IMAGE QUALITY VS RATE OPTIMIZED CODING OF WARPS FOR VIEW SYNTHESIS IN 3D VIDEO APPLICATIONS

Nikolce Stefanoski[1], *Manuel Lang*[1,2], *and Aljoscha Smolic*[1]

[1]Disney Research Zürich [2]ETH Zürich

ABSTRACT

In this paper, a method for efficient warp coding is presented. Coded warps are transmitted together with stereo or multi-view video to enable additional functionalities at the receiver-side through image-domain-warping, e.g. depth adaption for stereo displays or support of multi-view autostereoscopic displays. Warp coding is performed by partitioning warps using a resolution pyramid and predictively exploiting intra and inter partition dependencies. View synthesis is employed within the coding loop to control the overall coding process, i.e. to evaluate the contribution of coded partitions to the synthesis quality. It is shown that coded warps represent a practically negligible portion of about 3.6% of the overall (video+warp) bit rate. Furthermore, it is shown that a transmission of warps leads to a reduction of synthesis time up to a factor of 8 in comparison to a fully automatic receiver-side view synthesis which uses only decoded video as input.

Index Terms— warp coding, image-domain-warping, view synthesis, depth adaptation, 3D displays

1. INTRODUCTION

Content production workflows, distribution channels, data compression techniques, and display technology for stereoscopic 3D video is constantly improving and adapted to enable a high quality experience, e.g. in the 3D cinema or on stereo displays in the home. However, the necessity to wear glasses is often regarded as a main obstacle of today's mainstream 3D display systems. Multi-view autostereoscopic displays (MAD) allow glasses free stereo viewing and support motion parallax viewing in a limited range. In contrast to stereo displays, MADs require not two but multiple different views as input. Research communities and standardization bodies investigate novel 3D video formats [1], which on the one hand are well compressible and on the other hand enable an efficient generation of novel views as required e.g. by MADs.

Recently, the Moving Pictures Experts Group (MPEG) issued a Call for Proposals (CfP) [2] with the goal to identify i) a 3D video format, ii) a corresponding efficient compression technology, and iii) a view synthesis technology which enables an efficient synthesis of new views based on the proposed 3D video format. The authors of this paper proposed a view synthesis method based on image-domain-warping (IDW) [3] in a joint proposal [4], which is compatible with existing stereo or multi-view video formats, i.e. it is able to synthesize new views using only stereo or multi-view video as input. In particular, to do the view synthesis, the IDW approach doesn't depend on auxiliary data like depth maps which are usually computed in a semi-automatic way at the encoder side. Instead, warps are computed reliably and completely automatically from the transmitted video data at the decoder and used for IDW-based view synthesis (Fig. 1a). A large subjective study coordinated by MPEG proved that multi-view video coding in combination with IDW leads to high quality synthesis results. Consequently, the proposal [4] was considered as one of the four winning proposals [6].

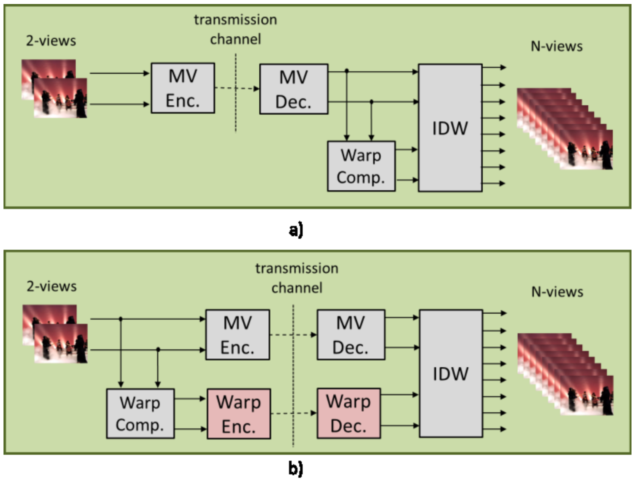


Fig. 1. Illustration of a system for stereo video transmission and subsequent IDW-based view synthesis with a) decoder-side warp computation and b) encoder-side warp computation.

In this paper, we i) present an image quality vs. rate optimized warp coding method, ii) analyze the additional bit rate which is needed for transmission of warps if they are computed at the encoder side (Fig. 1b), and iii) evaluate the reduction of the run-time of the view synthesis resulting from a shift of the warp computation from the decoder to the encoder side.

2. IMAGE-DOMAIN-WARPING

The IDW-component presented in Fig. 1 uses two images I_{left} and I_{right} and two warps as input. Depending on the position of the view to be synthesized two dedicated warps W_{left} and W_{right} are derived from the input warps [4]. A new image I_{synth} is then synthesized according to

$$I_{\text{synth}} = I_{\text{mask}} \circ \Psi(I_{\text{left}}, W_{\text{left}}) + (\mathbf{1} - I_{\text{mask}}) \circ \Psi(I_{\text{right}}, W_{\text{right}})$$

where I_{mask} is a binary mask or an alpha mask with values in [0,1], operator ∘ represents a component-wise multiplication, and operator

$$\Psi(I, W)[i, j] := I(W^{-1}[i, j]) \qquad (1)$$

warps image I with an 2D image warp W [7]. A similar synthesis approach is used with multi-view input. Obviously, a use of distorted warps would have an impact on the quality of synthesized images. This is taken into account in the coding scheme presented in the next section.

3. IMAGE QUALITY VS RATE OPTIMIZED WARP CODING

The coder successively encodes the warps of all time instances and views. Left view and right view warps are encoded separately and multiplexed into a single bit stream. Without loss of generality, the coding of the warp sequence assigned to the left view is described in the following. We denote the warp at time instant f as W^f. Each warp W^f is represented as a regular quad grid (Fig. 2) with fixed resolution $M \times N$ where each node of the grid is indexed with integer coordinates i, j and has a 2D location $W^f[i, j] \in \mathbb{R}^2$ assigned.

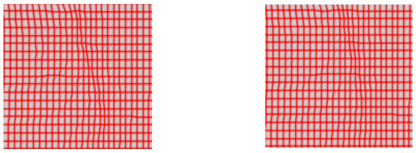


Fig. 2. Illustration of parts of temporally consecutive warps.

To efficiently exploit spatio-temporal dependencies, the coding scheme shown in Fig. 3 is used. First, each warp is spatially partitioned using a quincunx resolution pyramid. Each partition is then predictively encoded using a closed loop DPCM in combination with a spatio-temporal predictor. Finally, CABAC [8] is employed for entropy coding. This coding scheme is based on our previous work [5]. We extend this coding scheme by a Coder Control, which adjusts per frame the quantization step size and the number of partitions to be encoded. It has the goal to achieve the best compromise between number of bits needed for coding a warp and quality of the image which is synthesized by IDW using a corresponding reconstructed warp.

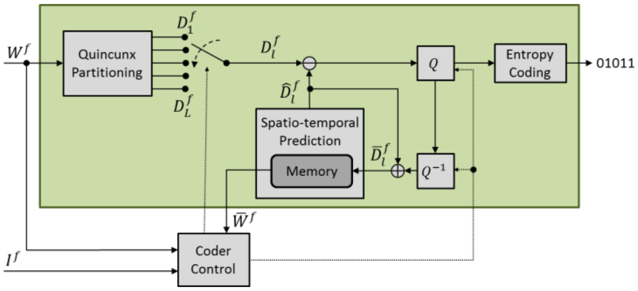
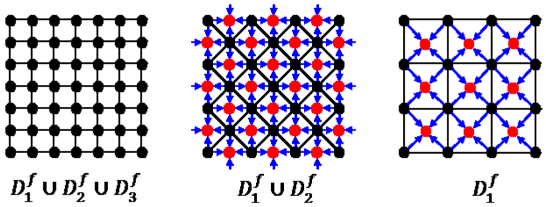


Fig. 3. Block-diagram of warp coder.

3.1. Spatial partitioning

The set of 2D locations of each warp W^f is partitioned into groups of locations (GOLs) using a quincunx resolution pyramid (Fig. 4, black nodes). The lowest resolution grid of the resolution pyramid and the difference sets between successive resolutions specify a partitioning of the locations of W^f into GOLs D_l^f. Locations of each GOL are then successively encoded from the top D_1^f to the bottom D_L^f of the pyramid, where L denotes the total number of GOLs. Thereby, the quincunx pyramid guarantees that for each interior location $W^f[i, j] \in D_l^f$ always four neighbors are located in $\bigcup_{k=1}^{l-1} D_l^f$, which can be used for intra warp prediction.



$D_1^f \cup D_2^f \cup D_3^f$ $D_1^f \cup D_2^f$ D_1^f

Fig. 4. Illustration of resolution pyramid, groups of locations, and intra warp prediction.

3.2. Prediction and quantization

Locations of each GOL D_l^f are coded in a predictive way using already coded and reconstructed locations $\overline{W}^f[i, j]$ of GOLs D_k^f with $k < l$ of the same time instant f and GOLs D_k^r of other already coded time instants r with $k \leq l$. Similar as in video coding standards, three warp coding types and corresponding prediction modes are supported: INTRA, INTER_P and INTER_B. The INTRA mode computes $\widehat{W}^f[i, j]_{\text{INTRA}}$ as the centroid of its reconstructed neighbors. In Fig. 4, the process of intra prediction is indicated with blue arrows. Modes INTER_P and INTER_B are defined as

$$\widehat{W}^f[i, j]_{\text{INTER\_P(r)}} = \widehat{W}^f[i, j]_{\text{INTRA}} + \overline{W}^r[i, j] - \widehat{W}^r[i, j]_{\text{INTRA}}$$
$$\widehat{W}^f[i, j]_{\text{INTER\_B(r,s)}} = \alpha\widehat{W}^f[i, j]_{\text{INTER\_P(r)}} + (1\text{-}\alpha)\widehat{W}^f[i, j]_{\text{INTER\_P(s)}}$$

where $\alpha = |f - s|/|r - s|$. Prediction errors $W[i, j] - \widehat{W}^f[i, j]$ are then uniformly quantized and entropy coded.

3.3. Coder control

Coder control (CC) determines for each time instant f, the quantization step size $\tilde{\Delta}$ and the total number of GOLs $\tilde{L}$ to be encoded. Hence, instead of coding all GOLs per time instant, CC can decide to encode only the first $\tilde{L} \in \{0, \dots, L\}$ GOLs if that gives the best compromise in terms of bit rate and image quality. Locations of not encoded GOLs are always reconstructed by employing the prediction mode assigned to the current time instant and assuming zero valued residuals. Note that if $\tilde{L} = 0$ the coding of the complete warp is skipped. To determine the parameter set $(\tilde{\Delta}, \tilde{L})$ for a time instant, we assume that the encoding decisions up to this time instant are given and maximize the objective function

$$Q\big(W, \tilde{L}, \tilde{\Delta}\big) - \lambda\, R(W, \tilde{L}, \tilde{\Delta}) \qquad (2)$$

where $\tilde{L} \in \{0, \dots, L\}$ and $\tilde{\Delta} \in [\Delta, \Delta - 0.2, \Delta + 0.2]$. Here Q computes the peak-signal-to-noise-ratio (PSNR) between the images $\Psi(I, W)$ and $\Psi(I, \overline{W})$ (Eq. 1), which are obtained by in-loop view synthesis with the original and the reconstructed warp respectively, while rate R represents the number of bits required for coding the first $\tilde{L}$ GOLs with step size $\tilde{\Delta}$. The use of a ± 0.2 variation in step size of the quantizer is motivated by experimental results. The Lagrangian multiplier λ [9] represents the slope of the image quality vs. warp rate function (QRF) $Q(R)$ which is obtained by maximizing (2) and varying λ under the assumption that the QRF is continuous [10]. Consequently, a maximization of (2) prevents the coding of GOLs which don't lead to an appropriate increase in image quality in relation to the necessary increase in bit rate, where λ controls this relation. The approach guarantees that the maximum view synthesis quality is achieved with the bits spent for a warp with the assumption of given coding decisions up to the current time instant. Similar objective functions are also used in the area of video coding.

4. EXPERIMENTAL RESULTS

4.1. Warp coding results

Warp coding experiments have been performed using the warps which were computed in the proposal [4]. In the proposal, 2-view video was encoded using a multi-view video coder [11], while warps were computed at the decoder side and used for IDW-based view synthesis. Instead of computing warps at the decoder, the warps are coded with the approach presented in Section 3 to enable the application scenario shown in Fig. 1b. For coding, hierarchical group of warps structures of size 12 and 15 are used to enable a random access at each 0.5 seconds for the 25 and 30 Hz sequences, respectively, as it was specified in the CfP [2]. Fig. 5 shows the impact of the bit rate used for coding the warps on the image quality of the synthesized views. The image quality is measured between the synthesis results obtained with i) the original warps and ii) the coded and reconstructed warps.

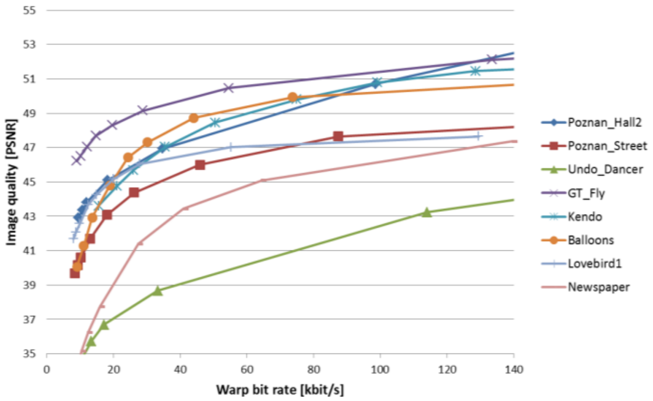


Fig. 5. Warp rate vs. image quality curves.

Each curve is obtained by coding with a fixed quantization parameter $\Delta = 0.5$ and by varying the Lagrange parameter

$$\lambda \in \{0.038, 0.035, 0.03, 0.025, 0.02, 0.015, 0.01, 0.005\}.$$

Informal viewing showed that visually lossless quality was achieved at a PSNR of about 45dB for almost all sequences. In the case of the Undo_Dancer sequence, which is an animated sequence containing large high frequency patterns, a lower PSNR of about 39 dB was perceived as visually lossless. Table 1 shows i) the exact rates of the warps at visually lossless quality as well as ii) the rates of the coded 2-view videos at the highest rate points as they were submitted in [4] as answer to the CfP. Obviously, the warp bit-rate represents only 3.6% on average of the bit rate needed for transmitting the warps and the video together. Note that in the other three of the four winning proposals of the CfP [2], depth maps were part of the 3D video format besides the 2-view video data. Thereby the coded depth maps represented a portion between 5.8% and 22.1% of the total bit rate in the respective proposals.

Table 1. Warp rate to total rate (video + warps) ratios.

Sequence	Warp		Video		Warp rate to
name	PSNR [dB]	Rate [kbit/s]	PSNR [dB]	Rate [kbit/s]	total rate ratio
Poznan_Hall2	45.1	18.2	41.4	520	3.4%
Poznan_Street	44.4	26.2	37.5	1307	2.0%
Undo_Dancer	38.7	33.3	32.6	998	3.2%
GT_Fly	48.3	19.6	36.4	1098	1.8%
Kendo	47.0	35.5	42.1	690	4.9%
Balloons	46.4	24.4	41.9	800	3.0%
Lovebird1	45.2	20.0	38.7	828	2.4%
Newspaper	45.1	64.5	38.6	719	8.2%
Average					3.6%

4.2. Synthesis time measurements

We compare the run-times needed by the IDW-based view synthesis algorithm when applied in the application scenarios shown in Fig. 1, i.e. with decoder-side and with encoder-side warp calculation. Furthermore, the IDW-run-time is compared of the run-time of the MPEG View Synthesis Reference Software 3.5 (VSRS 3.5) [12]. Note that VSRS 3.5 performs depth-image-based rendering

(DIBR) [13][14] and requires besides image data also depth maps as input. Table 2 shows run-times and timing-ratios for two cases, i) sequential, where 28 views are computed sequentially, and ii) parallel, where it is assumed that all 28 views can be computed in parallel. Time measurements for the sequential case are conducted on a PC based on Intel Core i7-920, 4 x 2.67GHz, and 12 GB, while only one CPU core is used. The times for the parallel case are derived from the sequential case. In the following, we assume that if warp calculation is performed at the encoder-side then warps are either pre-computed offline (like it is done with depth maps for DIBR) or computed in real-time with better or dedicated hardware.

Table 2. Comparison of run-times needed for the synthesis of 28-views from stereo image input expressed in seconds.

	A) IDW with enc.-side warp calc.	B) IDW with dec.-side warp calc.	C) VSRS time	B/A	C/A
sequential	15.57	19.88	80.21	1.3	5.2
parallel	0.56	4.86	2.86	8.7	5.2

According to the previous assumptions, a shift of the warp calculation from the decoder to the encoder reduces the required run-time for the synthesis by the time needed for the warp calculation. This shift of complexity to the encoder side leads to a reduction of the run-time of the IDW-based view synthesis by a factor of 1.3 and 8.7 for the sequential and parallel case, respectively. In comparison to VSRS, IDW-based view synthesis can be performed 5.2 faster for both, the sequential and the parallel case. The factors in both cases are the same since the run-time of both synthesis approaches, IDW with encoder-side warp-calculation and VSRS scale with the number of output views.

We would like to emphasize that warp calculation can be performed completely automatically in contrast to the usual depth estimation process. We strongly believe that with dedicated hardware real-time warp computation and IDW [15] is possible.

5. CONCLUSION

We presented a warp coding method which is optimized to achieve the highest view synthesis quality with the bit budget spent per warp. Furthermore, we showed that coded warps represent a practically negligible portion of about 3.6% on average of the overall (video+warp) bit rate. It was also shown that a transmission of warps leads to a reduction of synthesis time i) up to a factor of 8 in comparison to a fully automatic receiver-side view synthesis which is possible with the image-domain-warping approach, and ii) by a factor of about 5 in comparison to a view synthesis based on depth-image-based rendering. In our future work, we plan to analyze the inter-parameter dependencies in the space $(\tilde{\Delta}, \tilde{L}, \lambda)$ and the parameter dependencies between different time instances to further increase the warp coding efficiency.

6. REFERENCES

[1] A. Smolic, K. Müller, P. Merkle, A. Vetro, “Development of a new MPEG standard for advanced 3D video applications”, ISPA 2009. Salzburg, Austria, Sep. 2009.

[2] ISO/IEC MPEG, “Call for Proposals on 3D Video Coding Technology,” MPEG N12036, March 2011.

[3] M. Farre, O. Wang, M. Lang, N. Stefanoski, A. Hornung, A. Smolic, “Automatic Content Creation for Multiview Autostereoscopic Displays Using Image Domain Warping”, Hot3D Workshop 2011, Barcelona, Spain, July 2011.

[4] N. Stefanoski, P. Espinosa, O. Wang, M. Lang, A. Smolic, S. Bosse, M. Farre, K. Müller, H. Schwarz, M. Winken, T. Wiegand, “Description of 3D Video Coding Technology Proposal by Disney Research Zurich and Fraunhofer HHI”, MPEG, Doc. M22668, Geneva, Switzerland, Nov. 2011.

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

[6] ISO/IEC MPEG, “Report of Subjective Test Results from the Call for Proposals on 3D Video Coding,” MPEG N12347, 2012.

[7] G. Wolberg,“Digital Image Warping”, IEEE Computer Society Press, Los Alamitos, Calif., 1990.

[8] 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.

[9] Hugh Everett III,“Generalized Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources”, Operations Research, Vol. 11, No. 3, pp. 399-417, 1963.

[10] T. Wiegand, B. Girod, “Lagrange multiplier selection in hybrid video coder control”, ICIP 2001, Thessaloniki, Greece, Oct. 2001.

[11] H. Schwarz, C. Bartnik, S. Bosse, H. Brust, T. Hinz, H. Lakshman, D. Marpe, P. Merkle, K. Müller, H, Rhee, G. Tech, M. Winken, and T. Wiegand, “3D Video Coding Using Advanced Prediction, Depth Modeling, and Encoder Control Methods”, under submission at Picture Coding Symposium 2012, Krakow, Poland, May 2012.

[12] Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, “Reference softwares for depth estimation and view synthesis,” ISO/IEC JTC1/SC29/WG11, Archamps, France, Tech. Rep. M15377, Apr. 2008.

[13] L. Yu, T. Masayuki, Y. Zhao, C. Zhu, “3D-TV System with Depth-Image-Based Rendering: Architecture, Techniques and Challenges”, Springer New York, 1st Edition, 2012.

[14] Christoph Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV", Proc. SPIE 5291, 93, 2004.

[15] P. Greisen, M. Schaffner, S. Heinzle, M. Runo, A. Smolic, A. Burg, H. Kaeslin, M. Gross, “Analysis and VLSI Implementation of EWA Rendering for Real-time HD Video Applications”, under submission at IEEE Transactions on Circuits and Systems for Video Technology.