3D VIDEO OBJECTS AT SCALABLE LEVELS OF QUALITY

Christian Weigel, Marco Rittermann

Institute of Media Technology, Technische Universität Ilmenau
Postfach 10 05 65, 98684, Ilmenau, Germany
phone: + (49) 3677-69 2757, fax: + (49) 3677-69 1255, email: christian.weigel@tu-ilmenau.de
web: www.tu-ilmenau.de/mt

ABSTRACT

In this paper we present an approach for the generation and coding of 3D video objects where the quality is scalable in a definable manner. At first a production chain for the generation and display of 3D video objects based on image based rendering (IBR) methods is described. Starting with this specific generation chain, issues of applying a scalable coding framework for 3D video objects are discussed. By developing a common model of generation a theoretical approach is introduced and basic experiments are presented. For the comparison and the validation of the proposed methodology a quality metric (3DVQM) is utilized and explained further.

1. INTRODUCTION AND MOTIVATION

Most methods for the generation of natural free viewpoint video rely on image-based rendering (IBR) techniques using either no geometry, implicit geometry, or explicit geometry. An excellent survey of these techniques and first approaches for the coding of different representations can be found in [1]. 3D video objects are a subgroup of these techniques. We refer to them as the "observation of a time variable three dimensional object with free choice of the viewpoint". In this definition the emphasis lies on the term "object", i. e. representations of whole scenes like panoramic views are not within the scope of the definition. In our approach a method with implicit geometry usage is employed. The generation of virtual views of the 3D video object relies on a three step morphing algorithm as proposed in [2].

Regardless of the generation method, the amount of data required to generate seamless virtual views is immense compared to conventional two dimensional video. Therefore, efficient coding of these data is essential. With respect to the versatile kinds of transmission channels and differences in the performance of various playback devices, the coding method which is used must be scalable. The new type of representation offers new methods but also introduces new problems for scalable coding which will be discussed in the next sections.

2. A GENERATION METHOD FOR 3D VIDEO OBJECTS

The basic principle of the generation process of 3D video objects as developed by us is depicted in Fig. 1. In our system, the object of interest is recorded in front of a blue screen by a multi-camera setup. Currently we are using an uncalibrated, horizontal, and convergent setup with six cameras. The angle between the cameras is typically chosen to be 15 degrees. To obtain the mere object information, the next step is a segmentation process which is done by chroma keying. In order to get the geometric relations among each pair of cameras, pixel correspondences need to be assessed. For this purpose, tools have been developed for both manual and automatic search of correspondences [3]. While the captured image material is used for the manual search, an additional calibration sequence that contains a checker board is required for the automatic search. Once the correspondences are found, the fundamental matrices for each pair of cameras are estimated using the RANSAC algorithm [4]. According to these matrices scanlines for each pair of images are calculated. Then, the input material is rectified to horizontally align the scanlines. From these images disparity values are obtained by subdividing each scanline into runs. Subsequently, a perline block-matching search for correspondences is performed. The disparity values are error corrected by removing vertical portions that were introduced by the previously applied rectification process. Finally, outliers are detected and eliminated by comparing the disparity values with the values in adjacent scanlines. After this preprocessing stage the view synthesis algorithm can be initiated. To obtain a high quality virtual view, several optimization steps are applied [5].

3. SCALABLE CODING ISSUES

3.1 Requirements

The generation process as described in the previous section is not designed for the coding of 3D video objects. It rather represents a basis for a common approach of a scalable coding framework. Furthermore it is very useful for measurements regarding the quality of 3D video objects since it allows for the selective manipulation of a number of parameters within the complete generation chain.

The importance for coding of 3D video objects, regardless which generation method is used, is obviously. The amount of data that is required to generate arbitrary views is immense, e. g. for a viewing area of a 75 degrees sector of a circle with only horizontal movement, six times more data in comparison to 2D video is required[1]. In comparison to conventional 2D video coding new approaches for coding are needed which is, for instance, shown by recently conducted standardization activities [6]. Due to the large variety of possible generation methods a common model for scalable coding would be useful in order to develop new methods.

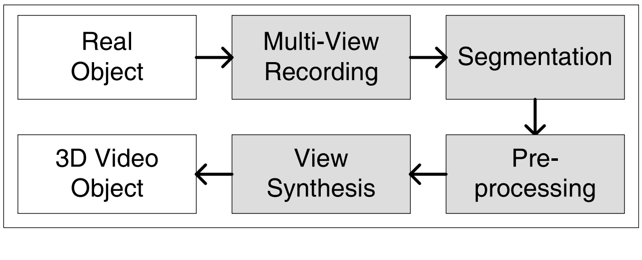

Figure 1: Example of a generation chain for 3D video objects

[1] The value was subjectively evaluated. Six cameras in a 15 degree setup were used.

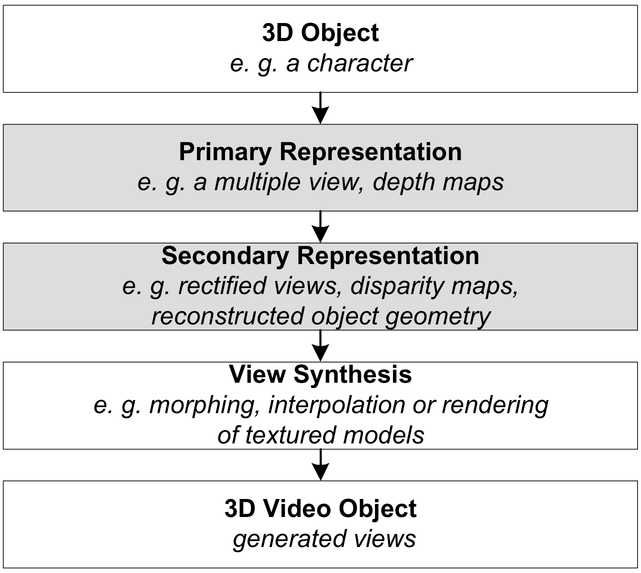
3D Object
e. g. a character
Primary Representation
e. g. a multiple view, depth maps
Secondary Representation
e. g. rectified views, disparity maps, reconstructed object geometry
View Synthesis
e. g. morphing, interpolation or rendering of textured models
3D Video Object
generated views

Figure 2: Model of the acquisition of 3D video objects

3.2 Developing a Common Model

A first model for the acquisition of 3D video objects was developed in [11]. The necessity for this model was given by the development of an objective quality metric for 3D video objects. It was developed in order to constitute a common model for all types of 3D video objects. The basic elements of the model together with some examples are depicted in Fig. 2. The most important issue in this model is the introduction of terms for specific representations. The *primary representation* contains all recordable information, i. e. everything that is captured. The *secondary representation* comprises these information and/or information extracted from it after a subsequent processing stage.

Since the model for acquisition does not take into account any type of coding, it needs to be extended. The performance of a 2D video coding method and thus the yielded coding gain is determined by the comparison of the quality of representations prior to and past the coding/decoding step[2]. In this paper the decoded representation is termed as *target representation*. The terminology used for the coding of 2D video is illustrated in Fig. 3(a). Applying this model to the model of acquisition of 3D video objects leads to a model of the coding of 3D video objects. This model is illustrated in solid lines in Fig. 3(b). In addition to the model of 2D video coding two more stages are introduced. These are a preprocessing step and the view synthesis. This apparently simple extension leads to a number of questions regarding the coding and scalability of 3D video objects.

3.3 A Theoretical Approach

The introduction of the secondary representation as well as the two more stages of processing raise a number of questions. These questions and their answers establish a theoretical basis for the development of a common scalable coding framework for 3D video objects. The questions can be separated into two basic groups: questions regarding the coding itself (1-4) and questions regarding the scalable coding (5-7):

1. What are the requirements for the coding of the new representations?

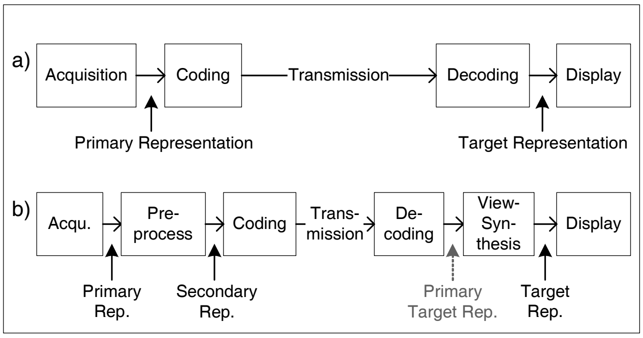

Figure 3: Model of coding for (a) 2D videos and (b) 3D video objects

2. Which kinds of coding are adequate for the different kinds of representations?
3. How can the quality of the coding be evaluated by objective metrics?
4. Does the coding influence the quality of the view synthesis?
5. How are the minimum and complete primary represenation defined, i. e. how can the base layer be determined and what is the "maximum" basis for scalable coding?
6. Which parameters are subjects of scalability?
7. If the answer to question four is "yes" – How can these relations be incorporated into a scalability methodology?

Several approaches have already been made to answer question one. The requirements are mostly the same as they are for 2D video coding, although the complexity and thus the computational demands are higher. An exemplary overview can be found in [7]. Regarding question two, there are also first approaches [8]-[10]. Due to the fact that each of these approaches is designed for a specific kind of secondary representation[3] they raise the question for a more common approach. Comparing the techniques, it can be observed that all primary representations are basically of the same type. Therefore the coding of this representation with a common approach would be useful. On the other hand, the efficiency of coding raises with the transformation to the secondary representation. This tradeoff between universality vs. efficiency is essential for 3D video object coding and must be incorporated into a common scalability methodology.

The answer to question three is an important part for the design of the whole system. Due to the introduction of the view synthesis step, the primary and the target representation are not comparable the way they are in 2D video coding. There is almost an infinite number of possible target representations at any instance of time. Therefore traditional comparison methods are not applicable due to missing pixel references. A solution is the introduction of an additional representation, the *primary target representation.* Depending on the method of coding this representation is comparable to the secondary representation with common 2D video methods. Anyway, for new types of (coded) side information, new kinds of quality assessment and comparison are required, depending on the type of side information. The crucial drawback of the introduction of the primary target representation for coding quality assessment is revealed by the answer to question four.

If there is a connection between coding and the quality obtained by the view synthesis, the approach outlined in the previous paragraph fails. Then, the comparison of the secondary representation and the primary target representation

[2] Usually an objective metric as PSNR or specific features like block artefacts are used for comparison.

[3] E.g. Rayspace, model-based with explicit geometry, LDI

does not yield the final result. A quality assessment of the target representation must be performed. Therefore, a new objective quality metric is required. Such a metric was introduced in [11] and is explained in Section 4. The method was developed, among others, using 3D video objects generated by the production chain introduced in Section 2. First experiments with this specific generation method have shown that there is a connection between coding and view synthesis. Although the 3DVQ does not allow for comparison with the secondary representation, it represents a first, useful tool for the development of 3D video object coding algorithms.

The fifth question is an important question for coding and in particular for scalability. In 2D video coding the base layer is usually created by reduction of the temporal, spatial, or SNR quality. 3D video objects offer more possibilities. Here, the quality can be additionally determined by the degree of freedom, i. e. the number and distribution of possible viewpoints. The number of samples required for a complete degree of freedom depends on the method of view synthesis and is subject of current research activities [1]. Our experiments with the generation methods described in Section 2 have shown that along with the number of samples there are a lot of additional parameters that influence the quality of the final 3D video object. We have investigated the quality with different angles between the cameras. Subjective assessments yield an optimal angle of 15 degrees in a horizontal camera setup.

The statements of the previous paragraphs yield the answer to question six. The following parameters are suited for scalable coding of 3D video objects:

- The “classical” parameters, i. e. time and spatial resolution, quantization.
- The number of samples to be coded. This can be interpreted as a new type of spatial resolution.
- Parameters that arise from inter-sample coding.
- Parameters that are used to code side information.

Along with these parameters, the answer to the last question is of particular interest. If there is a connection between coding and view synthesis, the view synthesis is a subject of scalability because then the target representation is the representation to be evaluated. From this point of view even more parameters for scalability are usable in future coding techniques.

4. MEASURING THE QUALITY OF 3D VIDEO OBJECTS

4.1 Objectives of Measurement

As mentioned in the previous section, the quality of the generated 3D video objects at the target representation has to be determined. Due to the expenditure of subjective assessment an objective assessment is necessary. 3D video objects have to be compared and their subjective quality has to be predicted. In order to adapt algorithms of scalability it is more useful to measure certain quality features instead of an overall quality. Therefore effects on quality caused by the algorithms can be determined more exactly.

4.2 Assessment of 3D Video Objects

In [11] and [12] a methodology for the assessment of 3D video objects irrespective of their generation was shown. Using another (*ideal*) object as ground truth a 3D video object quality metric 3DVQM based on regression to quality features (distortions of shape, local distortions caused by e. g. occlusions, static or dynamic deviations of perspective etc.) has been modelled. Because of the regression-based method this 3DVQM can be adopted to a certain kind of 3D video object and a certain quality feature. This has to be done by extended subjective tests (based on methods according to [13]). Once the 3DVQM is adapted it can be used to fix criteria for the different layers.

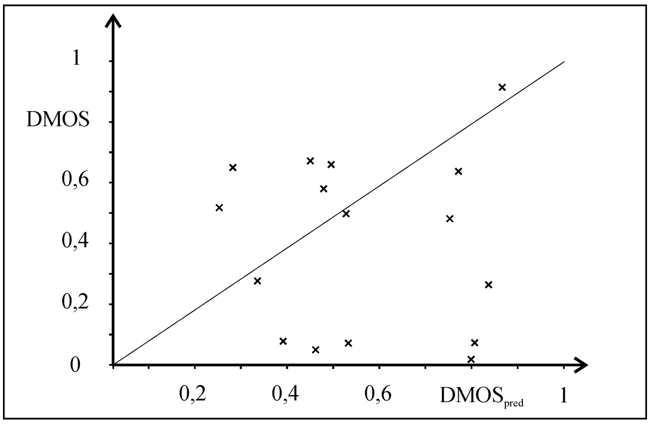

Figure 4: Scatter diagram: DMOS and its prediction

4.3 Example of Quality Assessment

In the following example the gradation of quality of 3D video objects was determined. These objects were generated by the image-based method explained in Section 2. In Fig. 4 the correlation between the subjective assessment (differential mean opinion score *DMOS*) and its prediction by the 3DVQM is shown. In this example only one extended subjective test was necessary to achieve a Spearman correlation of 0.45 for the overall quality. Actually, the Pearson correlation of some single quality features is more than 0.8.

5. BASIC EXPERIMENTS

The theoretical considerations and the availability of the 3DVQM reveal a number of possibilities for practical work. We decided to investigate the relation between coding and view synthesis first. The aim was to find relationships between the primary target representation and the 3DVQM of the primary representation. We employed coding techniques well known from the two dimensional video domain. Since the measurements aimed at basic coherences, we further reduced the complexity by using simulcast coding, e. g. every sequence of the primary representation is coded separately. We used sequences of a 15 degree camera setup as input (4:2:0, 768×576, 4 s). The first coder used was the H.264 reference software at Main Profile, the second the MPEG-4 Part 2 reference coder using the greyscale shape mode. After decoding, a view synthesis as described in Section 2 was carried out for two different paths. The quality of the results was then estimated with the 3DVQM (see Section 4).

The results given in Table 1 and 2 show the relation between the PSNR and the total bitrate of all input streams vs. 3DVQM for some examples[4]. It is obvious that the quality of the 3D video object increases with the bitrate of the coded primary representation[5]. We found two reasons for this. Firstly, some attributes and methods that have been used to calculate the 3DVQM are also applied in 2D video quality metrics, e. g. the standard deviation of the spatial information. Thus, the distortions of the texture of the primary target representation are propagated during the view synthesis and are incorporated in the 3DVQM. Secondly, and more important, the coding artefacts cause errors in the process of view snythesis and the reconstruction of the

[4] The bitrate of H.264 coding is specified without the alpha channel because this channel was not coded but used directly.

[5] The lower the values of the 3DVQM, the higher the quality.

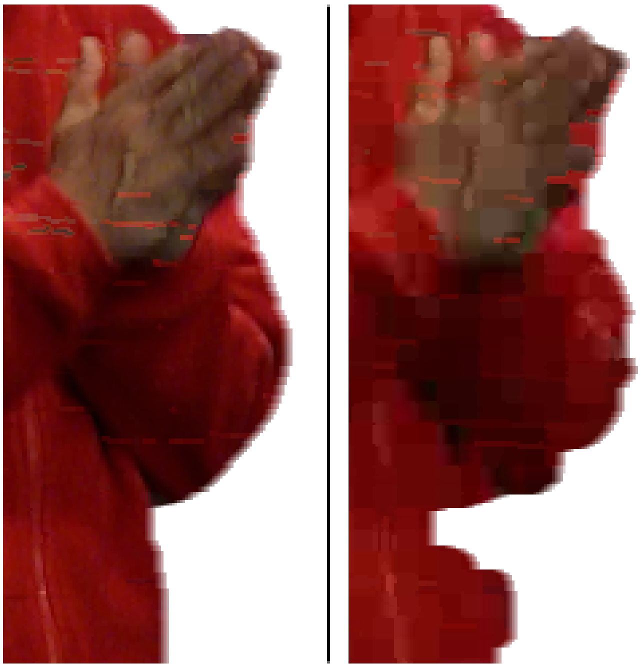

Figure 5: Distortion of image and silhouette caused by coding artefacts (Left: reference 3DVO, Right: 3DVO reconstructed from Seq. 1f)

virtual view is distorted. An example for this are block artefacts causing wrong point correspondences and thus wrong disparity values which results in more distortion than the block artefacts themselves. In Fig. 5 an example of lost correspondences and wrong silhouette reconstruction is given. Such distortions are of high relevance since they are very well perceptible. To apply scalable coding in a controllable manner the challenge is to parameterize the observed effects objectively.

6. CONCLUSION AND FUTURE WORK

The theoretical fundamentals presented in this paper constitute a common approach for the scalable coding of 3D video objects. Together with the development of a common quality metric a basis for practical work is given. First experiments with 3D video objects generated by the method described in this paper revealed relations between coding and view synthesis. It was demonstrated that the quality of the primary target representation influences the quality of the view synthesis in different ways.

In future work the relation between coding and view synthesis needs to be investigated further. A mathematical description is required to create a model that can be incorporated in a scalability framework. Furthermore the accuracy of the 3DVQM must be increased. The multiview coding needs to be enhanced in comparison to simulcast coding with respect to the new requirements for 3D video objects. This will be done firstly based on a specific kind of view synthesis method and will become more common in future research.

7. ACKNOWLEDGMENTS

This work is supported by EC within FP6 under Grant 511568 with the acronym 3DTV.

	Seq. 1a	**Seq. 1b**	**Seq. 1c**
total bitrate (kbit/s)	12998.84	1747.11	466,01
average PSNR (dB)	48.53	43.32	37.59
3DVQM	**4,37**	**4,74**	**4,89**

Table 1: Comparison of the 3DVQM values for an input sequence at different PSNR and total bitrates in the primary target representation using the H.264 coder.(The lower the 3DVQM the better the quality.)

	Seq. 1d	**Seq. 1e**	**Seq. 1f**
total bitrate (kbit/s)	21648,52	3073,25	1651,47
average PSNR (dB)	42.90	36.61	33.97
3DVQM	**4,59**	**4,65**	**4,70**

Table 2: Comparison of the 3DVQM values for an input sequence at different PSNR and total bitrates in the primary target representation using MPEG-4 Part 2 shaped video coder.(The lower the 3DVQM the better the quality.)

REFERENCES

[1] H.-Y. Shum, S. B. Kang and S.-C. Chan, "Survey of image-based representations and compression techniques," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, pp. 1020–1037, Nov. 2003.

[2] S. M. Seitz, "Image-Based Transformation of Viewpoint and Scene Appearance," *PhD thesis*, pp. 24-26, USA, 1997.

[3] G. Horna and L. L. Kreibich, "3D-Videoobjektgenerierung mittels Multiview-Aufnahmen," Student research project, TU Ilmenau, Germany, 2004

[4] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Comm. of the ACM*, vol. 24, pp. 381–395, 1981.

[5] J. v. d. Haar, "Erweiterte Blickpunktwahl für die View-Synthese von 3D-Videoobjekten," Diploma Thesis, TU Ilmenau, Germany, 2004

[6] "Call for Evidence on Multi-View Video Coding," document N6720 MPEG Meeting, Palma de Mallorca, Spain, Oct., 2004.

[7] "Requirements on Multi-View Video Coding v.2," document N6834 MPEG Meeting, Palma de Mallorca, Spain Oct., 2004.

[8] H. Kimata, M. Kitahara, K. Kamikura and Y. Yashima, "Multi-View Video Coding Using Reference Picture Selection For Freeviewpoint Video Communication," *Proc. PCS2004*, San Francisco, USA, December, 2004.

[9] K. Müller, A. Smolic, P. Merkle, M. Kautzner, and T. Wiegand, "Coding of 3D Meshes and Video Textures for 3D Video Objects," *Proc. PCS2004*, San Francisco, USA, December, 2004.

[10] Y.-S. Ho, S.-U. Yoon and S.-Y. Kim, "Framework for Multi-view Video Coding using Layered Depth Image," ISO/IEC JTC1/SC29/WG11(MPEG), M11582, January, 2005.

[11] M. Rittermann, "Quality Assessment of 3D Video Objects", in *Proc. of ISCE'03*, Sydney, Australia, 2003.

[12] M. Rittermann, "A Proposal for the Quality Assessment of 3D Video Objects", in *Proceedings of the 5th WIAMIS*, Lisbon, Portugal, 2004.

[13] International Telecommunication Union (ITU), "ITU-T Recommendation P.910 – Subjective Video Quality Assessment Methods for Multimedia Applications", Recommendation, 1999.