NON-LINEAR WARPING AND WARP CODING FOR CONTENT-ADAPTIVE PREDICTION IN ADVANCED VIDEO CODING APPLICATIONS

Aljoscha Smolic, Yongzhe Wang, Nikolce Stefanoski, Manuel Lang, Alexander Hornung, and Markus Gross

Disney Research, Zurich
Clausiusstrasse 49, 8092 Zurich, Switzerland
smolic@disneyresearch.com

ABSTRACT

This paper presents a new concept for scalable video coding, which is content adaptive and art-directable. Video retargeting is applied to scale video between different resolutions and aspect ratios without introducing inacceptable distortions or cutting off content. The non-linear warping operations are integrated into a spatial scalability framework, which includes two new building blocks, i.e. non-linear warping prediction and warp coding. Efficient algorithms for both processes are presented, tested and optimized. The presented results indicate that our non-linear scaling and warp coding algorithms provide efficient performance compared to standard linear scaling methods. Further, our advanced scaling algorithms, i.e. EWA splatting in combination with backward mapping, may be very useful for linear scaling as well.

Index Terms— Scalable video coding, video retargeting, non-linear warping, prediction

1. INTRODUCTION

Scalable video coding (SVC) is an important technology for a variety of advanced functionalities. 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. The recent SVC standard provides these advanced functionalities at a limited additional cost in coding efficiency [1]. 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; however, algorithms are mainly optimized for dyadic scaling. Pan-scan and cropping operations between different aspect ratios are also supported by the SVC standard.

On the other hand, video retargeting is a technology that recently received a lot of attention. Video retargeting adapts given video to a target resolution and aspect ratio by involving non-linear scaling operations [2]. Visually important content is preserved while distortions are hidden in visually less important areas. Visual importance can be defined as combination of automatic saliency computation [3] and interactive user input if possible and necessary in a given application scenario [2]. Thus, video retargeting enables content-adaptive and art-directable downscaling and reformatting of video content to different target resolutions.

In this paper we present the concept for combination of both, SVC and content-adaptive video retargeting. This introduces the concept of art-directability to video coding. We show how non-linear scaling via image warping can be integrated into a spatial scalability framework. The two new building blocks that we focus on are coding of non-linear warping functions and non-linear warping for prediction. We also show that some algorithms used in retargeting, such as EWA splatting [4] and backward mapping based on spline interpolation may be beneficial for other aspects in scalable video coding as well.

This paper is organized as follows. Section 2 introduces the system concept for combination of SVC and video retargeting. Section 3 presents a new algorithm for coding of non-linear warping functions. Then section 4 outlines approaches for non-linear prediction. Section 5 presents results and evaluation, and finally, Section 6 concludes the paper and gives an outlook to future research.

2. SYSTEM CONCEPT

Figure 1 illustrates the principle of video retargeting as used in [2]. A source video with high resolution and 16:9 aspect ratio is reformatted to a lower resolution video with different aspect ratio 4:3. Inevitably information is lost in such an operation. Linear scaling is not an option since it would distort the content in an unacceptable way. Conventional methods use pan-scan and cropping to extract a corresponding window out of the high resolution video by cutting off the rest. Video retargeting instead computes a non-linear warp W from source to target resolution and aspect ratio that preserves important image regions (in this case the persons) and hides the inevitable distortions in less important areas (here background). No content is cut off. Downscaling involves alias free forward mapping from the source to the target pixel grid known as EWA splatting [4].

Figure 2 illustrates the concept of integrating video retargeting into a spatial scalability framework. The extended SVC encoder takes 3 inputs, the high resolution source video I_s, the retargeted video with lower resolution I_t

and different aspect ratio, and a sequence of non-linear warping functions W that define the non-linear relation between the videos. I_t is encoded as base layer using e.g. H.264/AVC. Additionally, W is encoded using a novel algorithm as presented in section 3. Both, I_t and W are decoded for prediction of I_s. This non-linear prediction is done in the novel block called inverse retargeting. Section 4 elaborates different ways to achieve this inverse retargeting operation to upscale the base layer video to the higher resolution. Then, only the difference between the prediction and the high resolution source has to be encoded as an enhancement layer. 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-directablity to video coding since it allows the content provider to control the appearance of the base layer video via interactive retargeting [2].

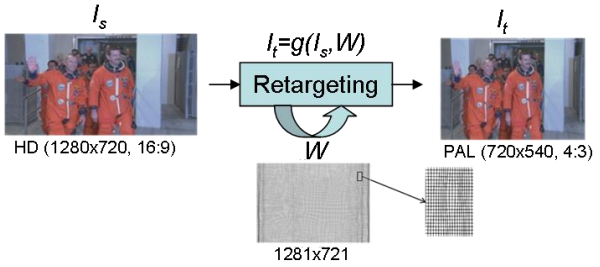

Fig. 1: Principle of video retargeting as used in [2].

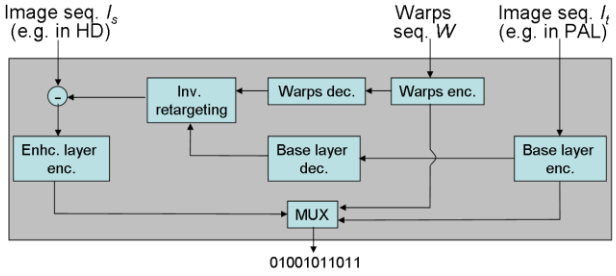

Fig. 2: Concept for spatial scalability based on non-linear video retargeting.

3. WARP CODING

Efficient coding and transmission of a sequence of non-linear warping functions is a first new building block in the concept illustrated in Figure 2. A warp sequence can be regarded as a sequence of regular quad grids with 2D vertex positions. During the last decade there has been intensive research in a related domain. In the domain of compression of 3D dynamic meshes, i.e. sequences of irregular triangle meshes with 3D vertex positions, efficient compression approached have been developed [5].

The warps encoder presented in this section encodes a sequence of warps W^f. Assume that warp W^f is describing the shape deformation of rectangular pixels of an HD image, which are mapped to a PAL image. Then, 2D positions $W^f[i,j]$, $W^f[i-1,j]$, $W^f[i-1,j-1]$, $W^f[i,j-1]$ describe the shape deformation of a rectangular quad (i,j), $(i-1,j)$, $(i-1,j-1)$, $(i,j-1)$ which corresponds to one pixel of an HD image. In particular, the number of quads of a warp W^f is equal to the number of pixels of an HD image, while the aspect ratio of the warp W^f corresponds to that of a PAL image.

Strong spatial and temporal dependencies exist within and between warps, respectively, which manifest in similarities of neighboring quads and their smooth changes in temporal direction. We exploit this redundancy for efficient compression using a predictive coding approach. A block diagram of the warps encoder is shown in Figure 3.

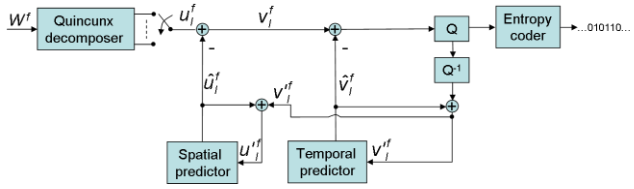

Fig. 3: Block diagram of the warps encoder.

The warps encoder consists of (i) a quincunx decomposer, which decomposes a warp W^f into spatial layers, (ii) a spatial and temporal predictor, which are part of two successive closed-loop DPCMs [6] and which have the aim to reduce spatial and temporal redundancy, and (iii) an entropy coder. Due to its high coding efficiency, we employ CABAC [7] for entropy coding.

Warps W^f are encoded in the same hierarchical frame coding order like video frames. Each warp W^f is first send to the quincunx decomposer, which decomposes the parameter domain of a warp W^f (the parameter domain corresponds to a regular quad grid), based on a quincunx resolution pyramid.

The lowest resolution quad grid of the resolution pyramid and the difference sets between successive resolutions specify a decomposition of the parameter domain of a warp into disjoint subsets D_l. Hence, 2D positions $W^f[i,j]$ with (i,j) in D_l are grouped into spatial layers, which are represented by vectors u^f_l of dimension $2|D_l|$. These vectors are successively encoded in a predictive way from low to high resolution l, using already encoded and again decoded vectors u'^f_k of lower spatial layers of the same frame f. Spatial predictor computes for each position $W^f[i,j]$ of spatial layer l a barycenter using already decoded neighboring positions of lower spatial layers. Barycenters specify a spatial prediction vector $\hat{u}^f_l$. Spatial prediction errors $v^f_l=u^f_l-\hat{u}^f_l$ are further temporally predicted based on decoded spatial prediction errors v'^r_l of reference frames r, which are selected according to the hierarchical frame coding order. Temporal prediction errors are then uniformly quantized and entropy encoded.

An alternative to the presented warp coding approach is to separate the 2D positions $W^f[i,j]=(W^f_x[i,j], W^f_y[i,j])$ by their x and y coordinates. These coordinates are quantized

and stored in two separate gray-scale images $D^f_x[i,j]$, $D^f_y[i,j]$ (Figure 4), which are then encoded with a video coder.

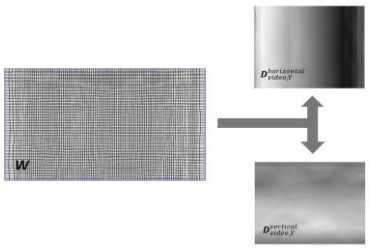

Fig. 4: Separation of warp into *x* and *y* components for encoding as video.

4. NON-LINEAR PREDICTION

Non-linear prediction of the high resolution source video I_s from the retargeted video I_t is the second new building block in the concept illustrated in Figure 2. This process as shown in Figure 5 may be called inverse retargeting. A point-based warp W_p (res. 1280x720) is first derived from the corresponding quad-based warp W (res. 1281x721), which describes pixel positions instead of pixel shapes, respectively. Given the point-based warp in the resolution of the high resolution source video, a corresponding position (x, y) in I_t can be computed for each pixel to be predicted in I_s. In general the position (x, y) will not be on the pixel raster in I_t. We therefore compute the predicted value $\hat{I}_s$ by spline interpolation [8].

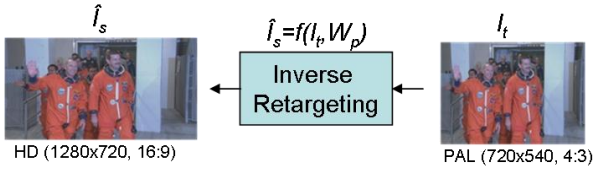

Fig. 5: Inverse retargeting using backward mapping.

Backward mapping is in general an appropriate approach for image interpolation. In our case this requires coding and transmission of the original warp function W_s in high resolution of I_s. An alternative is to compute an *inverse warp* W_t that would point from each pixel in I_t to a position in I_s. The advantage of such an approach is that the inverse warp function to be transmitted is of the lower resolution of I_t, which leads to a lower bitrate overhead. On the other hand, predicted pixel values $\hat{I}_s$ have to be computed by forward mapping in this case, which may reduce prediction accuracy compared to backward mapping. In consequence there is a trade-off to be considered between bitrate costs for warp coding and prediction accuracy. In the next section we present an experimental evaluation of this trade-off.

An inverse warp can be computed from a given warp W by deriving a continuous parameterization of the warp function. This can be done with help of generalized barycentric coordinates [9]. Forward mapping is then performed by deforming the shape of each pixel of I_T according to the inverse warp. This gives a continuous image having the aspect ratio of the source image (e.g. 16:9). A predicted image in source image resolution (e.g. HD resolution) can then be computed by sampling the continuous image with a Dirac-comb of appropriate resolution.

5. RESULTS

In this section we present experimental results evaluating 2 specific aspects. The first set of experiments investigates different options for warp coding and non-linear prediction as presented in the previous sections. The second set of experiments compares non-linear prediction based on retargeting with linear prediction as applied in the SVC standard.

5.1. Warp coding and non-linear prediction

In section 3 we presented 2 methods for warp encoding, quad-grid-based (QGB) and image-based (IMB). Section 4 described two ways to perform non-linear prediction via inverse retargeting, backward mapping (BWM) using a high resolution warp and forward mapping (FWM) using a low resolution warp. This gives 4 options for prediction, which we compared.

Figure 6 illustrates the results for test sequence Crew with resolutions of 1280x720 for I_s and 720x540 for I_t (computed using retargeting [2]). We encoded the original warp sequence (W_s of resolution I_s) and the inverted warp sequence (W_t of resolution I_t) using both presented methods at different bitrates. Then $\hat{I}_s$ was computed via backward mapping using W_s and forward mapping using W_t. Obviously, backward mapping outperforms forward mapping significantly in terms of prediction accuracy at a certain bitrate. Further, quad-grid-based warp encoding outperforms the image-based approach in the backward mapping case.

Apparently the quality saturates at relative low bitrates of e.g. 300 kbit/s for quad-grid-based warp coding with backward mapping. Such a bitrate is relatively low compared to the bitrate that is necessary to encode the corresponding video information.

5.2. Comparison to linear scaling

A second set of experiments was designed to compare the achievable prediction accuracy using non-linear warping with linear scaling as used in the SVC standard. For that we retargeted 5 HD test sequences to 720x540 and performed inverse retargeting based on backward mapping using the uncoded warp sequences. We then downscaled the same test sequences to the same base resolution using linear scaling as specified in the standard reference software "Joint Scalable Video Model" (JSVM) and upscaled them again using filters specified in the JSVM [10]. We then compared the

prediction accuracy in terms of PSNR between the predicted and the original sequence. In order to get a fair comparison we also performed linear downscaling but using the same scaling algorithms as in the retargeting case, i.e. EWA splatting for downsampling and backward mapping using spline interpolation for upsampling.

The results are shown in Figure 7. Apparently the non-linear retargeting approach outperforms linear scaling methods as specified in the JSVM in most cases. However, applying linear scaling in combination with advanced algorithms gives best results. This means that EWA splatting for downsampling and backward mapping using spline interpolation for upsampling is an attractive solution for spatial scalability in video coding in general. In any case the retargeting framework provides very good prediction performance compared to spatial scalability as available in standard SVC. Note that the base layer video is different in these experiments, i.e. in the reference case using linear scaling it is distorted in an unacceptable way.

Fig. 6: Prediction accuracy in terms of PSNR between I_s and $\hat{I}_s$ using different combinations of warp coding and inverse retargeting.

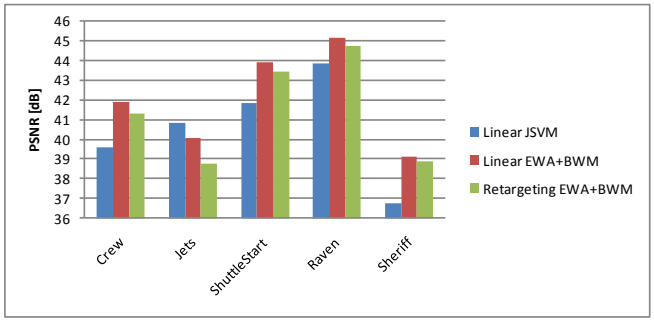

Fig. 7: Prediction accuracy using linear scaling and retargeting.

6. CONCLUSIONS AND FUTURE WORK

We presented a new approach to spatial scalability in video coding, which integrates content-adaptive and art-directable video retargeting. Non-linear scaling is performed via warping functions to support different aspect ratios without unacceptable image distortions. New building blocks in this concept are warp coding and non-linear prediction via inverse retargeting. We have shown that the combination of our new quad-grid-based warp coding algorithms with backward mapping based on spline interpolations leads to best prediction results. Comparison to linear scaling as defined in the SVC standard has proven very good prediction performance of our approach, while providing extended functionality. Moreover, our advanced scaling algorithms show superior performance for the linear case as well, i.e. there seems to be potential for broader usage in SVC.

Our future research will include integration of our concept for advanced spatial scalability into a full SVC framework such as the JSVM [10]. This will include generalization for non-linear prediction of residuals and motion data as used for inter-layer inter prediction in the SVC standard.

7. 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] 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.

[3] 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.

[4] 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.

[5] N. Stefanoski and J. Ostermann, "Spatially and temporally scalable compression of animated 3D meshes with MPEG-4/FAMC", Proc. ICIP 2008, San Diego, CA, USA, Oct. 12-15, 2008.

[6] N. S. Jayant, P. Noll, "Digital Coding of Waveforms", Prentice Hall, 1984.

[7] D. Marpe, H. Schwarz, and T. Wiegand, "Context-Based Adaptive Binary Arithmetic Coding in the H.264 / AVC Video Compression Standard", IEEE Trans. on CSVT, Vol. 13, No. 7, pp. 620-636, July 2003.

[8] C. de Boor, "A practical Guide to Splines", Springer-Verlag, 1978.

[9] M. Meyer, H. Lee, A. Barr and Mathieu Desbrun, "Generalized Barycentric Coordinates on Irregular Polygons", Journal of Graphics Tools, Vol. 7, pp. 13-22, 2002.

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