This material is published in the open archive of Mid Sweden University
DIVA http://miun.diva-portal.org
to ensure timely dissemination of scholarly and technical work. Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright. In most cases, these works may not be reposted without the explicit permission of the copyright holder.

Karlsson, L.S.; Sjostrom, M. , "Region-of-Interest 3D Video Coding Based on Depth Images," *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2008* , pp.141-144, 28-30 May 2008

http://dx.doi.org/10.1109/3DTV.2008.4547828

© 2008 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

REGION-OF-INTEREST 3D VIDEO CODING BASED ON DEPTH IMAGES

L. S. Karlsson, M. Sjöström

Mid Sweden University
Department of Information Technology and Media
SE-851 70 Sundsvall, Sweden

ABSTRACT

Three dimensional (3D) TV is becoming a mature technology due to the progress within areas such as display and network technology among others. However, 3D video demands a higher bandwidth in order to transmit the information needed to render or directly display several different views at the receiver. The 2D plus depth representation requires less bit rate than most 3D video representations, although the necessary views have to be rendered at the receiver. In this paper we propose to combine the 2D plus depth representation with region-of-interest (ROI) video coding to ensure a higher quality at parts of the sequence that are of interest to the viewer. These include objects close to the viewer as well as faces. This allows either the bit rate to be reduced by 12-28 % or the quality within the ROI to be increased by 0.57 - 1.5 dB, when a fixed bit rate is applied.

***Index Terms*—** Three-dimensional displays, Video coding

1. INTRODUCTION

Three-dimensional (3D) TV has been researched for decades, however, the progress within display and network technology as well as software has made commercial implementations possible for both large TV screens [1] and mobile phones [2]. Several techniques for providing a 3D experience to the viewer has been presented [3], including multiview, holograpy and volumetric. The multiview methods are considered the most promising as they provide all necessary depth cues. It also allows freedom of movement when enough views are displayed. However, the main disadvantage of multiview solutions is that each additional view gives a large increase in spatial resolution. Thus, the redundancy between the views must be exploited in the encoding to make real-time transmission possible. Especially for low bit rate applications such as mobile phones. This has been applied using several types of representations of the data and encoding. The suggested representations ranges from multiview, where all the views are transmitted as they were captured [4] to the 2D plus depth representation containing only one view and depth information [5]. Encoding is performed using existing standards including MPEG-C part 3 (ISO/IEC 23002-3) [6], that supports multiview coding (MVC) and 2D plus depth.

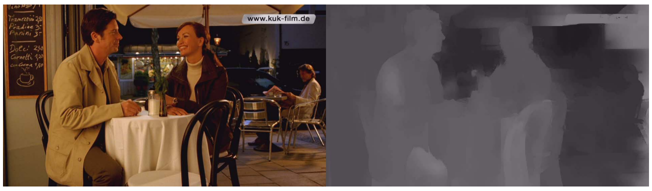


Fig. 1. Each frame in the 2D plus depth format consists of a monoscopic 2D view (left) and a depth image (right) that can be used to generate other views. [7]

In this paper we have focused on the 2D plus depth representation of the 3D video [5, 6]. (See figure 1.) This representation contains one monoscopic 2D color video sequence, representing one view, and a depth image sequence. The bit rate is substantially reduced compared to transmitting all of the views and it is directly compatibility with 2D video. However, the necessary views have to be rendered before they can be displayed and a part of the scene to render might be occluded in the monoscopic 2D image. The problem with occlusion may be solved by using layered depth-images [8], that allows more than one pair of color and depth values per pixel at the cost of a much more complicated realisation.

In previous approaches, the monoscopic 2D video sequence and the depth image sequence are compressed and transmitted separately. The monoscopic 2D video sequence is compressed using MPEG-2 in most cases [5, 9, 10] as it is the most common standard for broadcasting 2D TV. In the paper [9] by Grewatsch et al, the motion vectors from the encoding of the monoscopic 2D video sequence was used to encode the depth image sequence as well. However, Pouzarad et al. in [10] claims that these motion vectors are not suited for the sharp edges and distant objects present in the depth images. A method was suggested that detects these regions in the depth images and reestimates the motion vectors for them. Another approach is to simply use the H.264 (MPEG-4 AVC) standard. It was shown in [5] to be the more suitable standard for

Thanks to Swedish Graduate School in Telecommunications and the EU Objective 1- programme Södra Skogslän for funding.

978-1-4244-1755-1/08/$25.00 ©2008 IEEE

the type of data that is included in a depth image sequence.

In this paper we propose a method to increase the percieved quality in a 3D video sequence at low bit rates. This can be achieved by applying region-of-interest (ROI) coding to a 2D plus depth video sequence. ROI coding has been used for ordinary 2D video to increase the percieved quality at limited bit rates. This is achieved by increasing the quality in regions interesting to the viewer at the expense of a reduction of quality in the background [11]. In the proposed method the position of the ROI is determined using the information in the depth map. We also propose to add skin detection to the ROI detection. This ensures inclusion of faces in the ROIs.

The paper is organized as follows: In section 2 the 2D plus depth representation is described in detail. Section 3 gives a short introduction of ROI video coding and a description of the proposed algorithm. This is followed by experimental set up in section 4 and experimental results in section 5.

2. THE 2D PLUS DEPTH VIDEO SEQUENCE

The 2D plus depth video sequence representation presented by Fehn in [5] consists of a monoscopic 2D color video sequence and a depth image sequence. The monoscopic 2D video sequence represents one view of the scene it is created from. The depth image then contains depth information with respect to the view represented by the monoscopic 2D image. (See figure 1.)

The depth image data of each pixel (m, n) in the 2D plus depth representation is presented by an 8-bit graylevel depth value ν where graylevel 0 represents the farthest value and 255 the closest. The graylevel depth value ν is related to the real metric depth value Z by,

$$Z = Z_{far} + \nu \cdot \frac{Z_{near} - Z_{far}}{255} \quad \text{with} \quad \nu \in [0, ..., 255]\,, \tag{1}$$

where Z_{far} denotes the far clipping plane and Z_{near} the near clipping plane.

3. ROI CODING BASED OF 2D PLUS DEPTH VIDEO

ROI coding has been applied to increase the percieved quality in 2D video [11]. A ROI is either determined using general visual cues or based on an application. For example, faces are interesting to the viewer in a video conferencing application. Once the correct position of the ROI is determined this can be used to re-allocate bits from the background to the ROI in order to increase the perceived quality. The bit-allocation can be performed independently of codec (pre-processing) or by controlling the parameters within the codec. In [12], an ROI based approach was applied, where the quantization parameters of the MPEG-2 encoder was controlled by the information recieved from the rendering of an artificial 3D scene. Hence, this was based on the known position of the viewer.

We propose that ROI coding is used to improve the percieved quality of the 3D video sequences. This can be achieved by applying ROI video coding to the monoscopic 2D video part of the 2D plus depth representation. Two ROI detection methods have been employed. The ROI is firstly considered to contain the information classified as being close to the viewer (ROI detection method I). Secondly, additional content, such as faces, is also considered important to the viewer and therefore the depth and face cues are combined (ROI detection method II). After the position of the ROI has been determined, we apply the spatio-temporal filter in [11] to achieve an increase in perceived quality without increasing the bit rate.

3.1. ROI detection method I

The depth image sequence indicates how the monoscopic 2D video sequences should be rendered in order to generate the necessary views. The depth image sequence indicates how close to the viewer the information in each pixel will appear. Thus, by extracting the ROI based on depth values, the content close to the viewer will be given a higher priority. We propose that the ROI is determined using a threshold of the depth values A_D that classifies a certain amount of the pixels as ROI pixels. The threshold is adapted to the scene statistics as the depth content varies highly from scene to scene and camera movement within a scene. This is achieved by calculating the probability $p(\nu)$ that a pixel has the depth value ν to determine A_D. It is given by

$$p(\nu) = \frac{H(\lceil \nu/a \rceil)}{M \cdot N} \tag{2}$$

where $H(\lceil \nu/a \rceil)$ is the histogram of depth value ν for a frame of size $M \times N$ and binsize a. The depth threshold A_D is then defined by

$$A_D = \{\min A; \sum_{d=0}^{d=A} p(\lceil \nu/a \rceil) \geq 1/3\}$$

Based on this threshold, the position of the ROI is defined for each frame f by a binary detection map

$$B_{DM,depth}^{(m,n)} = \begin{cases} 1, & \text{if } \lceil I^{f,(m,n)}/a \rceil \geq A_D \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

for each pixel (m, n).

3.2. ROI detection method II

The sequences contain talking heads and thus this ROI detection method is determined using both skin-color detection (to detect the face) and depth detection (See section 3.1). The parametric model in [13] was used for the skin-color detection, and the binary detection map, $B_{DM,skin}$, was extracted using experimentally determined thresholds. The skin detection map, $B_{DM,skin}$, can be joined with the binary depth detection map, $B_{DM,depth}$, into one combined detection map $B_{DM,comb} = B_{DM,depth}$ OR $B_{DM,skin}$ for each frame.

3.3. Spatio-temporal filter

The spatio-temporal (SPTP) filter has been show to successfully re-allocate bits from the background to the ROI in [11]. The filtering is performed in two steps. The first step is the calculation of the quality map Q describing the position of the ROI and the distance to its border. The second step is the spatio-temporal filtering that is performed based on this quality map. Further details and equations can be found in [11].

4. EXPERIMENTAL SETUP

The proposed algorithms were evaluated using two 2D plus depth sequences extracted from the *Cebit 2006* sequence [7] (Full resolution $M \times N = 960 \times 540$, low resolution $M \times N = 480 \times 270$ and framerate 25 fps). Their depth images were estimated from stereo video. The first test sequence *cafe* contains a talking couple in a cafe and a waiter. The second, *2man*, contains two men. One is a securityguard at the airport using a metaldetector on the other.

The two described types of ROI detection were applied to the monoscopic 2D sequences. The same parameters were used for the quality map Q and the SPTP filter as in the tests in [11].

The sequences filtered using SPTP and the two types of ROI detection, were encoded using the H.264 codec, JM 10.1 [14] for the High Profile. In addition the original is encoded using this encoder and used as a reference in the tests. Two performance measures were used: The bit rate in the case of a fixed quantization parameter, $Qp = 28$ and the PSNR of the ROI when a target bit rate is applied. The average PSNR of the ROI of the intensity component Y for frame f and the total number of frames F is defined as

$$PSNR_{ROI,Avg} = \frac{1}{F}\sum_{f=1}^{F} 10\log_{10}\frac{255^2}{MSE_{ROI}^{(f)}}$$

$$MSE_{ROI}^{(f)} = E_{ROI}\{(Y^{(f,(m,n))} - \hat{Y}^{(f,(m,n))})^2\}.$$

5. RESULTS

The performance of Method I is always better than method II considering $PSNR_{ROI,Avg}$ and bit rate. This is due to the larger background in Method I. The percieved quality of the sequences depend on the ROI detection method. This can be seen in figure 3 showing a frame from the original and the pre-processed *cafe* sequences ($M \times N = 480 \times 270$) that was coded using a fixed bit rate.

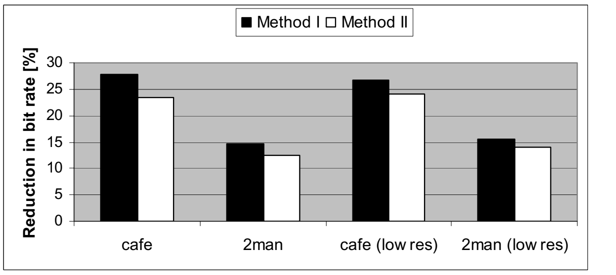


Fig. 2. The decrease in bitrate in percent acheived by using the two ROI encoding methods compared to encoding the original sequence.

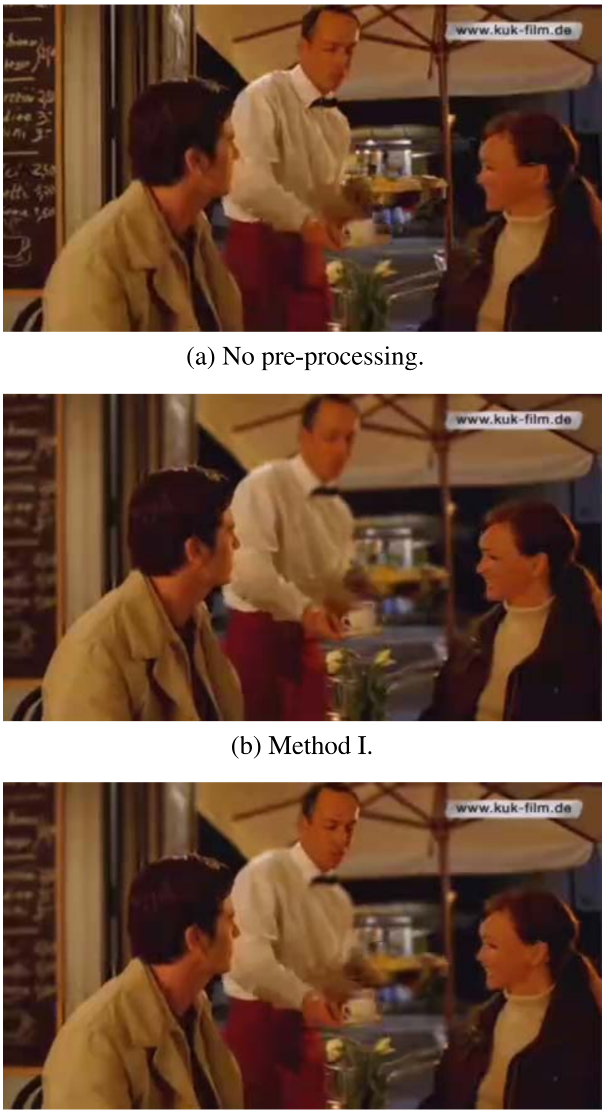
Method II has a better perceived quality than method I and the original sequence encoded without pre-processing. The faces of the people in this sequence are important to the viewer, since the people are communicating with each other. Method I fails to include the face of the waiter (in the background) within the ROIs. Method II, on the other hand, inludes all three faces (in addition to other objects close to the viewer) within the ROIs. This results in a higher quality of all faces for method II and therefore a better perceived quality in the whole sequence.

The ROI video coding methods applied in this paper save about 24 - 28 % of the bit rate for the *cafe* sequence (See figure 2) at both resolutions. However, the background of the *2man* sequence contains less information and movement. Therefore less bits are saved (12 - 15 %) compared to the *cafe* sequence.

The $PSNR_{ROI,Avg}$ is increased by at least 1.1 dB for the *cafe* sequence with a preserved bit rate (See figure 4.), for both Method I and Method II. The *2man* sequence gives a smaller increase of 0.57 dB, again due to its background. The result is similar for the low resolution version of the video sequences.

6. CONCLUSIONS

3D TV gives an increased visual experience, but at the expense of a higher bit rate as each additional view increases the amount of information to transmit. Particulary in mobile phones there is a limitiation of avaliable bit rate. The 2D plus depth representation demands less bit rate to transmit than for example multiview. In order to decrease the bit rate further, we propose to use ROI coding on the 2D monoscopic images in the video sequence. We propose in this paper to use the depth image, possibly along with facial regions to define the ROI such that regions close to the viewer and facial regions obtain higher quality. With the methods proposed in this paper, the $PSNR_{ROI,Avg}$ is increased by 0.57 - 1.5 dB depending on the background content for a fixed bit rate. In addition, there is a decrease in bit rate by 12-28 % for fixed encoder parameters.



(a) No pre-processing.

(b) Method I.

(c) Method II.

Fig. 3. Frame 188 of the *cafe* sequence ($M \times N = 480 \times 270$) coded using H.264 at 150 kbps using either no pre-processing or the two ROI detection methods and the spatio-temporal filter.

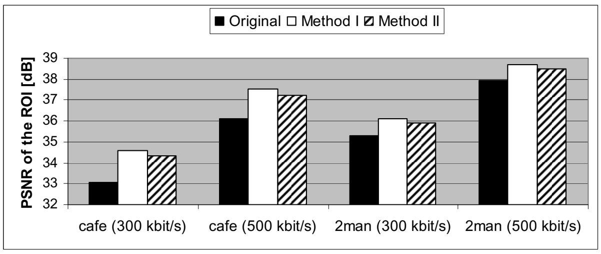


Fig. 4. The $PSNR_{ROI,Avg}$ of the *cafe* and *2man* sequences considering both encoding of the orginal sequences and the filtered sequences.

7. REFERENCES

[1] Phillips Research Press Information, "Philips 3D information display solution adds extra dimension to in-store messaging," http://www.research.philips.com/newscenter/archive/, September 2005.

[2] J. Harrold and G.J. Woodgate, "Autostereoscopic display technology for mobile 3dtv applications," in *Stereoscopic Displays and Applications XVIII*. SPIE, 2007.

[3] J.-Y. Son, B. Javidi, and K.-D. Kwack, "Methods for displaying three-dimensional images," *Proceedings of the IEEE*, vol. 94, pp. 502–523, March 2006.

[4] P. Merkle, K. Muller, and T. Wiegand, "Efficient compression of multiview video exploiting inter-view dependencies based on h.264/mpeg3-avc," in *ICME*. IEEE, 2006, pp. 1717–1720.

[5] C. Fehn, "Depth-imag-based rendering (dibr), compression and transmission for a new approach on 3d-tv," in *Stereoscopic Displays and Virtual Reality Systems XI*. SPIE, 2004, pp. 93–104.

[6] ISO/IEC 23003-3, "Mpeg-c part 3: Representation of auxiliary video and supplemental information," October 2007.

[7] WOWvx, ," http://www.wowvx.com/video.html.

[8] J. Duan and J. Li, "Compression of layered depth image," *IEEE Transactions on Image Processing*, vol. 12, pp. 365–370, March 2003.

[9] S. Grewatsch and E. Muller, "Sharing of motion vectors in 3d video coding," in *ICIP*. IEEE, 2004, pp. 3271–3274.

[10] M.T. Pourazad, P. Nasipoulos, and R.K. Ward, "An h.264 based video encoding scheme for 3d tv," in *EUSIPCO*. EURASIP, 2006.

[11] Linda S. Karlsson, "Spatio-Temporal Pre-Processing Methods for Region-of-Interest Video Coding," Licenciate Thesis No. 21 (2007), Dept. of Information Technology and Media, Mid Sweden University, Sundsvall, http://urn.kb.se/resolve?urn=urn:nbn:se:miun:diva-51.

[12] E. Masala and D. Quaglia, "Perceptually optimized mpeg compression of synthetic vide sequences," in *ICIP*, 2005, vol. 1, pp. 601–604.

[13] Y.-X. Lv, Z.-Q. Liu, and X.-H. Zhu, "Real-time face detection based on skin-color model and morphological filters," in *ICMLC*, 2003, pp. 3203–3207.

[14] JM 10.1, ," http://iphome.hhi.de/suehring/tml.