Warped-Skip Mode for 3D Video Coding

Fabian Jäger and Christian Feldmann
Institut für Nachrichtentechnik
RWTH Aachen University
52056 Aachen, GERMANY
{jaeger,feldmann}@ient.rwth-aachen.de

***Abstract*—Upcoming display technologies like auto stereoscopic displays require synthesis of virtual viewpoints to allow for depth impression without requiring the viewer to wear glasses. View synthesis relies on the availability of depth information at the receiver side to warp the recorded video data to arbitrary viewpoints. This depth data contains information about the 3D position of recorded pixel values and therefore inherently also contains the pixel-wise disparity information between multiple views to be coded. In this paper we propose to use this depth information also for compression of multi view video+depth sequences. By warping the base view's frame of a multi view video to the position of the enhancement views' camera a prediction signal is formed, which can either be used for the proposed warped skip mode or as a prediction signal to reduce the residual's energy.**

I. Introduction

Recent developments in the field of 3D video coding show the advantage of adding depth information to conventional video data to allow for synthesis of arbitrary viewpoints at the receiver side. View synthesis can be used in many different ways. The primary application for view synthesis in terms of 3D video is to target auto stereoscopic displays, which require many (more than 20) different views of the same scene to be displayed at the same time. As it is currently infeasible to encode and transmit that many views at a decent bit rate and quality, 3D video introduces the concept of sending a reduced number of frames (currently 2 or 3) and synthesizing the remaining viewpoints based on the accompanying depth data. The MPEG 3D Video Adhoc Group issued a Call for Proposals on 3D Video Coding Technologies [1] in March 2011 to seek for new ways of coding multi view video+depth (MVD) data, which also shows the necessity for new, innovative coding approaches for this special type of data.

Previous work in this field by Merkle et al. [2] showed that multi view video+depth (MVD) data can be coded by using the MVC coding standard [3], which exploits inter-view dependencies for combined temporal/inter-view prediction. In their coding scheme video and depth data is coded independently, but each type of data uses inter-view prediction to increase coding efficiency. Instead of using motion vectors as with temporal prediction, MVC uses disparity vectors to express the location of the prediction signal in another view's coded frame. The rest of the coding concept is the same as in single view video coding with motion compensation. That means, that a motion vector and a residual signal is formed for a frame using temporal prediction. It was shown that depth data can also be efficiently encoded by means of inter-view prediction. Later, Morvan et al., Shimizu et al. and Yamamoto et al. [4], [5], [6] extended the approach of independent MVC encoding of video and depth data by introducing a new prediction signal based on a synthesized frame. By utilizing the base view's compressed depth data to warp its video+depth frame to the enhancement view's camera position an additional prediction signal is constructed. It was shown that this additional prediction signal reduces the bit rate to encode the enhancement views by a small percentage. Their approaches treated the warped prediction signal the same way as conventional temporal prediction signals are treated in motion compensation. Since the warping process is imperfect, it does not result in a pixel-precise prediction for the enhancement view. This means that there are still disparity vectors and a residual to be encoded into the bitstream for every block. Due to this design approach, previously proposed warped prediction did not yield a high coding efficiency gain compared to conventional multi view coding.

In this paper an alternative coding approach for MVD is proposed, which also uses a warped version of the base view's video+depth as a prediction signal for encoding the enhancement views. In our approach we assume that the warping process is not pixel-precise, but still contains almost all the information of the enhancement view. To reduce the amount of bits for coding the enhancement view while having the same visual quality in all coded views, we introduce a new coding mode, which is called *warped-skip mode*. This new mode allows to skip certain image regions based on the warped prediction signal. For those regions no additional information, like disparity vectors or a residual, is encoded. The mode decision, whether to select the warped-skip mode or not, cannot be based on a rate-distortion optimization (RDO) as the warping process using compressed depth maps produces non-pixel-precise images. An RDO-based decision would only seldomly decide for the warped-skip mode, although the visual impression of the warped signal would allow to use this mode. Therefore we introduce a non-RDO-based decision criterion based on an occupancy mask. This mask represents the reliability of the warped and inpainted base view and allows for a more efficient warped-skip mode decision.

The remainder of this paper is organized as follows. Section II introduces the algorithm for forming the warped prediction signal before we describe the new mode decision process in Section III. Afterwards, experimental results are shown in Section IV, before we conclude this paper with a summary and an outlook on future extension.

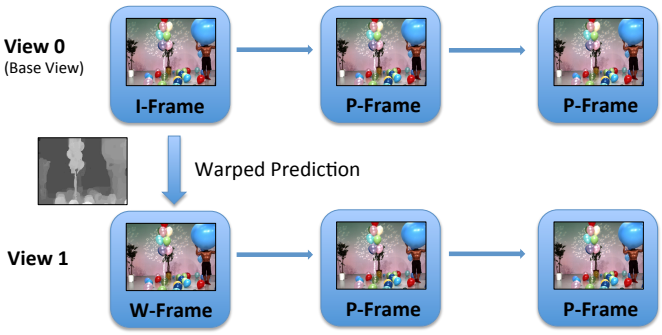


Fig. 1. Prediction structure for the proposed coding approach using warped prediction to reduce the bit rate of enhancement views' key pictures. The required depth map for the warping process is encoded separately and afterwards incorporated into the coding process for the video data.

II. Warped Prediction

In the proposed coding scenario we assume that the depth sequences of all views are already encoded by using an MVC approach as previously described. When encoding the texture video data the reconstructed depth maps are used to form the warped prediction signal. As the warped prediction signal is based on the video plus depth frame of the sequences base view, the texture video of the base view needs to be encoded in a first step and reconstructed for the warping process. Previous experiments comparing inter-view and temporal dependencies of multi view video showed that temporal prediction yields better performance than inter-view prediction. Consequently we use warped prediction only for key pictures and especially to replace intra-coded pictures of the enhancement views with pictures relying on warped prediction. These pictures are therefore called *W-Frames* throughout the rest of this paper. In Figure 1 the general prediction structure is illustrated. For simplification an IPPP coding structure is assumed in this figure, but the concept can easily be applied to hierarchical B-Frames or other coding structures.

A. Warping Process

After having reconstructed the base view's compressed video and depth data of a key frame, this information is utilized to generate warped representations of this view for every enhancement view to be encoded. This is accomplished by taking camera parameters into account, which describe the focal length (f), nearest and farthest depth value (z_{near}, z_{far}) and the baseline distance between the base view's camera and the enhancement view's camera (b). Based on these camera parameters an integer pixel shift (disparity) for every possible depth value is computed. For every pixel of the reconstructed base view the pixel shift in respect to the current depth value is applied and leads to the warped version of the base view as seen from the additional view's camera. The disparities mapping depth map values d to disparity values disp are pre-computed as follows:

$$z(d) = \frac{1}{\frac{d}{d_{max}} \cdot \left(\frac{1}{z_{near}} - \frac{1}{z_{far}}\right) + \frac{1}{z_{far}}} \tag{1}$$

and

$$\mathrm{disp}(z) = \frac{f}{z} \cdot b \tag{2}$$

where d_{max} is the maximum disparity value depending on the number of bits used for depth information (e.g. 255 for 8bit).

As we assume cameras with coplanar image planes and with camera centers, which are aligned on a 1D line in 3D space, we don't have to consider rotations and vertical pixel shifts between the cameras. Consequently, the warped horizontal pixel position x' can be computed by adding the value $disp$ to the position x in the base view. The depth value $d(x)$ can be read from the reconstructed depth map of the base view. Applying this pixel shift to all pixels in the base view while preserving correct depth ordering (pixels closer to the camera occlude pixels, which are farther away) results in the targeted warped prediction signal for the corresponding enhancement view.

B. Occupancy Mask

When computing the warped view, an occupancy mask is also generated, which describes for each enhancement view and each pixel position whether the pixel could be warped from the base view or not. It is represented by a floating point matrix of the same dimensions as the video frame. Before the warping process starts, the occupancy mask $P(x, y)$ is initialized with zeros, which means that all pixels are unknown in the enhancement view's warped prediction signal. Each pixel position (x, y) being successfully warped from the base view's pixels by following the above formulas, gets an occupancy value of 1.0 in the occupancy mask. After the warping process for the enhancement view is finished, the occupancy mask can be described as follows:

$$P(x,y) = \begin{cases} 1.0, & \text{if successfully warped,} \\ 0.0, & \text{if unknown.} \end{cases} \tag{3}$$

After the warping process a simple inpainting algorithm is applied to the unknown pixels. This algorithm searches to the left and to the right of the unknown pixel for known pixels and fills all unknown pixels in between with the pixel value of the closest background position. The occupancy mask is also modified for these unpainted pixels following equation (4)

$$P(x,y) = \max(1.0 - r \cdot \min(k_l, k_r), 0) \tag{4}$$

where r is the attenuation rate and k_l and k_r are the distances to the next known pixel on the left and on the right of the previously unknown pixel. Typical values for r are in the range from 0.1 to 0.2, which means that pixels inpainted from a known pixel more than 5 or 10 pixels away are treated the same as unknown pixels.

III. Mode Decision

As previously stated, the mode decision process needs to be adjusted to utilize the warped prediction signal in an efficient way. The occupancy mask $P(x, y)$ describes the reliability of each pixel in the warped prediction signal and can therefore be used to decide whether to use the warped-skip mode or

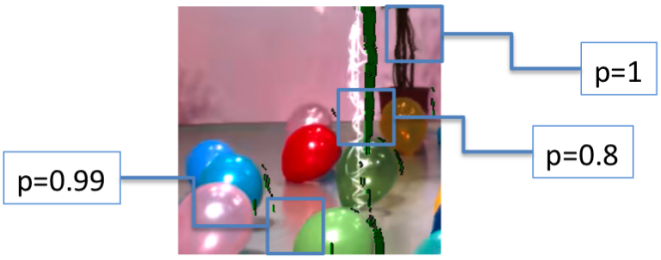


Fig. 2. Segment of the warped prediction signal for the Balloons sequence before the inpainting algorithm is applied. Disoccluded regions are drawn in dark green. The average occupancy $p(\mathrm{CU})$ for three different potential coding units is illustrated.

to encode the enhancement view's key frame conventionally. To incorporate the occupancy mask into the mode decision process the average occupancy for the current coding unit (CU) is computed by

$$p(\mathrm{CU}) = \frac{\sum_{(x,y)\in \mathrm{CU}} P(x,y)}{\mathrm{size}(\mathrm{CU})} \tag{5}$$

If the average occupancy for the current CU is above a defined threshold t_p, the current coding unit uses the proposed warped-skip mode and no further RDO-based mode decision is tested. In the other case, the warped-skip mode is not chosen and the coder can continue with its conventional RD-based mode decision. The previously constructed warped prediction signal is still available as a prediction signal for a disparity-compensated prediction mode including disparity vectors and a residual signal. Hence, the introduction of a warped-skip mode does not contradict the idea of having an additional warped prediction signal for conventional disparity/temporal prediction modes.

IV. Experimental Results

To measure the coding efficiency gained by the proposed warped-skip mode, we modified the current test software (HM 3.3) for High Efficiency Video Coding [7] to either use the warped prediction as a reference image in the former I-Frames of the enhancement views or to use the proposed warped-skip mode based on the warped reference and the occupancy mask. In the following section only a 2 view case is discussed (base view + 1 enhancement view), but the concept can easily be extended to more enhancement views. In all the presented coding scenarios the depth maps of both views are encoded independently with HM 3.3 in simulcast mode.

Using the compressed base view as a reference image for the enhancement view is referred to as HMVC as this concept is very similar to the MVC coding standard. For completeness we also show results for encoding the two texture videos independently in simulcast mode with HM 3.3. As the PSNR values for both views are very similar when encoded with HMVC and HEVC simulcast, we only plotted the average PSNR to avoid confusing RD plots with too many curves. To illustrate the difference in terms of PSNR between the base view and the enhancement view when using the proposed method, both curves are plotted in this case.

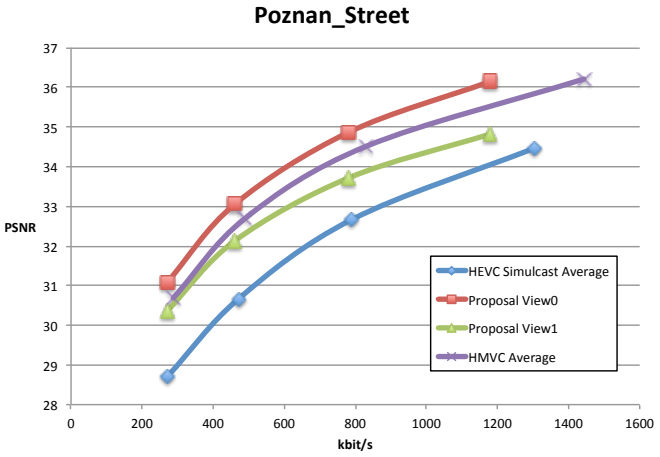


Fig. 3. Rate-distortion results for the sequence Poznan_Street (1920x1088 Pixels, 25 Frames per Second).

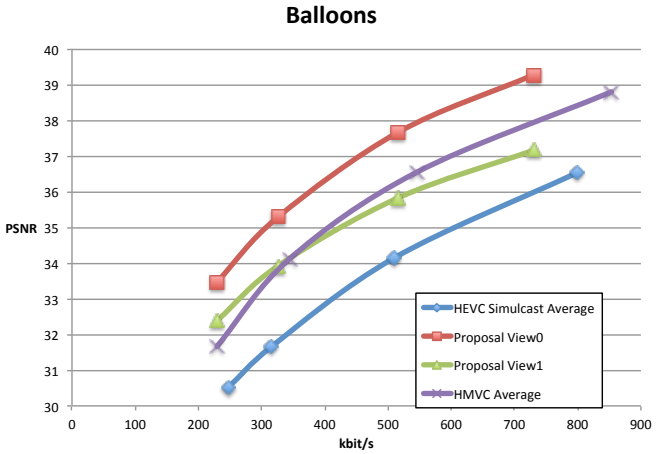


Fig. 4. Rate-distortion results for the sequence Balloons (1024x768 Pixels, 30 Frames per Second).

It is also important to note that the proposed method does not target RD-optimal reconstruction quality measured in PSNR, but high visual quality in 2D and especially in stereo viewing. The plotted bit rates include both, texture and depth data and the depth rate is about 10% of the overall bit rate.

For the sequence Poznan_Street there is a PSNR improvement of about 0.5dB compared to HMVC and 2.5dB to simulcast. This holds only for the base view as the enhancement view is between 0.1 and 0.8db below HMVC. But as most of the enhancement view's frame is just a shifted copy of the base view's picture, the visual quality is comparable to the base view's. As already mentioned in the introduction, objective quality metrics are not suited to measure the quality of a warped image as already small pixel shifts result in a high MSE, although these pixel shifts are visually not disturbing. As we are targeting visual quality the warped-skip mode is an efficient coding tool, which yields very low bit rates for the enhancement view while inheriting its visual quality from the base view.

For the sequence Balloons the curve of the proposed method's enhancement view and the average HMVC curve intersect. That means that for low bit rate scenarios the warped-skip mode performs even objectively better than the simple warped prediction due to its extremely low encoding costs. With the explanation given before, the visual quality for these sequences is even more superior compared to the warped prediction without warped-skip mode. The base view's PSNR values, which give a rough approximation of both views' visual quality, are between 1.2 and 2 dB better compared to HMVC. Compared to simulcast it yields an improvement in objective quality of up to 3.5 dB.

To give an impression of the visual quality of the proposed warped-skip mode, we show some exemplary results in Figures 5 and 6. It can be seen that the visual quality of the two views for the proposed method is almost the same, although the PSNR values are about 1.2 to 2.0 dB apart. This underlines the behavior of the proposed warped-skip mode as it targets visual quality instead of pixel-precise reconstruction.

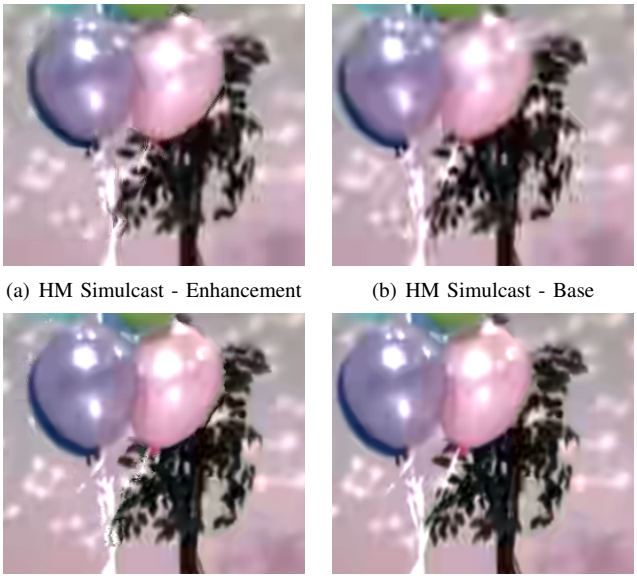
(a) HM Simulcast - Enhancement (b) HM Simulcast - Base

(c) Proposed - Enhancement (d) Proposed - Base

Fig. 5. The visual quality of the sequence Poznan_Street at 230 kbit/s for HM simulcast and for the proposed warped-skip mode. The right images show the base view's quality while the left images show the enhancement view's quality.

V. Conclusion

In this paper a novel coding mode for multi-view video+depth is proposed, which exploits inter-view dependencies by using warped prediction in combination with a new coding mode, the warped-skip mode. High visual quality in the enhancement views can be achieved by incorporating an occupancy mask into the mode decision process instead of using pure RDO for mode decision. The new coding mode does not require any additional information (besides the warped-skip flag) to be encoded for regions, which can be completely described by a warped prediction signal from the base view.



(a) HM Simulcast - Enhancement (b) HM Simulcast - Base

(c) Proposed - Enhancement (d) Proposed - Base

Fig. 6. The visual quality of the sequence Balloons at 320 kbit/s for HM simulcast and for the proposed warped-skip mode. The right images show the base view's quality while the left images show the enhancement view's quality.

As a fallback, the warped prediction signal can still be used together with disparity vectors and a residual signal if the warped-skip mode is not selected due to disocclusions during the warping process. The proposed coding method yields very low bit rates for the enhancement views while ensuring almost the same visual quality as the base view. Further improvements of this new approach could leverage more sophisticated inpainting methods for disoccluded regions. Moreover, pixel shifting artifacts may need to be addressed in future research activities, although they tend not to reduce the visual quality of the reconstructed video as severe as other compression artifacts do.

References

[1] MPEG Video and Requirement Groups, "Call for proposals on 3d video coding technology," MPEG output document N12036, Tech. Rep., March 2011.

[2] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, vol. 1. IEEE, 2007, pp. I–201.

[3] P. Merkle, K. Muller, A. Smolic, and T. Wiegand, "Efficient compression of multi-view video exploiting inter-view dependencies based on h. 264/mpeg4-avc," in *2006 IEEE International Conference on Multimedia and Expo*. IEEE, 2006, pp. 1717–1720.

[4] Y. Morvan and D. Farin, "System architecture for free-viewpoint video and 3d-tv," *Consumer Electronics, IEEE Transactions on*, vol. 54, no. 2, pp. 925–932, 2008.

[5] S. Shimizu, M. Kitahara, H. Kimata, K. Kamikura, and Y. Yashima, "View scalable multiview video coding using 3-d warping with depth map," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 11, pp. 1485–1495, 2007.

[6] K. Yamamoto, M. Kitahara, H. Kimata, T. Yendo, T. Fujii, M. Tanimoto, S. Shimizu, K. Kamikura, and Y. Yashima, "Multiview video coding using view interpolation and color correction," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 11, pp. 1436–1449, 2007.

[7] T. Wiegand, B. Bross, W.-J. Han, J.-R. Ohm, and G. J. Sullivan, "Working draft 3 of high-efficiency video coding (hevc)," Joint Collaborative Team on Video Coding (JCT-VC), Doc. JCTVC-C403, 2011.