created by linear scaling (LN BL). With both approaches, the same coding conditions are applied as reported in Section 4.1 .

In Fig. 6, we see a comparison of the operational RD curves obtained by SVC and SVC-CASS, respectively, showing the aggregate rate of the scalable bit stream and the enhancement layer PSNR. In addition, the fixed operational point of the corresponding base-layer coding (LN BL and RT BL) is shown.

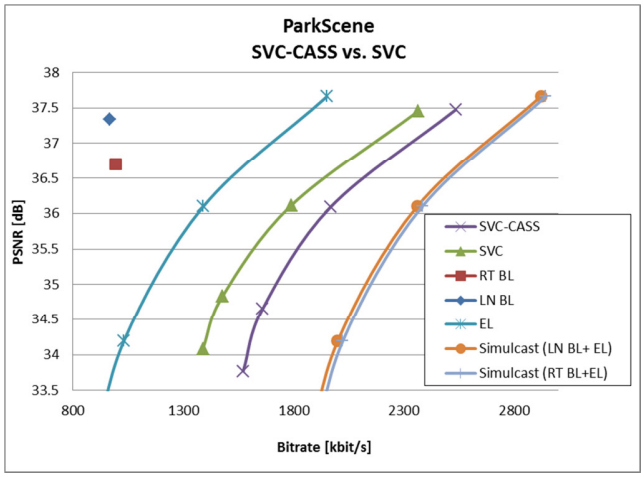

Fig. 6. Comparison of operational RD performance between SVC and SVC-CASS.

The results show that the base layers are encoded at a similar objective quality (a difference of 0.6 dB). In Fig. 4, second row, corresponding reconstructed frames of the ParkScene sequence are shown. Obviously, in RT BL salient regions are much less distorted, which leads to an overall higher semantic quality of the reconstructed image sequence. However, the encoding with SVC-CASS requires slightly more bits than SVC. In the area above 36 dB, a fixed overhead in bit rate is observed of about 180 kbps in comparison to SVC. At a quality of nearly 37.5 dB the bit rate overhead is about 7.2%. Of course, SVC-CASS has the advantage of encoding a base layer with higher semantic quality.

In Table 1, the bit rate for the base-layer and the aggregate rate of the SVC-CASS encoded scalable bit stream are shown for quality levels interesting for applications (around 37dB and 38dB, respectively). Obviously, an SD resolution base layer, which has been retargeted in a content-adaptive way, can be transmitted with SVC-CASS together with an HD resolution enhancement layer by introducing an average overhead in bit rate of 9.3% compared to SVC.

5. CONCLUSION

We presented an extension of SVC by Content-adaptive Spatial Scalability (CASS). In particular, we presented generalizations of the three inter-layer prediction tools of SVC, which draw information from an additionally encoded warping function, and demonstrated their efficiency. Furthermore, we demonstrated that with help of the CASS extension, spatial layers can be encoded, which have been adapted in a content-adaptive way to display configurations with different aspect rations. As a consequence, image sequences with higher semantic quality can be transmitted to the consumers in a scalable way. Finally, we showed that all this can be done by introducing an average overhead in bit rate of 9.3% compared to the SVC standard.

6. REFERENCES

[1] H. Schwarz, D. Marpe, T. Wiegand , “Overview of the Scalable Video Coding Extension of the H.264/AVC Standard,” IEEE Trans. Circuits Syst. Video Techn., pp. 1103-1120, Volume 17, Issue 9, Sept. 2007

[2] A Segall, G.J Sullivan, “Spatial Scalability Within the H.264/AVC Scalable Video Coding Extension,” IEEE Trans. Circuits Syst. Video Techn., pp. 1121 - 1135, Volume 17, Issue 9, Sept. 2007

[3] E. Francois, J. Viéron, S. Sun, G. J. Sullivan, “Extended spatial scalability: A generalization of spatial scalability for SVC extension of AVC/H.264,” Proc. PCS, Beijing, China, Apr. 2006.

[4] M. Rubinstein, D. Gutierrez, O. Sorkine and A. Shamir, “A Comparative Study of Image Retargeting,” ACM SIGGRAPH Asia 2010

[5] P. Krähenbühl, M. Lang, A. Hornung, and M. Gross, “A System for Retargeting of Streaming Video,” Proc. ACM SIGGRAPH Asia, Yokohama, Japan, December 16-19, 2009.

[6] T. Misu, Y. Matsuo, S. Sakaida, Y. Shishikui, E. Nakasu, “Novel Video Coding Paradigm with Reduction/Restoration Processes,” Proc. PCS, Nagoya, Japan, Dec. 7-10, 2010.

[7] A. Smolic, Y. Wang, N. Stefanoski, M. Lang, A. Hornung, M. H. Gross, “Non-linear warping and warp coding for content-adaptive prediction in advanced video coding applications,” Proc. ICIP, Hong Kong, China, 2010

[8] Y. Wang, N. Stefanoski, X. Fang, A. Smolic, “Content-Adaptive Spatial Scalability for Scalable Video Coding,” Proc. PCS, Nagoya, Japan, Dec. 7-10, 2010.