CORRELATION HISTOGRAM ANALYSIS OF DEPTH-ENHANCED 3D VIDEO CODING

Philipp Merkle[1], Jordi Bayo Singla[1], Karsten Müller[1], and Thomas Wiegand[1,2]

[1]Fraunhofer Institute for Telecommunications – Heinrich Hertz Institute, Berlin, Germany;
[2]Dep. of Telecommunication Systems, School of EE and CS, Technical University of Berlin, Germany

ABSTRACT

This paper introduces a correlation histogram method for analyzing the different components of depth-enhanced 3D video representations. Depth-enhanced 3D representations such as multi-view video plus depth consist of two components: video and depth map sequences. As depth maps represent the scene geometry, their characteristics differ from the video data. We present a comparative analysis that identifies the significant characteristics of the two components via correlation histograms. These characteristics are of special importance for compression. Modern video codecs like H.264/AVC are highly optimized to the statistical properties of natural video. Therefore the effect of compressing the two components using the MVC extension of H.264/AVC is evaluated in the second part of the analysis. The presented results show that correlation histograms are a powerful and well-suited method for analyzing the impact of processing on the characteristics of depth-enhanced 3D video.

Index Terms— 3D video, correlation histogram, multi-view video coding, video plus depth representation

1. INTRODUCTION

In 3D video two different views of the scene are shown to the eyes of a user. Currently, 3D video is emerging from 3D cinemas to home entertainment and mobile device applications. This requires efficient technologies for the whole 3D video processing chain, including content production, coding, transmission and display [1].

The 3D video representations can be grouped into video-only and depth-enhanced formats. Most popular video-only 3D formats include stereo video (SV) and multi-view video (MVV). They all have in common that they only consist of the video data of two or more camera views of the same scene. Efficient algorithms, standards, and techniques for end-to-end systems with video-only formats are available [2]. However, the video-only formats do not support 3D video on autostereoscopic displays (thus avoiding stereo glasses) and adaptation of the 3D impression to the actual display conditions. Depth-enhanced formats do support such advanced applications, as the depth or disparity information included in the representation enables the synthesis of virtual views via depth-image-based rendering (DIBR)[3].

The motivation for the presented work is to gain deeper insight into the specific characteristics of depth information. For this purpose a detailed analysis via correlation histograms is carried out. The paper is organized as follows: Section 2 introduces the depth-enhanced 3D representations. Section 3 describes the correlation histogram and coding methods used in the statistical analysis. Section 4 evaluates the statistical characteristics of uncompressed and compressed synthetic and natural test data.

2. DEPTH-ENHANCED 3D REPRESENTATIONS

A depth-enhanced 3D video representation consists of color or texture video data of one or more camera views of the same scene and additional depth information, which represents the spatial arrangement or 3D geometry of the scene. The most popular depth information format is sample-accurate depth map: for each sample in the color video sequences a depth value is given that provides the (relative) distance from the camera's image plane.

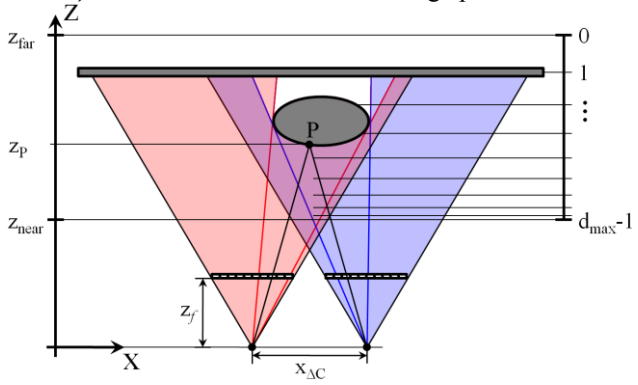

Fig. 1: Relation between 3D scene, multi-camera geometry and 2D depth maps.

The multi-view video plus depth (MVD) format is the most common depth-enhanced 3D representation, consisting of the video and depth sequences of one to N cameras (typically 2-3). In the context of 3D video processing and coding the depth information is treated as a sequence of luminance-only or grayscale images. Fig. 1 illustrates the 3D scene geometry and the assignment of depth values. For converting the real world or scene depth value z_P of a point P to the pixel value in the depth image d_P, the depth range of the scene needs to be limited to the range between the near and far clipping plane, z_{near} and z_{far}, respectively. In a next step, the continuous real world z values in the range

$[z_{near}, z_{far}]$ are quantized to the discrete depth values in the range $[0, d_{max} - 1]$. In Fig. 1 quantization with an inverse characteristic is illustrated. Such a quantization has the advantage that the foreground is quantized finer and the background coarser. Smaller quantization steps for objects that are closer to the camera's image plane preserve more detail in the foreground and less in the background.

One challenge associated with depth-enhanced 3D video representations is that the depth signal is typically not captured together with the video signal, as available sensors neither have a high enough resolution, nor an appropriate sensor range, nor the necessary frame rate. Hence, high-resolution depth information needs to be estimated from the multi-view video data. Depth estimation is based on the idea of deriving the disparity between samples, patches, or blocks of neighboring camera views by disparity matching algorithms [4]. Assuming a parallel camera setup, the relation between depth z_P and disparity δ_P of point P is as follows

$$\delta_P = \frac{z_f \cdot x_{\Delta C}}{z_P}$$

with z_f being the focal length and $x_{\Delta C}$ being the camera baseline (cp. Fig.1). The major difference between depth and disparity is that depth values are independent from the camera geometry and resolution, while disparity values depend on these parameters.

3. STATISITCAL ANALYSIS

The statistical characteristics of the video and depth signal differ significantly, which is particularly important when designing efficient coding algorithms for these signals. Hence, the correlation histogram (CH) method, introduced in the following section, will be used for a statistical analysis, in order to compare the characteristics of depth with video. The analysis will be carried out for uncompressed MVD data, evaluating the systematic differences, as well as for compressed MVD data, evaluating the impact of coding on the statistics. For the latter, the MVC extension of the H.264/AVC video coding standard is used, as briefly described in section 3.2.

3.1. Correlation Histograms

Image histograms are well-known as analysis and processing (e.g. color and contrast adjustment) tools for digital pictures [5]. In such a histogram a bin $H(k)$ contains the number of samples $p(u, v)$ with color k.

In the context of depth-enhanced 3D video, the relationship between corresponding samples in the video and depth signal (to which we refer as sample pairs) is analyzed. For this, the concept of image histograms is extended to so-called 2D or correlation histograms that use an array of bins $H(k, l)$. Here, each bin represents the number of specific sample pair combinations with values k and l (e.g. if $H(34,217) = 10$, an image or sequence contains 10 sample pairs with values $k = 34$ and $l = 217$, respectively). The statistical analysis covers the video luminance and the depth component; both are being represented with 8 bit per sample, which leads to an array of 256×256 bins. As illustrated in Fig. 2 (a), two types of CHs, namely temporal and spatial, are evaluated for the video and depth sequences of MVD data. After analyzing the complete sequence for temporal correlation, the bin $H_T(k, l)$ contains the number of sample pairs with

$$k = p(u, v, t),\ l = p(u, v, t-1).$$

For spatial correlation, the bin $H_S(k, l)$ contains the number of sample pairs with

$$k = p(u, v, t),\ l = \begin{cases} p(u-1, v, t) \\ p(u, v-1, t) \end{cases}.$$

By taking the value of horizontal as well as vertical neighboring samples into account, important features in both spatial directions are detected. For analyzing the characteristics of depth-enhanced 3D video, such CHs are superior to other methods (e.g. spectrum analysis), as they are able to well detect sharp edges and constant regions, which are both characteristical for depth maps.

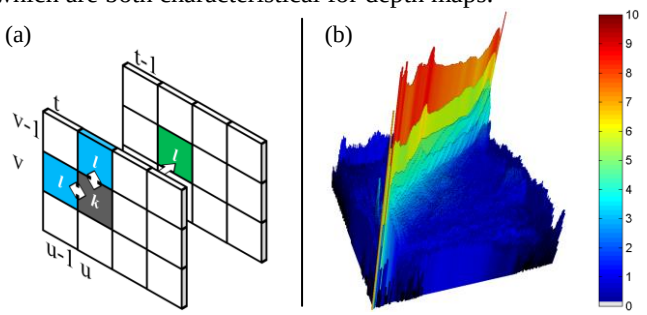

Fig. 2: (a) temporal (green) and spatial (blue) pixel pair relations for correlation analysis, (b) sample correlation histogram rendered as 3D columns and according color map (logarithmic scale).

3.2. Compression

In the multi-view video coding (MVC) extension [6], the multi-frame motion compensation method of H.264/AVC is extended in a way that a picture can have temporal as well as inter-view reference pictures for motion- and disparity-compensated prediction, respectively. Advantages of MVC are the improved coding efficiency and backward compatibility to H.264/AVC. For that, the first view is compressed using a profile conforming to H.264/AVC without multiview capability. The remaining views are coded using MVC and typically require 30-50% less bit rate than H.264/AVC simulcast coding at the same quality.

4. RESULTS

Five MVD data sets have been analyzed, with one synthetic and four natural sequences. The synthetic sequence (*Storyteller*) is very important for the analysis, as it contains synthetic "true" depth information. The depth for the natural sequences has been estimated from the color information with various methods, but the achieved depth values can only be regarded as an approximation of the "true" depth.

All presented results in Fig. 3 are normalized, in order to compensate for possible differences in the test data sets regarding resolution, number of frames, and number of

cameras. Furthermore the bin values are assigned to the colors with logarithmic scale for better differentiation, as illustrated in Fig. 2 (b). The analysis results in Fig. 3 show a different type of CH in each row (a)-(f) and a different MVD data set in each column. Rows (a)-(c) contain the CHs for the video component: temporal and spatial CHs for original video data in Fig. 3 (a) and (b), respectively, and spatial CHs for compressed video data in Fig. 3 (c). In contrast, rows (d)-(f) contain the CHs for the depth component, following the same order as for video: temporal and spatial CHs for original depth in Fig. 3 (d) and (e), respectively, and spatial CHs for compressed depth in Fig. 3 (f). Note, that the data for Fig. 3 (c) and (f) was coded with MVC at a very low bit rate.

One important feature of CHs is, that bins near the diagonal represent flat and bins far from the diagonal steep pixel transitions. Regarding the video CHs, the expected characteristic of a compact distribution around the diagonal can be observed. In contrast to that the depth results reveal, that the depth values of the individual data sets have been achieved with different methods and formats. Especially for *Newspaper* only a limited number of the available 256 depth values is used, while for *Horse* only a limited range of depth values occurs.

Comparing the spatial CHs for original video in Fig. 3 (b) with the according depth results in Fig. 3 (e) turns out, that they are significantly different: depth histograms are much more frayed with isolated areas, representing depth edges between foreground and background objects. Regarding the temporal CHs for original video in Fig. 3 (a) and original depth in Fig. 3 (d), basically the same characteristic differences as for spatial correlation can be observed. However, one important issue is highlighted by the temporal CHs for depth: the histogram of the synthetic sequence is significantly different from those of the natural sequences, although all of them have a static background. The reason is that temporal consistency is a problem for most depth estimation algorithms. The histograms reflect this by being less concentrated to the diagonal and by being frayed in less discrete patches or isolated areas.

The effect of coding on the video and depth characteristics is analyzed by comparing the spatial CHs in Fig. 3 (c) and (f) with the according results for original data. As expected, no significant difference can be observed for video, only a higher concentration to the diagonal and a reduction of high frequencies. These effects can also be observed for the depth coding results, but MVC has a much stronger impact here, as the original depth maps have a highly non-continuous characteristic: the reduction of high frequencies results in continuously shaped histograms with significantly different characteristics than those of uncompressed depth data.

5. SUMMARY AND CONCLUSION

This paper presented the correlation histogram method for analyzing depth-enhanced 3D video representations. Starting from an evaluation of the CHs for uncompressed MVD data, it has been shown that the video and depth component have significantly different characteristics, reflecting that video represents the color texture and depth the 3D geometry of the scene. Thus, the spatial CHs for color have a continuous region of values around the diagonal, while the histograms for depth have a high concentration on the diagonal, caused by constant depth within objects and additional frayed and discrete regions, resulting from sharp edges between objects. The temporal CHs identified, that depth estimated from natural video data lacks of temporal consistency. This also affects the compression efficiency of predictive coding algorithms. The analysis of the CHs for compressed video and depth evidently showed, that H.264/AVC-based video coding is well-applicable to natural video, but not to depth. The most important characteristic of depth maps, namely sharp edges between objects, is not preserved. Thus, the major requirement for depth coding algorithms is compression without changing these characteristics significantly. This can be visualized well by CHs.

In conclusion, correlation histograms are a powerful method for analyzing the characteristics of depth-enhanced 3D video. They are especially well-suited for evaluating the impact of different processing and coding algorithms.

ACKNOWLEDGMENT

This work was supported in part by EC within FP7 under Grant No. 213349 (with the acronym 3DPHONE).

We would like to thank the Interactive Visual Media Group of Microsoft Research (USA) for providing the *Ballet* and *Breakdancers* data sets, the KUK Filmproduktion GmbH (Germany) for providing the *Horse* data set, the ETRI (Korea) for providing the *Newspaper* data set, and Momentum DMT (Turkey) for providing the *Storyteller* data set.

REFERENCES

[1] P. Merkle, K. Müller, and T. Wiegand, "3D Video: Acquisition, Coding, and Display," *Proc. IEEE Intermational Conference on Consumer Electronics (ICCE '10)*, Las Vegas, USA, January 2010.

[2] A. Smolic, K. Müller, P. Merkle, P. Kauff, and T. Wiegand, "An Overview of Available and Emerging 3D Video Formats", *Proc. Picture Coding Symp. (PCS'09)*, Chicago, USA, May 2009.

[3] A. Vetro, S. Yea, and A. Smolic, "Towards a 3D Video Format for Auto-Stereoscopic Displays," *Proc. SPIE Conf. Applications of Digital Image Processing XXXI*, Vol. 7073, September 2008.

[4] R. Szeliski and P. Golland, "Stereo matching with transparency and matting," *Int. Journal on Computer Vision*, vol. 32(1), pp. 45–61, 1999.

[5] R. Gonzalez and R. Woods, *Digital Image Processing*, Prentice Hall, 2nd Ed., p. 88-108, 2002.

[6] Y. Chen, Y.-K. Wang, K. Ugur, M. Hannuksela, J. Lainema, and M. Gabbouj, "The Emerging MVC Standard for 3D Video Services", EURASIP Journal on Advances in Signal Processing, Vol. 2009, No. 1, January 2009.

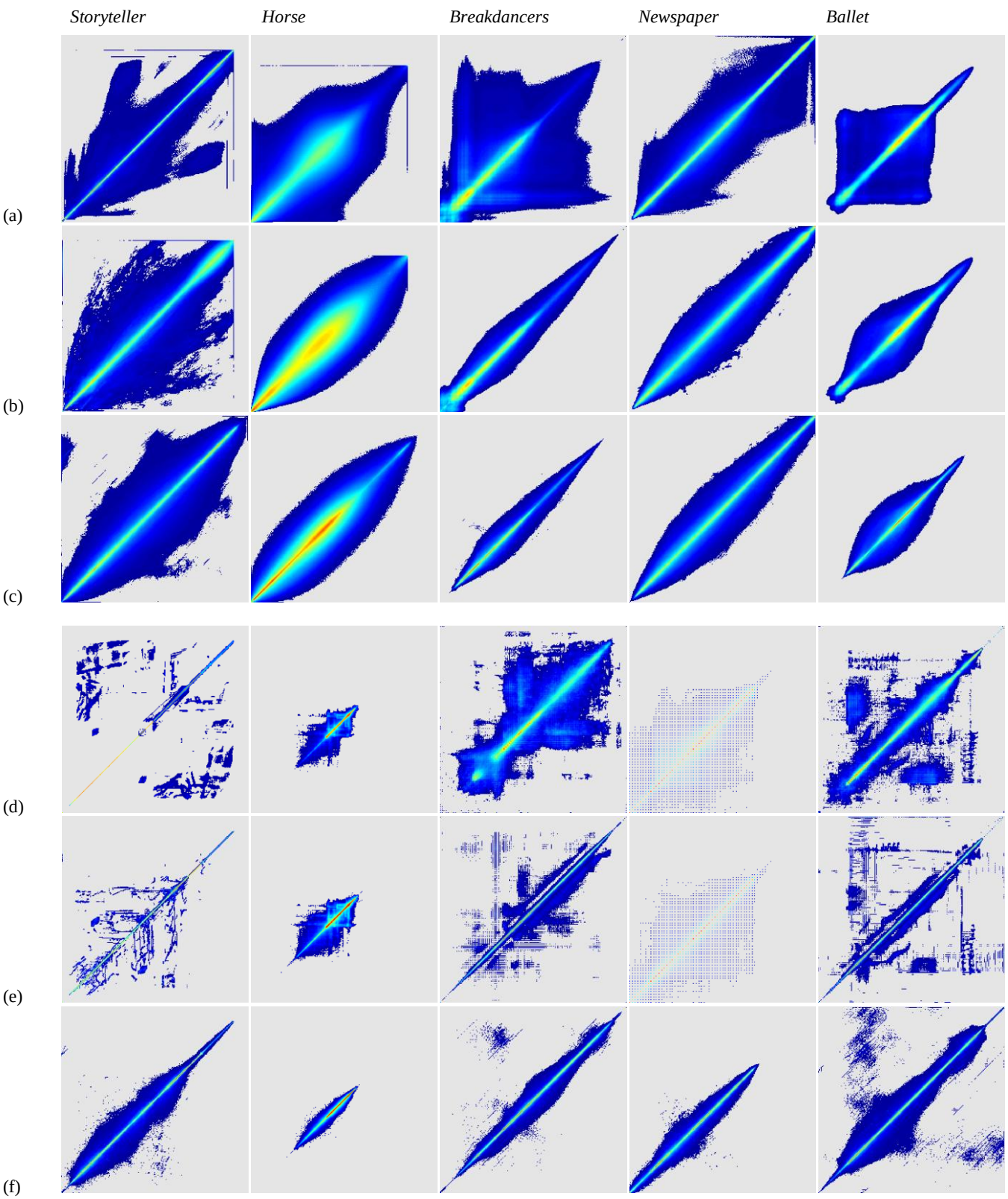

Fig. 3: Equalized, logarithmic correlation histogram results for five MVD sequences, with colors from blue for very low to red for very high values: (a)-(c) for video (luminance component) and (d)-(f) for depth; (a) and (d) for temporal correlation from original data; (b) and (e) for spatial correlation from original data; (c) and (f) for spatial correlation from compressed data at a low bit rate.