Efficient Transmission of 3D Video Using MPEG-4 AVC/H.264 Compression Technology

Marek Domański, Krzysztof Klimaszewski,
Olgierd Stankiewicz, Jakub Stankowski, Krzysztof Wegner,

Chair of Multimedia Telecommunications and Microelectronics,
Poznań University of Technology,
Polanka 3, 61131 Poznań, Poland
{domanski, kklima, ostankiewicz, jstankowski, kwegner} @ et.put.poznan.pl,

Abstract. At a receiver terminal, true 3D video provides ability to watch views selected from a large number of available views. Such ability is needed for the forthcoming 3D video applications like free-view television, autostereoscopic displays etc. Delivery of many views through communication channels is a challenging problem that has to be solved in the near future. In this paper, we study delivery of real 3D video using the state-of-the-art compression technology. Considered are the issues related to 3D video data model as well its application for generation of video from arbitrary virtual viewpoint. Included are respective experimental results.

Keywords: 3D video, transmission, AVC, depth map, video compression.

1 Introduction

Recently, 3D video has gained a lot of attention both in research and in industry [1]. For 3D video, among open and challenging problems, there is the problem of transmission of 3D video over communications networks. In this paper, we consider transmission and compression of 3D video.

Currently very many possible application scenarios are considered, therefore even the name ‘3D video” is understood in various ways. Some popular understanding of this name refers even to stereoscopic video. For the state-of-the-art video compression standard MPEG-4 AVC/H.264 [2], there already exists Stereo High Profile that provides efficient technology for compression of stereoscopic video [3]. Therefore, we consider 3D video services that require simultaneous pictures from several viewpoints. For example, currently under promising development there are autostereoscopic displays that provide glassless stereoscopic perception. Another promising 3D video application is Free-viewpoint Television (FTV) that provides an ability for a viewer to freely navigate through a 3D scene.

In the above mentioned applications, the receiver has an ability to produce pictures that correspond to many viewpoints. The number of viewpoints may vary but already for autosterescopic displays, the reasonable number will probably exceed 30 in the next future. Obviously simulcast transmission of 30 or more video streams would not be practical. Therefore, developed are compression technologies that exploit mutual redundancy that exists between videos from the neighboring viewpoints. The overview of current research results may be found in [4-6]. Among available techniques, Multiview Video Coding (MVC) has been already standardized as a part of MPEG-4/AVC [2]. This technique outperforms simulcast MPEG-4 AVC/H.264 coding of several viewpoint-video sequences by 20-30% [6].

Here, we are going to report the results obtained in Chair of Multimedia Telecommunications and Microelectronics, Poznań University of Technology, Poznań, Poland. The goal of the research was to find methodology for usage of MPEG-4 AVC/H.264 Multiview Video Coding (MVC) for efficient 3D video compression.

2 Scenarios for 3D video transmission

As already discussed, in most 3D video applications, the crucial issue is receiver ability to generate video related to many possible viewpoints. Therefore, it is enough to transmit some views only. The others may be generated from the very limited set of the delivered views. It is expected that already 3 views are enough to synthesize the other views for a stereoscopic display. For such virtual-view generation, some results of 3D scene analysis are need. Usually this 3D scene analysis is done by stereoscopic depth estimation that needs substantial computational efforts. Depth estimation may be done either in the receiver or in the transmitter (Fig.1). Because of complexity of depth estimation, the latter scenario seems to be more realistic. Therefore, let assume that depth is estimated at the transmitter side of the system, and depth maps are transmitted together with respective video sequences (see Fig. 1, lower scheme). In the receiver, both video and depth information are used to synthesize views that are needed but not transmitted. There exists a question how to use MVC for depth map transmission. This question will be answered further in this paper.

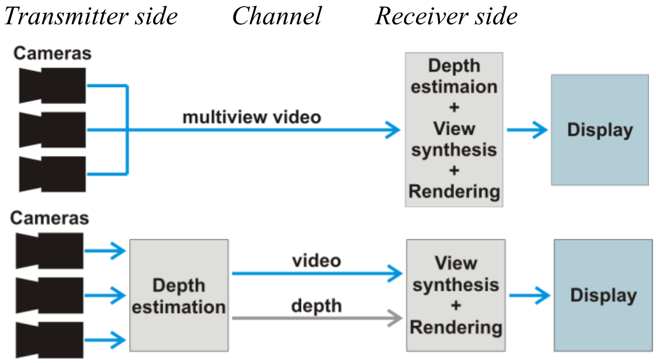

Fig. 1. Two scenarios of 3D video transmission.

3 Video acquisition and correction

3D video is acquired by several carefully synchronized cameras. Among various possible geometrical alignments of cameras, we consider cameras mounted with parallel optical axes. Such setup is appropriate for autostereoscopic displays.

Unfortunately, image sensors and optical systems in the cameras, even from the same production batch, can differ in orientation and position significantly. It is also very difficult to ensure exact camera body alignment. Therefore, cameras exhibit different intrinsic and extrinsic parameters and acquired images have to be rectified in a way that simulates the ideal camera positioning. The purpose of rectification is to produce artificial views that would be captured by hypothetical cameras with parallel camera axes, identical intrinsic parameters, and camera centers positioned along a straight line with all the image horizontal borders being parallel to the line of camera centers.

For a pair of cameras, among many rectification techniques that from [6] is widely used. Recently, this technique has been generalized for an arbitrary number of camera [7]. Here, we briefly summarize this technique for 3 cameras that are not ideally aligned (Fig. 2).

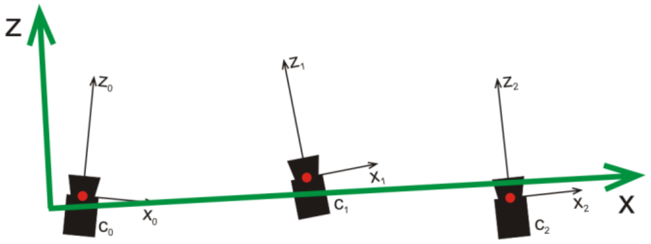

Fig. 2. Unideal alignment of 3 cameras.

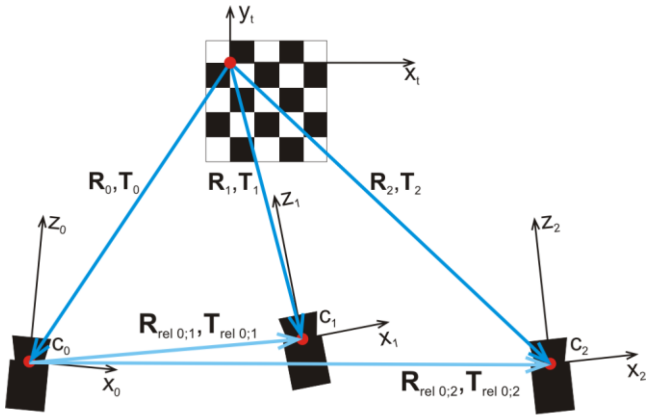

Fig. 3. Calculation of relative rotation ($\mathbf{R}_{rel}$) and relative translation ($\mathbf{T}_{rel}$).

Firstly, we calibrate each camera independently, using a high-contrast chessboard calibration pattern (Fig.3). This operation is performed in order to obtain accurate intrinsic matrix and distortion coefficients for each camera.

The goal of the next step is to estimate the rotation matrix **R** and translation vector **T** for each camera. These extrinsic parameters are calculated using already known cameras intrinsic parameters as well as data obtained from chessboard pattern images. After that, calculated are relative rotation ($\mathbf{R}_{rel}$) and relative translation ($\mathbf{T}_{rel}$) between camera 0 and the other cameras (Fig. 3).

In final step, all image data are transformed to a common coordinate system (Fig. 4).

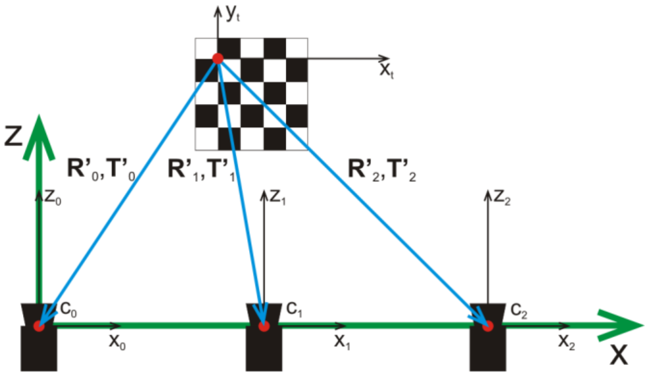

Fig. 4. Output of the final rectification step.

Moreover, for some camera systems, additional color correction may be necessary. The appropriate techniques are described in the references.

4 Experimental system

Research on 3D video compression yields usage of special experimental systems that are needed to acquire and display multiview video (Fig.5). Here, we briefly describe a system built in Chair of Multimedia Telecommunications and Microelectronics, Poznań University of Technology, Poznań, Poland [9]. The system includes also some computers for picture acquisition, correction, compression, decompression, synthesis and rendering. Instead of autostereoscopic display our system includes polarization stereoscopic display and projection system. In the system, 9 views may be acquired simultaneously. Some of these views are used only as reference in order to measure quality of the synthesized (virtual) views.

For this system, Canon XH-G1 HDTV (1920×1080) cameras (Fig. 6a) have been chosen because of their good ability to provide exact synchronization. Special controller has been built in order to control the camera system via LANC interface. Uncompressed video from individual cameras is acquired via HD SDI coaxial cables and DeckLink HD video grabbers into SSD memory in computers. The camera system is mounted on a special movable camera rig (Fig. 6b).

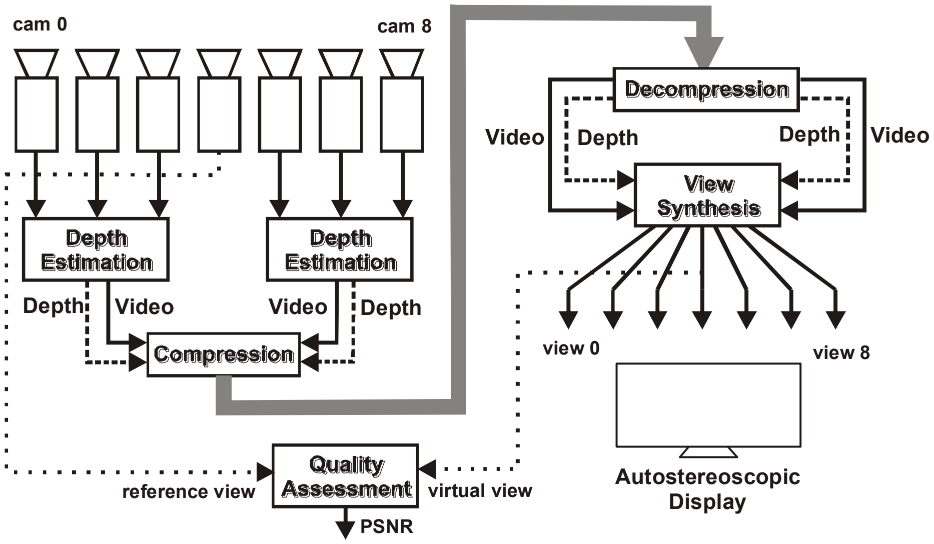

Fig. 5. Structure of the experimental system.

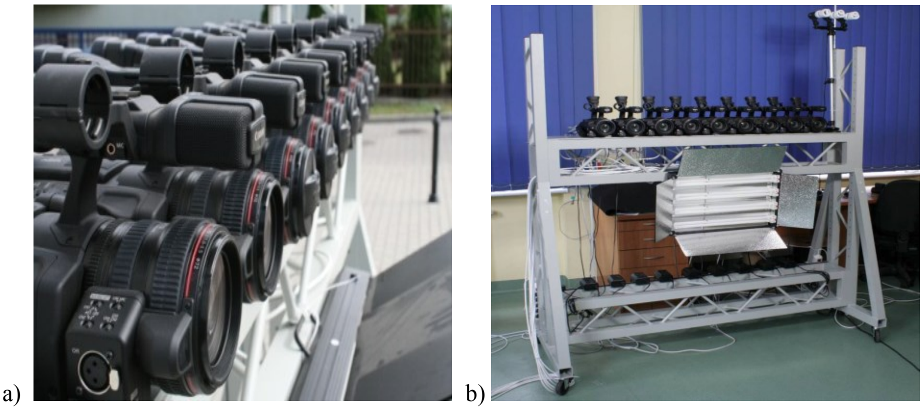

Fig. 6. The camera system (a) and the camera rig (b).

5 Depth estimation

As already mentioned in Section 2, considered are the systems with stereoscopic depth analysis on the transmitter side (Fig. 1, lower scheme). The output of this analysis is provided in the form of so called depth maps. A depth map is an image composed of distances from camera to points of the scene. Here, we consider systems where depth maps are calculated from input video, through estimation of disparities between neighboring views.

For 3D video transmission systems, we propose a depth map estimation algorithm (Fig. 7) that consists of: noise reduction technique, state-of-the-art disparity estimation (belief propagation - BP) and refinement (middle-level hypothesis - MLH) techniques and disparity to depth conversion.

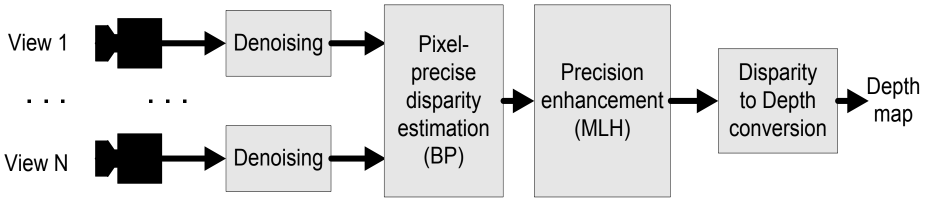
View 1
. . .
. . .
View N
Denoising
Denoising
Pixel-precise disparity estimation (BP)
Precision enhancement (MLH)
Disparity to Depth conversion
Depth map

Fig. 7. Scheme of depth estimation employed in our system.

Each view is independently denoised. This is attained by identification and linear filtration (in time domain) of still/background regions. Thanks to that, temporal consistency of 3D scene representation is assured. For all views, the denoised video is then fed into disparity estimation block.

Disparity is estimated by matching of the provided views. We have proposed to use Belief-Propagation (BP) algorithm [10] that is a smoothness and cost optimization technique. Smoothness term is related to continuity of the 3D model. Cost term is related to similarity between pixels in the matched images. Due to computational complexity reasons, BP algorithm is used only for full-pixel-precise estimation of disparity. Higher disparity precision is attained by use of a refinement technique [11]. Finally, disparities are converted to a resultant depth map.

Some experiments have been performed to assess quality of above mentioned depth estimation algorithm. Table 1 presents the results of quality evaluation done over four test sequences that are used by ISO/IEC MPEG group [12, 13]. The depth map quality was assessed by comparison of the original view from a real camera with the respective virtual view that was synthesized with use of uncompressed 3D video. Note quite moderate decrease of subjective quality due to depth map estimation and view synthesis. Nevertheless this quality deterioration of synthesized views is an important constrain that limits the final quality of decompressed and synthesized video. Therefore, further research is still needed in order to improve depth maps that are used in transmission of compressed 3D video [3-6].

Table 1. Subjective and objective quality of view synthesized with use of uncompressed 3D video stream versus the original view.

sequence	MOS		ΔMOS	PSNR [dB]
	original	synthesized		
Poznan_Steet	9.63	6.71	-2.92	35.39
Poznan_Carpark	9.11	6.24	-2.87	31.21
Book arrival	9.73	5.47	-4.26	36.23
Alt Moabit	8.62	6.11	-2.51	35.51

6 Video compression

In the most probable scenario, video will be transmitted from several cameras together with the respective depth maps, as already mentioned in Section 2 and Fig. 1 (lower scheme). Assume that we have to transmit video and the corresponding depth map for *N* real viewpoints. The state-of-the-art technique to compress multiview video is Multiview Video Coding (MVC) [3] that is a formal part of MPEG-4 AVC/H.264 video compression standard [2]. Therefore, we are going to consider this technology to compress video and the corresponding depth maps for *N* real viewpoints. An open issue is the best way of inclusion the depth information into visual data transmitted and compressed using MPEG-4 AVC / H.264 possibly with its extension MVC (i.e. High Multiview Profile).

Our analysis implies that there exist three basic schemes of joint compression of video and depth using MPEG-4 AVC / H.264 (Fig. 8).

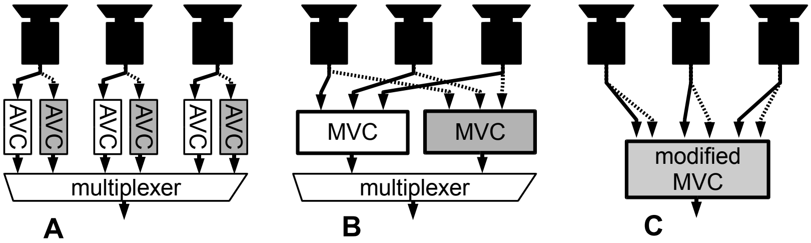

Fig. 8. Three different scenarios for multiview video and depth map compression. Solid lines correspond to video data, dotted lines correspond to depth map data.

Scheme A

In the simplest solution, all video sequences from real cameras are compressed separately by a video encoder, for example an AVC encoder. Also, each depth map is compressed separately (Fig. 8 A). There are several advantages of this scenario. The first is the possibility to implement encoders for all views in parallel. The second advantage is the usage of a well-known technology. In this scenario, the main disadvantage, however, is that the encoder cannot exploit similarities between neighboring views. Therefore, this variant of compression exhibits the lowest compression efficiency among the variants considered.

Scheme B

As the bandwidth consumption for multiview video transmission is large, even slight improvement in compression efficiency can result in a considerable decrease of required total bandwidth. As already mentioned, in order to decrease total bandwidth consumption, a new standard of multiview video compression was established, namely MPEG-4 AVC / H.264 Annex H Multiview Video Codec (VC) [2]. MVC exploits similarities between views and provides means to transmit additional data required for correct processing of video at the receiver side. Increase of coding efficiency can reach several tens percent as compared to separate coding of each view (simulcast) like in Fig. 8A.

The use of MVC to compress video and depth maps is presented in Figure 8 B. In the most straightforward version, there are two separate MVC encoders, one for compression of views and another one for depth maps. Both encoders exploit mutual similarities between views and depth maps. Application of two separate encoders, for multiview video and for multiview depth is still quite efficient as video and depth differ significantly by many aspects. Therefore, the separate encoders may well adapt to individual properties of video and depth.

As already discussed, only few views and the corresponding depth maps are transmitted. Usually they correspond to quite distant viewpoints. Therefore mutual correlation of video or depth is usually lower as compared to classic multiview video sequences.

Schemes A and B use compression techniques that are already standardized and may be adopted to 3D video.

Scheme C

The third approach (Fig. 8 C) makes use of modified (nonstandard) MPEG-4 MVC. Here, video and depth maps are encoded jointly, from all cameras. In his case, the syntax of MVC bitstream has to be modified slightly in order to embed depth information. In that way, the headers and control data may be shared by video and its depth data. Nevertheless, this approach is still under research and the efficient bitstream syntax is not defined.

For all the above mentioned three cases, there is a need to divide bitrate between video data and depth maps. There exists the optimal bitrate division ratio that results in the best synthesized video quality for a given total bitrate. Here, we are going to deal with this issue. Such an optimum division needs to be found in order to provide the best synthesized view quality. The issue of bitrate allocation between the views and the depth maps was already investigated in [16], but the results obtained correspond to local optimization of the depth quantization parameter. Here, a more global approach is considered.

In this paper, we consider the case of two separate MVC encoders employed for views and depth maps (Fig. 8B). The encoders are fully compliant with MPEG-4 MVC video compression standard. The bitrates produced by the two encoders are controlled by the respective quantization parameters: quantization index for view compression QP and quantization index for depth map compression QD.

In the experiments, for every possible pair of parameters from the eligible range from 10 to 51, we have performed the compression, the virtual view synthesis and the calculation of the virtual view quality in terms of PSNR in comparison to the real reference view (Fig. 5). JMVC reference software [17] and VSRS [14] view synthesis reference software were used for three standard test sequences (Table 2).

Table 2. Parameters of sequences used in the experiment

	Champagnetower	Bookarrival	Newspaper
camera views used	37, 41 39 as a reference	4, 6 5 as a reference	4, 6 5 as a reference
resolution	1280×960	1024×768	1024×768
frames per second	29.41	16.67	30

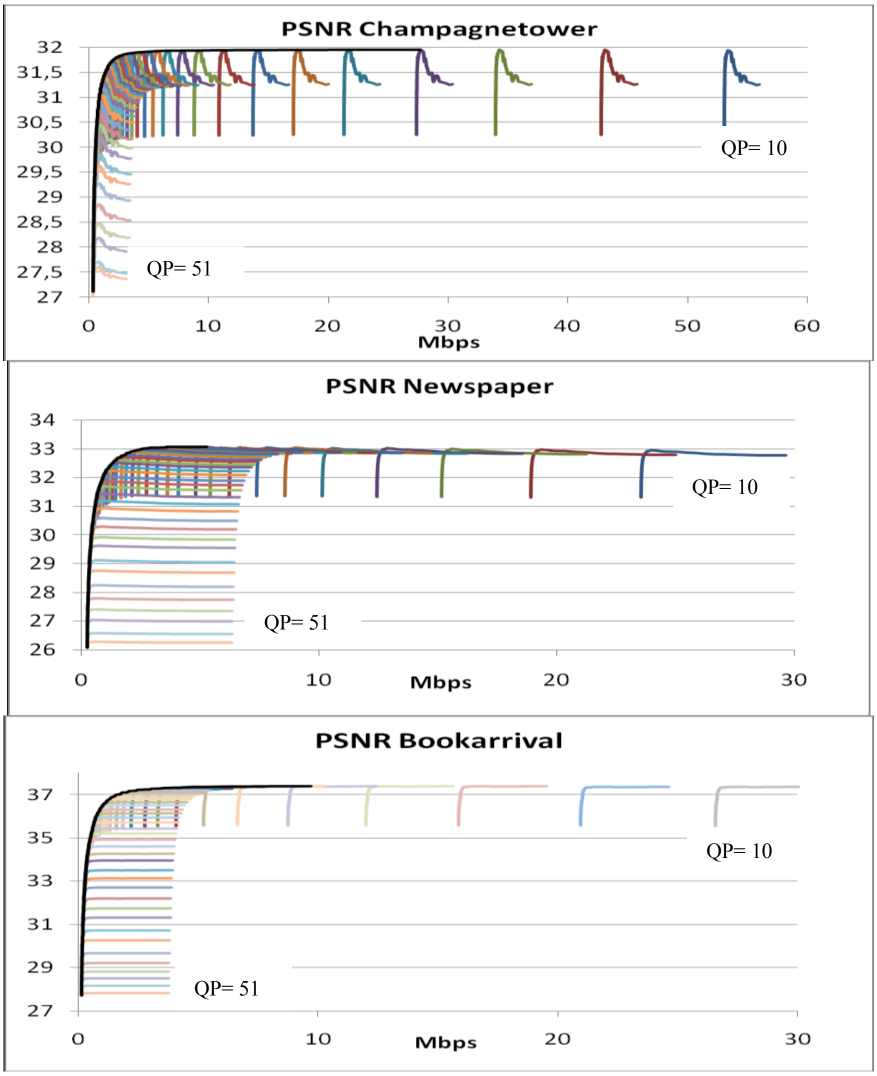

Fig. 9. Quality (dB of PSNR) of synthesized view for three different test sequences as a function of total bitrate for views and depth maps.

The experimental results are presented on the graphs in Figure 9. On the graphs, each line corresponds to results obtained for a single QP index and variable QD index. It is noticeable that there exists a line of maximal performance for a given total bitrate. On the graphs, this line is outlined by a solid black line. It preserves the similar shape for all three sequences.

For all three cases following observations hold:

- Several QP and QD pairs can give similar bitrate values but there is an optimal pair of QP and QD indexes that provides the highest quality of the synthesized view.

- An increase of bitrate above a certain value does not cause any increase of the synthesized view quality. For all three sequences an increase of total bitrate above 5 Mbps does not cause any significant improvement in the synthesized view quality.
- For sequences with lower quality depth maps, like Champagnetower or Newspaper, an increase of the bitrate may lead to a decrease of synthesized view quality, while there is no such effect for sequences with higher quality depth map (e.g. Bookarrival). This phenomenon is caused by errors in depth map that are removed in the process of compression with higher QD index. Decreasing QD increases the bitrate but also leads to preservation of depth artifacts.

For each sequence an optimal (in terms of synthesized view quality) path can be found in the space of QP and QD indices (Fig. 10). This optimal path corresponds to an increase or decrease of the total bitrate. It can be seen that quality of a synthesized view depends significantly on QP index value, while the QD index has much smaller influence on the synthesized view quality. The optimal relation between QD and QP may be very roughly approximated by a line with a slope of approximately 1.2.

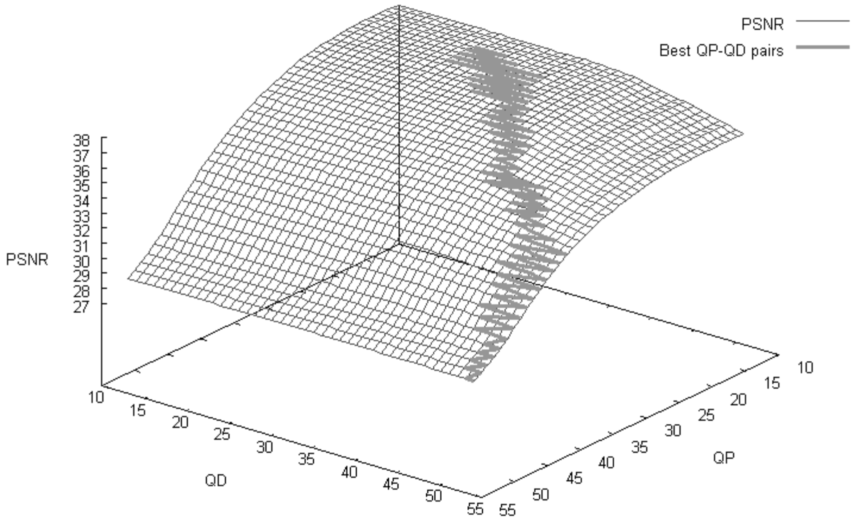

Fig. 10. Optimal path in the QP-QD space that preserves the optimal quality of synthesized view for a given bitrate.

7 View synthesis

In compression systems, two view synthesis techniques may be used. The first was implemented in Chair of Multimedia Telecommunications and Microelectronics, Poznań University of Technology [18]. Another technique is that currently used by

MPEG [14]. Both provide similar results in the sense of quality of synthesized pictures.

The first technique exploits two views and two depth map in order to overcome problems related to occlusions. This technique has its real-time implementation for low-resolution video.

In the experiments reported, the technique from [14] was used for the sake of reproducibility of the results.

8 Conclusions

The main conclusion is that, currently, the depth map quality is insufficient for compression tasks. Currently, the quality of the estimated depth maps is the main constrain for the quality of video reconstructed in the decoder. For wide range of depth-map bitrates, the estimated errors dominate over compression errors. These conclusions have been obtained for two state-of-the-art depth estimation techniques [10,15].

If the depth estimation problem would be solved, MPEG-4 MVC may be efficiently used for 3D video compression. Some minor modifications of the bitstream semantics and syntax may further improve compression performance. These modifications would allow to apply the compression scheme from Fig. 8C.

Other important issues are related to real-time implementations. In particular, the most severe problems are again related to the depth estimation. Also real-time video rectification would need some a priori preparations.

Nevertheless, the above consideration indicate that Multiview High Profile of MPEG-4 AVC / H.264 [2,3] is a good starting point to develop the compression standard for 3D video. For such a standard, backward compatibility with MPEG-4 AVC / H.264 would be a huge advantage.

Acknowledgment. This work was supported by the public funds as a research project.

References

1. Special Issue on Advances in 3-Dimensional Television and Video, Signal Processing: Image Communications, vol. 24, issues 1+2, pp. 1-133 (2009)
2. International Standard ISO/IEC 14496-10:2009, Information technology — Coding of Audio-Visual Objects, Part 10, Advanced Video Coding, 5th Ed. (2009)
3. ISO/IEC 14496-10: 2009/FDAM 1: 2009(E), Information technology — Coding of Audio-Visual Objects — Part 10: Advanced Video Coding, Amendment 1: Constrained Baseline Profile, Stereo High Profile and Frame Packing Arrangement SEI Message, ISO/IEC JTC1/SC29/WG11 (MPEG) Doc. 10701, London (2009)
4. Tech, G., Smolic, A., Brust, H., Merkle, P., Dix, K., Wang, Y., Mueller, K., Wiegand, T.: Optimization and Comparision of Coding Algorithms for Mobile 3DTV, 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, Potsdam (2009)

5. Smolic, A.; Mueller, K.; Merkle, P.; Kauff, P.; Wiegand, T.: An Overview of Available and Emerging 3D Video Formats and Depth Enhanced Stereo as Efficient Generic Solution, Picture Coding Symposium 2009, PCS 2009 (2009)
6. Pei-Kuei Tsung; Li-Fu Ding; Wei-Yin Chen; Tzu-Der Chuang; Yu-Han Chen; Pai-Heng Hsiao; Shao-Yi Chien; Liang-Gee Chen: Video Encoder Design for High-Definition 3D Video Communication Systems, IEEE Communications Magazine, vol.48, pp. 76-86 (2010)
7. Zhang, Z.: A Flexible New Technique for Camera Calibration, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 11, pp. 1330–1334 (2000)
8. Stankowski, J., Klimaszewski, K.: Rectification Algorithm for Parallel Multi-Camera Setup, International Conference Computer Vision and Graphics, ICCVG 2010, submitted for publication in Lecture Notes on Computer Science (2010)
9. Klimaszewski, K., Stankiewicz, O., Stankowski, J., Wegner, K., Domański, M.: Przygotowanie Wielowidokowych Sekwencji Wizyjnych dla Badań nad Telewizją Trójwymiarową, (in Polish) submitted for publication in: Krajowa Konferencja Radiokomunikacji, Radiofonii i Telewizji, KKRRiT, 2010 and in: Przegląd Telekomunikacyjny (2010)
10. Stankiewicz, O., Wegner K.: Depth Map Estimation Software version 3, ISO/IEC JTC1/SC29/WG11 (MPEG) Doc. M15540, Hannover (2008)
11. Stankiewicz, O., Wegner, K., Domański, M.: Stereoscopic Depth Refinement by Mid-Level Hypothesis, IEEE International Conference on Multimedia & Expo, ICME 2010, Singapore (2010), to be published
12. Domański, M., Grajek, T., Klimaszewski, K., Kurc, M., Stankiewicz, O., Stankowski, J., Wegner, K.: Poznan Multiview Video Test Sequences and Camera Parameters", ISO/IEC JTC1/SC29/WG11 (MPEG) Doc. M17050, Xian (2009)
13. Feldmann, I., Mueller, M., Zilly, F., Tanger, R., Mueller, K., Smolic, A., Kauff, P., Wiegand, T.: HHI Test Material for 3D Video, ISO/IEC JTC1/SC29/WG11 (MPEG) Doc. M15413, Archamps (2008)
14. Tanimoto M., Fujii T., Suzuki K., Fukushima N., Mori Y.: Reference Softwares for Depth Estimation and View Synthesis, ISO/IEC JTC1/SC29/WG11 (MPEG) Doc. M15377, Archamps (2008)
15. Tanimoto M., Fujii T., Tehrani M.P., Wildeboer M.: Depth Estimation Reference Software (DERS) 4.0, ISO/IEC JTC1/SC29/WG11, (MPEG) Doc. M16605, London (2009)
16. Klimaszewski K., Wegner K., Domański M.: Influence of Distortions Introduced by Compression on Quality of View Synthesis in Multiview systems, 3DTV-Conference 2009 The True Vision Capture, Transmission and Display of 3D Video, Potsdam (2009)
17. Chen Y., Pandit P., Yea S., Lim C.S.: Draft Reference Software for MVC, Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG , ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, Doc. JVT-AE207, London (2009)
18. Domański, M., Gotfryd, M., Wegner, K.: View Synthesis for Multiview Video Transmission, The 2009 International Conference on Image Processing, Computer Vision, and Pattern Recognition IPCV'09, Las Vegas (2009)