CONTENT-ADAPTIVE SPATIAL SCALABILITY FOR SCALABLE VIDEO CODING

Yongzhe Wang[1], *Nikolce Stefanoski*[2], *Xiangzhong Fang*[1], *and Aljoscha Smolic*[2]

1 Shanghai Jiao Tong University 2 Disney Research, Zurich

ABSTRACT

This paper presents an enhancement of the SVC extension of the H.264/AVC standard by content-adaptive spatial scalability (CASS). CASS introduces a novel functionality which is important for high quality content distribution. The video streams (spatial layers), which are used as input to the encoder, are created by content-adaptive and art-directable retargeting of existing high resolution video. Video is retargeted to resolutions and aspect ratios which are mainly dictated by target display devices. Thereby no content is cut off, but visually important content is preserved at the expense of a non-linear distortion of visually unimportant areas. The non-linear dependencies between such video streams are efficiently exploited by CASS for scalable coding. This is achieved by integrating warping-based non-linear texture prediction and warp coding into the SVC framework. The results indicate high prediction accuracy of non-linear predictors and high compression efficiency with limited increase in bit rate and complexity compared to the standard SVC.

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

1. INTRODUCTION

In Scalable video coding (SVC), different instantiations of the same video sequence in terms of temporal and spatial resolution and quality can be reconstructed from the same scalable bitstream. Only the corresponding portions of the bitstream need to be accessed and decoded [1]. The need for scalability is motivated by the resolution diversity of current display devices as well as the diverse, limited, and time-varying channel capacity [2].

Spatial scalability was so far defined via linear scaling operations between different resolutions. Changes of aspect ratios are possible via different scaling in both dimensions (Fig. 1a,b); however, algorithms are mainly optimized for dyadic scaling. Pan-scan and cropping operations between different aspect ratios are also supported by the SVC standard.

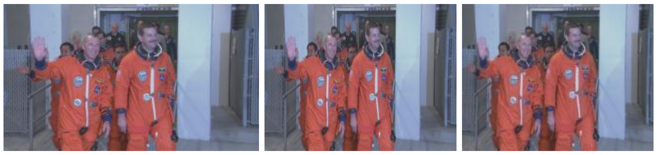

Fig. 1: Retargeting of a frame having a) 16:9 (HD) aspect ratio to a frame having 4:3 (PAL) aspect ratio using b) linear scaling and c) video retargeting (non-linear scaling).

Video retargeting is another technology that recently received a lot of attention. Given video is adapted to a different target resolution and aspect ratio by involving non-linear scaling operations [3]. This is done in a way that visually important content is preserved while distortions are hidden in visually less important areas (Fig. 1a,c). Visual importance can be defined as combination of automatic saliency computation [4] and interactive user input if possible and necessary in a given application scenario [3]. Thus, video retargeting enables content-adaptive and art-directable scaling and reformatting of video to different target resolutions and aspect ratios.

A concept for combination of both, SVC and content-adaptive video retargeting has been recently presented [5]. In this paper, we present content-adaptive spatial scalability (CASS) as a new approach for scalable video coding. CASS is realized a) by integrating warping-based non-linear texture prediction into SVC, which is used to exploit non-linear inter-layer dependencies and b) by encoding a sequence of warping functions as side-information.

This paper is organized as follows. Section 2 presents the concept of the CASS that introduces a novel type of scalability into SVC. Section 3 describes the proposed filter extension for non-linear inter-layer texture prediction. Then section 4 provides experimental results and evaluation against the conventional spatial scalability based on linear scaling. Finally section 5 concludes the paper and gives an outlook to future research.

2. CONTENT-ADAPTIVE SPATIAL SCALABILITY

Consider two successive spatial layers, a base layer (BL) with horizontal and vertical dimensions w_{BL} and h_{BL} (e.g. 720x540) and an enhancement layer (EL) with dimensions w_{EL} and h_{EL} (e.g. 1280x720). As illustrated in Figure 2, BL image sequence I_{BL} is generated by the video retargeting as used in [3] from EL image sequence I_{EL}, together with a sequence of non-linear warping functions W which describes the shape deformation of each pixel of the I_{EL} when mapped to the I_{BL}. The aspect ratios w_{EL}/h_{EL} and w_{BL}/h_{BL} of image sequences I_{EL} and I_{BL}, respectively, are not necessarily the same. The conventional methods either distort the content in an unacceptable way by linear scaling or keep the scaling ratios in both dimensions by cropping parts of EL. In comparison, using video retargeting for BL generation is a superior way since it preserves high level semantic information, like the aspect ratio of faces at the expense of distortions in the background.

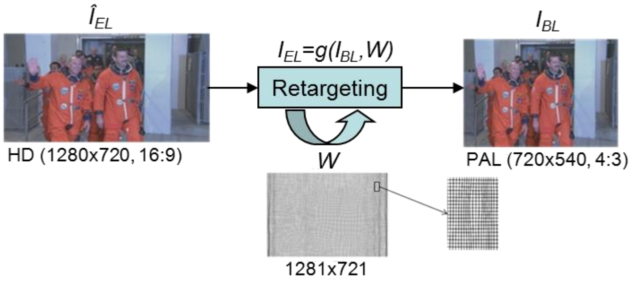

Fig. 2: BL generation using video retargeting [3]

Figure 3 illustrates an extended SVC encoder that takes 3 inputs, the high resolution source video I_{EL} as EL, the retargeted lower resolution video I_{BL} as BL and W_p as side-information, where W_p is a point-based warp (res. 1280x720), which describes a mapping between pixel positions in I_{EL} and sub-pixel positions in I_{BL}. Two new blocks have been integrated into the framework of the current SVC standard to enable the CASS, i.e. *warp coding* and *content-adaptive spatial (CAS) prediction*. For warp coding, a novel algorithm has been proposed to encode the warps efficiently by exploiting spatial and temporal dependencies existing within and between warps [5]. The non-linear inter-layer texture prediction is done in the novel block called CAS prediction. It up-samples the retargeted BL I_{BL} to higher resolution EL $\hat{I}_{EL}$ so that only the difference needs to be transmitted. Base layer bits, encoded warps and enhancement layer bits are multiplexed into a scalable bitstream, which can be partially accessed, transmitted, decoded, etc.

This concept is content-adaptive since it favors visually important image regions over less important areas. It further introduces art-directability to video coding since it allows the content provider to control the appearance of the base layer video via interactive retargeting [3].

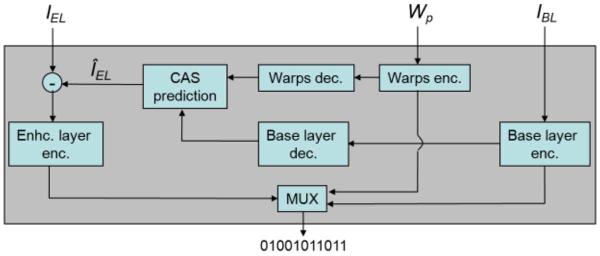

Fig. 3: Block diagram of an extended SVC encoder for content-adaptive spatial scalability (CASS)

3. NON-LINEAR INTER-LAYER TEXTURE PREDICTION

The process of non-linear prediction of the high resolution source video I_{EL} based on the retargeted video I_{BL} and the point warp W_p is performed in the block CAS prediction (Fig. 3). Both forward mapping and backward mapping approaches have been tested and it is reported that the latter outperforms the former in terms of prediction accuracy at a certain warp bitrate [5]. Actually, the backward mapping approach for inter-layer texture prediction corresponds to the coarse-to-fine projection design of the SVC standard. The mapping or projection process consists of first projecting the sample grid of the finer level to the coarser level of the pyramid and then using this projection to propagate data from the coarser level to the finer level [2]. In the case of inter-layer texture prediction in the SVC standard, the linear up-sampling process is separable, i.e. I_{BL} can be up-sampled first horizontally then vertically by applying one-dimensional interpolation filters. Horizontally, for each pixel in $\hat{I}_{EL}'$ (res. 1280x540) the corresponding sample position in I_{BL} is first calculated to 1/16th sample position increments. Then a one-dimensional 16-phase 4-tap cubic spline interpolation filter is applied to the luma component. The same is performed vertically from $\hat{I}_{EL}'$ to get $\hat{I}_{EL}$. For the chroma component, a simple bilinear filter is separated into a one-dimensional 16-phase 2-tap linear filter both horizontally and vertically. The filter design shows good performance with minimal complexity [6].

In order to keep the features of the interpolation filters in the enhanced SVC design while enabling CAS prediction, we apply a backward mapping approach and derive the up-sampling process in a non-separable way, i.e. we loop over all pixels directly in $\hat{I}_{EL}$ for both luma and chroma components. For the luma component, consider a pixel in $\hat{I}_{EL}$ at position *(a, b)*. The corresponding sample position in I_{BL} is read directly from the decoded position warp $(x, y)=W_p(a, b)$ and cast to 1/16th sample precision. As illustrated in Figure 5, $\hat{I}_{EL}(a, b)=I$ is then calculated using the neighboring 16 pixels of position (x, y) for the luma component. First, four vertical intermediate sample values E, F, G and H are interpolated from the four neighboring samples in the vertical direction. For example, F is interpolated from A, B, C and D using the one-dimensional 4-tap cubic spline filter the phase selection of which is determined by the vertical position of F (i.e. y). Then I is interpolated by E, F, G and H. The phase is determined by x.

For chroma components, considering the chroma samples are not co-located with luma sample positions, proper calculation of positions as well as phase shift of the interpolation filter is needed according to the format of the source video. The interpolation process for chroma components is similar to that for the luma component except that the simple bilinear filter is used instead to reduce the complexity.

One important issue is that the CAS prediction performed in the encoding and decoding process needs to match the downscaling operations used for video retargeting. Otherwise, it would result in reduced compression efficiency or visual artifacts. Downscaling in video retargeting involves alias free forward mapping from the source to the target pixel grid, which is known as EWA splatting [7]. The effectiveness of combining EWA splatting and the extended up-sampling filter is analyzed in section 4.

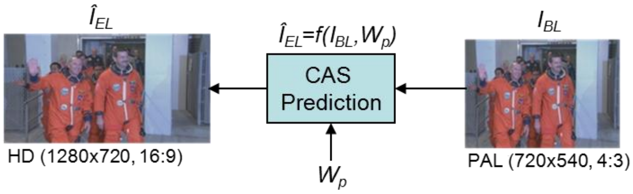

Fig. 4: Inverse retargeting using backward mapping

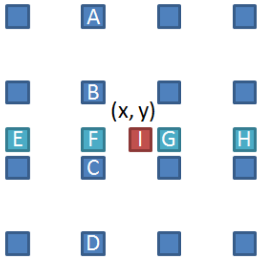

Fig. 5: Interpolation of a sample value at the sub-pel position *(x, y)* for luma component in BL

4. RESULTS

In this section we present experimental results evaluating two specific aspects: prediction quality and the overall encoding efficiency using inter-layer texture prediction for CASS.

4.1. Prediction quality

This set of experiments is designed to evaluate our non-linear inter-layer texture prediction quality for CASS by comparing to the conventional linear scaling approach. For that we retargeted five HD 1280x720 test sequences to 720x540 as BL (RT BL) and performed inverse retargeting using the up-sampling filters described in section 3. The warp sequences used in the inverse retargeting are either uncoded or coded at maxSL 13, the parameter that controls the trade-off between the bitrate and the quality of the coded warp. As explained in [5], with the increase of the warp bitrate, the prediction quality is also increasing due to the higher quality of the warp but saturated at relative low bitrates typically less than 300 kbit/s. The bitrates of the warp sequences encoded at maxSL 13 for all five test sequences are between 145 kbit/s and 228 kbit/s, which are relatively low compared to the bitrate that is necessary to encode the corresponding video. For the linear (LN) case, the BL is generated by the down-sampling filter as defined in JSVM [8]. It is then up-sampled again by the 4-tap cubic spline interpolation filter as defined in the SVC standard.

The results are shown in Figure 6. Apparently non-linear prediction using our proposed methods outperforms the conventional linear scaling methods in most cases, for both coded and uncoded warps. Even if warp is coded at a relative low bitrate, the prediction quality loss is typically less than 1dB compared to prediction from the uncoded warp. This proves the effectiveness of our overall solution for the non-linear inter-layer texture prediction, i.e. EWA splatting for non-linear down-sampling, backward mapping with extended non-separable two-dimensional interpolation for non-linear up-sampling, and the warp coding. Note that LN BL is distorted in an unacceptable way while RT BL takes the content-adaptivity and the art-directability into account (Fig. 1).

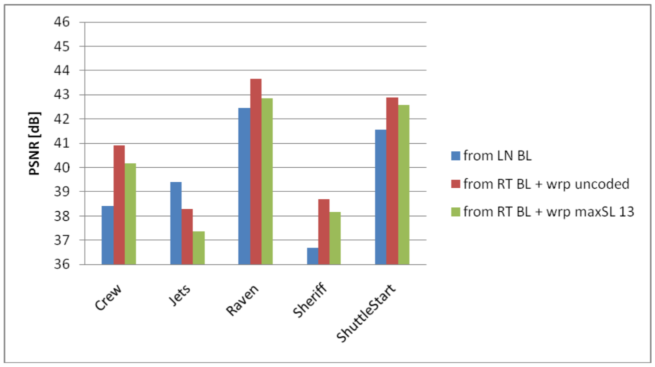

Fig. 6: Texture prediction quality using linear scaling and retargeting

4.2. Coding efficiency

The effectiveness of our non-linear inter-layer texture prediction techniques for CASS has been evaluated in this set of experiments. This was done for both intra only coding and inter coding, however we focused here on texture up-sampling only first for both coding structures. Four different solutions are compared:

1. Scalable coding using CAS prediction; BL is retargeted from EL (label *RT ILP(I)*).

2. Scalable coding using conventional spatial scalability; BL is linear down-scaled (label *LN (ILP(I))*).

3. Single layer coding: EL is encoded using SVC single- layer coding mode (label *EL single layer*).

4. Simulcast coding: BL and EL are encoded using SVC single-layer coding mode (labels *LN* resp. *RT Simulcast*).

Results are depicted for two sequences, Crew (Fig. 7, 8) and ShuttleStart (Fig. 9). In Figure 7 and Figure 9, simulations have been carried out using intra only coding while in Figure 8 GOP 16 is used with hierarchical B frames, BL is encoded at a fixed QP while EL is encoded at different QPs given this BL, both at 60 frames/second.

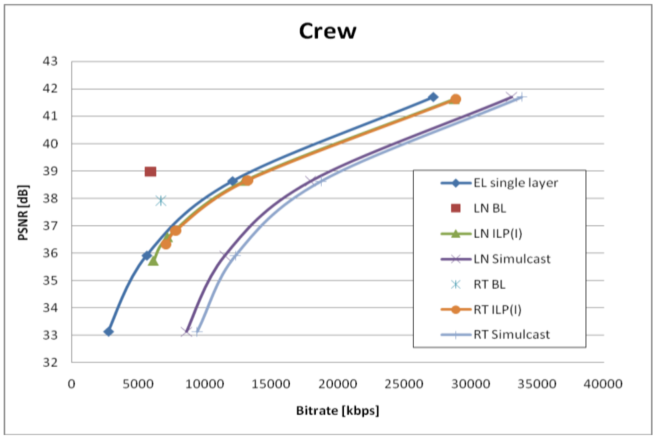

Fig. 7: RD curves for Crew sequence with intra only coding

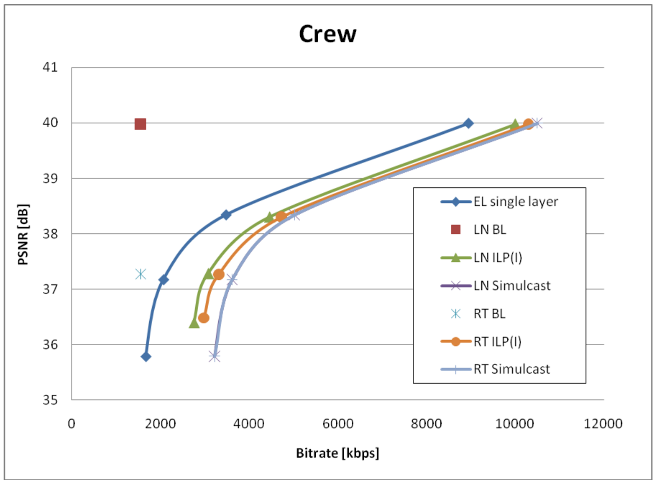

Fig. 8: RD curves for Crew sequence with GOP 16 coding

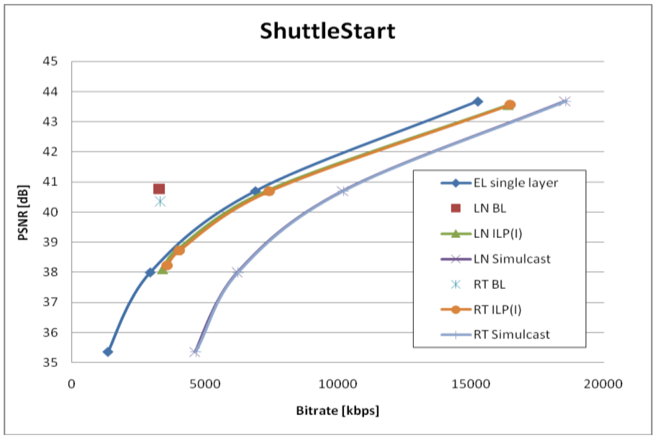

Fig. 9: RD curves for ShuttleStart sequence with intra only coding

The results show that RT BL is more expensive to encode than LN BL. However, it also preserves visually important content better since I_{BL} is created in a content-adaptive and art-directable way by video retargeting. The subjective quality is therefore better. Note that LN BL is in principle unusable if there is a strong deviation between the aspect ratios of I_{BL} and I_{EL}. Then, the enhanced detail in important areas in RT BL also helps to predict these areas better for the RT EL.

Table 1 compares coding efficiency between CASS and conventional spatial scalability for intra only coding using the improved BD-PSNR model proposed by Bjontegaard [9]. Bitrates are compared at high, low, and average range. The overall coding efficiency of CASS is comparable to that of the conventional linear scalability despite the fact that RT BL is more expensive to encode and warps costs extra bitrate. This is due to the better prediction. Subjective quality of RT EL is even better compared to LN EL, because visually important areas are better reconstructed.

Inevitably, the complexity of the decoding process increases using CASS, but only to a limited extent. The average decoding processing time per frame using CASS is 1.46 s in comparison to 0.89 s using the standard SVC reference decoder on a normal PC with a T9300 CPU.

	% Bit high	% Bit low	% Bit
Crew	1.190	3.799	1.945
Jets	0.442	3.804	1.518
Raven	0.346	3.513	0.775
Sheriff	-1.411	-0.188	-1.327
ShuttleStart	0.999	2.642	1.538
Average	0.313	2.714	0.890

Table 1: Bitrate change of CASS

5. CONCLUSIONS AND FUTURE WORK

We introduced a new content-adaptive spatial scalability (CASS) concept to SVC, which integrates content-adaptive and art-directable video retargeting. Different spatial layers are generated using non-linear scaling operations to support different aspect ratios without unacceptable image distortions. Inter-layer dependencies are exploited by integrating warping-based non-linear texture prediction into the SVC framework. The non-linear inter-layer texture prediction is achieved by extending the up-sampling process in a non-separable way with the help of a sequence of warping functions which is encoded as side-information. A comparison to the linear scaling approach has proven a higher prediction accuracy as well as a comparable overall coding efficiency while providing extended functionality at the cost of limited complexity increase. Subjective quality of RT BL and EL is superior due to content-adaptive coding better preserving visually important areas.

Our future research will include generalizing non-linear inter-layer prediction for motion and residual data to complete the full non-linear SVC framework.

6. REFERENCES

[1] H. Schwarz, D. Marpe, and T. Wiegand, “Overview of the Scalable Video Coding Extension of the H.264/AVC Standard”, IEEE Trans. on CSVT, Vol. 17, No. 9, September 2007.

[2] A. Segall, and G. J. Sullivan, “Spatial scalability”, IEEE Trans. on CSVT, Vol. 17, No. 9, September 2007.

[3] 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.

[4] C. Guo, Q. Ma, and L. Zhang, “Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform”, Proc. CVPR 2008, Anchorage, AL, USA, June 24-28, 2008.

[5] A. Smolic, Y. Wang, N. Stefanoski, M. Lang, A. Hornung and M. Gross, “Non-linear Warping and Warp Coding for Content-Adaptive Prediction in Advanced Video Coding Applications”, Proc. ICIP 2010, Hong Kong, China, September 26-29, 2010.

[6] S. Sun, “Upsampling Filter Design with Cubic Splines”, Joint Video Team, Doc. JVT-S016, Mar.-Apr. 2006.

[7] M. Zwicker, H. Pfister, J. Van Baar, and M. Gross, “EWA Splatting”, IEEE Trans. on Visualization and Computer Graphics, Vol. 8, No. 3, July-September 2002.

[8] J. Reichel, H. Schwarz, and M. Wien, “Joint Scalable Video Model 11 (JSVM 11)”, Joint Video Team, Doc. JVT-X202, Jul. 2007.

[9] G. Bjontegaard, “Improvements of the BD-PSNR model”, ITU-T SG 16/Q 6 (VCEG), Doc. VCEG-AI11, Jul., 2008.