PLANAR 3D SCENE REPRESENTATIONS FOR DEPTH COMPRESSION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY
BURAK OĞUZ ÖZKALAYCI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

FEBRUARY 2014

Approval of the thesis:

PLANAR 3D SCENE REPRESENTATIONS FOR DEPTH COMPRESSION

submitted by **BURAK OĞUZ ÖZKALAYCI** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Gönül Turhan Sayan
Head of Department, **Electrical and Electronics Engineering**

Prof. Dr. A. Aydın Alatan
Supervisor, **Electrical and Electronics Eng. Dept., METU**

Examining Committee Members:

Prof. Dr. Gözde Bozdağı Akar
Electrical and Electronics Engineering Dept., METU

Prof. Dr. A. Aydın Alatan
Electrical and Electronics Engineering Dept., METU

Prof. Dr. Levent Onural
Electrical and Electronics Engineering Dept., Bilkent University

Assoc. Prof. Dr. Çağatay Candan
Electrical and Electronics Engineering Dept., METU

Assist. Prof. Dr. Fatih Kamışlı
Electrical and Electronics Engineering Dept., METU

Date: **February 6, 2014**

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: BURAK OĞUZ ÖZKALAYCI

Signature :

ABSTRACT

PLANAR 3D SCENE REPRESENTATIONS FOR DEPTH COMPRESSION

Özkalaycı, Burak Oğuz
Ph.D., Department of Electrical and Electronics Engineering
Supervisor : Prof. Dr. A. Aydın Alatan

February 2014, 167 pages

The recent invasion of stereoscopic 3D television technologies is expected to be followed by autostereoscopic and holographic technologies. Glasses-free multiple stereoscopic pair displaying capabilities of these technologies will advance the 3D experience. The prospective 3D format to create the multiple views for such displays is Multiview Video plus Depth (MVD) format based on the Depth Image Based Rendering (DIBR) techniques. The depth modality of the MVD format is an active research area whose main objective is to develop DIBR friendly efficient compression methods.

As a part this research, the thesis proposes novel 3D planar-based depth representations. The planar approximation of the stereo depth images is formulated as an energy-based co-segmentation problem by a Markov Random Field model. The energy terms of this problem are designed to mimic the rate-distortion tradeoff for a depth compression application. A heuristic algorithm is developed for practical utilization of the proposed planar approximations in stereo depth compression. The co-segmented regions are also represented as layered planar structures forming a novel single referenced MVD format.

The proposed planar based depth compression solutions are compared against the state-of-the art image/video and MVD compression standards. The compression performances are analyzed for depth reconstruction and novel view rendering by DIBR techniques. All the experiments are performed with the ground truth texture of the MVD data, since the scope of the thesis is limited with the depth modality. The visual and objective evaluations show that the proposed planar representations are promising for efficient depth compression with artifact-free novel view rendering. As a remarkable contribution, the proposed layered planar MVD representation also brings the depth perception quality considerations in the MVD compression schemes by decoupling the texture and geometry to a wide extent.

Keywords: 3DTV, MVD, DIBR, depth compression, energy based co-segmentation, model fitting

ÖZ

DERİNLİK SIKIŞTIRILMASI İÇİN DÜZLEMSEL 3B SAHNE GÖSTERİMLERİ

Özkalaycı, Burak Oğuz
Doktora, Elektrik ve Elektronik Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. A. Aydın Alatan

Şubat 2014 , 167 sayfa

Yakın geçmişte yaşanan 3B televizyon istilasının ardından otosteroskopik ve holografik teknolojilerin benzer şekilde yaygınlaşacağı tahmin edilmektedir. Bu teknolojilerin özel bir gözlüğe ihtiyaç duymadan aynı anda çok sayıda görüş açısını izleyiciye sunabilmesi sayesinde 3B gerçeklik hissiyatı zenginleşecektir. İzleyiciye farklı görüş açılarının sağlanması için en olası 3B veri formatı, Derinlik Görüntüsü Temelli Resmetme (DIBR) tekniklerine dayanan Çok-görüntülü Video artı Derinlik (MVD) formatıdır. MVD formatının içeriğindeki derinlik görüntülerinin DIBR dostu etkin sıkıştırılması, aktif bir araştırma konusudur.

Bu araştırmanın bir parçası olarak tez, yeni bir 3B düzlemsel derinlik gösterimi önermektedir. Önerilen stereo derinlik görüntülerinin düzlemsel kestirim formülasyonu, Markov Rasgele Alanlar modellemesi yardımı ile enerji tabanlı bir birlikte-bölütleme problemi olarak düşünülmüştür. Problemin enerji terimleri, derinlik görüntülerinin sıkıştırılmasındaki hız-distorsiyon takasını taklit edecek şekilde tasarlanmıştır. Önerilen düzlemsel kestirimin stereo derinlik sıkıştırmasında pratik kullanımı için buluşsal bir algoritma geliştirilmiştir. Birlikte-bölütlenen bölgeler ayrıca katmanlı düzlemsel yapılar şeklinde ifade edilerek yeni tek referanslı bir MVD gösterimi oluşturulmuştur.

Önerilen düzlemsel tabanlı derinlik sıkıştırma çözümleri, en gelişkin teknolojinin görüntü/video ve MVD sıkıştırma standartları ile karşılaştırılmıştır. Sıkıştırma performansları, derinlik geri çatımı ve DIBR teknikleri ile yeni görüntü resmetme başlıklarında değerlendirilmiştir. Tezin kapsamı derinlik kipi ile sınırlandığından dolayı tüm deneyler MVD verilerinin orijinal dokuları ile yapılmıştır. Görsel ve objektif değerlendirmeler, önerilen düzlemsel gösterimlerin etkin derinlik sıkıştırma ile doğal yeni görüntü resmetme için umut verici olduğunu göstermektedir. Önerilen tabakalı düzlemsel MVD gösterimi doku ve geometrideki bozulmaları büyük ölçüde birbirinden bağımsız hale getirmektedir. Bu sayede kayda değer bir katkı olarak tabakalı düzlemsel MVD gösterimi, derinlik algı kalitesini dikkate alan MVD sıkıştırma yaklaşımları beraberinde düşündürmektedir.

Anahtar Kelimeler: 3DTV, MVD, DIBR, derinlik sıkıştırması, enerji tabanlı birlikte-bölütleme, modele uydurma

to Özlem

ACKNOWLEDGEMENTS

I found myself to be a lucky one since I have been surrounded with special people at home, at school, at work all the time. I believe I am an outcome of these fruitful relations. First of all, I make a broad appreciation for all those people who build the one I know as me. Then I want to mention some of them who took significant roles in my doctoral studies.

I express my deep and sincere gratitude to my supervisor, Prof. Aydın Alatan. Without his guidance, stimulation and encouragement I cannot imagine to finalize my research with a thesis like this. His confidence, friendship and positive attitude made it all possible. I appreciate all his contributions of time, ideas and motivation for this study.

I thank my thesis progress committee members Prof. Levent Onural and Prof. Gözde Bozdağı Akar, for their valuable suggestions and motivations. Every meeting refreshed my courage by their feedbacks to find my research track.

I thank Cevahir Çığla and Emrah Taşlı for joining me in this journey. The Ankara office of Vestek R&D was incredible and I know it is once in a life time experience. I am very glad to share all the joy and pain with them.

An important milestone of the Ph.D work was the qualification exam but our team turned it into a total fun. I thank Ahmet Saraçoğlu and Serdar Gedik for this joyful teamwork.

My dear friend, Engin Tola's contributions to my thesis and study are quite concrete and valuable. He has the merit that I can articulate my ideas in a speed of ligthning. I thank him sincerely for his friendship, understanding and prolific ideas.

Lastly I thank my wife, Özlem, for her love, support and understanding. She has been always by my side and been successful to make me smile.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGEMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xiv

LIST OF FIGURES xv

LIST OF ABBREVIATIONS xxiii

CHAPTERS

1 INTRODUCTION 1

- 1.1 Stereoscopic and 3D Video 1
- 1.2 3DV System Architecture 2
  - 1.2.1 3DV Acquisition and Displays 2
  - 1.2.2 3DV Representation and Compression 5
    - 1.2.2.1 Conventional Stereo 5
    - 1.2.2.2 Multiview Video 6
    - 1.2.2.3 Video plus Depth 9

1.2.2.4 Multiview Video plus Depth 10
1.2.3 3DV Rendering 12
1.3 Problem Statement and Scope of the Thesis 13
1.3.1 Outline of the thesis 14
2 PLANAR STEREO DEPTH REPRESENTATION 17
2.1 Depth Image Representations: Related Literature 17
2.2 Motivation for the Proposed Representation 21
2.3 Energy Based Formulation of the Proposed Representation 23
2.3.1 MRF Modelling 24
2.3.2 Optimization of MRF Energy 27
2.3.3 Continuum of labels 28
2.4 Planar Model Fitting Examples 31
3 STEREO DEPTH COMPRESSION BASED ON PLANAR REPRESENTATION 39
3.1 Energy Design for Compression Applications 40
3.1.1 Pareto Optimality in Rate-Distortion 41
3.1.2 Proposed Algorithm to Weight Costs 44
3.2 Encoding of Proposed Planar Representation 45
3.3 Planar Models as Depth Prediction 48
3.4 Experiments 50
3.4.1 Speedup of Planar Model Fitting 51
3.4.2 Comparative Planar Compression Experiments 53

3.4.3 Comparative Planar Prediction Experiments . 72

4 PLANAR LAYERED MVD REPRESENTATION 87

4.1 MVD Representations in the Litearture 88

4.2 Proposed Planar Layered MVD Representation 90

4.3 Experiments 95

4.3.1 Visual Assessment 100

4.3.2 Depth Perception Comparison 120

5 CONCLUSIONS AND FUTURE WORK 135

5.1 Summary of the Thesis 135

5.2 Conclusions 137

5.3 Future Work 139

REFERENCES 143

APPENDICES

A GRAPH CUT ALGORITHM 155

A.0.1 MRF optimization with label costs 157

B MIDDLEBURY CAMERA CALIBRATION MATRICES 161

C HEVC-3D RENDERER CONFIGURATION 163

CURRICULUM VITAE 165

LIST OF TABLES

TABLES

Table 2.1 *Middlebury* dataset. 32

Table 2.2 The number of models and reconstruction accuracies of the depth images for various cost weightings. 36

Table 4.1 Objective evaluation of novel view renderings by PSNR and SSIM measures for geometric distortions resulting in horizontal shifts of one pixel at most. 120

Table 4.2 PSNR and SSIM values of the novel view rendering results obtained for *Art* dataset are tabulated. 124

Table 4.3 PSNR and SSIM values of the novel view rendering results obtained for *Moebius* dataset are tabulated. 124

Table 4.4 PSNR and SSIM values of the novel view rendering results obtained for *Books* dataset are tabulated. 124

Table C.1 Parameter values of HEVC-3D renderer. 163

LIST OF FIGURES

FIGURES

Figure 1.1 3DV system illustration 3

Figure 1.2 An illustration of autostereoscopic display. 4

Figure 1.3 The recommended Group of Pictures structure in H.264/MVC. The inter-view prediction directions are marked as red arrows. . . . 7

Figure 1.4 In H.264/MVC, the bitrate increase has a linear characteristics similar to the simulcast encoding of the views. 7

Figure 1.5 A multiview video encoder, which explicitly utilizes the scene geometry. 8

Figure 1.6 A video plus depth example. 9

Figure 1.7 Requirments of the call for proposal on 3D Video coding by MPEG community. 11

Figure 1.8 Geometry based classification of the IBR methods. 12

Figure 2.1 Original depth image (left), and its DCT-based approximation by H.264 18

Figure 2.2 The edge information encoded by a Shape-Adaptive DWT approach 19

Figure 2.3 Platelet representation of a depth image 20

Figure 2.4 An example for segmentation based depth image representation 21

Figure 2.5 Factor graph representation of the MRF model. 25

Figure 2.6 The bold and dark connections represent the 3D neighborhood of a pixel according to the MRF model. 27

Figure 2.7 The progress of the modified PEARL algorithm in fitting planar models to *Moebius* dataset. Left and right columns are the planar depth reconstruction results according to the labeling images given in the middle columns. The number of planar models increase from top to bottom. 35

Figure 2.8 The planar solutions obtained for *Moebius* dataset by different cost weighting combinations. For each sub figure, the planar reconstruction is given at the top and its labelling image is given at the bottom. See Table 2.2 for their numerical details. 36

Figure 2.9 The planar depth image reconstruction examples of *Middlebury* dataset. 37

Figure 2.10 The planar model labeling images of the reconstructed depth images given in Figure 2.9. 38

Figure 3.1 Example of a Pareto curve in 2-dimensional objective space with a feasible set of $\mathcal{C}$. 42

Figure 3.2 The planar depth encoder's byte-stream package definition. . 47

Figure 3.3 *Top row:* The labeling image given at left side encodes the 9 planar model assignments. The byte packed version of the labeling image is given at right side. *Bottom row:* A zoomed detail of the same region for the labeling and byte packed labeling images from left to right. Red boxes show the zoomed region. (Images are histogram equalized for better visualization.) 48

Figure 3.4 The planar prediction based depth encoder's byte-stream package definition. 51

Figure 3.5 The labeling and the corresonding depth reconstruction of *Teddy* dataset is given in the top and bottom rows, respectively. From left to right columns, the solution for the downscaled model, initialization of the full scale model and the final solution of the full scale model are shown. 52

Figure 3.6 Compression performances of the lossless coding tools for the labeling image given at the left side. 54

Figure 3.7 Visual comparison of the depth compression results in depth reconstruction and novel view rendering. PSNR (dB) and SSIM scores are also stated consecutively. The rate of the compression methods from top to bottom are equal to 0.0547, 0.0647, 0.0601, 0.0652, 0.0673 bits per pixel. 58

Figure 3.8 Compression experiments comparing the mean PSNR values of the *reconstructed stereo depth images*. 59

Figure 3.9 Compression experiments comparing the mean PSNR values of the *reconstructed stereo depth images*. 60

Figure 3.10 Compression experiments comparing the mean PSNR values of the *reconstructed stereo depth images*. 61

Figure 3.11 Compression experiments comparing the mean SSIM scores of the *reconstructed stereo depth images*. 62

Figure 3.12 Compression experiments comparing the mean SSIM scores of the *reconstructed stereo depth images*. 63

Figure 3.13 Compression experiments comparing the mean SSIM scores of the *reconstructed stereo depth images*. 64

Figure 3.14 Compression experiments comparing the PSNR values of the *novel view renderings* obtained by the reconstructed stereo depth images. 66

Figure 3.15 Compression experiments comparing the PSNR values of the *novel view renderings* obtained by the reconstructed stereo depth images. 67

Figure 3.16 Compression experiments comparing the PSNR values of the *novel view renderings* obtained by the reconstructed stereo depth images. 68

Figure 3.17 Compression experiments comparing the SSIM scores of the *novel view renderings* obtained by the reconstructed stereo depth images. 69

Figure 3.18 Compression experiments comparing the SSIM scores of the *novel view renderings* obtained by the reconstructed stereo depth images. 70

Figure 3.19 Compression experiments comparing the SSIM scores of the *novel view renderings* obtained by the reconstructed stereo depth images. 71

Figure 3.20 Mean PSNR values of the *depth reconstruction* results obtained by JPEG 2000 and planar prediction with JPEG 2000 residual coding. 74

Figure 3.21 Mean PSNR values of the *depth reconstruction* results obtained by JPEG 2000 and planar prediction with JPEG 2000 residual coding. 75

Figure 3.22 Mean PSNR values of the *depth reconstruction* results obtained by JPEG 2000 and planar prediction with JPEG 2000 residual coding. 76

Figure 3.23 PSNR values of the *novel view rendering* results obtained by the depth reconstructions of JPEG 2000 and planar prediction with JPEG 2000 residual coding. 77

Figure 3.24 PSNR values of the *novel view rendering* results obtained by the depth reconstructions of JPEG 2000 and planar prediction with JPEG 2000 residual coding. 78

Figure 3.25 PSNR values of the *novel view rendering* results obtained by the depth reconstructions of JPEG 2000 and planar prediction with JPEG 2000 residual coding. 79

Figure 3.26 Mean PSNR values of the *depth reconstruction* results obtained by HEVC variants and planar prediction with HEVC in residual coding. 80

Figure 3.27 Mean PSNR values of the *depth reconstruction* results obtained by HEVC variants and planar prediction with HEVC in residual coding. 81

Figure 3.28 Mean PSNR values of the *depth reconstruction* results obtained by HEVC variants and planar prediction with HEVC in residual coding. 82

Figure 3.29 PSNR values of the *novel view rendering* results obtained by the depth reconstructions of HEVC variants and planar prediction with HEVC in residual coding. 83

Figure 3.30 PSNR values of the *novel view rendering* results obtained by the depth reconstructions of HEVC variants and planar prediction with HEVC in residual coding. 84

Figure 3.31 PSNR values of the *novel view rendering* results obtained by the depth reconstructions of HEVC variants and planar prediction with HEVC in residual coding. 85

Figure 4.1 Examples of MVD representations in the literature. 90

Figure 4.2 A planar layer extraction example for *Art* dataset. Left and right colums are the texture and planar labelings of the stereo pair. Mid column is the texture and the approximated planar depth of the extracted layer. 91

Figure 4.3 A 3D illustration of the planar layered MVD representation of *Art* dataset. Its corresponding layer information in 2D image planes are presented in Figure 4.4. In order to visualize the 3D extent of the planar layers, various camera views are rendered. 92

Figure 4.4 A planar layered MVD representation example for *Art* dataset. The top two rows represents the MVD data in its raw format. The bottom two rows are the texture images of the layers and their depth approximations according to the planar model assignments. 93

Figure 4.5 The left and right images are the labelings obtained by a planar representation of the stereo depth images. The shape of the extracted layers are bit-plane coded in the middle image. The number of planar models is 8 and histogram equalization is applied to labeling images for better visualization. 95

Figure 4.6 The layered planar MVD encoder's byte-stream package definition. 95

Figure 4.7 The payload of the texture information in bit per pixel for the MVD representations in the raw and the proposed layered format. The payloads are obtained by lossless compression of the texture images by PNG encoder. 97

Figure 4.8 The results of experiment, comparing the novel view rendering PSNR values obtained by the proposed and the conventional MVD compression schemes. 102

Figure 4.9 The results of experiment, comparing the novel view rendering PSNR values obtained by the proposed and the conventional MVD compression schemes. 103

Figure 4.10 The results of experiment, comparing the novel view rendering PSNR values obtained by the proposed and the conventional MVD compression schemes. 104

Figure 4.11 The results of experiment, comparing the novel view rendering PSNR values obtained by the proposed and the conventional MVD compression schemes. 105

Figure 4.12 The results of experiment, comparing the novel view rendering PSNR values obtained by the proposed and the conventional MVD compression schemes. 106

Figure 4.13 The results of experiment, comparing the novel view rendering PSNR values obtained by the proposed and the conventional MVD compression schemes. 107

Figure 4.14 The results of experiment, comparing the novel view rendering PSNR values obtained by the proposed and the conventional MVD compression schemes. 108

Figure 4.15 The results of experiment, comparing the novel view rendering PSNR values obtained by the proposed and the conventional MVD compression schemes. 109

Figure 4.16 The results of experiment, comparing the novel view rendering SSIM scores obtained by the proposed and the conventional MVD compression schemes. 110

Figure 4.17 The results of experiment, comparing the novel view rendering SSIM scores obtained by the proposed and the conventional MVD compression schemes. 111

Figure 4.18 The results of experiment, comparing the novel view rendering SSIM scores obtained by the proposed and the conventional MVD compression schemes. 112

Figure 4.19 The results of experiment, comparing the novel view rendering SSIM scores obtained by the proposed and the conventional MVD compression schemes. 113

Figure 4.20 The results of experiment, comparing the novel view rendering SSIM scores obtained by the proposed and the conventional MVD compression schemes. 114

Figure 4.21 The results of experiment, comparing the novel view rendering SSIM scores obtained by the proposed and the conventional MVD compression schemes. 115

Figure 4.22 The results of experiment, comparing the novel view rendering SSIM scores obtained by the proposed and the conventional MVD compression schemes. 116

Figure 4.23 The results of experiment, comparing the novel view rendering SSIM scores obtained by the proposed and the conventional MVD compression schemes. 117

Figure 4.24 Top to bottom; texture and planar approximation of the 4-layered representation of *Aloe* dataset and its novel view rendering results. The distances between the novel views are the quarters of the baseline and starts from the left view of the raw MVD data. The red arrows highlight the artifacts due to layer boundary errors. 118

Figure 4.25 Visual comparison of novel view rendering results obtained for the *Art* dataset by different MVD compression schemes and the corresponding ground truth camera views. 121

Figure 4.26 Visual comparison of novel view rendering results obtained for the *Moebius* dataset by different MVD compression schemes and the corresponding ground truth camera views. 122

Figure 4.27 Visual comparison of novel view rendering results obtained for the *Books* dataset by different MVD compression schemes and the corresponding ground truth camera views. 123

Figure 4.28 The comparison of the depth perception of the *Art* dataset for different MVD compression schemes. The views from left to right are located at the left quarter and in the middle of the baseline. The anaglyph technique is applied by red-cyan encoding. Best viewed in digital copy with a proper zoom in. 129

Figure 4.29 The comparison of the depth perception of the *Moebius* dataset for different MVD compression schemes. The views from left to right are located in the middle and at the right quarter of the baseline. The anaglyph technique is applied by red-cyan encoding. Best viewed in digital copy with a proper zoom in. 130

Figure 4.30 The comparison of the depth perception of the *Books* dataset for different MVD compression schemes. The views from left to right are located at the left quarter and in the middle of the baseline. The anaglyph technique is applied by red-cyan encoding. Best viewed in digital copy with a proper zoom in. 131

Figure 4.31 The depth perception comparison of the layered planar MVD under extreme geometric distortion (4 planar model) with the uncompressed raw MVD format for the *Aloe* dataset. The left column shows the rendering results for the proposed layered planar MVD representation. The anaglyph technique is applied by red-cyan encoding. Best viewed in digital copy with a proper zoom in. 132

Figure 4.32 Top to bottom, novel view rendering results of raw and proposed MVD representations for the depth images approximated by 4 planar models. Left to right, novel views positioned at the 0.25, 0.5 and 0.75 of the baseline distance. 133

Figure A.1 A graph construction example for two random variables with unary and pairwise energies. 156

Figure A.2 Possible s-t cuts for two random variable case. 156

Figure A.3 GC based MRF update rule examples. From left to right; initial labeling map, $\alpha - \beta$ swap move and α-expansion move. 157

Figure A.4 Graph construction for a label cost. 158

Figure A.5 A graph construction example for an α-expansion move with labeling cost. The current labeling of the 1-D field of 6 variable is given at the top. The auxiliary nodes and the edges for the label cost are highlighted. 159

LIST OF ABBREVIATIONS

3DTV	3 Dimensional Television
3DV	3 Dimensional Video
bpp	Bits Per Pixel
DC	Direct Current
DCT	Discrete Cosine Transform
DES	Depth Enhanced Stereo
DWT	Discrete Wavelet Transform
DIBR	Depth Image Based Rendering
HVS	Human Vision System
IBR	Image Based Rendering
ITU	International Telecommunication Union
ITU-T	ITU - Telecommunication Standardization Sector
JCT-VC	Joint Collaborative Team on Video Coding
LDI	Layered Depth Image
MOS	Mean Opinion Score
MPEG	Moving Picture Experts Group
MRF	Markov Random Field
MVC	Multiview Video Coding
MVD	Multiview Video plus Depth
MVP	Multiview Profile
PEARL	Propose Expand And Re-estimate Labels
PNG	Portable Network Graphics
PSNR	Peak Signal to Noise Ratio
SSIM	Structural Similarity
VCEG	Video Coding Experts Group

CHAPTER 1

INTRODUCTION

Today 3D became the buzz word of the entertainment, multimedia and consumer electronics sectors. The revenue success of the 3D theaters triggered the invasion of the 3D technologies to our daily lives. As the first step, the stereoscopic video added the dimension of depth to the human perception by providing the left-right eye view separation. However, the forthcoming 3D technology, which is mentioned as 3D Video (3DV) in general, promises much more realistic perception of the scene by providing interactivity of the viewer with the displayed scene. However, this interactivity of 3DV comes with a r/evolutionary changes in the infrastructure of the monoscopic/stereoscopic video systems. Hence, 3DV systems, which contain scene acquisition, representation, compression, transmission, rendering and displaying steps, need to be reconsidered as a whole; each step contains its own problems and various dependencies to other steps.

1.1 Stereoscopic and 3D Video

The principals of the stereo and 3D video are both based on the stereopsis property of the human visual system (HVS). Stereopsis is the process of visual perception leading to the perception of depth from the two slightly different projections of the world onto the retinas of the two eyes. In stereoscopic and 3DV systems, the left and right eyes of the observer are exposed to slightly different views of the scene. The relative differences in two retinal images, that are called *retinal disparity*, create the depth perception. The two slightly different views of the same view just make up the stereoscopic video content.

The shortcomings of the stereoscopic video come out with the rivalry between the monocular and binocular depth perception cues in HVS. For instance, the movement of the observer might break down the reality of the scene, since the observer's left and right eye views do not change according to his/her movement. Instead the same left and right eye view during the movement creates a shearing-like effect on the observer's perception.

At this point, 3DV systems promise to increase the sense of reality by the interactivity of the observer with the scene. This interactivity is simply achieved by the capability of providing various left and right eye view couples of the scene in accordance with the observer's actions, commands, movements or positions. The interaction between the observer and the 3DV system might be active or passive. In the active interaction case the 3DV system tracks the observer's position or commands and provides the corresponding left and right eye view pair to the observer. In the passive case, the 3DV system provides all the possible view combinations to observer's space and the observer's views change as he/she navigates.

1.2 3DV System Architecture

In Figure 1.1, a schematic of a feasible 3DV system architecture is illustrated. The acquisition, displaying, representation/compression, and rendering units of the 3DV systems in the literature will be explained in detail in the forthcoming subsections.

1.2.1 3DV Acquisition and Displays

In the most general form of 3DV systems, the scene of interest is captured by a synchronized camera array. In order to provide the end user of the 3DV system to navigate in the scene space realistically, the plenoptic function (light field) of the scene should be reconstructed from these samples. Chai et. al. [2] analyzed the sufficient number of samples to reconstruct the plenoptic function in theory and concluded that spectral support of the plenoptic function is bounded only

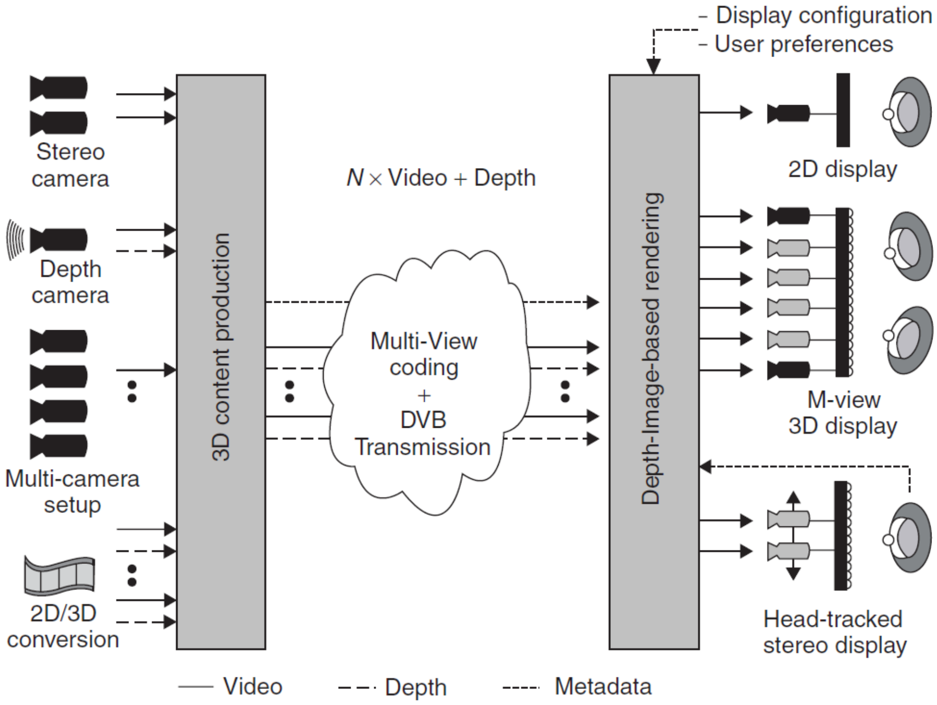


Figure 1.1: 3DV system illustration (Reprinted with permission. Copyright John Wiley & Sons, Inc. 2010 [1]).

by the minimum and maximum depth of scene objects. However in practice, the factors, like scene geometry, texture, object and camera motions make it hard to arrange a sufficient sampling arrangement all the time. Hence, to oversample the plenoptic function is one of the practical solutions. Another practical approach is to utilize the scene geometry information in the plenoptic function reconstruction. Since the depth of the scene objects constrains the plenoptic function, fewer view samples with the scene geometry information might suffice to reconstruct the plenoptic function.

The aforementioned two alternatives of plenoptic reconstruction constitute the tradeoff between the scene geometry information and the number of views of the scene; this is a very well-known tradeoff in Image Based Rendering (IBR) techniques. According to IBR, the same quality of novel view rendering can be obtained by fewer images with more knowledge of the scene geometry [3]. The same tradeoff appears in the 3D content acquisition for representing the 3D scene. The 3D scene can be represented with more views of the scene and less knowledge of the scene geometry or fewer views and more knowledge of the scene geometry.

However, in general the acquisition of the scene geometry information is not as trivial as capturing different views of the scene. There are commercial time-of-flight cameras for capturing the depth of the scene; however, they have synchronization and resolution problems compared to conventional cameras. The conventional acquisition of the scene geometry is achieved in a depth estimation framework which is a vastly studied topic and known as *stereo correspondence problem* [4] in computer vision. The pinhole camera model and projective geometry constraints are typically utilized on multiple views of the scene for the solution of this problem. A detailed taxonomy of the stereo correspondence algorithms is given in [4]. It should also be mentioned that the state-of-the-art stereo correspondence algorithms require considerable computational effort and still error prone due to the ill-posed nature of the problem. Hence, the tradeoff between the scene geometry information and the number of views of the scene becomes a computational capability tradeoff of the 3DV acquisition system. For a survey of capturing technologies for 3D video applications, the interested readers should refer to [5].

The principle of the 3DV displays is mainly based on separating the left and right eye views of the observer. The widely available 3D displays provide the

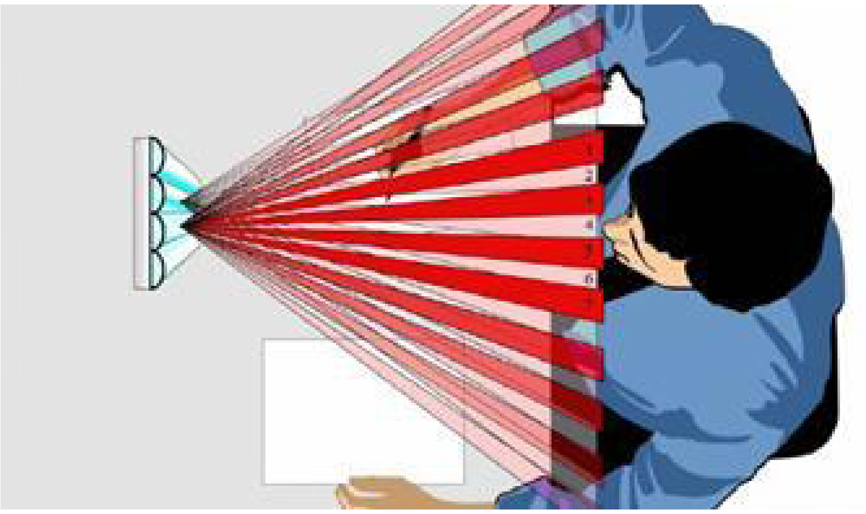


Figure 1.2: An illustration of autostereoscopic display (Reprinted with permission. Copyright John Wiley & Sons, Inc. 2010 [1]).

stereo view separation with the help of special glasses. The *active shutter glasses* technique receives a synchronization signal from the display and toggles the left and right glass transparency in order to expose the stereo image pair to right and left eyes of the viewer separately. Another glasses-based approach, called *passive glasses*, differentiate the stereo pair by different polarization properties maintained by the light projector or the retarder in front of the display panel. The glasses-based approaches are limited to conventional stereo video applications or head tracking based free view TV applications. Hence, the 3D perception is limited to two views of the scene or to a single user in the latter case.

There are also glasses-free 3D display technologies, namely autostereoscopic and holographic displays. The autostereoscopic displays can allocate the viewing zones of the display to different views of the scene as shown in Figure 1.2. By proper positioning of the observers, the 3D perception can be achieved by these displays. In case of holographic displays, the observers can be positioned freely, since holographic displays generate the light field of the scene at the display plane. However, the mentioned glasses free technologies are not mature for consumer applications yet. For the details of the 3D display technologies one may examine the survey in [6].

1.2.2 3DV Representation and Compression

The acquired 3DV data should be represented in a compression and rendering efficient way in order to maintain the channel transmission capabilities and create a realistic 3D impression for the end user. Regarding these challenges, the suggested 3DV representations in the literature evolved from stereo view towards view and geometry hybrid solutions. For a detailed survey of representation and compression approaches in 3D TV applications, it is advised to refer [7] and [8].

1.2.2.1 Conventional Stereo

The conventional stereo video is composed of left and right eye views of the scene. This primitive 3DV representation has no geometry information about the scene.

Rendering novel views of the scene is not considered in this representation; hence, user interaction with 3D scene is out-of-scope. However, in order to enhance the depth perception of the end user and lower the eye strain implications of a stereo content, there are some rendering application proposals for conventional stereo video. For instance, Lang et. al. [9] propose a nonlinear disparity mapping by an image warping in the guidance of sparse depth estimates of feature points in the scene.

For the compression of stereoscopic video, the inter-view redundancies are also utilized in addition to conventional temporal and spatial redundancies of a standalone video stream. For backward compatibility of the stereoscopic video stream, a Multi-View Profile (MVP) is specified by ITU-T [10]. In MVP, the left eye view is coded as MPEG-2 main profile bit stream, whereas the right eye view is coded as an enhancement layer. In the encoding of enhancement layer, inter-view redundancies are exploited by using the decoded left view.

1.2.2.2 Multiview Video

In order to give the observer the freedom to navigate to different views of the scene, the number of the views are increased in the multiview video case. The number of the views might range from 5 views to hundreds. Increasing the number of the views increases the rendering capabilities of the end user. The user is able to navigate in the 3D scene with a wider range and higher visual quality by increasing the number of the views it receives in multiview video. However, increasing the number of views in a multiview video also increases the compression cost. In the worst of simulcast transmission of multiview video, the bandwidth requirement increase linearly with the number of views; the resultant large bandwidth requirement cannot be handled by present infrastructure. Hence, for practical applications, the number of the views of the multiview video does not in general exceed 10 and much more efficient compression techniques than simulcast are utilized.

In multiview video, in addition to spatial and temporal statistical redundancies of conventional video, geometrical redundancies exist between the views. The

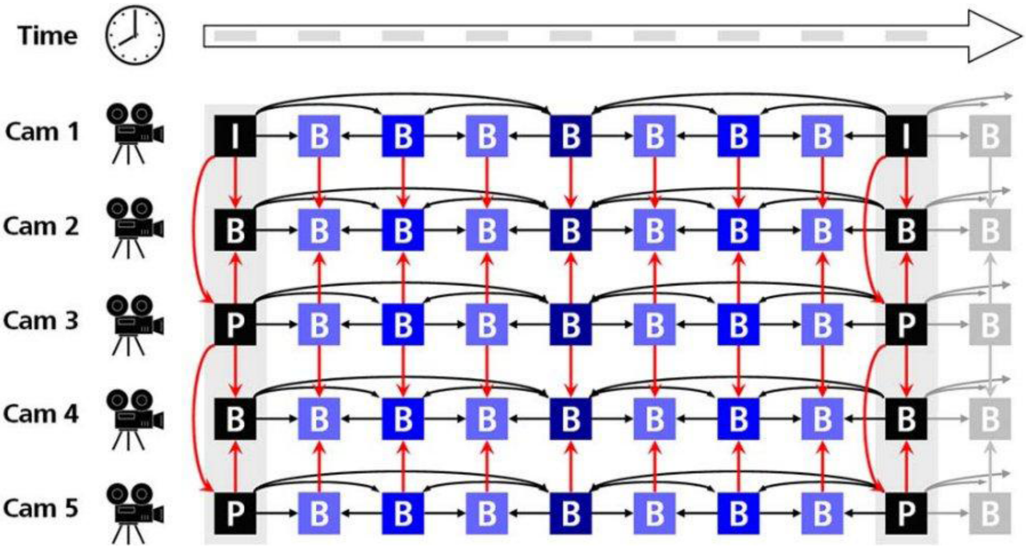


Figure 1.3: The recommended Group of Pictures structure in H.264/MVC. The inter-view prediction directions are marked as red arrows (Reprinted with permission. Copyright Elsevier 2008 [11]).

efficient compression methods for multiview video in the literature all exploit this inter-view geometric redundancy. One of the state-of-the-art compression methods is the amendment to H.264/MPEG-4 AVC video compression standard, called Multiview Video Coding (MVC) [13]. MVC is designed to be backward compatible with H.264/AVC, while exploiting the interview geometrical redun-

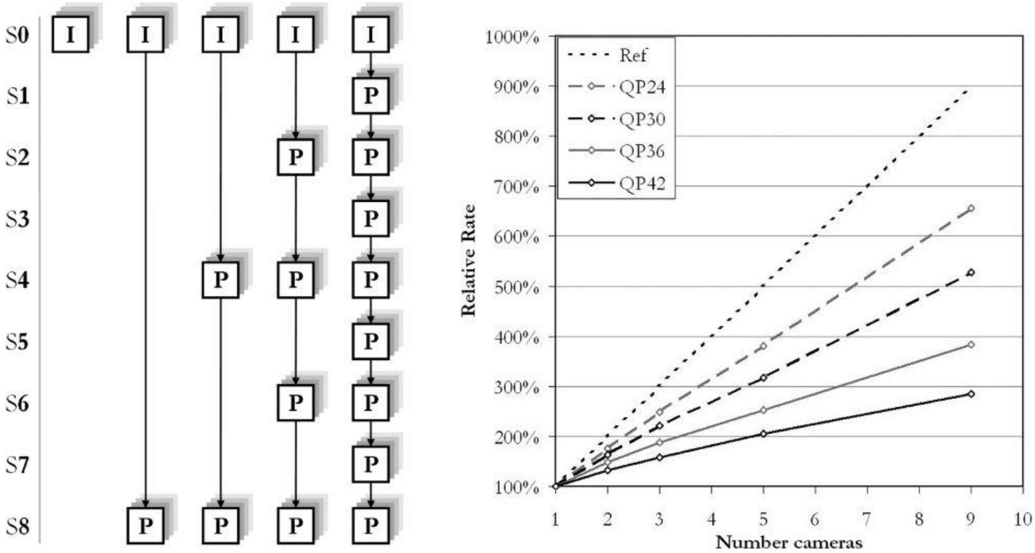


Figure 1.4: In H.264/MVC, the bitrate increase has a linear characteristics like the simulcast encoding of the views (Reprinted with permission. Copyright IEEE 2007 [12]).

dancies. In MVC, spatio-temporal prediction based on hierarchical B pictures is introduced for exploiting the inter-view redundancies [12]. However, MVC compression method still has a linear bitrate increasing characteristics for the increasing number of views as shown in Figure 1.4 [12].

In the literature, there are other multiview video compression approaches which exploit the geometric redundancies more explicitly by using the epipolar constraints over the camera setup. Their common approach for multiview video compression is to use novel view prediction routines. The geometric constraints utilized in the novel view prediction might be loose to just derive pairwise rendering optimal disparity maps [15], or very strict to estimate the whole 3D scene geometry [16], [14], [17]. For these approaches, it can be said that they acquire the scene geometry information, which is not explicitly available in the multiview video representation, to some extent, in order to make an efficient prediction during compression. An exemplary outline of a backward compatible scalable multiview video compression method is illustrated in Figure 1.5. In comparison to MVC approach, the geometry constrained novel view prediction approach brings computational burden at the encoder/decoder side [18] or rate increase

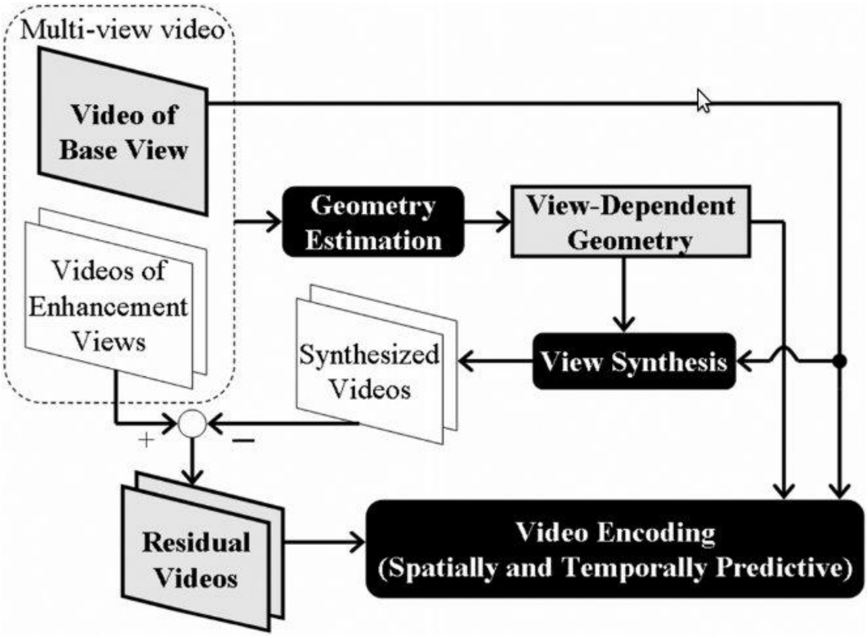


Figure 1.5: A multiview video encoder, which explicitly utilizes the scene geometry. (Reprinted with permission. Copyright IEEE 2007 [14].)

by transmitting geometry related auxiliary information extracted/estimated at the encoder side [15], [14].

1.2.2.3 Video plus Depth

Since the depth perception of the conventional stereo video depends also on the display size and the observer's distance to display [19], a video plus depth representation of conventional stereo video is proposed in ATTEST project [20] for adaptive stereo video rendering. The target of the ATTEST project was to design a backward compatible, flexible and modular broadcast 3DTV system. The backward compatibility of the proposed system, which is standardized as MPEG-C part 3 and H.264 Auxiliary Picture Syntax, is implemented by transmitting the depth video as an auxiliary data for the conventional 2D video [21]. From monoscopic video and associated per pixel depth information, novel views of the 3D scene are synthesized by Depth Image Based Rendering (DIBR) methods in order to create the stereo view pairs. Hence, the view plus depth representation provides virtual stereo camera to be arranged optimally for the display and observer distance at the end user side. An instance from *Interview* (view plus depth) data is shown in Figure 1.6.

In addition to provide adaptive rendering capabilities at the end user side, the video plus depth data is a more compressible representation than the left-right pair of a stereo view. However, the weak point of the view plus depth representation is its proneness to rendering artifacts especially around the occlusion

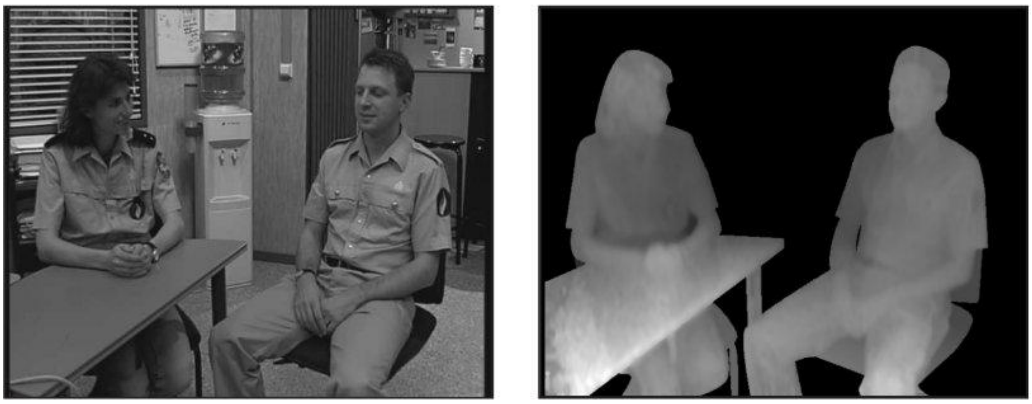


Figure 1.6: A video plus depth example.

regions. In stereo view synthesis, the camera position of the monoscopic video is assumed to be between the novel views. Hence, increasing the depth difference between the objects in the scene and increasing the novel camera distance to monoscopic camera enlarges the regions to be rendered in the novel views but occluded in the center view. In order to handle this occlusion problem, smoothing the depth maps is a common approach [22],[23]. However, depth smoothing still cannot avoid occlusion related rendering artifacts totally. Hence, the novel view rendering capabilities of the view plus depth representation are very limited due to the occlusion phenomenon.

1.2.2.4 Multiview Video plus Depth

In order to tackle the rendering limitations of the video plus depth format, utilization of multiple video plus depth data, which is called Multiview Video plus Depth (MVD), is proposed in the literature [24]. Since the occluded regions in one of the view are visible in some of the other views, high quality, occlusion free novel view rendering is possible by MVD format. By a proper arrangement of the camera rig orientation, the MVD representation is also capable of rendering the continuous trajectory of the novel views between the captured views, that seems to provide the desired interaction of 3DV applications.

ISO Moving Picture Experts Group (MPEG) standardization body issued a "Call for Proposals on 3D Video Coding Technology" document [25] in order to set the standard for the 3DV transmission, and the main exploration experiments of the proposal are performed in 2 and 3 video plus depth data sets. The exploration experiments show that the multiple depth information might provide high quality novel view rendering while keeping the number of videos as low as 2 or 3 [26]. Hence the 3DV proposal of the MPEG community is expected to provide increased rendering capabilities at low bit rates as shown in Figure 1.7. One of the main goals of the proposal is to decouple the rate and the number of output views, and hence, the rate required for transmitting the 3DV format could be based only on the transmission constraints [27]. The scene geometry information included in the 3DV representation decouples the bandwith requirements from the rendering capabilities.

A straightforward and backward compatible approach for the compression of MVD data is to compress the color/texture videos and depth videos independently by an MVC compression scheme. However, more efficient MVD compression methods are also proposed in the literature; some of these methods share motion vectors between texture and depth [28], transmits a base view and disoccluded texture regions [29], preprocess depth images [30], and encodes depth images with a non-DCT transforms [31].

The depth modality of the MVD data brings a new optimality question. The rate-distortion performance of the conventional video coding systems are usually measured using the Peak Signal to Noise Ratio (PSNR) metric [32]. However, the depth modality is an auxiliary information, and it is not observed directly by the end user. The depth maps of a MVD data are utilized in the DIBR module of a 3DV system; hence, a novel view rendering distortion based metric is the desired metric for the depth map coding of MVD data. There are novel view rendering distortion metric proposals in the literature for rate-distortion performance of depth coding [33],[34] and bit allocation between texture and depth coding [35].

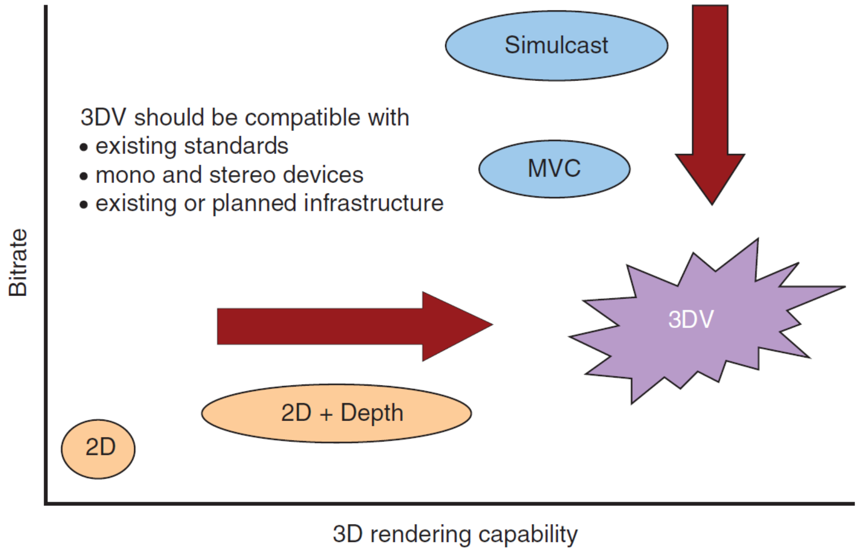

Simulcast
3DV should be compatible with
• existing standards
• mono and stereo devices
• existing or planned infrastructure
MVC
Bitrate
3DV
2D + Depth
2D
3D rendering capability

Figure 1.7: Requirments of the call for proposal on 3D Video coding by MPEG community (Reprinted with permission. Copyright John Wiley & Sons, Inc. 2010 [1]).

1.2.3 3DV Rendering

The aim of 3D video rendering is to provide high quality, natural views of the scene. The number of the views to be rendered might be one for displays of head tracking systems or around a hundred for autostereoscopic/holographic displays. In order to provide the scene views, 3DV systems utilize Image Based Rendering (IBR) techniques. In the literature, IBR is the general title of view rendering methods based on captured videos or photos. IBR techniques are classified as *pure image based*, *implicit geometry based* and *explicit geometry based* approaches in [3]. The tradeoff between the number of views and the geometry information for 3D scene representation again plays the main role in this classification. According to Figure 1.8, the IBR techniques utilizes less number of images towards the *more geometry* direction for high quality view rendering.

The rendering quality of the pure image based techniques are satisfactory with high number of 2D images but not practical for 3DV transmission systems. The methods in the implicit geometry category use the information of camera positions and epipolar relations between the views, but the rendering quality is not reliable, since underlying assumptions can be violated easily. The approaches, which utilize explicit geometry, are the most popular methods for 3DV systems to provide high rendering quality with a feasible transmission cost. Especially the Depth Image Based Rendering (DIBR) methods are under consideration for standardization within the 3D community [25]. In DIBR, the depth information

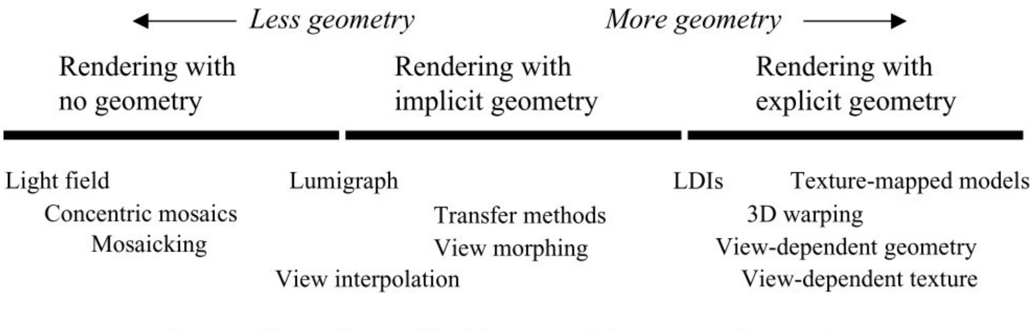


Figure 1.8: Geometry based classification of the IBR methods (Reprinted with permission. Copyright SPIE 2000 [3]).

about the views makes it possible to 3D warp the pixels to image plane of the desired novel view.

1.3 Problem Statement and Scope of the Thesis

The desired properties of the forthcoming 3D technology is to be realistic and interactive. These desired properties should be satisfied according to feasible infrastructures for a realization in the near future. Hence, for the 3D video applications, the main problem is to provide high quality 3D rendering capabilities to the end users through a limited capacity transmission channel. The MVD data format is the latest response of the 3D community for the forthcoming 3DV applications.

The multiview depth content and its DIBR based utilization are the main key topics for the exploration of the MVD format. The depth modality in MVD presents two fundamental differences against the video modality. The first one is the characteristics of a typical depth image which is much smoother than the characteristics of a typical conventional image. The intensity values of a depth image smoothly vary along a scene object due to the solid nature of the object. Hence, the depth images can be considered as piecewise continuous signals. However, the color intensities of an object might change abruptly in a small neighborhood. The second fundamental difference is the utilization of the depth modality in MVD is indirect in comparison to video modality. The depth images are not displayed to the end user of the 3D application but their DIBR results are displayed. The nonlinearities of the DIBR method makes the depth distortion analysis much more complicated than the conventional distortion analysis of a video.

These fundamental differences result in performance degradations, when the depth modality is handled in a manner similar to a conventional video. As an example, the depth compression experiments performed in [11] show that the DCT based depth compression deteriorates the depth boundaries and results in severe rendering artifacts. Hence, the depth modality of MVD should be reconsidered regarding its piecewise smooth properties and targeted 3D rendering applications. The quest for an appropriate depth image representation for efficient compression and high quality 3D rendering is an active research area. A novel stereo depth representation based on planar models and extends it to a planar layered MVD representation is proposed in this work. Although the applications of MVD data are the main focus of this work, the texture information of the MVD is beyond the scope of this thesis.

The original contributions of this thesis are: i) The depth images of the MVD data are handled in a unified 3D planar co-segmentation setup. ii) The proposed co-segmentation is formulated in an energy based approach which introduces the rate distortion objectives to the segmentation based depth representations in the literature. iii) An algorithm, which can be used in a lossy depth compression framework, is proposed to constrain the solutions of the planar representations. iv) The proposed planar representation and its acquisition algorithm are also utilized for a novel layered MVD representation which can decouple the texture and geometric distortions to a wide extent for novel view rendering applications.

1.3.1 Outline of the thesis

In Chapter 2, the proposed co-segmentation based representation of stereo depth images will be introduced. The energy based formulation of the co-segmentation problem will be stated by a Markov random field model. The optimization problem to obtain the maximum a posteriori estimate will be solved by a graph cut based expectation maximization like algorithm, which is slightly modified in order to exploit smooth depth image characteristics.

In Chapter 3, the energy based formulation of the planar representation will be reconsidered for rate distortion optimality properties. The concept of Pareto optimality will be discussed and a heuristic algorithm will be developed to obtain different Pareto optimal solutions which will make the planar depth compression realizable for different rate constraints. The pure planar and planar prediction with residual coding versions of the depth compression experiments will be conducted in comparison to state of the art solutions in the literature.

In Chapter 4, the multi-reference based MVD representation will be converted to a single reference based MVD representation with the guidance of planar representations obtained in the previous chapters. The compression performance of the novel layered representation will be studied in novel view rendering applications in comparison to state of the art MVD compression techniques. The different novel view rendering characteristics of the proposed MVD representation will be visually compared and discussed for possible depth perception distortion considerations in MVD compression.

Chapter 5 will summarize the proposed planar representation based stereo depth and MVD compression techniques with highlighting their advantages and disadvantages against other techniques in the literature. The conclusions of the thesis will be provided based on the novel contributions and their possible utilizations and advancements in future works.

CHAPTER 2

PLANAR STEREO DEPTH REPRESENTATION

The raw format of depth images is that of a single channel images whose intensity value encodes the depth of a 3D point in the scene. In general, single byte precision is used to represent the disparity between a stereo pair or the depth (z-value) of the point in 3D space. Range cameras and stereo disparity estimation techniques are well known depth image sources for 3D applications.

The objects in 3D space have continuous surfaces, in general. This property of a 3D object expresses itself in depth images as smooth variations inside the object region. However, sharp discontinuities might occur across the object boundaries. Therefore, depth images can be represented as piecewise smooth functions, in general.

In Section 2.1, the depth image representations in the literature for 3D applications will be briefly summarized. Then, the motivations of the proposed representation will be explained in Section 2.2, and in Section 2.3, its energy based formulation and solution will be explained in detail. The chapter will end with various examples of the proposed stereo depth representation in Section 2.4.

2.1 Depth Image Representations: Related Literature

The problem of depth image representations for 3DV applications can be rephrased as *"what is the optimal approximation model of the depth images for efficient compression and artifact free DIBR rendering?"*. While the piecewise smooth-

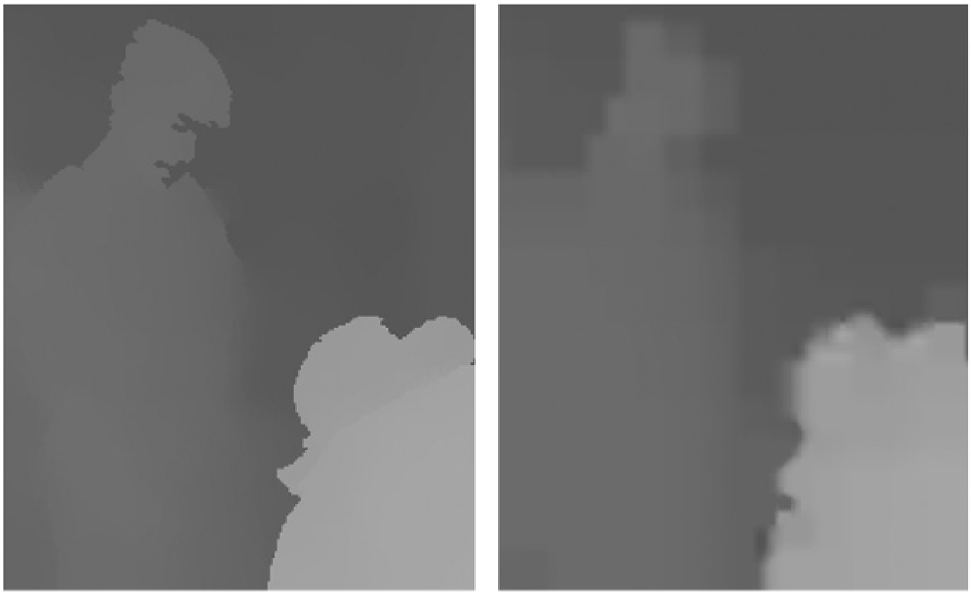


Figure 2.1: Original depth image (left), and its DCT-based approximation by H.264 (right). Note the blurring on the depth boundaries for the DCT-based approximation (Reprinted with permission. Copyright Elsevier 2008 [11]).

ness models for the depth images plays an important role in the predictability of the depth images, the artifact free DIBR rendering is usually obtained by sharp depth discontinuities at object boundaries [24]. As a conventional example, DCT based representation can exploit the piecewise smoothness of the depth images but can not conserve the sharp depth discontinuities as shown in Figure 2.1.

In [36] and [37], arbitrary shape adaptive lifting-based wavelet transforms are proposed for depth image approximation. The proposed lifting operations avoid filtering across the edges and decrease the number of significant high frequency wavelet coefficients around the edges. The reduction of the high frequency coefficients also reduces the ringing artifacts around the edges, and hence, preserve the sharpness of the edges. The experiments performed in [36] and [37] indicate that the bit-rate reductions maintained by the proposed lifting schemes are much more than the bit-rate required to encode the required edge information. An exemplary edge information of a depth map is shown in Figure 2.2. In [37], the filters used in lifting operations are designed to be optimal for piecewise planar images. A similar approach, which again requires edge information, is proposed in [38], and replaces DWT with a graph based transform. A recent study, advanced these shape adaptive DWT approaches to a scalable architecture both

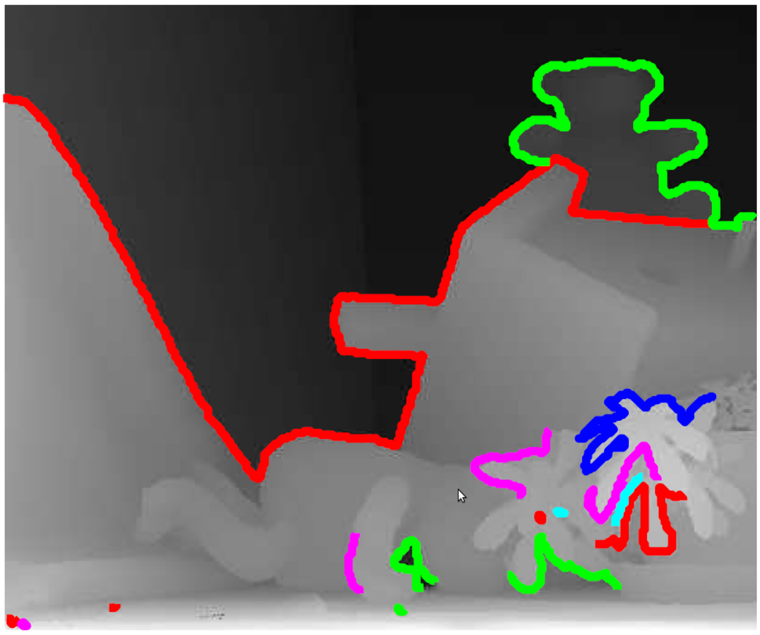


Figure 2.2: The edge information encoded by a Shape-Adaptive DWT approach (Reprinted with permission. Copyright IEEE 2008 [36]).

for depth and edge representation in a rate-distortion optimization friendly way [39].

In order to efficiently exploit the piecewise linear characteristics of the depth images, a platelet based transformation is proposed in [11]. A quad-tree guided refinement procedure is utilized for platelet based representation of the depth image and an exemplary representation is given in Figure 2.3. In [11], the performance of the platelet approach is also compared with the Intra mode of H.264 and H.264/MVC using PSNR values of depth images and rendered novel views. In these experiments, the platelet approach performed the worst on depth map reconstruction, whereas performed the best on novel view rendering. In [40], the platelet representation is extended with some contour prediction methods exploiting the neighboring context and corresponding edge contents of the color image.

In [41], the DCT transform based coding artifacts around the depth edges are proposed to be suppressed by a sparsity-based in-loop de-artifacting filter. The piecewise smoothness of the depth image characteristics are modeled by the assumption that the depth images are representable with a sparse set of coef-

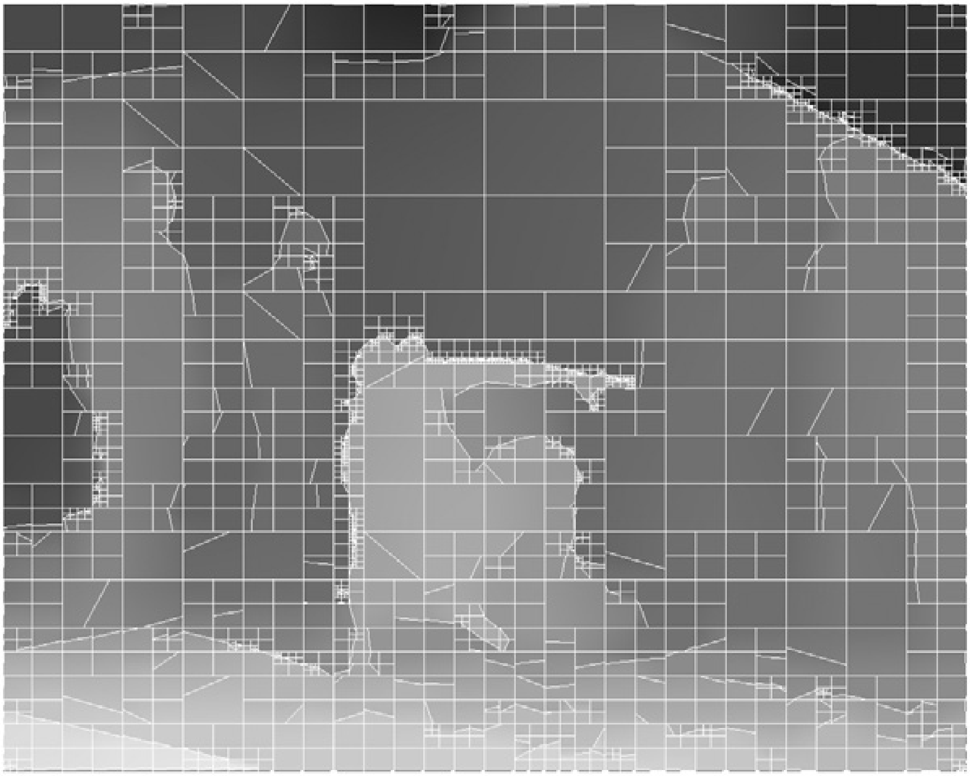


Figure 2.3: Platelet representation of a depth image (Reprinted with permission. Copyright Elsevier 2008 [11]).

ficients in an over-complete set of transforms. The coefficients of the decoded depth image in the over-complete set of transforms are thresholded for denoising at first. Then, the final depth image is obtained by a weighted combination of all denoised depth images in each transform domain. The weighting operation favors the sparser representations among the denoised depth images. The experiments in [41] indicate that the proposed sparsity-based in-loop de-artifacting filter both increases the PSNR performance of the reconstructed depth image and the rendered novel views.

Segmentation based representations for depth images are also proposed in the literature as shown in Figure 2.4. In [42], the depth image is segmented into a desired number of regions and each region is represented with its shape and mean depth value. A down-scaled proxy of the residual depth image is used to represent the depth variations in each region. In a similar fashion, the method in [43] represents each segmented region with its shape and a linear depth model. As an extreme example of segmentation, a piecewise constant model is utilized

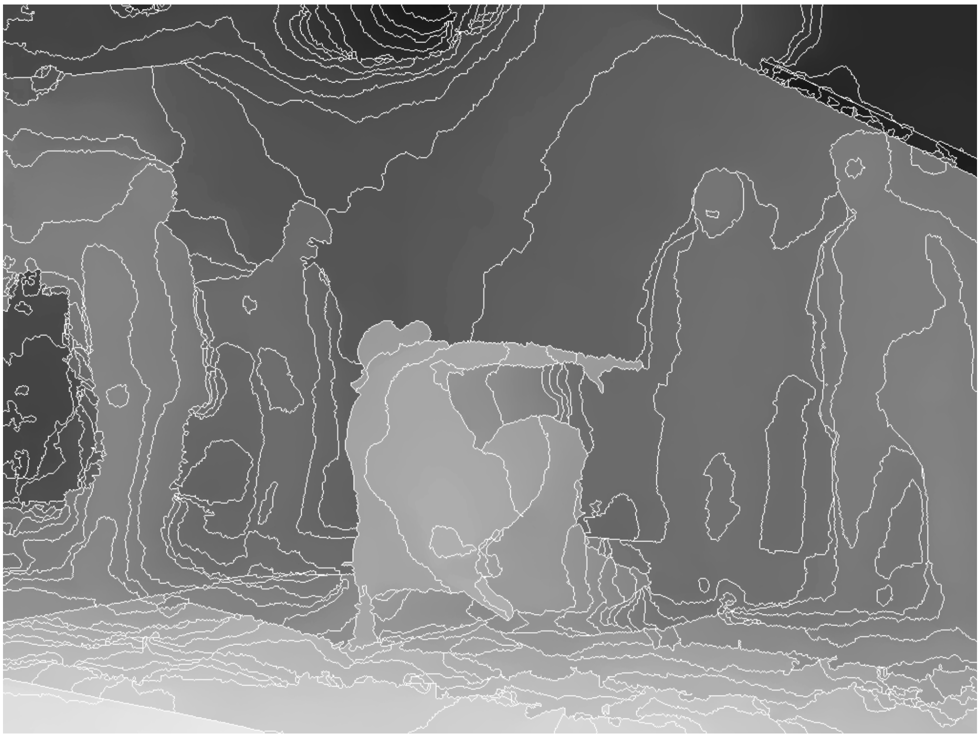


Figure 2.4: An example for segmentation based depth image representation (Reprinted with permission. Copyright IEEE 2011 [43]).

in [44] for lossless compression of depth images.

The depth image representations in [45] and [46], mimics the piecewise smooth characteristics in a diffusion scheme. Regularly or wisely sampled sparse set of depth values are densified to cover the whole depth image according to a diffusion equation. In order to avoid diffusion across the depth boundaries, region shapes or edge information has again been included in the representation.

2.2 Motivation for the Proposed Representation

All the aforementioned depth image representations in the literature concentrate on conserving the sharp depth discontinuities for high quality DIBR renderings. While some of them use blocks as representation units to be compatible with the conventional video coding standards, such as [11], some others prefer arbitrary shaped regions as the representation units, as in [43]. One of the ultimate motivations of the proposed depth image representation is to utilize scene objects as representation units. The object based approaches in video coding is known to have broad possibilities in interaction and manipulation while maintaining an efficient compression [47]. However, the extraction of semantically meaningful video object segments is still an open research area. Fortunately, the depth modality in MVD format provides valuable information related with the main subject of the depth representation problem; i.e. the depth discontinuities along the object boundaries. Hence, the thesis argues that the depth representation units should coincide with the semantic object boundaries as much as possible to provide novel potential 3D applications and interactions.

Piecewise smoothness of the depth images is mostly exploited with constant or planar models in the literature. Planar models are also widely used for object segmentation and stereo reconstruction in computer vision [48],[49],[50],[51],[52]. Relying on these studies, the proposed representation should make a planarity assumption on the scene objects in order to obtain object-like depth representation segments. While the aforementioned depth representations utilized the planar depth modelling in 2D image plane, the proposed approach should utilize a planar modelling in 3D space where the scene objects exist. According to the pinhole camera model, the correspondence between the image plane point and the 3D space point is a non-linear relation. Hence, the proposed representation slightly differs from others and has a motivation to represent depth images in a 3D enviroment.

Except a recent study [39], the aforementioned depth representations, which explicitly extract the depth discontinuities to be preserved, achieve the depth boundary extraction in a pre-processing step. In general it is difficult to formulate the rate-distortion objectives in such preprocessing steps. Hence, this kind of disjoint designs makes any optimality arguments questionable. The proposed approach also aims to unify the rate-distortion objectives with the extraction of object-like representation units.

2.3 Energy Based Formulation of the Proposed Representation

Without loss of generality, the depth representation problem will be considered for the MVD case throughout the rest of the thesis. The raw input format of the depth modality is assumed to be multiview depth images. As long as the camera calibration and the depth image of a view are known, the depth values for the image pixels can be back-projected to 3D space. The proposed planar representation aims to fit 3D planes to this given 3D point cloud, which is obtained by back-projecting the pixels of all views. The planar model assignment to a 3D point also associates that geometric model with a 2D image pixel by the one-to-one correspondence between 3D points and image pixels. Hence, the proposed planar model assignments can also be considered as co-segmentation of multiple depth images, i.e. provide joint segmentation and parameter estimation.

This co-segmentation problem is formulated in an energy minimization framework. Three main cost terms are utilized in order to satisfy three objectives. The first one is the reconstruction error of the depth images which is called as *data cost term*, i.e. the geometric distortion of the representation. Minimizing the geometric distortion increases the loyalty of the representation to raw data and indirectly increases the rendering quality. The second cost term, which is called *smoothness cost term*, favors the same planar model assignments in every local neighborhood of the image. Minimizing smoothness term provides smooth shaped and well-connected planar assignments. Finally, the last cost term is the *label cost term* which favors the use of a minimum number of planar models in the representation. Utilization of the *minimum description length principle* [53] helps in clustering the similar planar models under a single assignment which will hopefully coincide with scene object geometries.

The last two cost terms, smoothness and label terms favor the representation to be efficiently compressible by enforcing a minimum number of models in well-shaped regions. Combination of these terms with the minimum geometric distortion term, provides the desired tradeoff between the rate and distortion for the proposed representation [54].

The mathematical formulation of these cost terms is designed as a Markov Random Field (MRF) which is an effective way of modelling spatial dependencies in images. The overall energy to be minimized for planar representation of depth images is given as,

$$E(f) = \sum_{p} \mathcal{D}(f_p) + \sum_{p,q \in \mathcal{N}} \mathcal{V}(f_p, f_q) + \sum_{m \in \mathcal{M}} \delta_m(f). \tag{2.1}$$

The details of Markov random field, f, is introduced in the next section.

2.3.1 MRF Modelling

Let $\mathcal{M}$ be the set of all possible planar models in 3D space, and m_i be an arbitrary indexing for the elements of $\mathcal{M}$. According to the proposed depth representation, every pixel of a depth image, p, will be assigned to a planar model $m_{i(p)}$. Then, f becomes a labelling image whose pixel values, f_p, are the index values of the planar model assignments, $i(p)$. Let the observed depth values of each pixel be denoted as d_p and $\mathcal{N}$ be a neighborhood relation over the pixels.

By the MRF model, f_p variables are regarded as a family of random variables whose joint probability density function is defined according to total clique potentials as,

$$p(f) = \frac{e^{-E(f)}}{Z} \tag{2.2}$$

$$\text{where} \qquad Z = \sum_{f \in \mathcal{F}} e^{-E(f)} \tag{2.3}$$

is called *partition function*, which normalizes the joint probability function to sum up to 1. According to (2.2), the configurations with smaller total clique potentials are more probable.

The unary and pairwise potentials are the most widely used clique potentials, since there are many mature inference techniques for these formulations. By the recent tools developed for the MRF inference problem, higher order potentials are also utilized in order to enrich the relations in the model [55]. The factor

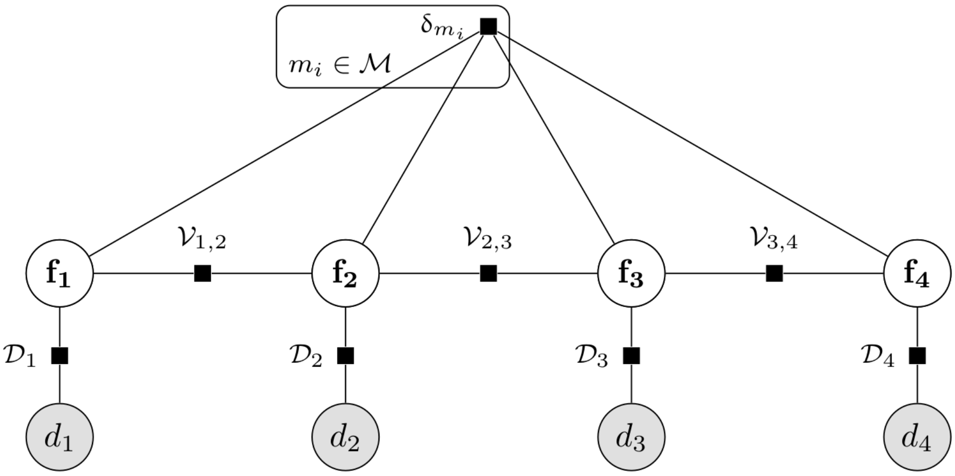


Figure 2.5: Factor graph representation of the MRF model.

graph representation of the proposed representation's MRF model with unary, pairwise and higher order potentials, noted as $\mathcal{D}, \mathcal{V}, \delta$ respectively, is given in Figure 2.5. The resultant posterior probability density function of the field in factorized form is also given as,

$$p(f|d) = \frac{1}{Z(d)} \prod_{p} e^{-\mathcal{D}_p} \prod_{p,q \in \mathcal{N}} e^{-\mathcal{V}_{p,q}} \prod_{m_i \in \mathcal{M}} e^{-\delta_{m_i}} \ , \tag{2.4}$$

$$= \frac{1}{Z(d)} \ e^{-\sum_p \mathcal{D}_p - \sum_{p,q \in \mathcal{N}} \mathcal{V}_{p,q} - \sum_{m_i \in \mathcal{M}} \delta_{m_i}} \ , \tag{2.5}$$

$$= \frac{1}{Z(d)} \ e^{-E(f,d)} \ , \tag{2.6}$$

where $$E(f,d) = \sum_{p} \mathcal{D}_p + \sum_{p,q \in \mathcal{N}} \mathcal{V}_{p,q} + \sum_{m_i \in \mathcal{M}} \delta_{m_i} \ . \tag{2.7}$$

According to the pinhole camera model, a labelling value f_p, defines a depth value, $\hat{d}_p$, for the pixel p by intersecting the projection ray of the pixel with the plane m_{f_p} in 3D space. With respect to this definition, the unary potential of a pixel, $\mathcal{D}$ is set to distortion of the representation as,

$$\mathcal{D}_p = \lambda_{\mathcal{D}} |d_p - \hat{d}_p| \ . \tag{2.8}$$

The pairwise potentials, given in (2.9), are designed as Potts model which favors the piecewise constant configurations, as,

$$\mathcal{V}_{p,q} = \begin{cases} 0, & \text{if } f_p = f_q \\ \lambda_{\mathcal{V}}, & \text{otherwise} \end{cases} . \tag{2.9}$$

With this potential, the model differences regarded identically without considering the similarity of planar models. The higher order potentials, δ_{m_i}, behave like binary flags for the existence of an assignment of the planar model, m_i, in the representation, as,

$$\delta_{m_i} = \begin{cases} \lambda_{\delta}, & \text{if } \exists p : f_p = i \\ 0, & \text{otherwise} \end{cases} . \tag{2.10}$$

Hence the sum of all higher order potentials is linearly proportional to the number of planar models utilized in the representation. For a given field configuration, the utilization of a planar model can be evaluated by checking all the random variables of the field. The potential's dependency to all field variables makes it a higher order one.

Although the geometric distortions, modeled in unary potentials, result in rendering artifacts indirectly due to DIBR techniques, the rendering distortions are discarded in the model for simplicity and generality of the depth representation. In a model which considers including the color images of the views, the unary potentials can be improved by rendering distortions.

The pairwise and higher order potentials are defined to be homogenous, i.e. they do not change spatially or according to a specific planar model. In case any a priori information is known about the depth boundaries or the planar geometry of the scene, they can be integrated into the formulation by adapting corresponding potentials.

Although the illustration in Figure 2.5 is one dimensional for visualization concerns, the neighborhood system for pairwise potentials, $\mathcal{N}$, is constructed in 3D space. Let $W_{i,j}$ be the 3D warping function, which maps a pixel on i^{th} view to its stereo correspondence in j^{th} view. Assume the views, I_i, are indexed according to their positions from left to right. The neighborhood of a pixel, $\mathcal{N}_p$ is defined as,

$$\mathcal{N}_p = \mathcal{N}_p^8 \cup \{W_{i,i+1}(p),\ W_{i-1,i}^{-1}(p)\,|\,p \in I_i\}\ , \tag{2.11}$$

where
$$\mathcal{N}_p^8 = \{q\,|\,p, q \in I_i,\ \|p-q\|_\infty \leq 1,\ p \neq q\}\ , \tag{2.12}$$

i.e., the $\mathcal{N}_p^8$ is the well-known 8-neighborhood of the pixel p on its image plane. The second set on the right hand side of (2.11) defines the 3D neighborhoods between views; this is crucial to obtain a coherent co-segmentation of depth images for multiviews. An illustration of this neighborhood is presented in Figure 2.6.

2.3.2 Optimization of MRF Energy

According to given depth images, d, the most probable configuration of the MRF, f, is the Maximum A Posterior (MAP) estimate, as,

$$\begin{aligned} f^* &= \arg\max_f p(f|d)\ , & (2.13)\\ &= \arg\max_f \frac{1}{Z(d)}\, e^{-E(f,d)}\ , & (2.14)\\ &= \arg\min_f E(f,d)\ . & (2.15) \end{aligned}$$

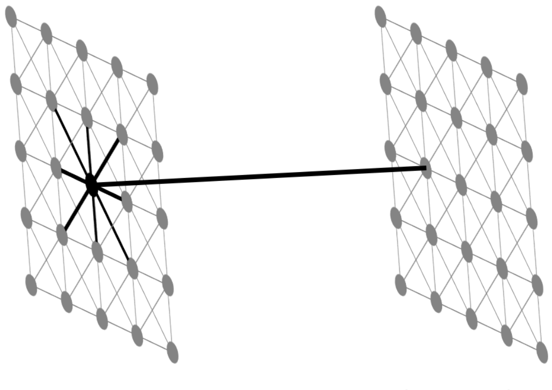


Figure 2.6: The bold and dark connections represent the 3D neighborhood of a pixel according to the MRF model.

To find the MAP estimate for the MRF model becomes an optimization problem to minimize the energy terms given in (2.7). Exact solutions for some special cases and approximate solutions are available in the literature. Greedy algorithms [56],[57],[58] Viterbi-like message passing algorithms [59] and variational approaches [60] are some of them worth mentioning. Among them, the graph cut (GC) algorithm is the dominant approach in the computer vision community due to its efficiency [55]. The fundamentals of the GC algorithm is briefed in Appendix A.

2.3.3 Continuum of labels

Assume the origin of the 3D space is set as one of the camera center of a view in the MVD dataset; this is a widely used convention. All the planar surfaces in the field of view of this camera projects onto an area of its image plane. The only exception is the planar surfaces crossing the origin, that are projected to a line on the image plane. Discarding such cases as they do not have an integrable area of the image plane, the set of planar models, $\mathcal{M}$, to be utilized in the proposed depth representation can be parametrized as vectors in $\mathbb{R}^3$ as,

$$\mathcal{M} = \{ m = (a, b, c) \mid ax + by + cz + 1 = 0 \} \ . \tag{2.16}$$

Ideally, the energy of the MRF model given in (2.7) should be solved for all possible planar models, m in $\mathcal{M}$. However, as an efficient MRF energy minimizer, graph cut is a combinatorial algorithm which works on a finite set of labels (models in the current case). Hence the continuum of the model parameters should be efficiently explored for the proposed MRF-based approach.

In [61], Isack and Boykov introduced an energy minimization based geometric multi-model fitting algorithm, which is called *Propose Expand And Re-estimate Labels* (PEARL). Different than other multi-model fitting algorithms in the literature, such as [62],[63],[64]; PEARL algorithm simultaneously assigns and prunes models by minimizing a MRF-based energy with label costs similar to (2.7). The continuum of the parameter space is sampled efficiently in Expectation-Maximization (EM)-like cycles. The pseudo code of PEARL algorithm is given

Algorithm 2.1 PEARL algorithm [61]

```
Given a dataset, $d$, on a field, $f$, with a neighborhood system, $\mathcal{N}$.
    Propose:
 1: At initialization, set $i$=0
 2: Sample initial set of models, $\mathcal{M}_0$, by fitting geometric models to randomly
    sampled data points, $d_p$'s
 3: (optional) Add a model, $\emptyset$, to represent outliers
 4: *(optional for $i>0$) Sample more or merge/split current models in $\mathcal{M}_i$
    Expand:
 5: Solve the model assignment problem by running $\alpha$-expansion for the energy
    given in (2.7) and for $\alpha \in \mathcal{M}_i$
 6: If the energy does not decrease, stop
    Re-estimate Labels:
 7: Update the inlier model $m \in \mathcal{M}_i$ with the one minimizing the fitting error
    to data points assigned to that model
 8: Discard all models with no inlier assignments
 9: Set $i = (i + 1)$, go to step 2 (or optional to *)
```

in Algorithm 2.1.

The energy based formulation in the expand step of PEARL avoids the consecutive assignments and makes a competition among the available models to enlarge their support regions on the given dataset. In comparison to consecutively assigning models to remaining outlier dataset like in [65], the unified approach is more robust to noisy datasets as shown in [61]. Different than the mixture of models and K-means algorithms, the spatial and global regularization cost terms in the energy formulation also handles the number of models to be utilized in the solution, and prunes the set of models inherently. The PEARL steps of expand and re-estimate labels are the analogues to the expectation and maximization steps of the EM algorithm, respectively.

The original PEARL algorithm is designed to handle model fitting problem under noisy datasets. It starts with an excessive number of models in the initial set and updates the ones used in the inlier assignments. The models, which labeled no inlier data point, are discarded for the next cycles. Hence, the number of models in the candidate model set of each iteration, $\mathcal{M}_i$, has a trend to decrease in the original PEARL algorithm [61].

Algorithm 2.2 Modified PEARL algorithm to fit planar models to depth images.

```
Given a dataset, $d$, on a field, $f$, with a neighborhood system, $\mathcal{N}$.
    Propose:
1:  At initialization, set $i$=0
2:  Sample initial set of models, $\mathcal{M}_0$, by fitting geometric models to randomly
    sampled data points, $d_p$'s
    Expand:
3:  Solve the model assignment problem by running $\alpha$-expansion for the energy
    given in (2.7) and for $\alpha \in \mathcal{M}_i$
4:  If the energy does not decrease, stop
    Re-estimate Labels:
5:  Update the inlier model $m \in \mathcal{M}_i$ with the one minimizing the fitting error
    to data points assigned to that model
6:  Discard all models with no inlier assignments
7:  Add a model minimizing the fitting error to each connected regions according
    to current labeling of the field
8:  Set $i = (i + 1)$, go to step 2
```

The PEARL algorithm is modified for the proposed planar depth representation in order to efficiently exploit the characteristics of the problem. In the proposed planar representation of depth images, the given depth dataset is considered to be noise free with a data loyal perspective (although the conventional depth acquisition methods as mentioned in Section 1.2.1 might have various and specific noise characteristics, they are out of the scope of the thesis). Hence, the outlier model is discarded in the modified PEARL algorithm. In order to find good models for every depth data point, the trend to decrease in the number of candidate models is changed into a trend to increase during iterations.

The new planar models are appended to the candidate set at the end of each iteration until the fitting error is smaller than some predefined threshold or the number of the models in the candidate set hits the maximum allowed value. Different than the original PEARL algorithm, new models are not sampled randomly, but by exploiting the spatial smoothness of the depth modality.

Since planar models and depth images are spatially smooth, the fitting error for a given model should also be spatially smooth. According to this observation, in case of no proper planar model is available in the candidate set for a regular surface in the scene, the corresponding model assignments can not change abruptly due to regularization terms of the utilized energy formulation. This situation means that erroneous model assignments for depth images are also spatially smooth. In order to sample good models for erroneous regions, the connected components of the MRF field with respect to the given model assignment can be utilized efficiently.

The pseudo code of the modified PEARL algorithm to obtain planar representation of depth images is given in Algorithm 2.2. The details and the progress of the algorithm are explained with examples in the next section.

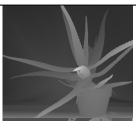
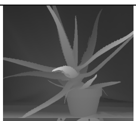
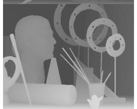
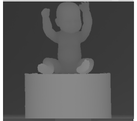
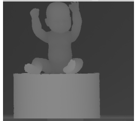
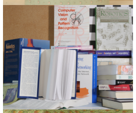
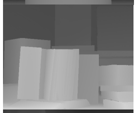
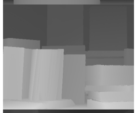
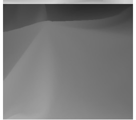
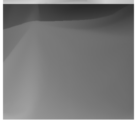
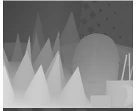
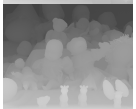
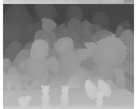
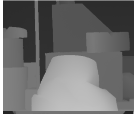
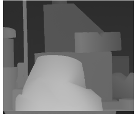
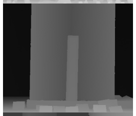
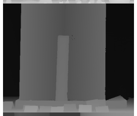
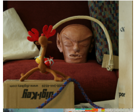
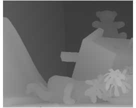
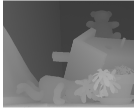
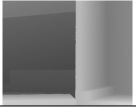
2.4 Planar Model Fitting Examples

In the planar representation experiments, the *Middlebury* stereo dataset [66] is utilized. *Middlebury* dataset is composed of multiples of horizontally shifted camera views. The disparity maps of the two views are also provided in each set. Since the internal and external camera parameters are not provided, a generic camera parameter construction method is utilized (see Appendix B) to create the corresponding 3D coordinate system. *Middlebury* dataset utilized in the experiments are summarized in Table 2.1.

For planar model fitting, the energy function given in (2.7) is minimized by the modified PEARL algorithm. All experiments are initiated by eight 3D planar models. For each of them, randomly sampled triplets from the 3D point cloud of the dataset is used to determine their model parameters. As long as the sampled triplets are not collinear in 3D space, they define a plane equation.

The progress of the planar models utilized in the representation is shown for the *Moebius* dataset in Figure 2.7 where the two columns in the middle are the planar model labeling (assignment) images of the stereo depth pair and their corresponding planar reconstruction are given on the left and right columns. Three of the randomly sampled planar models are utilized in the first iteration and corresponding connected components of the labelings initiated the planar model sampling properly. The number of the planar models utilized in the planar

Table 2.1: *Middlebury* dataset [66].

Name	Left view	Right view	Left depth	Right depth
Aloe				
Art				
Baby				
Books				
Cloth				
Cones				
Dolls				
Lampshade				
Laundry				
Moebius				
Monopoly				
Plastic				
Reindeer				
Teddy				
Wood				

representation given in Figure 2.7 increases up to 15 and finally converged to a solution with 14 planar models. The mean PSNR of the reconstructed depth images for the given example is 36.45dB. The labeling images obtained from the stereo pair images are quite coherent due to co-segmentation like MRF based modelling which defined the pairwise neighborhoods of the random variable over the point cloud in 3D space.

The effect of cost weighting values for the acquired planar representation is given in visually and numerically in Figure 2.8 and Table 2.2, respectively. The results show that by tuning the weighting factors of the data, smoothness and label cost terms, planar approximations with different number of models at different reconstruction quality can be obtained. The comparative analysis of the solutions visually and numerically shows that increasing the weight of the data cost term decreases the reconstruction error and increases the number of planar models utilized. Increase in the weight of the regularization costs, i.e. smoothness and labelling costs, results in increased reconstruction error and decreased number of planar models. Hence, the introduced Algorithm 2.2 can be regarded as a satisfactory realization of the energy-based formulation of the planar representation of the stereo depth images.

The local and global characteristics of the smoothness and labeling costs can also be distinguished by the given results. In comparison to solution in Figure 2.8e, the solution in Figure 2.8d is obtained by increasing the weights of data and labeling costs 10 and 100 fold, respectively. Figure 2.8d has fewer number of planar models that shows that the increase in labelling cost is effective in decreasing the number of planar models. In addition to this increase in global smoothness the decreasing of the reconstruction error is also achieved with the higher data cost weight. However, this result is obtained by the diminishing effect of smoothness costs that resulted in a pathcy labeling image due to weak local smoothness constraints.

The planar representation examples for the left view of the *Middlebury* dataset is given in Figure 2.9 and their corresponding labelling images are given in Figure 2.10. Based on the given solutions, the proposed planar representation can be regarded as capable of extracting the descriptive object boundaries in general. *Plastic* and *Wood* datasets are almost perfectly recovered by the planar models, since the scene mostly consists of planar objects. However, the *Dolls* and *Cloth* datasets are not efficient examples of planar representation, since labeling images misses some descriptive object boundaries or introduces artificial boundaries.

Overall the proposed MRF energy based solution to planar representation of stereo depth images is effective in depth representation with clear object boundary definitions. The energy based formulation is responsive to obtain planar representations with specific properties by manipulating the cost weights. These aspects of the proposed planar approach will be utilized in the forthcoming chapters for an efficient depth compression method.

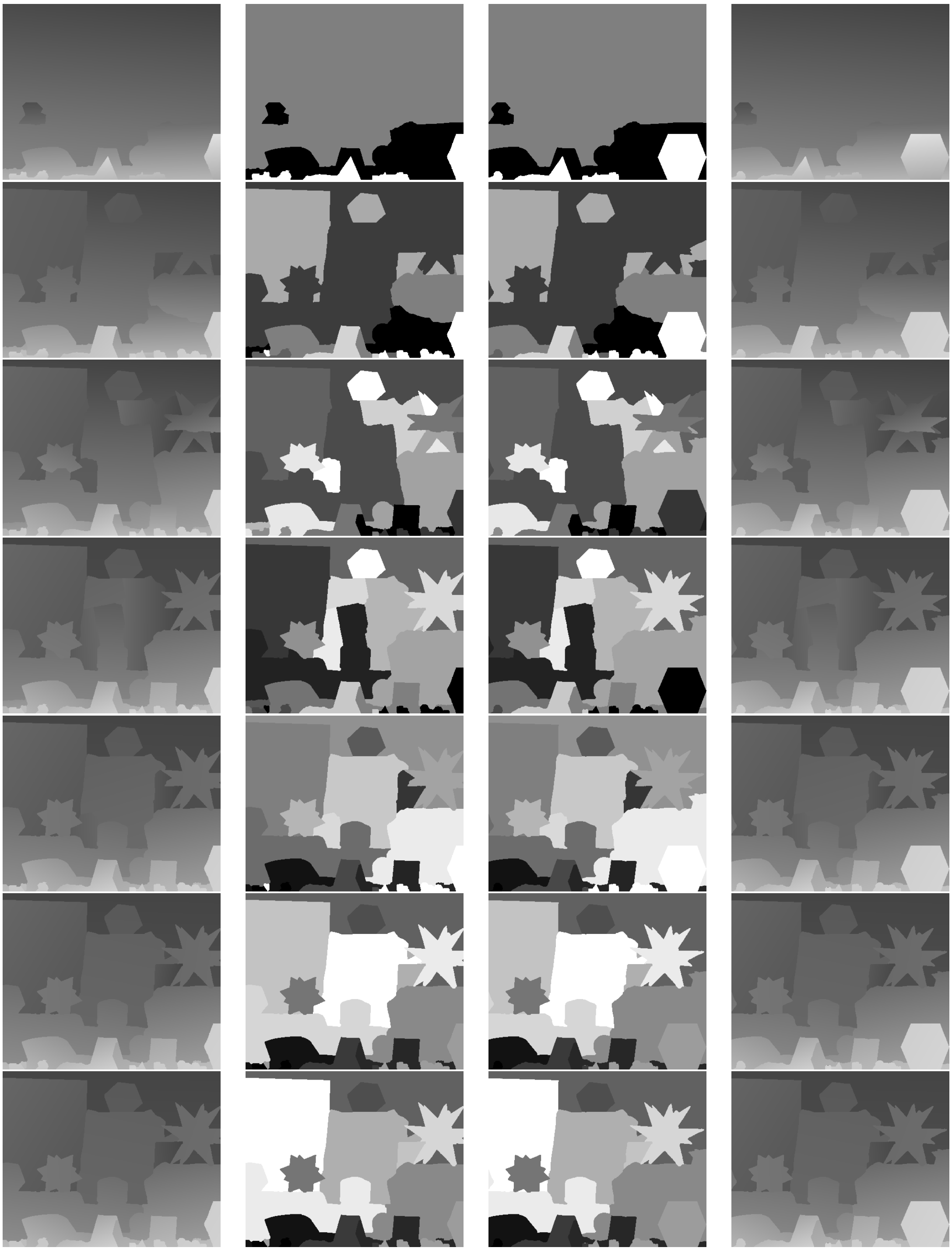
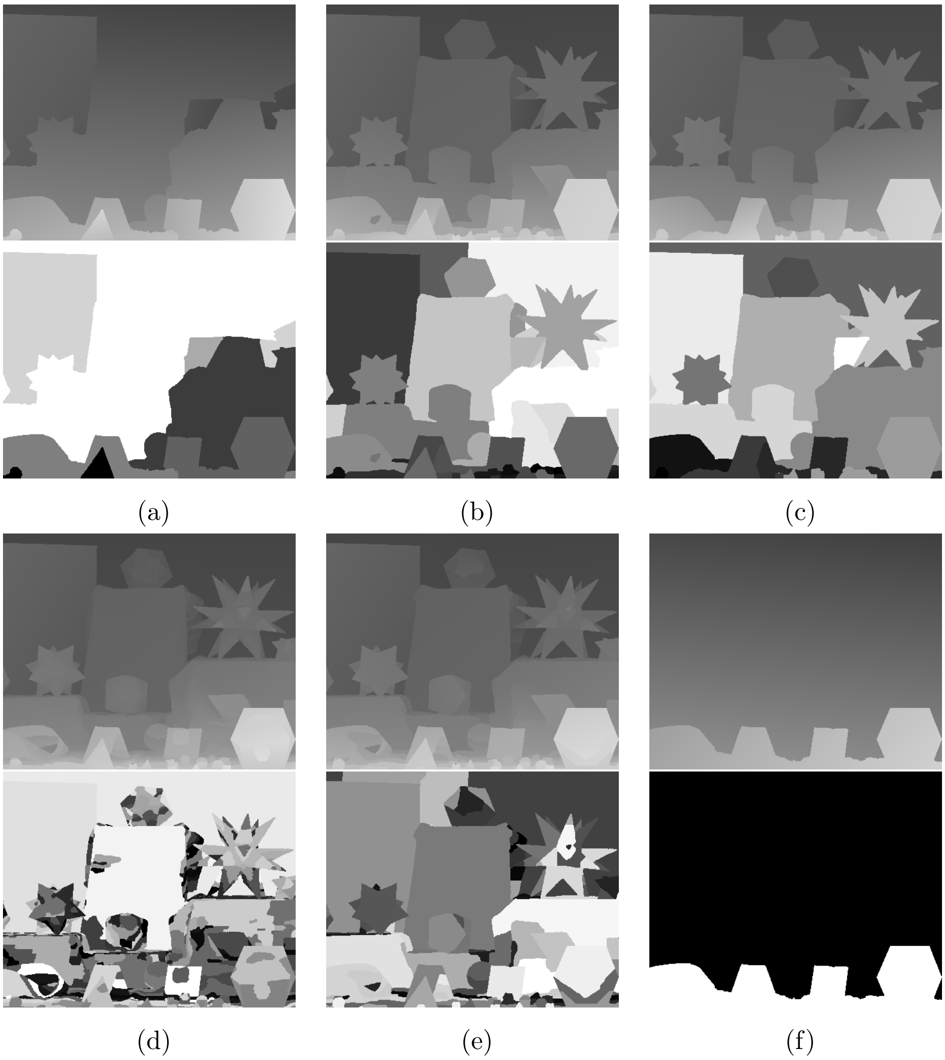


Figure 2.7: The progress of the modified PEARL algorithm in fitting planar models to *Moebius* dataset. Left and right columns are the planar depth reconstruction results according to the labeling images given in the middle columns. The number of planar models increase from top to bottom.



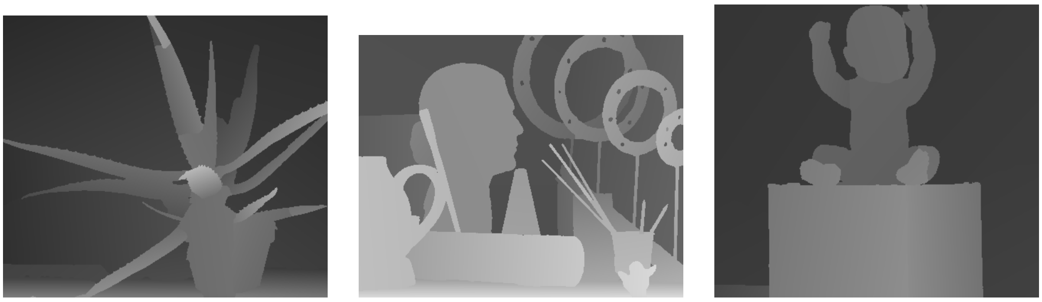
(a) (b) (c)

(d) (e) (f)

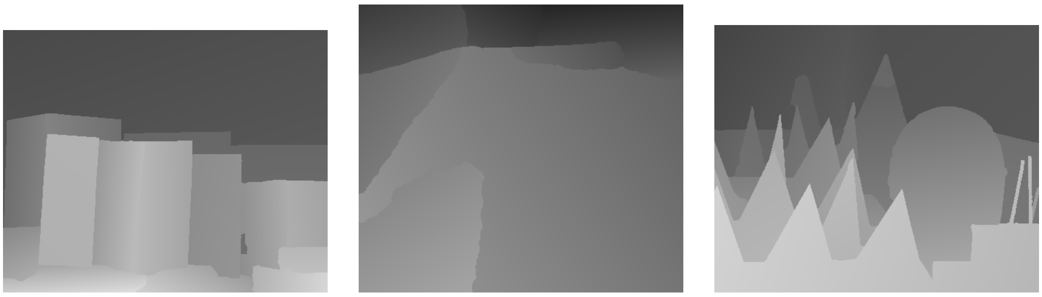
Figure 2.8: The planar solutions obtained for *Moebius* dataset by different cost weighting combinations. For each sub figure, the planar reconstruction is given at the top and its labelling image is given at the bottom. See Table 2.2 for their numerical details.

Table 2.2: The number of models and reconstruction accuracies of the depth images for various cost weightings.

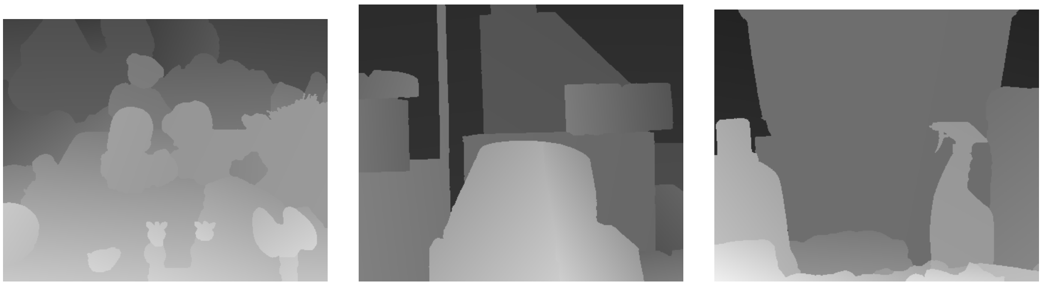
$\lambda_{\mathcal{D}}$	$\lambda_{\mathcal{V}}$	λ_{δ}	# Planar Models	PSNR (dB)	Figure
10	20	10^4	7	29.95	2.8a
20	10	10^4	23	39.40	2.8b
20	20	10^4	14	36.45	2.8c
100	1	10^6	26	46.27	2.8d
10	1	10^4	36	45.05	2.8e
10	20	10^5	2	26.64	2.8f



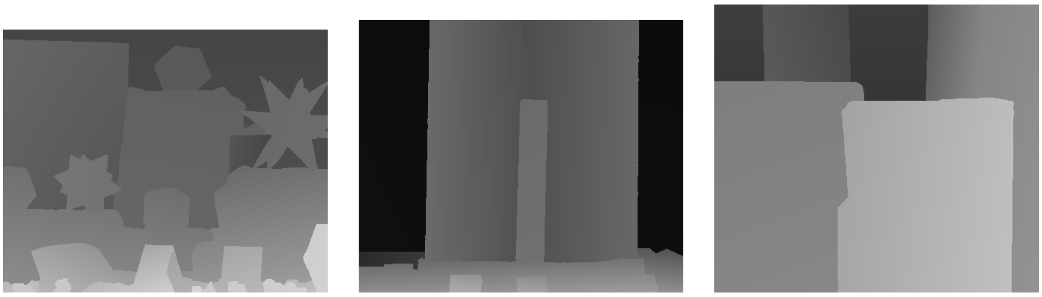
(a) Aloe (b) Art (c) Baby



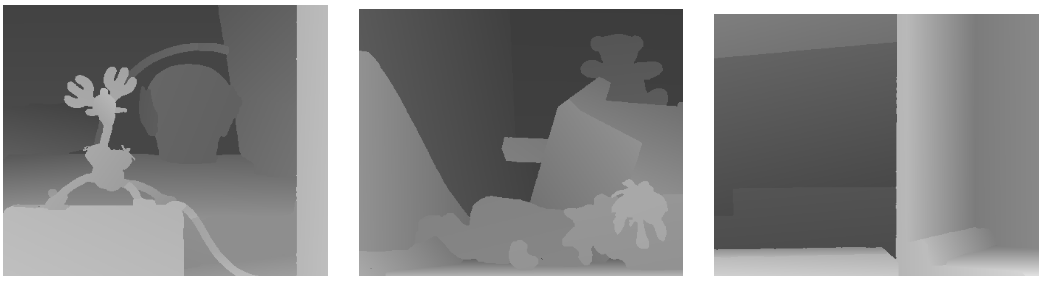
(d) Books (e) Cloth (f) Cones



(g) Dolls (h) Lampshade (i) Laundry

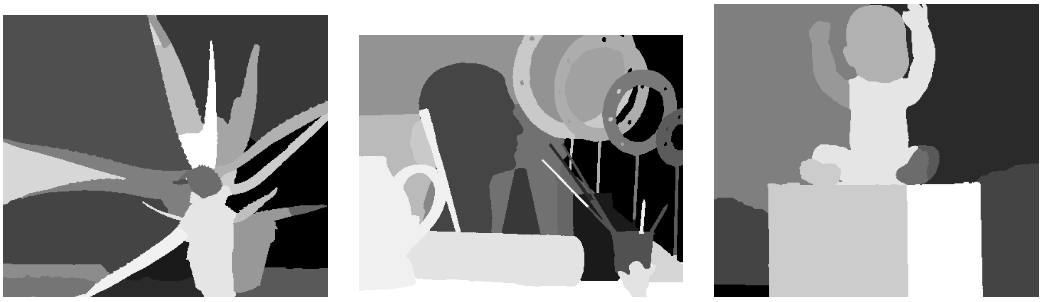


(j) Moebius (k) Monopoly (l) Plastic

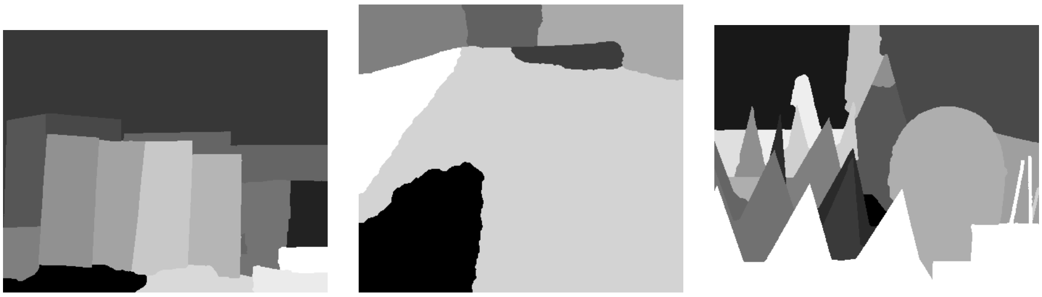


(m) Reindeer (n) Teddy (o) Wood

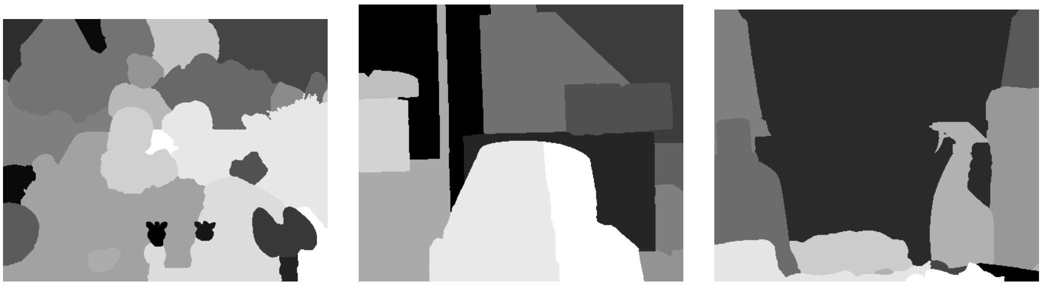
Figure 2.9: The planar depth image reconstruction examples of *Middlebury* dataset.



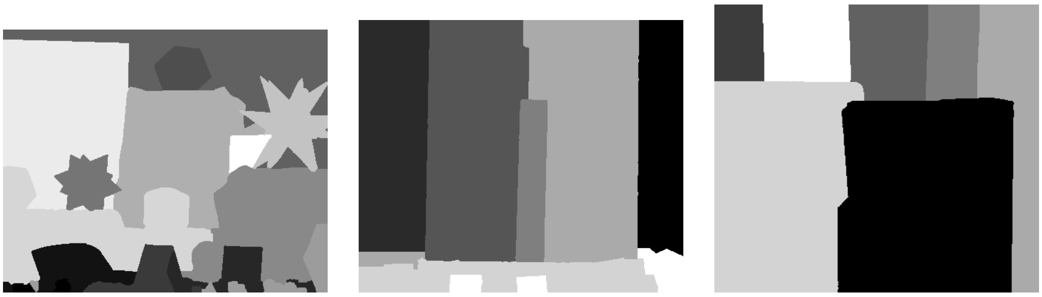
(a) Aloe (b) Art (c) Baby



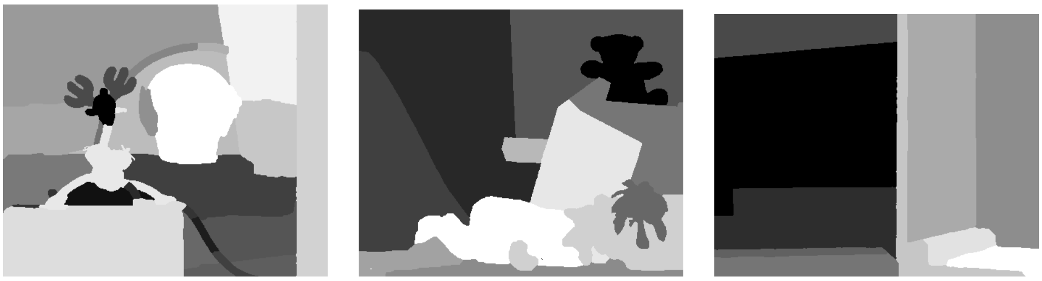
(d) Books (e) Cloth (f) Cones



(g) Dolls (h) Lampshade (i) Laundry



(j) Moebius (k) Monopoly (l) Plastic



(m) Reindeer (n) Teddy (o) Wood

Figure 2.10: The planar model labeling images of the reconstructed depth images given in Figure 2.9.

CHAPTER 3

STEREO DEPTH COMPRESSION BASED ON PLANAR REPRESENTATION

MVD data format for the forthcoming 3D applications made the depth compression problem for high quality novel view rendering a recent research area. The different statistical characteristics and DIBR-based utilization of the depth data brought unconventional image/video coding approaches to the literature. Non-rectangular or even arbitrary shaped coding units are utilized in the depth image representation as mentioned in Section 2.1. All these representations are motivated by the paramount observation that for an efficient depth compression with high quality novel view rendering results, sharp depth discontinuities at object boundaries should be preserved [11].

The introduced planar representation in the previous chapter will be considered as a depth compression tool in this chapter. The next section will introduce an MRF energy design method for a representation with the desired number of planar models. Then the obtained planar reconstruction of the depth image will be used as a depth prediction method and a residual coder will be integrated into the proposed depth coding approach. Lastly the compression experiments will be provided in comparison to conventional state of the art image/video coding methods.

3.1 Energy Design for Compression Applications

The proposed planar representation of depth images can be regarded as a depth compression tool. As long as the camera calibration of the 3D setup is available at the receiver side, the planar approximations of the depth images can be transmitted by encoding all the utilized planar model parameters and their corresponding labeling images. While the planar approximation introduces the distortion of the proposed lossy depth compression, bits required to encode the planar model parameters and their labeling images become the rate of the compression.

The planar representation experiments in Section 2.4 showed that the λ values determine the main characteristics of the obtained solution by weighting the data fitting and regularization energy terms. The MRF energy formulation given in (2.7)-(2.10) is rewritten below as a weighted summation of energy terms by emphasizing the weighting coefficients of the energy terms:

$$
\begin{aligned}
E(f,d) &= \lambda_{\mathcal{D}} \sum_{p} \mathcal{D}_p + \lambda_{\mathcal{V}} \sum_{p,q \in \mathcal{N}} \mathcal{V}_{p,q} + \lambda_{\delta} \sum_{m_i \in \mathcal{M}} \delta_{m_i} \ , && (3.1) \\
\mathcal{D}_p &= |d_p - \hat{d}_p| \ , && (3.2) \\
\mathcal{V}_{p,q} &= \begin{cases} 0, & \text{if } f_p = f_q \\ 1, & \text{otherwise} \end{cases} , && (3.3) \\
\delta_{m_i} &= \begin{cases} 1, & \text{if } \exists p : f_p = i \\ 0, & \text{otherwise} \end{cases} . && (3.4)
\end{aligned}
$$

The well known rate-distortion tradeoff in lossy data compression presents itself as a tradeoff between the data costs and regularization costs, smoothness and label costs, in the proposed MRF based formulation [54]. In order to decrease the distortion of the planar approximation, the planar model assignments should be specialized to each region; this may result in utilization of more planar models and locally more dynamic assignment maps. On the other hand, when fewer number of planar models are utilized, the assignment maps might extend and get smoother in spatial domain, whereas fitting error of the planar models could be greater.

By definition, the data cost term of the proposed MRF-based model is equal to the distortion objective of the planar reconstruction. By designing the sum of smoothness and label cost terms to be equal to the rate needed to encode the utilized planar model parameters and their assignment maps, it is possible to obtain rate-distortion optimal realization of planar reconstruction by minimizing the MRF energy. The label cost terms can easily be defined as the rate cost of encoding a planar model, but then the smoothness cost should be defined to be the rate cost of encoding the labeling images. Since efficient shape or image encoders utilize the context in a complex way, it is difficult to represent the bit costs of that coding algorithm in pairwise energy terms of the smoothness cost.

Although it is challenging to design exact rate-distortion objective function in the proposed MRF model, the combination of smoothness and label costs may be considered as a proxy for the rate objective. This is a legitimate assumption, since for the set of all natural depth images, a representation with fewer number of planar models and smoother labeling images will have a smaller entropy which results in decreasing the rate of any regular lossless compression algorithm.

In this manner, the MRF energy also becomes a proxy formulation of the rate-distortion optimization problem for the planar depth reconstruction. In this perspective, the planar depth reconstructions obtained by minimizing the MRF energy given in (3.1) are related to rate-distortion optimality in some sense. Hence different than other arbitrary shape based depth representation approaches [46],[43],[42], the proposed optimization scheme extracts the region shapes in considering the rate-distortion optimality objective, indirectly.

3.1.1 Pareto Optimality in Rate-Distortion

The implicit relation between the regularization costs and the rate can be formulated as a multi-objective or a vector optimization problem of finding rate-distortion optimal settings. In multiobjective optimization, the problem is stated as desirable and in general conflicting objectives but their detailed combination for the main problem is unknown. The conflicting nature between the objectives avoids all the minimum objectives to be satisfied at the same feasible solution. A

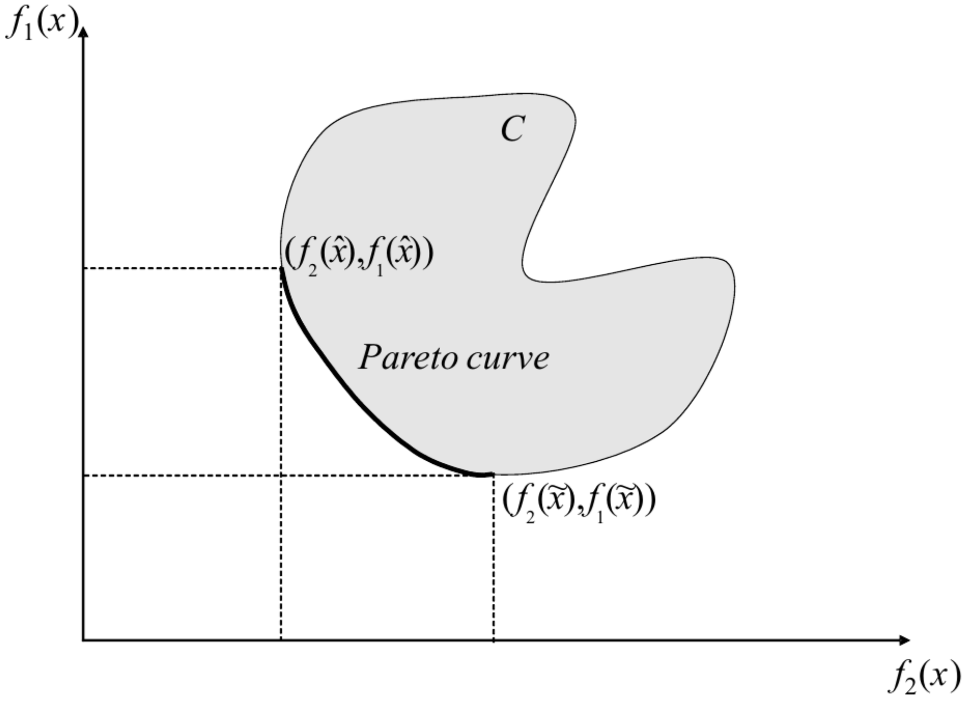

$f_1(x)$
C
$(f_2(\hat{x}),f_1(\hat{x}))$
Pareto curve
$(f_2(\tilde{x}),f_1(\tilde{x}))$
$f_2(x)$

Figure 3.1: Example of a Pareto curve in 2-dimensional objective space with a feasible set of $\mathcal{C}$ (Reprinted with permission. Copyright Springer 2008 [67]).

solution whose none of the objectives can be improved without degrading some of the other objectives is called a *Pareto optimal solution* [67]. In Figure 3.1, the image of the feasible set in the objective space is shown and the set of Pareto curve/surface is illustrated.

The main goal of the multiobjective optimization is to obtain the Pareto optimal surface and guide the decision maker in selecting the favored solution among them. Since the multiobjective optimization problems are in general computationally challenging and expensive, the exact and complete set of Pareto optimal solutions are not attainable in general. A survey of the approaches to approximate the solution set of the multiobjective optimization problems is given in [68].

One of the simplest approaches is to combine the objectives by positive weights to a single objective function, and it is known as the scalarization method. It is proven that the optimal solutions of the scalarized problem with positive weights are always Pareto optimal and under convexity assumptions of the feasible set, all Pareto optimal solutions are optimal solutions of scalarized problems with some positive weights [69].

There are two technical shortcomings of the scalarization method. The first one is the fact that the uniform sampling of weightings does not sample the Pareto surface uniformly in general, and the second one is that the scalarization method can not provide Pareto optimal solutions on the non-convex part of the Pareto surface [67].

In addition to these drawbacks, there are also practical difficulties in determining the proper weights to sample the relevant portion of the Pareto surface for the decision maker. The weights of the objectives do not necessarily correspond directly to the relative importance of the objective functions. The decision maker might not know how to change the weights to consistently change the solution. These possible difficulties make it difficult to develop an (heuristic) algorithm to manipulate the weights to reach a satisfactory region of the Pareto surface [67].

In multiobjective optimization perspective, the MRF formulation of the proposed planar representation, (3.1), is a scalarization instance as,

$$E(f,d) = \lambda_{\mathcal{D}} E_{\mathcal{D}} + \lambda_{\mathcal{V}} E_{\mathcal{V}} + \lambda_{\delta} E_{\delta} \tag{3.5}$$

where the multiobjective problem is stated as,

$$\min[E_{\mathcal{D}}(f,d), \quad E_{\mathcal{V}}(f,d), \quad E_{\delta}(f,d)] \ . \tag{3.6}$$

Since the data, smoothness and label costs are desirable objectives of rate-distortion optimality, a solution minimizing the (3.5) for any positive weighting is a Pareto optimal solution in the rate-distortion sense. In the next section, a heuristic algorithm will be introduced to determine the appropriate weightings to obtain a desired solution from the Pareto surface of the multiobjective problem.

3.1.2 Proposed Algorithm to Weight Costs

For a given encoder, an ideal compression instance can be obtained from the rate-distortion optimal solution set by constraining the rate. The corresponding Lagrangian formulation of the rate-distortion, *R*-*D*, is given as,

$$\min J, \text{ where } \quad J = D + \lambda R \ . \tag{3.7}$$

In practice, although exact rate-distortion optimization is infeasible, the given Lagrangian function is utilized in the decisions of an encoder in order to satisfy the physical constraints, such as the channel capacity [70].

The weight of the data cost with respect to the regularization costs, smoothness and label cost, provides the similar rate-distortion optimization tool of an encoder for the proposed planar representation based compression. However, different than (3.7), the MRF energy formulation given in (3.1) has two degrees of freedom between the cost weights.

The two degrees of freedom of the cost weights are determined by a heuristic algorithm which iteratively updates them towards to a favorable solution of the decision maker. Since one of the motivations in representing the depth images by planar models is to extract object-like representation units, the decision maker of the multiobjective problem is designed to select the minimum distortion solutions from the Pareto surface with a constraint on the maximum number of planar models utilized in the solution. According to such decision maker design, the set of selected Pareto optimal solutions might provide the characteristics of the planar representation for varying number of models.

Since the nature of the problem under data and regularization costs mimics the rate distortion trends, the changes in weightings result in a predictable direction of change in the solution as exemplified in Table 2.2. The developed heuristic algorithm simply tunes the weight of data or the label cost for a fixed smoothness cost weight. If the current solution violates the maximum number of planar model constraint, the label cost weight is increased or the data cost weight is decreased. Otherwise, the data cost weight is increased or the label cost is decreased in order to utilize more planar models for a better data fitting.

The scalarization of the objective costs can be minimized approximately by the PEARL algorithm described in the Section 2.3.3. In the straightforward approach, the computational burden of finding the desired Pareto optimal solution will be higher, since PEARL algorithm should be executed for each weight assignment. In order to avoid utilization of PEARL algorithm multiple times, the weight updating heuristic is integrated into the iterations of the PEARL algorithm in the proposed algorithm.

The pseudo-code of the algorithm to obtain a solution of at most n planar models with the minimum distortion is given in Algorithm 3.1. The algorithm starts with a relaxation part (the first while loop of the algorithm) which aims to sample appropriate planar models for depth images. In this part, the labeling cost is not utilized and the data cost is increased until the number of planar models utilized in the MRF assignment is 3 times the targeted number, n. After obtaining the excessive number of planar models, the labeling cost is included in the formulation and it is increased until the number of utilized planar models satisfy the constraint. Before each PEARL iteration the MRF assignment, planar models and weightings are backed up in order to reverse the last PEARL update. If a MRF assignment satisfies the constraint on the number of utilized planar models, then the upper bound on the labeling cost weight λ_δ is updated with the current weight. In order to find a solution with a smaller distortion, the backed up MRF assignment, planar models and weights are loaded to try a smaller label cost weight λ_δ between the upper and lower bound. According to the number of utilized models in the new MRF assignment, the upper or lower bound of label cost weight is updated for fine tuning. Among the MRF assignments obtained during the fine tuning, the one with the minimum distortion is kept as the solution. The algorithm stops when the gap between the upper and lower bounds of the label cost weight get smaller than a predefined threshold.

3.2 Encoding of Proposed Planar Representation

In order to reconstruct the stereo depth images which are represented with the proposed planar approximation, the camera calibration information, the utilized

Algorithm 3.1 Modified PEARL with weight updates

```
$E(f,d) = \lambda_{\mathcal{D}} \sum_p \mathcal{D}_p + \lambda_{\mathcal{V}} \sum_{p,q \in \mathcal{N}} \mathcal{V}_{p,q} + \lambda_\delta \sum_{m_i \in \mathcal{M}} \delta_{m_i}$
$n_f$ := Number of planar models utilized in $f$
$e_f$ := Total depth distortion of planar reconstruction w.r.t. $f$
$M$ := set of candidate planar models
procedure Fit Planar Models($E(f,d), n$)
    $\hat{f} \leftarrow \emptyset$ , $\hat{M} \leftarrow \emptyset$ and $e_{\hat{f}} \leftarrow \infty$
    $(\lambda_{\mathcal{D}}, \lambda_{\mathcal{V}}, \lambda_\delta) \leftarrow (1,1,0)$
    $M \leftarrow$ randomly sample $2n$ planar models
    $f \leftarrow$ random labeling
    while $n_f < 3n$ and $e_f > \varepsilon_e$ do
        $f_{back} \leftarrow f$ and $M_{back} \leftarrow M$                ▷ Save a restore point
        Update $f, n_f, e_f, M$ by a PEARL iteration
        $\lambda_{\mathcal{D}} \leftarrow 2\lambda_{\mathcal{D}}$
    end while
    $\lambda_\delta \leftarrow 1$ and $\lambda_\delta^{min} \leftarrow 1$
    while $n_f > n$ do
        $f_{back} \leftarrow f$ and $M_{back} \leftarrow M$
        Update $f, n_f, e_f, M$ by a PEARL iteration
        if $n_f \leq n$ then
            $\lambda_\delta^{max} \leftarrow \lambda_\delta$
            if $e_f < e_{\hat{f}}$ then
                $\hat{f} \leftarrow f$ and $\hat{M} \leftarrow M$
            end if
            $\lambda_\delta \leftarrow (\lambda_\delta^{min} + \lambda_\delta^{max})/2$
            $f \leftarrow f_{back}$ and $M \leftarrow M_{back}$
        else
            $\lambda_\delta^{min} \leftarrow \lambda_\delta$
            $\lambda_\delta \leftarrow 4\lambda_\delta$
        end if
        if $\lambda_\delta^{max} - \lambda_\delta^{min} < \varepsilon_{\lambda_\delta}$ then
            return $\hat{f}$ and $\hat{M}$
        end if
    end while
end procedure
```

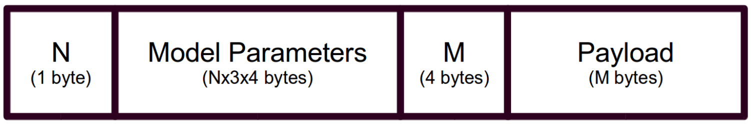


Figure 3.2: The planar depth encoder's byte-stream package definition.

planar model parameters and assignment maps of these models should be known. The byte-stream package for the proposed planar encoder is given in Figure 3.2. The camera calibration information is discarded in the bit stream, since it is considered to be a part of the configuration of a 3D application or system. However, for the case of fronto-parallel stereo-view setup, such as *Middlebury* dataset, the calibration information can be encoded in a single number representing the maximum disparity value in pixel unit, by the 3D space generation method given in Appendix B.

The number of the utilized planar models, N, is encoded by a single byte and it is followed by $3N$ floating points for encoding parameters of planar models as given in (2.16). Then, an unsigned integer number encodes the size of the payload encoding the assignment maps in bytes and it is followed by the payload. Such a stream definition is decodable by reading the bytes in appropriate groupings.

The compression of the assignment maps should be achieved in a lossless scheme. Since the assignment maps are piecewise constant images, they are efficiently compressible in general. Similar shape/boundary information utilized in depth compression is encoded by chain/crack codes [37],[44],[71], JBIG [72],[46] and PAQ [45] tools which are well-known in the data compression community. Any of these tools can be utilized in encoding the assignment maps. The common approach of all these methods is to predict the probability of the next coding unit in the stream according to its spatial context and then entropy code the stream by an arithmetic coder with the estimated probabilities [73].

Since the values of an assignment map vary between 0 and $N-1$ for a planar approximation utilizing N models, $\log_2 N$ bits are sufficient to encode an assigned value of a pixel. In raster scan order the values of consecutive pixels can be packed into a byte for a better compression efficiency. The number of pixels packed into a byte can be recovered at the decoder side according to the simple formula given in below:

$$n = \left\lfloor \frac{8}{\log_2 N} \right\rfloor \tag{3.8}$$

An example of an assignment map and its byte packed representation is given in Figure 3.3. The byte packing results in downscaling in the horizontal direction and introduces new intensity values at the boundaries of assignments as shown in the zoomed details in Figure 3.3.

3.3 Planar Models as Depth Prediction

Planar approximations might not be convenient and fully representative for an arbitrary scene geometry. In order to compensate the planar approximation errors, the residuals can be encoded up to a desired reconstruction quality or up to a rate constraint. In this perspective, the planar representation becomes a prediction tool of a residual coding approach.

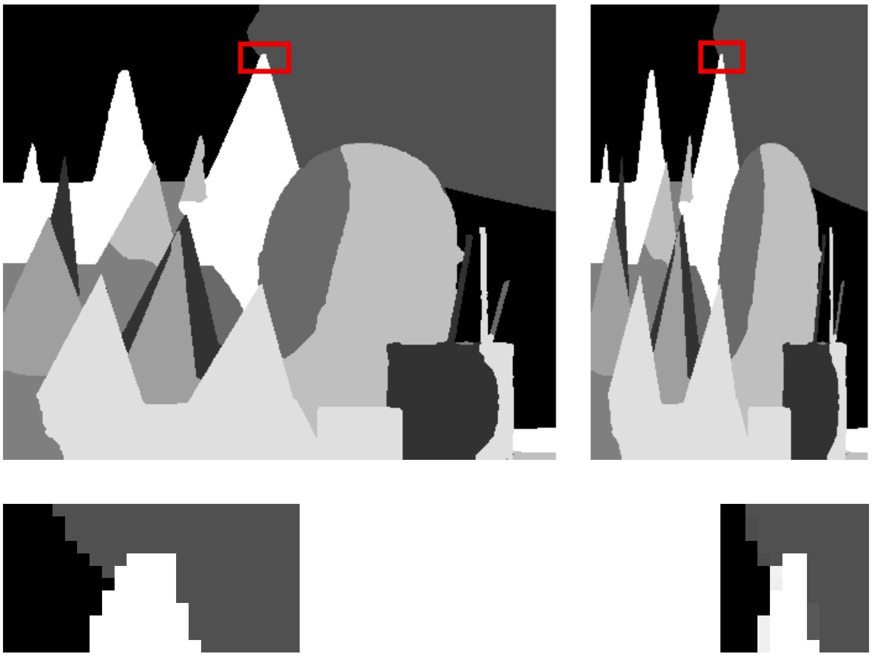


Figure 3.3: *Top row:* The labeling image given at left side encodes the 9 planar model assignments. The byte packed version of the labeling image is given at right side. *Bottom row:* A zoomed detail of the same region for the labeling and byte packed labeling images from left to right. Red boxes show the zoomed region. (Images are histogram equalized for better visualization.)

In [74], a similar MRF energy formulation with data, smoothness and label costs are utilized by Delong et al. in image compression for lossless and lossy cases. They hypothetically encode an image according to an intensity probability model which is a histogram or a Gaussian mixture model assigned by the PEARL algorithm. Their data cost term is the expected number of bits required to entropy code the intensities according to assigned probability model. The smoothness cost is the expected number of bits needed to encode the coding scheme changes during the traversal of the image pixels. And finally, the label cost is the number of bits needed to describe the coding scheme with respect to the assigned probability model.

The same interpretation of the energy terms for the proposed planar depth representation is,

$$E(f,d) = \sum_{p} -\log P(d_p|m_{f_p}) + \lambda_{\mathcal{V}} \sum_{p,q\in\mathcal{N}} \mathcal{V}_{p,q} + \lambda_{\delta} \sum_{m_i\in\mathcal{M}} \delta_{m_i} \tag{3.9}$$

$$\text{where} \qquad -\log P(d_p|m_{f_p}) = \lambda_{\mathcal{D}}|d_p - \hat{d}_p| \ . \tag{3.10}$$

In this perspective, the probability of a pixel depth value is modelled by a Laplacian distribution with a mean value that back-project the pixel to a 3D point on the plane of the corresonding model assignment.

In lossy compression case, Delong et al. replace the original image with its distorted version in the MRF formulation. The distorted version of the image is found by a rate-distortion optimality constraint solved iteratively by a coordinate descent optimization between MRF energy (equivalent to expected rate) and the distorted image.

According to Delong et al.'s formulation in [74], the energy cost minimized by the PEARL algorithm is equal to the expected number of bits needed in the lossless/lossy compression of the image by the corresponding model assignments. However, this hypothetical formulation does not construct an encoder and hence, the corresponding bit stream is not obtained. The expected bit costs for coding scheme changes and their description is also defined in an ad-hoc manner which does not have any relation to a realizable encoder. Another drawback of their model is that the entropy coding of intensity values does not utilize a context which is known to be a very efficient approach in image coding [75], [76]. The context-free modelling of the intensity probabilities decreases the compression performance of the hypothetical experiment severely in comparison to standard image compression techniques.

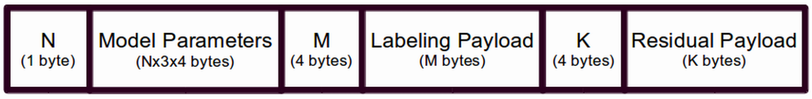
The hypothetical compression framework given in [74] is a unified formulation of the geometric model fitting and the residual coding steps. Hence, it can update its geometric model fitting solution by considering the residual coding part and vice versa. For a practical realization of the stereo depth encoder, this interdependency is discarded in the proposed approach with the residual coding. The planar approximation and residual coding are considered as two consecutive steps, i.e. the planar approximation results are considered to condition the residual coding.

Similar intra-depth prediction methods are proposed for HEVC standard to model the planar depth regions by linear or bi-linear 2D representations [77],[40]. However, beyond these 2D interpretations, the proposed planar representation models the planar regions in 3D space by the motivation of planar approximation of the scene geometry.

The bits required to encode the coding scheme definitions and their spatial support is realized by the stream package defined in the previous section. By appending the residual coding payload to this package definition as shown in Figure 3.4, the planar representation can be utilized as a prediction method in depth compression application.

3.4 Experiments

Since the main concern of the thesis is to investigate the possibilities of planar representations in depth compression for 3D applications, all the compression experiments in the rest of the thesis are performed by the freely available coding tools within the data/image/video compression communities. It is clear that the compression algorithms to be mentioned for planar models can be optimized by considering the distinct characteristics of the planar represented data, but



N (1 byte)	Model Parameters (Nx3x4 bytes)	M (4 bytes)	Labeling Payload (M bytes)	K (4 bytes)	Residual Payload (K bytes)

Figure 3.4: The planar prediction based depth encoder's byte-stream package definition.

these potential improvements in compression performance are discarded for the moment.

It is also important to notify that throughout the depth compression experiments in this thesis, the effects of texture compression on novel view rendering is discarded for all compared techniques by utilizing the original texture information for the stereo views. Since possible performance improvements in compression and novel view rendering are not considered by cooperating the texture information, the texture distortions are not considered as a variable during the evaluations.

In order to sweep the parameters of the experiments extensively, the modified PEARL algorithm with weight updates is speeded up by downscaling the stereo depth images. The details of the speedup are explained in the next subsection.

3.4.1 Speedup of Planar Model Fitting

Each PEARL update is a computationally demanding process. In addition to this, the relaxation part of the modified PEARL algorithm with weight updates increases the number of planar models in the candidate set that results in enormous memory needs. In order to ease these difficulties, the number of the variables in the MRF model is decreased by downscaling the stereo depth images. S times downscaling is applied to guarantee the maximum width and height of the depth images to be smaller or equal to 240 and 120 pixels, respectively. The nearest neighbor downscaling [78] is preferred to avoid creating non-existing 3D points on the object borders by a smooth interpolation operator.

By scaling the internal camera calibration matrices, the 3D points corresponding to downscaled depth images are back projected into the same 3D space defined by the full scale depth images. Based on this fact, the speedup in planar model fitting is obtained by finding the planar models that can efficiently approximate the scene geometry, according to the downscaled depth images.

The solution obtained by Algorithm 3.1 for the downscaled MRF model, $\hat{f}_S, \hat{M}$, is used as an initialization for the full scale MRF modelling. The model assignments of the full scale field are obtained by the nearest neighbor upscaling of the labeling solution, $\hat{f}_S$. The weights of the cost terms obtained in the downscaled MRF energy are also scaled as,

$$(\lambda_{\mathcal{D}},\, \lambda_{\mathcal{V}},\, \lambda_{\delta}) = (\lambda^S_{\mathcal{D}},\, S\lambda^S_{\mathcal{V}},\, S^2\lambda^S_{\delta})\ , \tag{3.11}$$

in order to mimic the same scalarization of the objectives for the full scale MRF energy.

The reason for the weight scaling defined in (3.11) can be explained by counting the number of terms in the summations of each cost terms. The number of

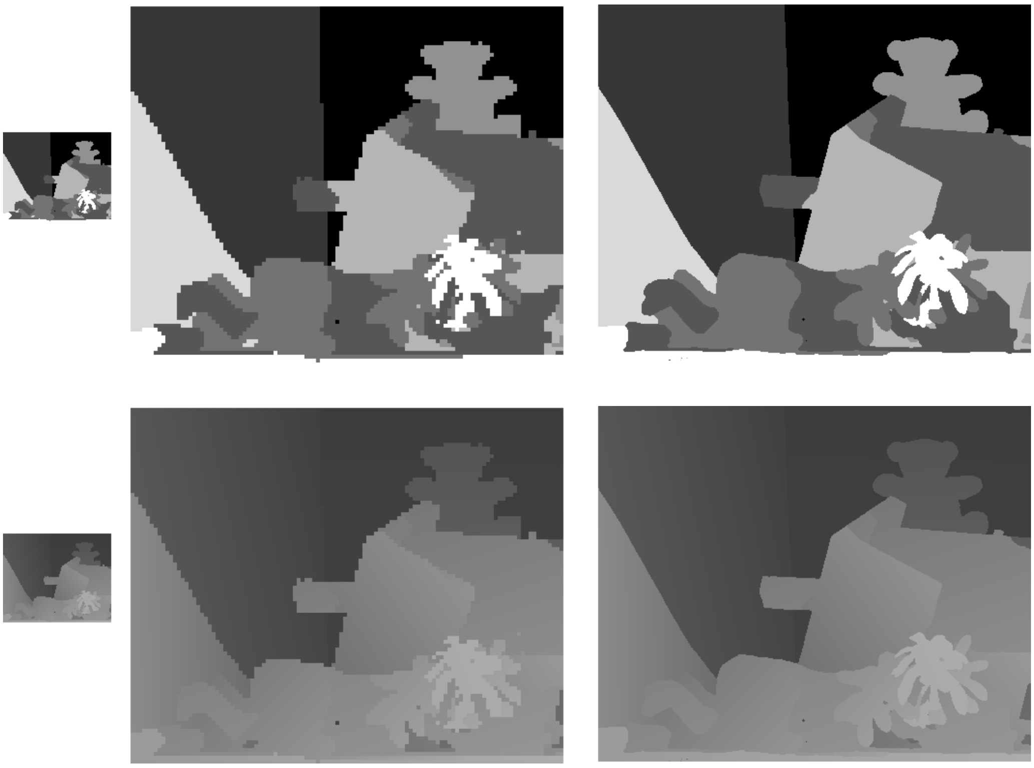


Figure 3.5: The labeling and the corresonding depth reconstruction of *Teddy* dataset is given in the top and bottom rows, respectively. From left to right columns, the solution for the downscaled model, initialization of the full scale model and the final solution of the full scale model are shown.

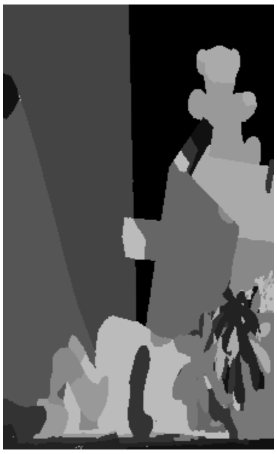
labeling cost terms is equal in downscaled and full scale formulations since the same planar models are utilized in both cases. The number of data cost terms is scaled with the number of pixels, i.e. proportional to increase in area, which is the square of the scale, S^2. The number of non-zero smoothness cost terms only occur at the boundaries of the labeling regions, hence they increase proportional to the increase in the circumference of the regions which is in the order of scale, S. Hence, the weight scaling in (3.11) compensates the changes in the scalarization of the objectives by upscaling the model.

To avoid any PEARL update in the full scale model, the planar models in the candidate set, $\hat{M}$, are fixed and a final minimization of the MRF energy is achieved until the convergence of α-expansion moves [58] of the graph cut algorithm. An example illustrating the evolution of the solution from downscaled model to final solution in full scale is given in Figure 3.5.

3.4.2 Comparative Planar Compression Experiments

For the lossless compression of the labeling image, available compression tools are evaluated and the one with the best compression efficiency is selected without considering the computational time and memory usage of the algorithm. The results of the evaluated lossless compression tools for a representative labeling image is given in Figure 3.6. Based on these results, PAQ8 compression tool is selected as the encoder of the labeling images for the rest of the experiments.

In brief, PAQ8 compression tool uses a binary arithmetic coder which models the probabiltiy density of a bit by a weighted mixing of various context models. The weights of the context models are updated on the fly in order to adapt to data characteristics. The PAQ algorithm family is considered as an improved prediction by partial matching algorithm which is one of the best approaches in lossless natural language compression applications [79]. The details of the algorithm can be found in [80] and [81]. It is worth to mention that PAQ8 is one of the top 10 performers in the lossless photo compression benchmark available in [82] and [83].



Compression Tool	Number of Bytes
PAQ8 [80]	1561
GRALIC [82]	2266
JBIG [84]	3294
LZMA [85]	3732
PPM [79]	4161
PNG [86]	4567
ZIP [87]	4663
CALIC [75]	4708
X264 [88]	5050
JPEG-LS [89]	5474
RLE [90]	11942

Figure 3.6: Compression performances of the lossless coding tools for the labeling image given at the left side.

The depth compression experiments are conducted in comparison to two image/video compression standards, JPEG 2000 and HEVC. While JPEG 2000 is a mature DWT based image compression standard, HEVC is the state-of-the-art DCT based video compression standard whose extensions are still in progress.

Since HEVC is a video coding standard, its intra mode is employed during the stereo depth image compression experiments. In the intra mode of the HEVC, angular, planar and DC based methods are available for spatial prediction of block regions in raster scan order [91]. 2D and local nature of the planar intra prediction mode of the HEVC should be mentioned to underline the differences with the proposed planar representation.

In order to compare the state of the art stereo view compression techniques, the HEVC-MV [92] extension is utilized in the experiments. The stereo depth images are seeded to encoder as a conventional stereo image pair. While the left depth image is intra-coded, the right depth image is predictive coded by leveraging the spatial redundancies between the views.

In addition to multiview coding a state of the art 3D video coding technique, HEVC-3D, is also included in the experiments. The HEVC-3D extension is the prospective recommendation of the Joint Collaborative Team on Video Coding (JCT-VC) for the compression of MVD data format. In addition to prediction and transform tools available in HEVC standard, the HEVC-3D extension widens its block compression techniques with wedgelet and contour based representations for depth compression [93].

The two-view plus two-depth case is utilized for the experiments of the HEVC-3D encoder. As a complete approach for MVD compression, HEVC-3D standard can utilize the decoded texture information in the depth compression algorithms. By considering this property of the encoder, two separate experiments are designed for HEVC-3D. In the first one, the texture information of the views are discarded by providing totally zero intensity stereo pair for the two-view of the MVD data. In the second case, the original views are provided to encoder as texture information and they are encoded with the same quantization parameter used for the depth images. In both experiments, the payload of the depth compression by HEVC-3D extension is measured by considering only the streams corresponding to the depth images.

The proposed planar representation is utilized in stereo depth compression experiments as pure depth compression and depth prediction tools as introduced in Sections 3.2 and 3.3, respectively. The defined byte-stream packages of each case are generated and the distortion analysis of the experiments is achieved by decoding these packages. The parameter n of Algorithm 3.1 that constraints the maximum number of planar models to be utilized in the solution, is swept between 4 and 48. This wide range makes it possible to analyze the planar approximations of the depth images from high to low distortion cases. The depth compression results are evaluated objectively by the PSNR and SSIM index which is accepted as a more human perception friendly metric [94].

Since PEARL-based planar model fitting algorithm provides approximate solutions, the Pareto optimal curve can be approximated as the upper convex hull bound of the obtained solutions. In the figures, the solutions for the proposed planar layered representation are shown as point scatters and their upper convex hull bound as a Pareto optimal curve estimate.

The depth compression results for *Middlebury* dataset is presented in Figures 3.8 and 3.11 for PSNR and SSIM, respectively. Except *Cloth* dataset, the depth map compression performance on the Pareto curve estimate of the planar representation is superior than JPEG 2000 compression. *Cloth* dataset is the only stereo image set which contains a single object which does not have clear object boundaries, but smoothly deformed surfaces. The average depth compression performance of the planar representation over the estimated Pareto optimal curve is comparable with the HEVC compression.

The depth compression performance of the proposed planar representation surpasses the HEVC and even its MVC and 3D extensions for some of the datasets in *Middlebury* collection. The best case among this collection is *Art* dataset which has many clear object boundaries. The performance relation of the planar representation for clear object boundary cases can be realized by comparing *Aloe* and *Cloth* datasets. In *Aloe* dataset, a plant in a pot places in front of a scene covered with the same textile in a similar deformed geometry. The comparison shows that the worst case setup for the proposed planar depth compression is jumped to the second best case as a result of the experiments by inserting a dynamic shaped object into the scene.

This characteristic can be reasoned by the shape encoding of the planar model assignments. The analytic representation of the depth maps in planar models has almost no cost for compression but encoding their spatial support constitutes the main cost. For the scenes, which has very smooth depth variations and fuzzy discontinuities, the crisp planar model assignments introduce redundant, unnatural object boundaries. The enforcement of crisp boundaries results in the utilization of the shape encoder for an inappropriate case.

However, in case of sharp depth discontinuities in the scene, the smooth depth variations along the object surfaces are approximated by the planar models with a negligible cost. The explicit encoding of the discontinuities with the shape encoder takes the advantage over the block-wise transform based approaches, similar to aforementioned depth boundary encoding based approaches in the literature.

Other inefficient utilization examples of the shape encoder for depth boundaries are the cases where the scenes consist of few planar objects, such as *Plastic*, *Wood*, and *Monopoly*. The relaxation part of the planar model fitting Algorithm 3.1 samples excessive number of planar models and the final solution might end up with unnatural boundaries on the smooth surfaces possibly due to noise in the depth values. The performance drop of the planar depth compression at high rates is due to excessive number of planar model fitting for these datasets.

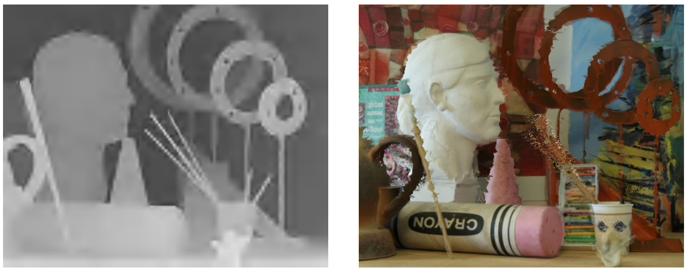
In Figure 3.7, the visual comparison of the planar approach with the other compression standards is presented for *Art* dataset. The compared results are obtained at similar bit rates.

SSIM performance of the planar approach in comparison to other compression standards is much better, as expected. The explicit encoding of the depth discontinuities favors the structural similarity of the reconstruction. However, the ringing artifacts of the conventional block-wise DWT and DCT based approaches might degrade structural similarity.

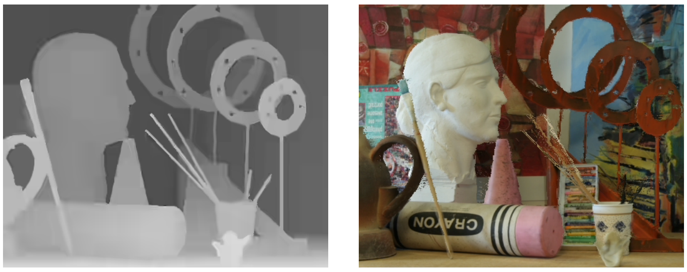
The comparative novel view rendering results of the depth compression methods are illustrated in Figures 3.14-3.17 for PSNR and SSIM, respectively. All the novel view rendering experiments utilized the ground truth texture information of the views. The available captured views at the midpoint of the side views are utilized as the ground truth images while measuring the PSNR and SSIM scores of the novel view rendering experiments.

PSNR scores of the novel view rendering experiments converge to an upper bound of the PSNR scores for the novel view rendering which utilizes the ground truth texture and depth information of the side views. A novel view rendering friendly efficient depth compression method should converge to this bound, as fast as possible.

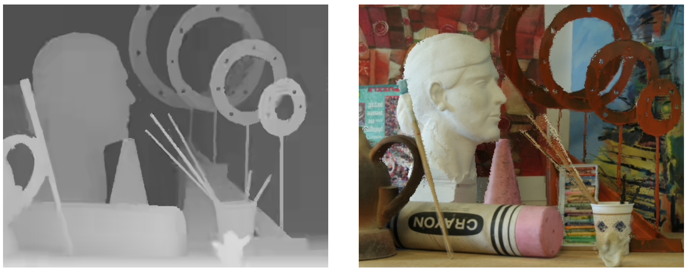
Similar to depth compression simulations, JPEG 2000 again performs the worst in novel view rendering. The convergence of the rendering quality for JPEG 2000 might not occur in the rate bounds of the experiments. In general, the performance of the planar approach is superior than JPEG 2000, except *Cloth* dataset; a similar result to depth compression comparison. The gap between the planar and the MVC and 3D extensions of the HEVC standards is smaller



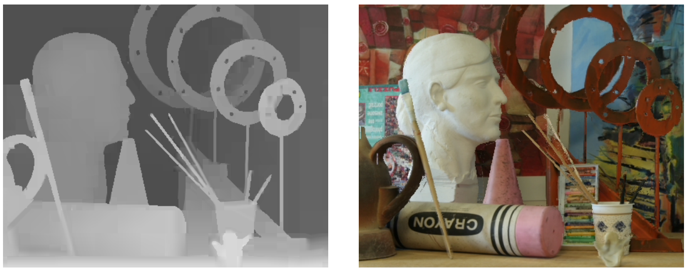
(a) JPEG2K: 29.59, 0.873 (b) JPEG2K: 26.80, 0.851



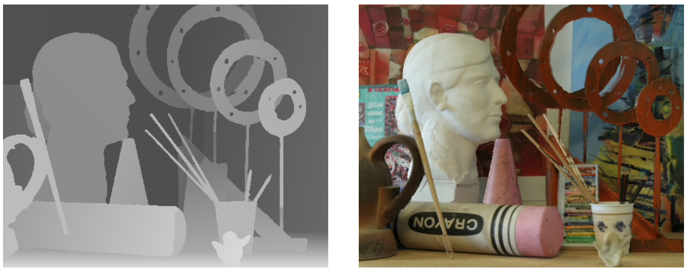
(c) HEVC: 34.42, 0.914 (d) HEVC: 29.20, 0.914



(e) HEVC-MV: 35.02, 0.957 (f) HEVC-MV: 29.64, 0.920

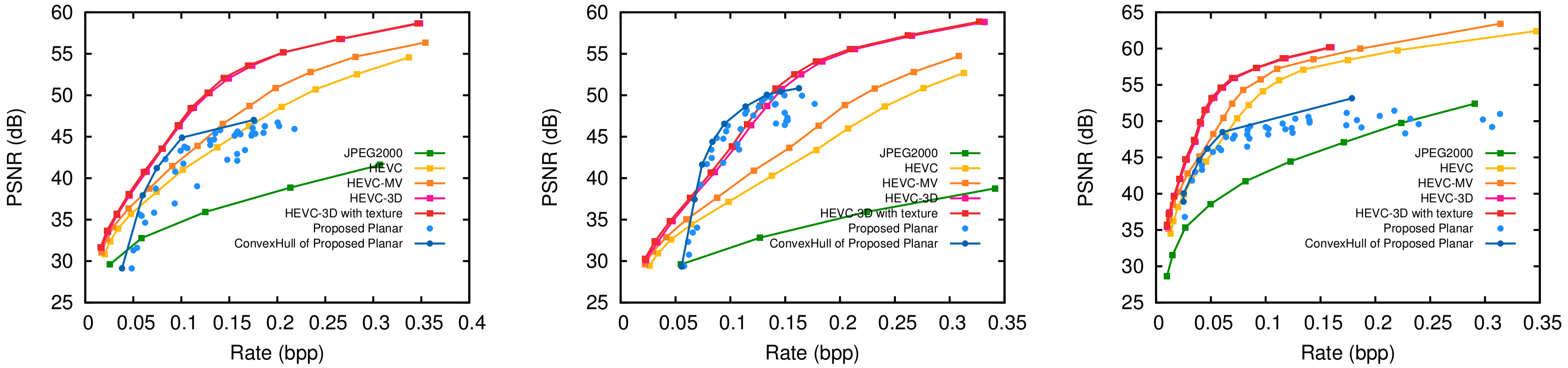


(g) HEVC-3D: 37.64, 0.972 (h) HEVC-3D: 31.66, 0.951

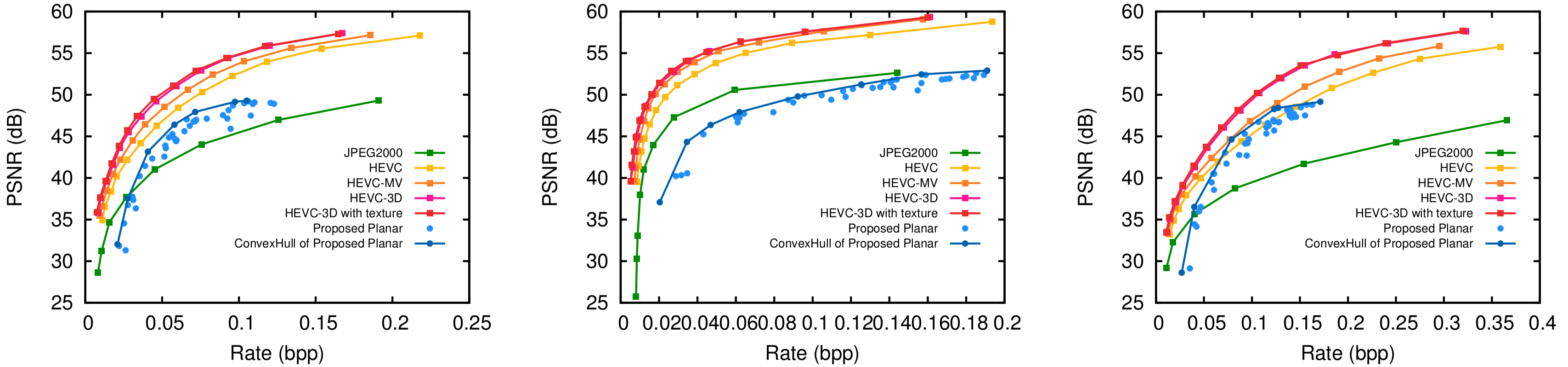


(i) Planar: 37.43, 0.989 (j) Planar: 30.19, 0.938

Figure 3.7: Visual comparison of the depth compression results in depth reconstruction and novel view rendering. PSNR (dB) and SSIM scores are also stated consecutively. The rate of the compression methods from top to bottom are equal to 0.0547, 0.0647, 0.0601, 0.0652, 0.0673 bits per pixel.

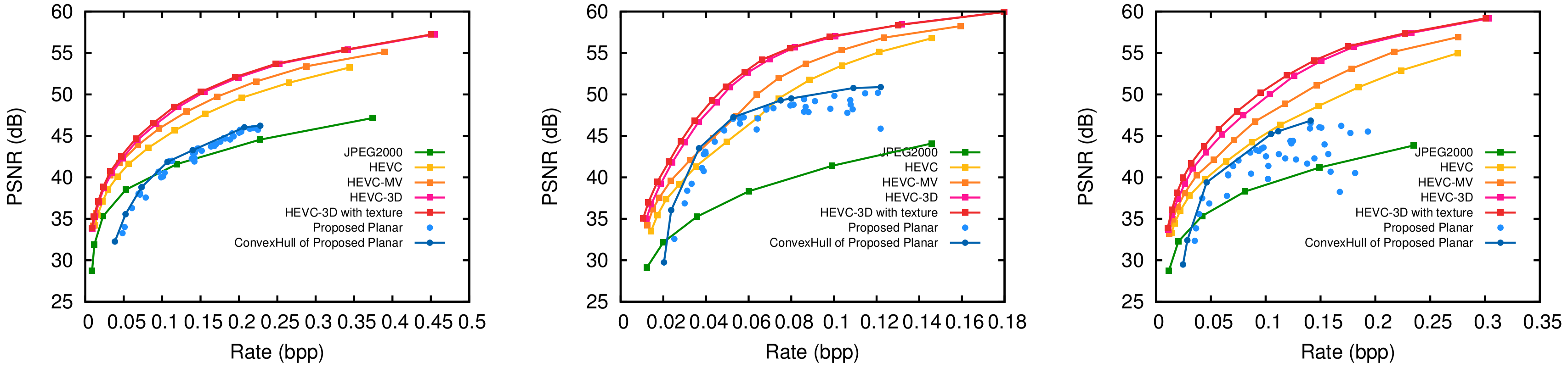


(a) Aloe (b) Art (c) Baby

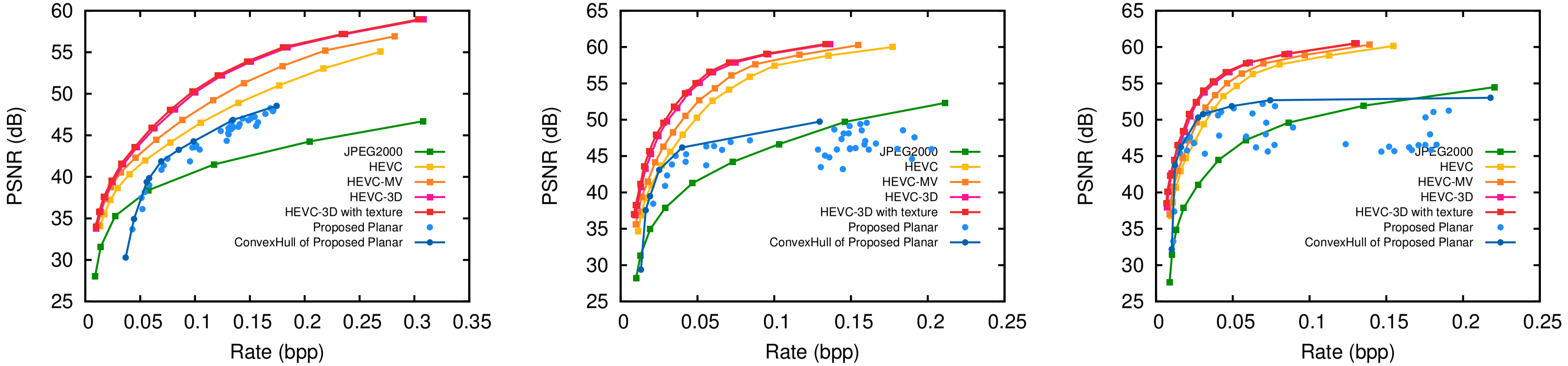


(d) Books (e) Cloth (f) Cones

Figure 3.8: Compression experiments comparing the mean PSNR values of the *reconstructed stereo depth images*.

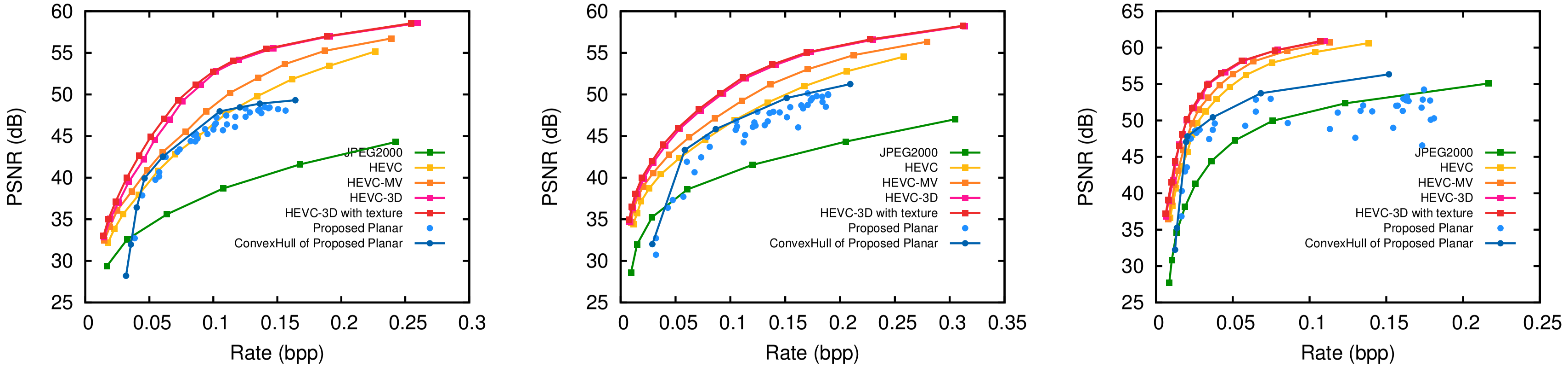


(a) Dolls (b) Lampshade (c) Laundry



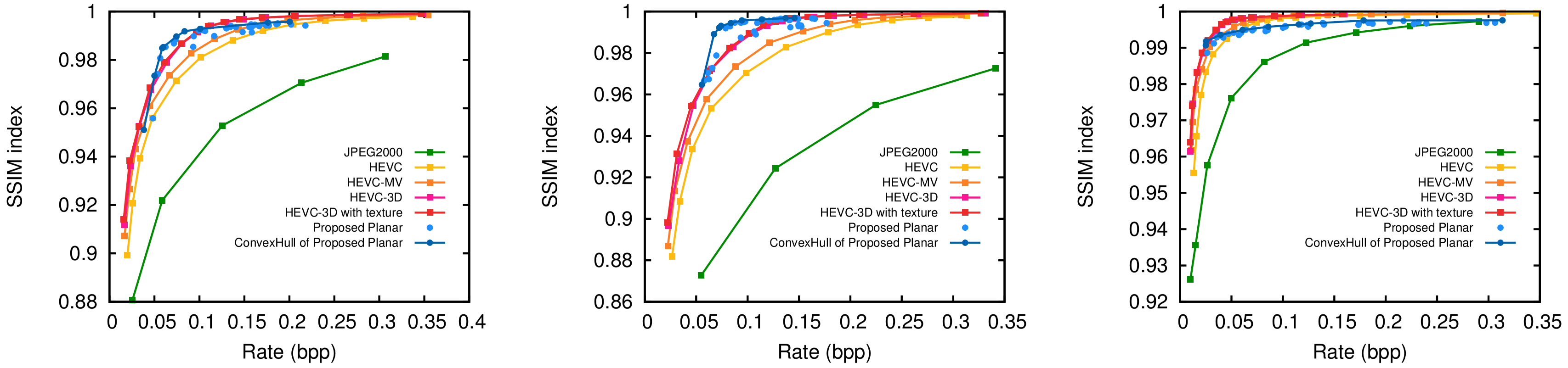
(d) Moebius (e) Monopoly (f) Plastic

Figure 3.9: Compression experiments comparing the mean PSNR values of the *reconstructed stereo depth images*.

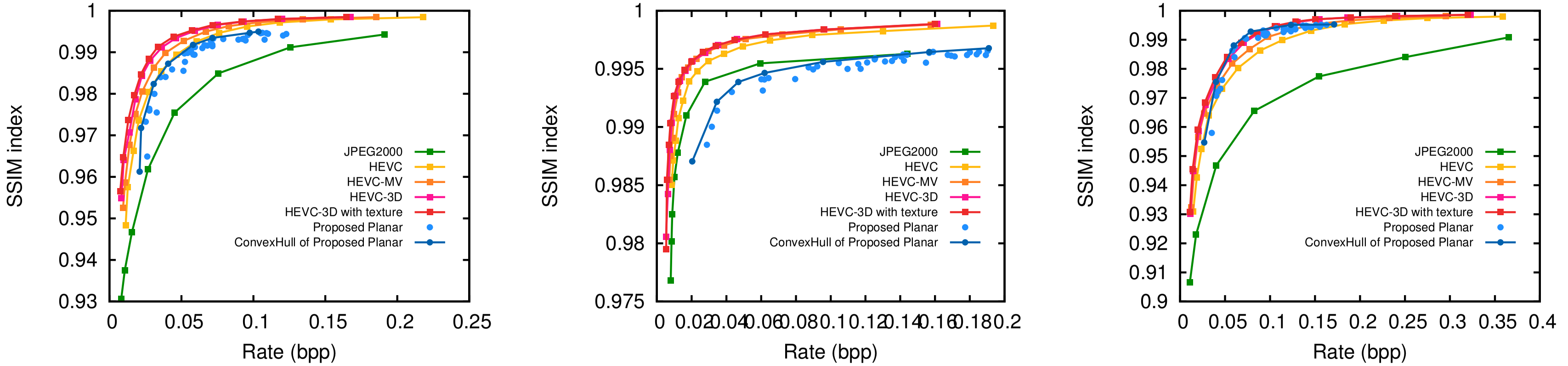


(a) Reindeer (b) Teddy (c) Wood

Figure 3.10: Compression experiments comparing the mean PSNR values of the *reconstructed stereo depth images*.

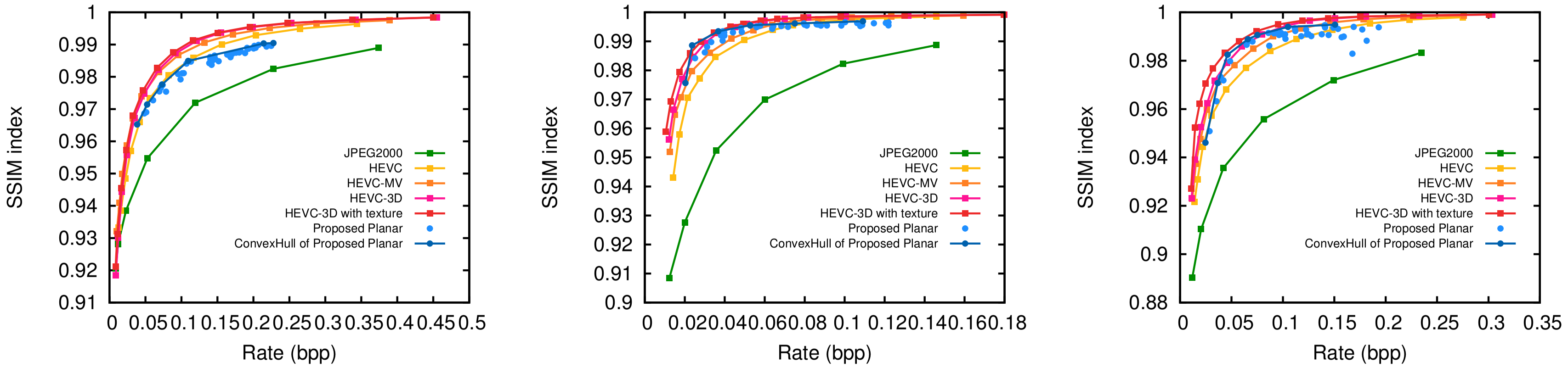


(a) Aloe (b) Art (c) Baby

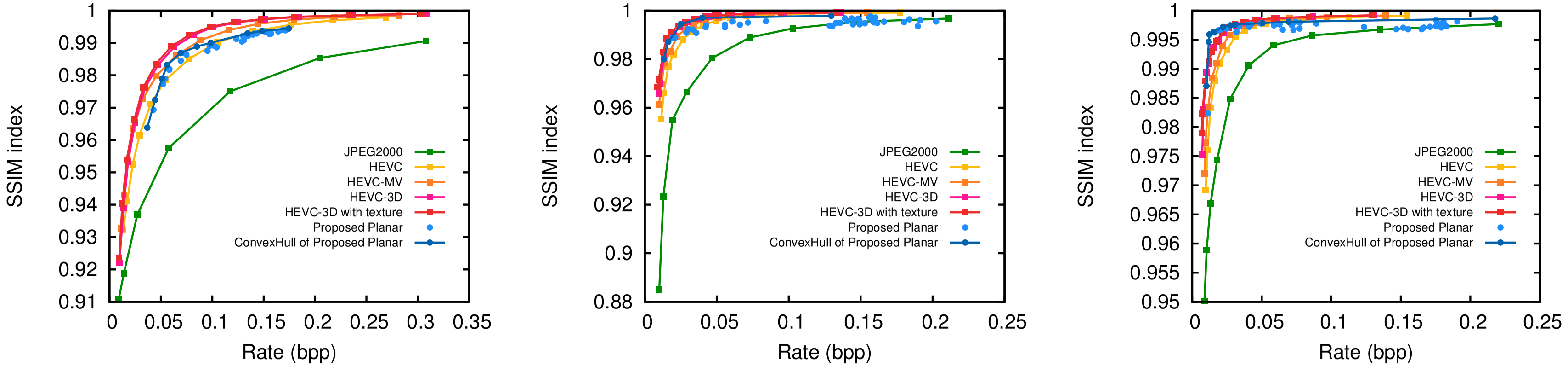


(d) Books (e) Cloth (f) Cones

Figure 3.11: Compression experiments comparing the mean SSIM scores of the *reconstructed stereo depth images*.

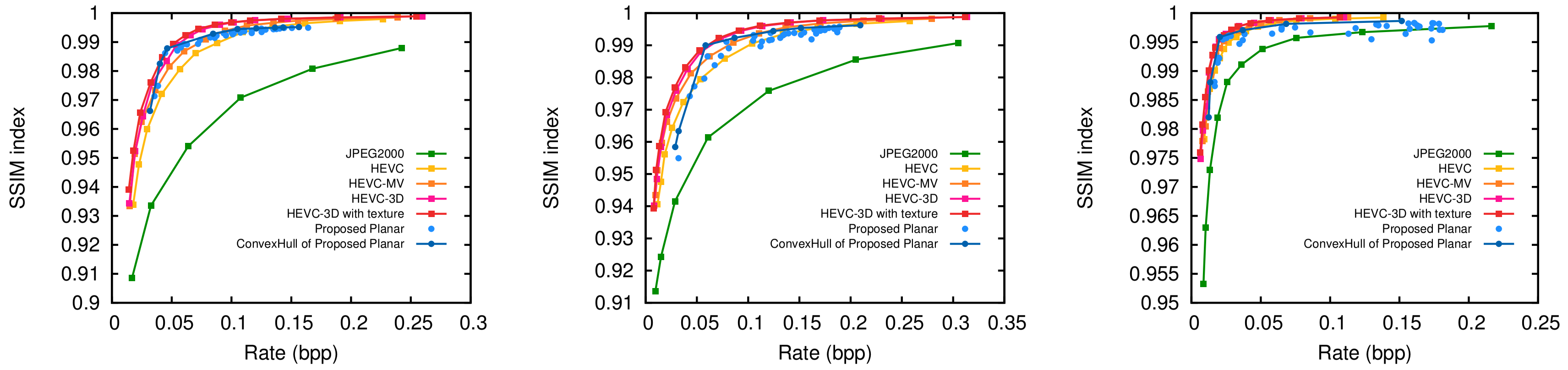


(a) Dolls (b) Lampshade (c) Laundry



(d) Moebius (e) Monopoly (f) Plastic

Figure 3.12: Compression experiments comparing the mean SSIM scores of the *reconstructed stereo depth images*.



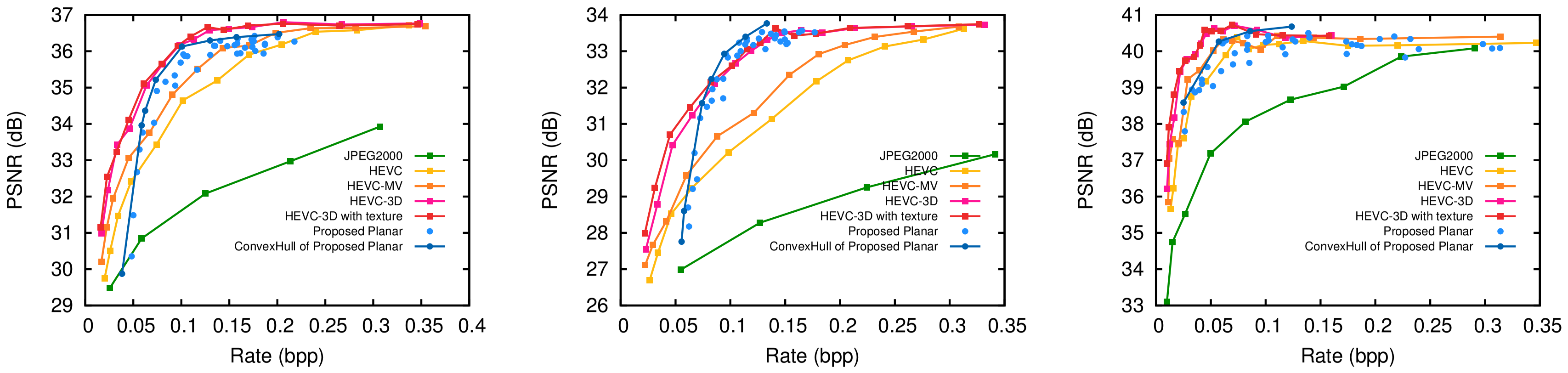
(a) Reindeer (b) Teddy (c) Wood

Figure 3.13: Compression experiments comparing the mean SSIM scores of the *reconstructed stereo depth images*.

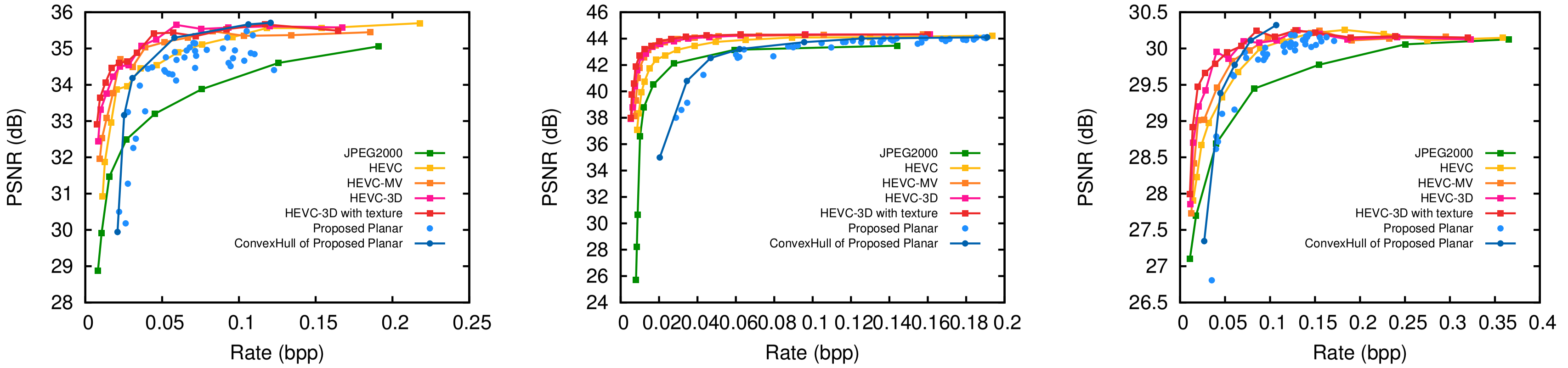
in novel view rendering. In general, the planar approach has better novel view rendering results than HEVC standard. These observations might be concluded as the inter-view prediction schemes utilized in the MVC and 3D extensions of the HEVC can be helpful in advancing the planar approach. However without these improvements the proposed planar depth compression approach is still comparable against the state-of-the-art MVD compression techniques in novel view rendering.

The novel view rendering performance of the planar approach should be mentioned, especially for the datasets with the planar objects in the scene. Comparing the PSNR plots of the depth reconstruction and novel view rendering results of the datasets, *Plastic*, *Wood*, and *Monopoly* shows that the depth reconstruction performance gap disappears during novel view rendering at high rates. For the planar scene cases, the planar depth compression converges to the upper bound of the novel view rendering at least as fast as the state-of-the-art MVD compression techniques. The inferior depth reconstruction performance at high rates does not affect the novel view rendering adversely. Based on these observations, the proposed representation should be regarded as an efficient depth compression method, especially for planar regions, as expected.

However, the more interesting results are obtained for the datasets *Art*, *Aloe*, and *Reindeer*. The novel view rendering performance of the planar approach is at least as good as the MVC extensions of the HEVC standard. The dominant characteristics of these datasets cannot be claimed to be planar, but their common property might be vital object boundaries for image content description. The higher performance on novel view rendering can be explained by the boundary description capability of the proposed planar representation. As mentioned before, for a satisfactory novel view rendering, the reconstruction of sharp depth boundaries is important for DIBR techniques. As long as the distortions due to planar approximation of the object surfaces are tolerable for DIBR, the proposed planar compression can benefit in rendering quality by maintaining clear depth boundaries.

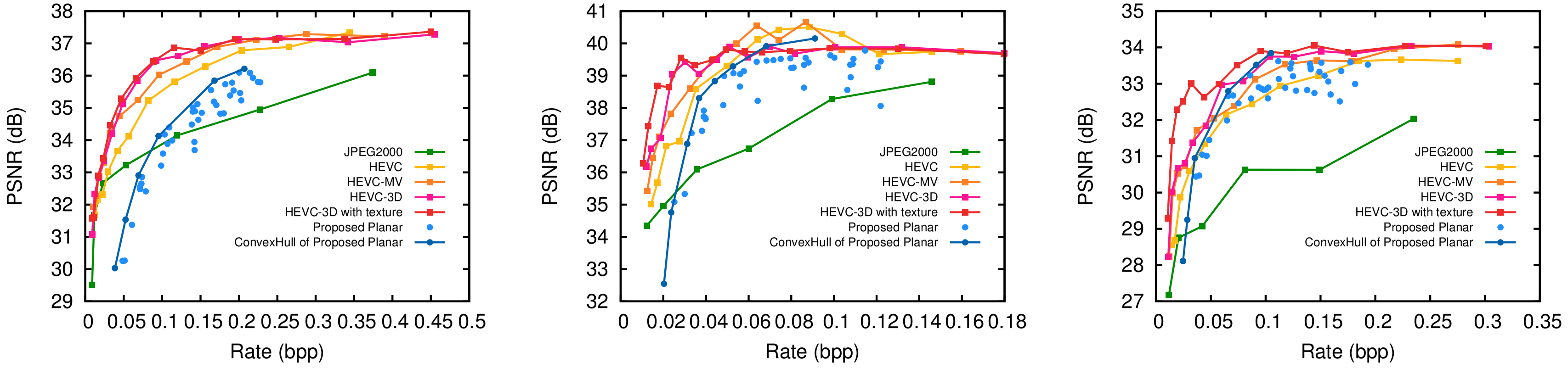


(a) Aloe (b) Art (c) Baby

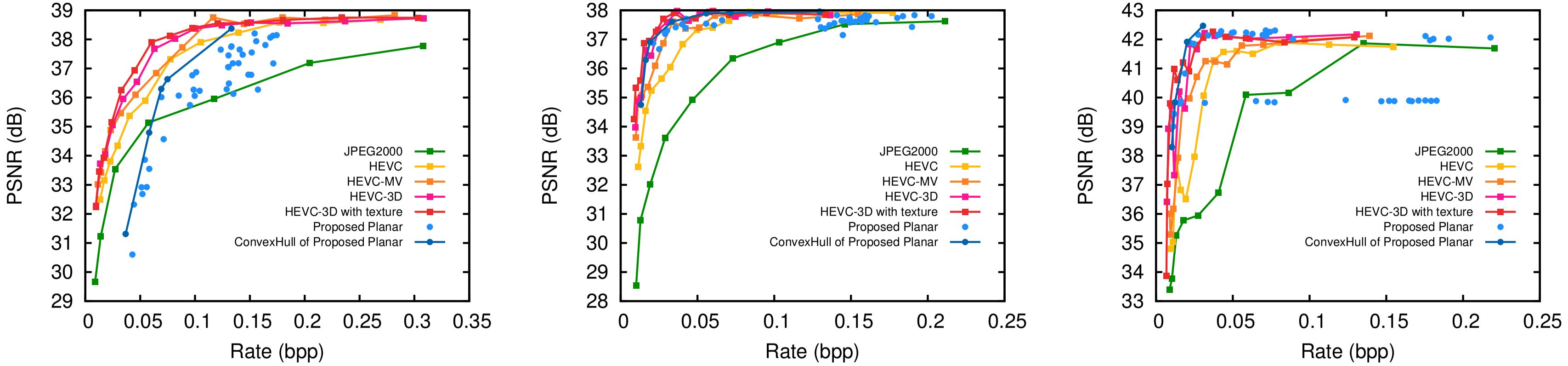


(d) Books (e) Cloth (f) Cones

Figure 3.14: Compression experiments comparing the PSNR values of the *novel view renderings* obtained by the reconstructed stereo depth images.

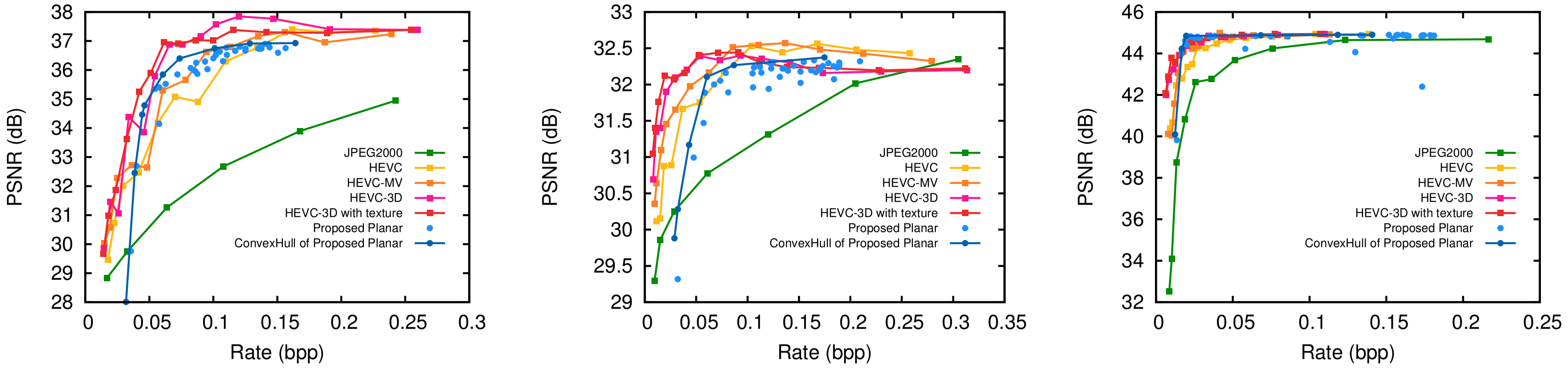


(a) Dolls (b) Lampshade (c) Laundry



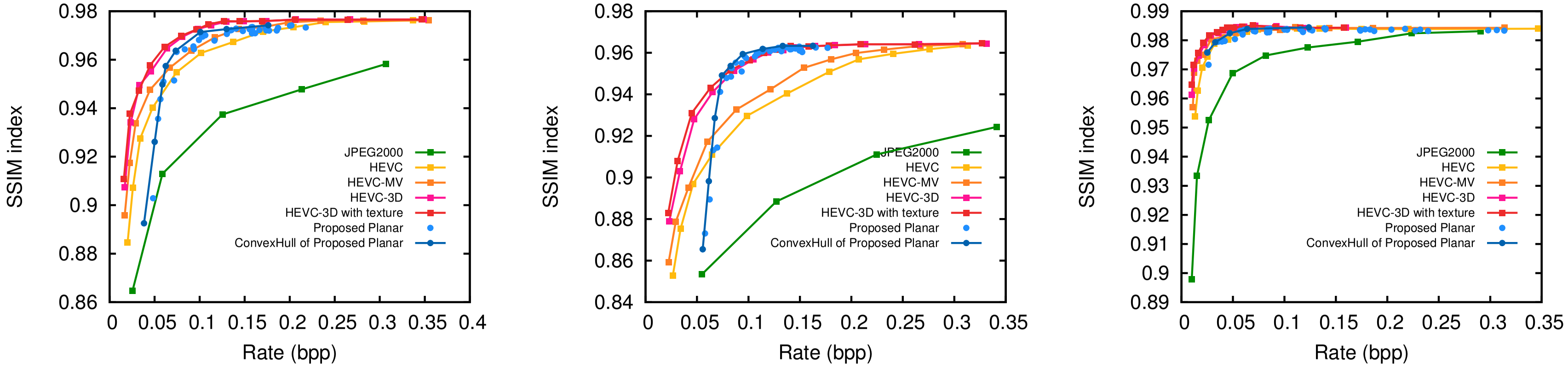
(d) Moebius (e) Monopoly (f) Plastic

Figure 3.15: Compression experiments comparing the PSNR values of the *novel view renderings* obtained by the reconstructed stereo depth images.

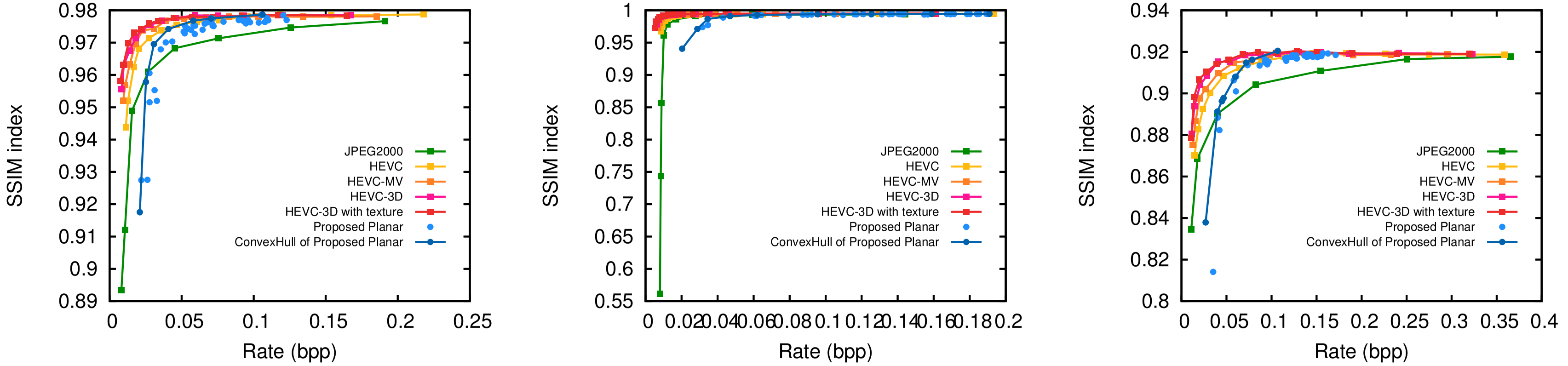


(a) Reindeer (b) Teddy (c) Wood

Figure 3.16: Compression experiments comparing the PSNR values of the *novel view renderings* obtained by the reconstructed stereo depth images.

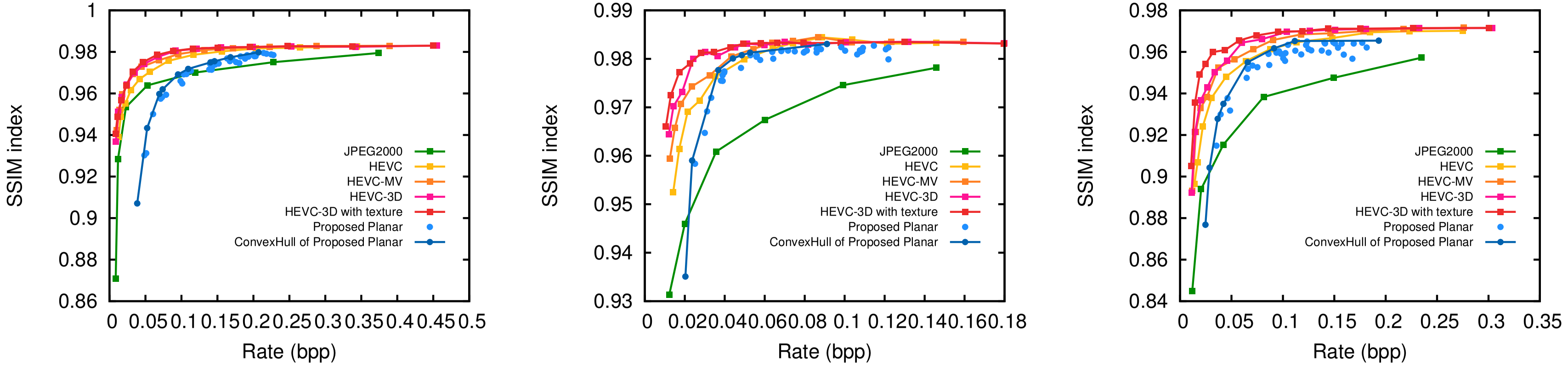


(a) Aloe (b) Art (c) Baby

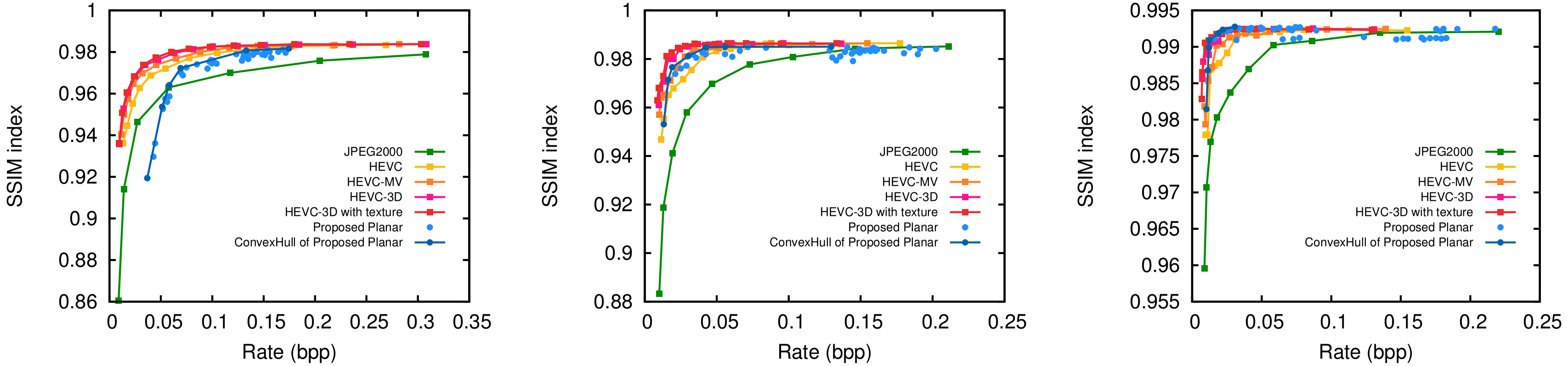


(d) Books (e) Cloth (f) Cones

Figure 3.17: Compression experiments comparing the SSIM scores of the *novel view renderings* obtained by the reconstructed stereo depth images.

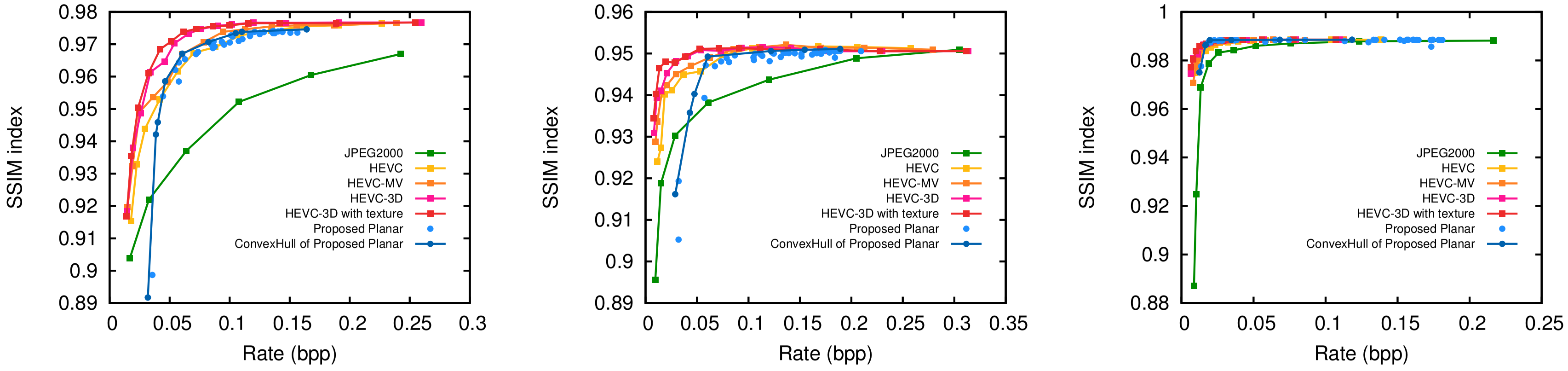


(a) Dolls (b) Lampshade (c) Laundry



(d) Moebius (e) Monopoly (f) Plastic

Figure 3.18: Compression experiments comparing the SSIM scores of the *novel view renderings* obtained by the reconstructed stereo depth images.



(a) Reindeer (b) Teddy (c) Wood

Figure 3.19: Compression experiments comparing the SSIM scores of the *novel view renderings* obtained by the reconstructed stereo depth images.

3.4.3 Comparative Planar Prediction Experiments

During these experiments, utilization of the planar representation of the depth images as a prediction method, the DWT and DCT based residual encoders are studied. The transform based residual encoding is a very popular technique in video compression community due to intra and inter prediction mechanisms [95],[96]. For implementation practicality, the JPEG 2000 and intra mode of HEVC are utilized as residual image encoders throughout the experiments for DWT and DCT, respectively.

The residual images are obtained according to proposed planar representation based approximations of the stereo depth pairs. The number of the planar models is constrained to be at most 10 in all prediction experiments. The residual image values theoretically can range between -255 and 255. They are represented as 9 bit unsigned images by adding an offset value of 255.

The byte-streams of the experiments utilizing the planar representation as a prediction tool are obtained as the byte package definition given in Figure 3.4. The payloads of the residual encoding are obtained by JPEG 2000 and HEVC for the DWT and DCT-based experiments, respectively. The quantization and target distortion parameters of the encoders are scanned uniformly to obtain a rate-distortion plot for the planar prediction scenario.

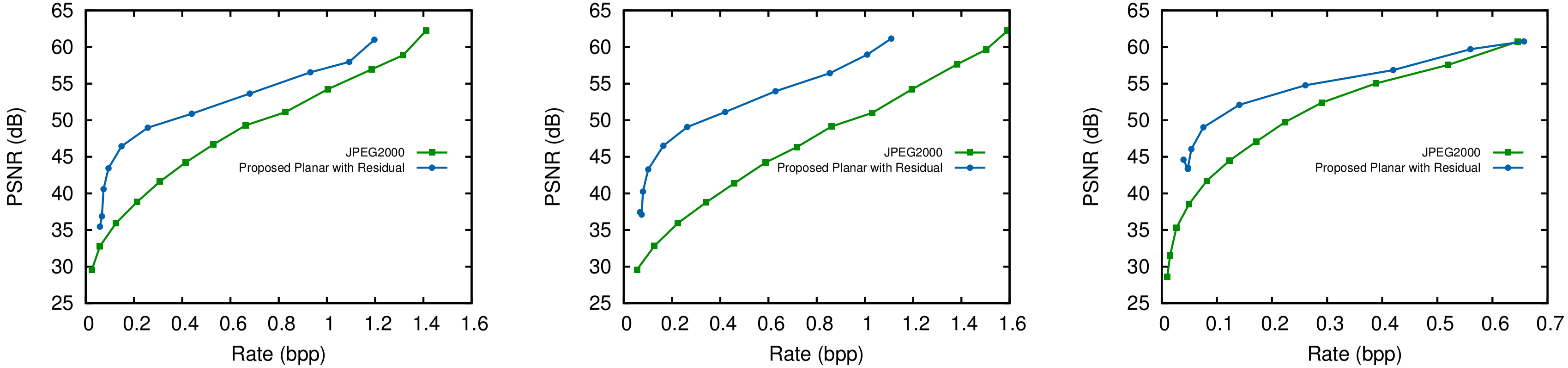
The results of the planar prediction experiments for depth compression and novel view rendering are given in Figures 3.20 and 3.23, respectively, for the DWT-based analysis. The planar prediction has a positive effect on depth compression for all the datasets, except *Cloth*. The novel view rendering is also affected positively in general by the base planar prediction. The distribution of the positive contribution of planar prediction over the datasets is similar to the results obtained in the pure planar compression experiments of the previous section.

The similar planar prediction experiments for DCT case are conducted by utilizing the HEVC standard in Figures 3.26 and 3.29. For the HEVC case it is difficult to say the planar prediction has a positive effect in all cases. However, the datasets on which the planar representation based compression has better results, the planar prediction also has a positive effect in depth compression and novel view rendering.

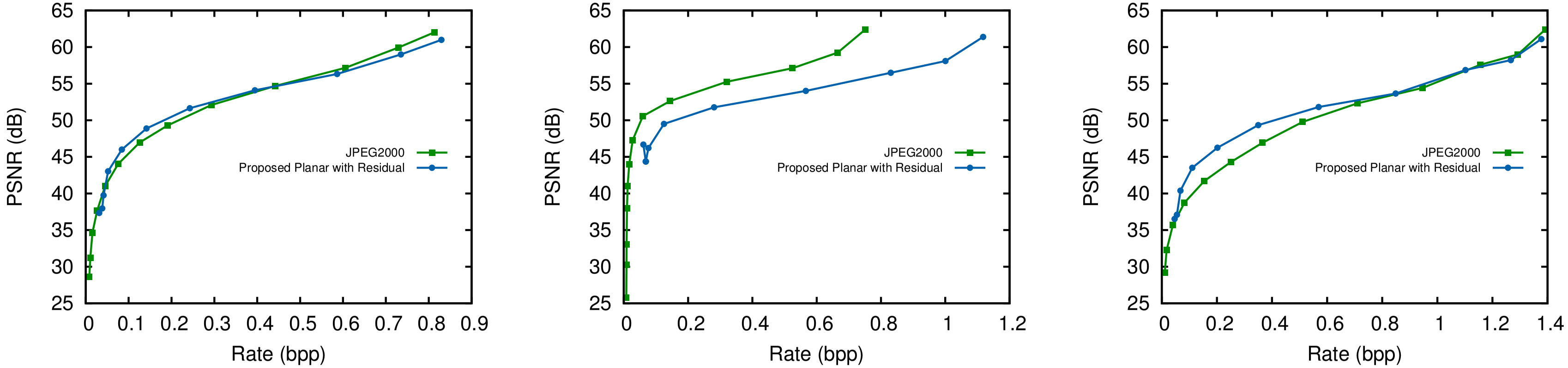
Woods dataset is an extreme example of the efficiency of the planar prediction in novel view rendering results. Since all the scene has a planar geometry in *Woods* dataset, the base planar prediction already provides a novel view rendering quality very close to the maximum possible quality obtained by increasing the compression rate. The residual coding can even degrade the depth reconstruction and novel view rendering quality by increasing the rate.

The DWT and DCT-based residual compression experiments are analyzed separately, in order to generalize the possible potentials of the proposed planar representation as a prediction tool for different residual transforming techniques. Since the residual images are provided similar to the conventional images to these transform based encoders, the error in the quantization step of the encoders can not be guaranteed to decrease the final reconstruction error, but the reconstruction error of the residual image. This leakage in the experiment is observed for the most of the datasets at the coarse quantization schemes especially for HEVC. Even the residual encoding of the experiments are not optimal, the results show the proposed planar prediction as a potential tool for efficient depth compression and novel view rendering.

The state-of-the-art compression tools, such as HEVC and its 3D extension, provides various prediction modes in order to be adaptive to the image/depth characteristics. Based on the superior results of the proposed planar prediction for some of the datasets, it can be considered as an efficient depth prediction mode candidate for these standards. The explicit boundary information provided by the proposed planar approach can also be utilized in possible redundancy and artifact removal techniques on the compression and rendering applications of the MVD data.

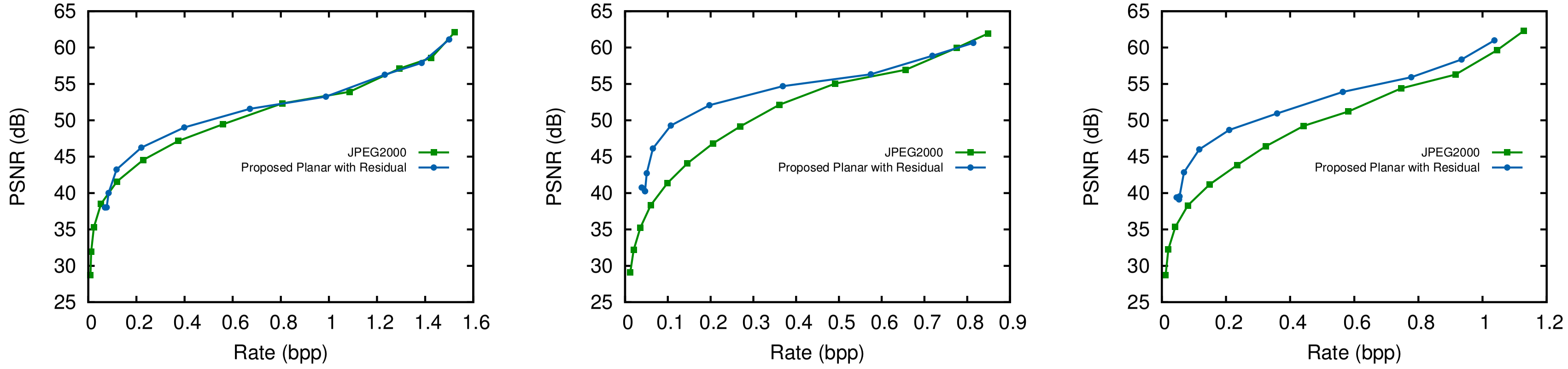


(a) Aloe (b) Art (c) Baby

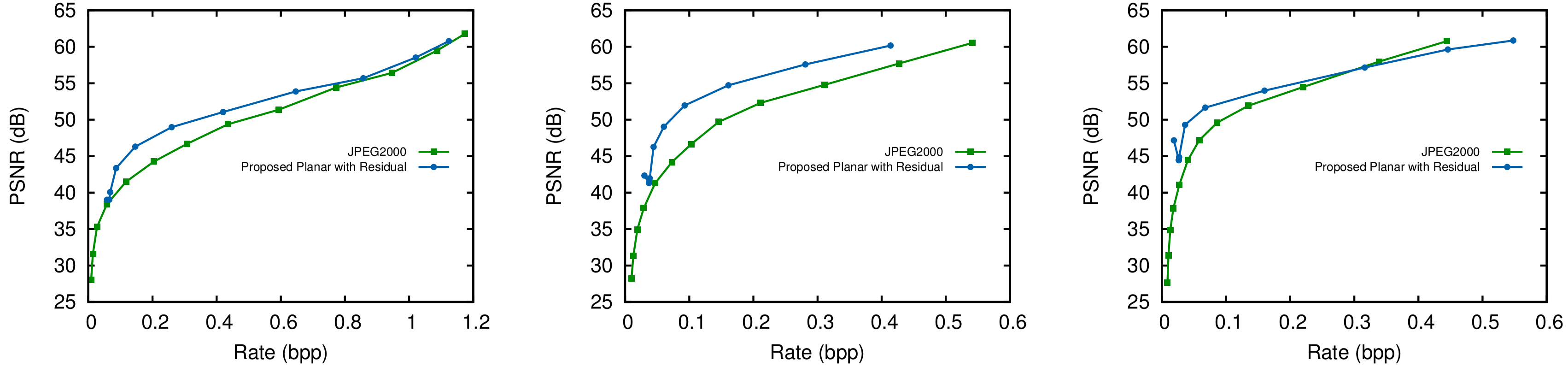


(d) Books (e) Cloth (f) Cones

Figure 3.20: Mean PSNR values of the *depth reconstruction* results obtained by JPEG 2000 and planar prediction with JPEG 2000 residual coding.

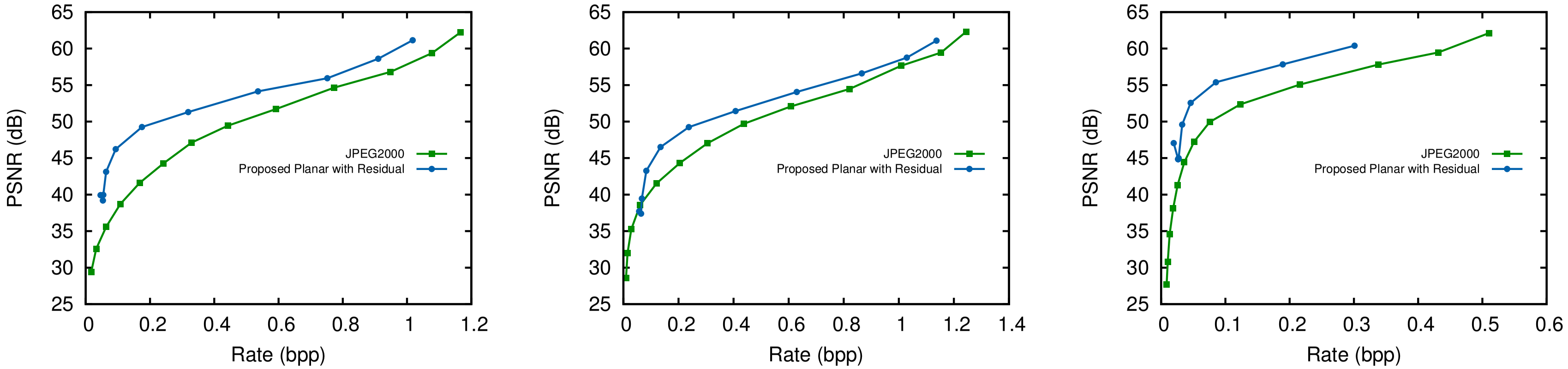


(a) Dolls (b) Lampshade (c) Laundry



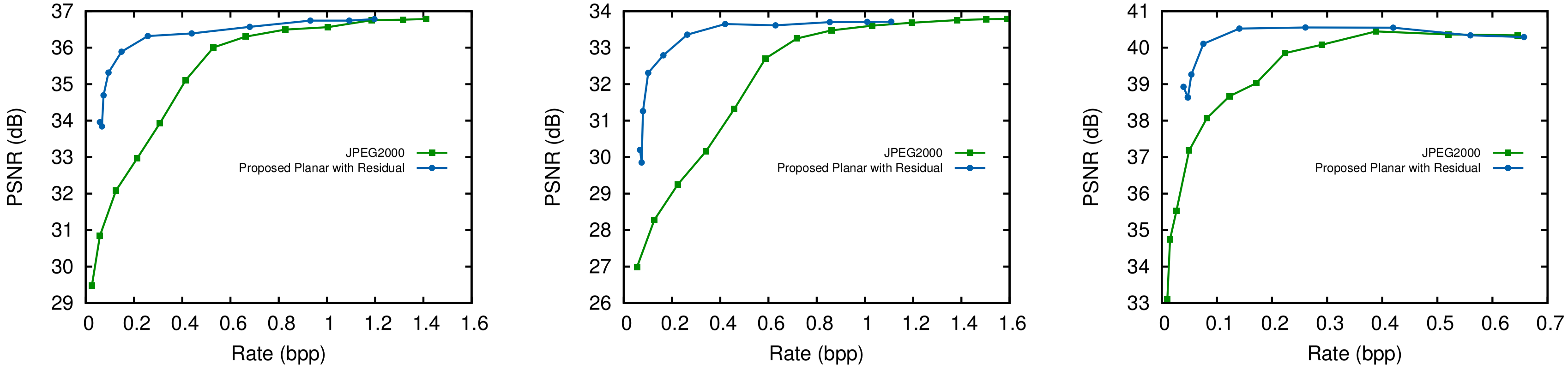
(d) Moebius (e) Monopoly (f) Plastic

Figure 3.21: Mean PSNR values of the *depth reconstruction* results obtained by JPEG 2000 and planar prediction with JPEG 2000 residual coding.

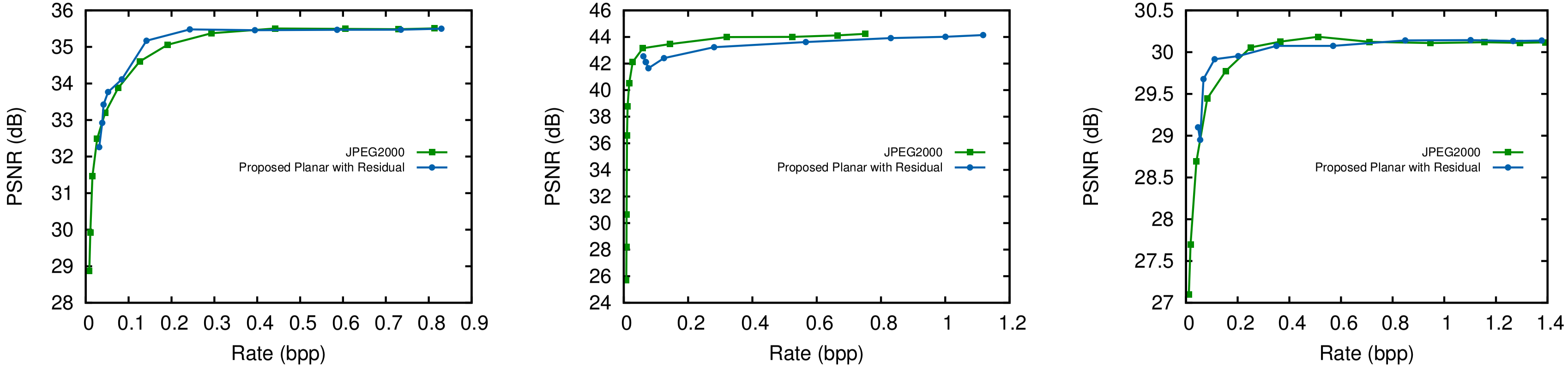


(a) Reindeer (b) Teddy (c) Wood

Figure 3.22: Mean PSNR values of the *depth reconstruction* results obtained by JPEG 2000 and planar prediction with JPEG 2000 residual coding.

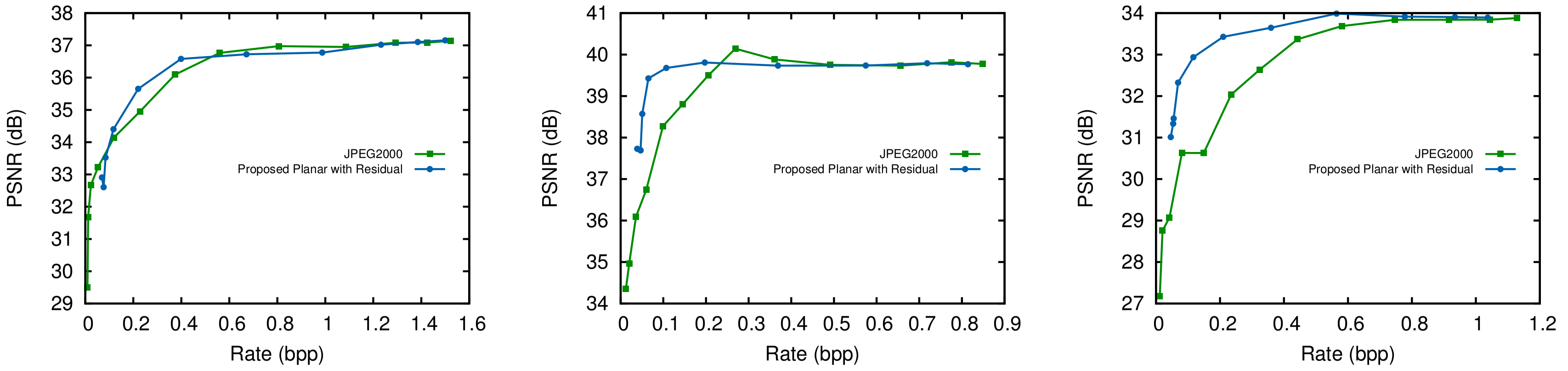


(a) Aloe (b) Art (c) Baby

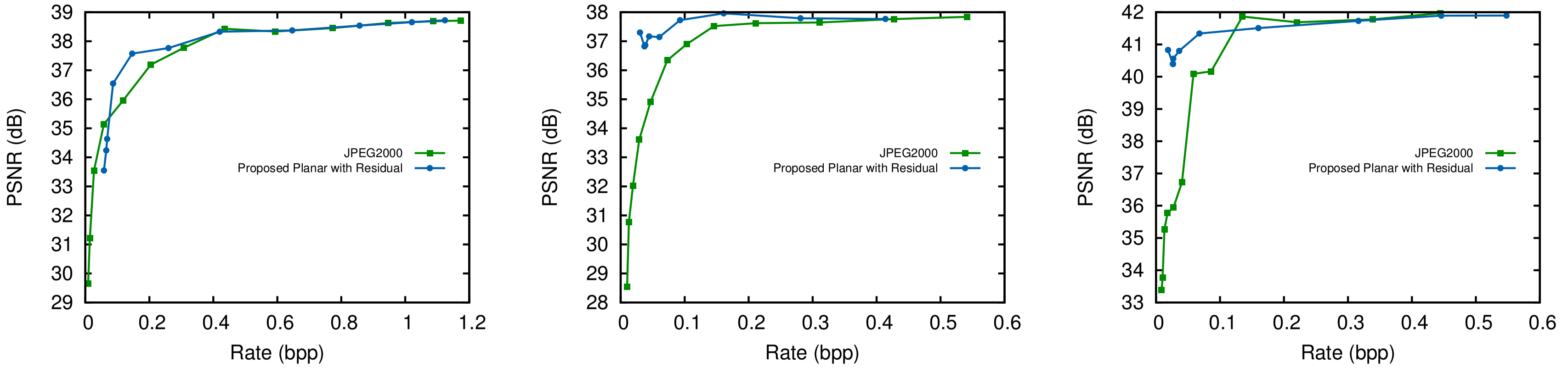


(d) Books (e) Cloth (f) Cones

Figure 3.23: PSNR values of the *novel view rendering* results obtained by the depth reconstructions of JPEG 2000 and planar prediction with JPEG 2000 residual coding.

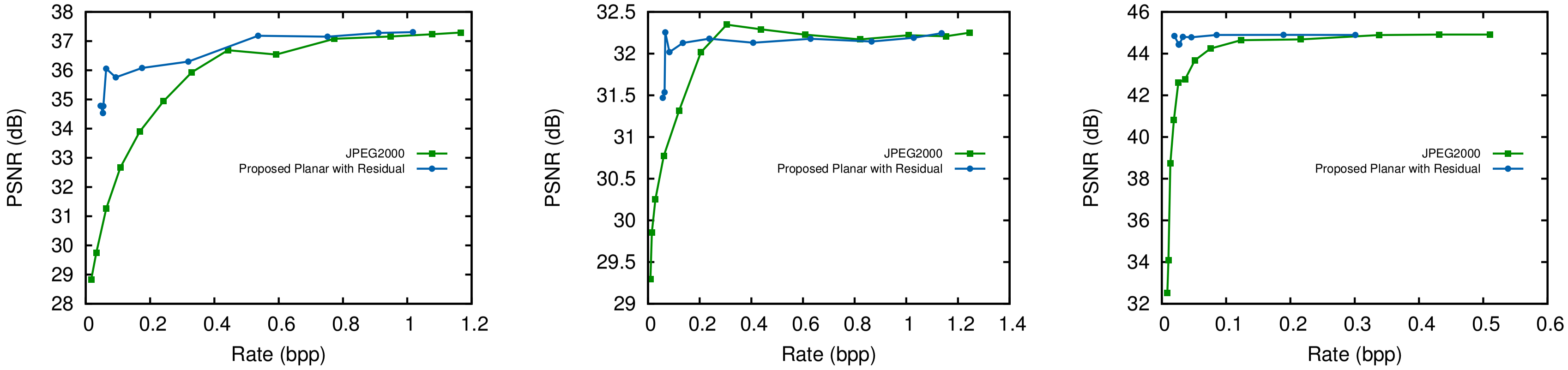


(a) Dolls (b) Lampshade (c) Laundry



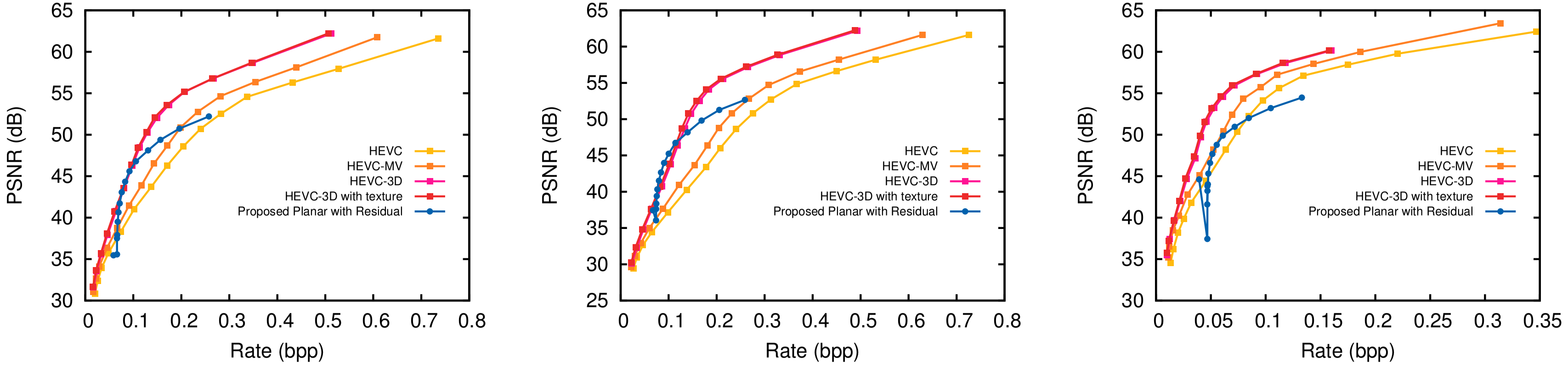
(d) Moebius (e) Monopoly (f) Plastic

Figure 3.24: PSNR values of the *novel view rendering* results obtained by the depth reconstructions of JPEG 2000 and planar prediction with JPEG 2000 residual coding.

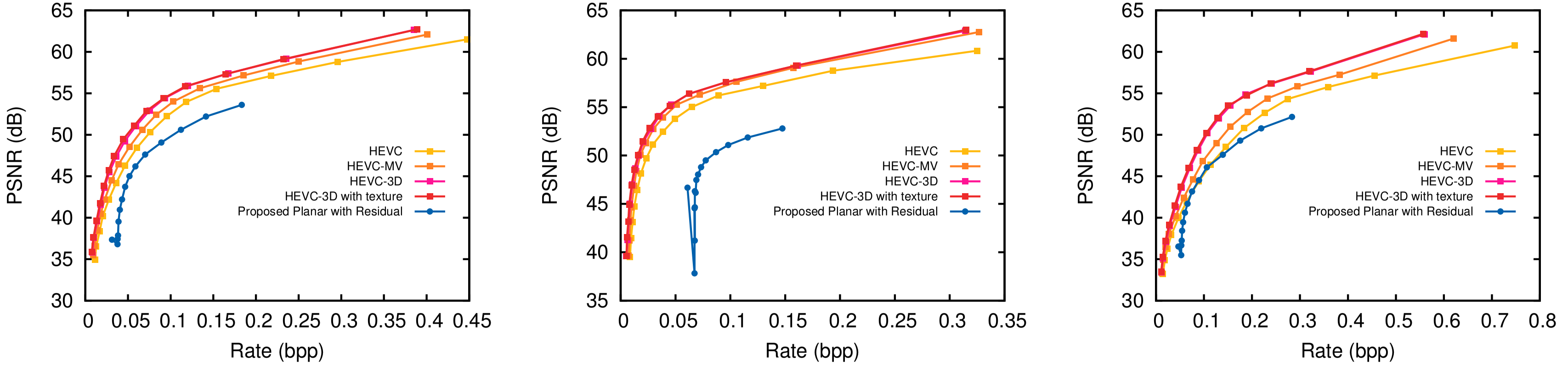


(a) Reindeer (b) Teddy (c) Wood

Figure 3.25: PSNR values of the *novel view rendering* results obtained by the depth reconstructions of JPEG 2000 and planar prediction with JPEG 2000 residual coding.

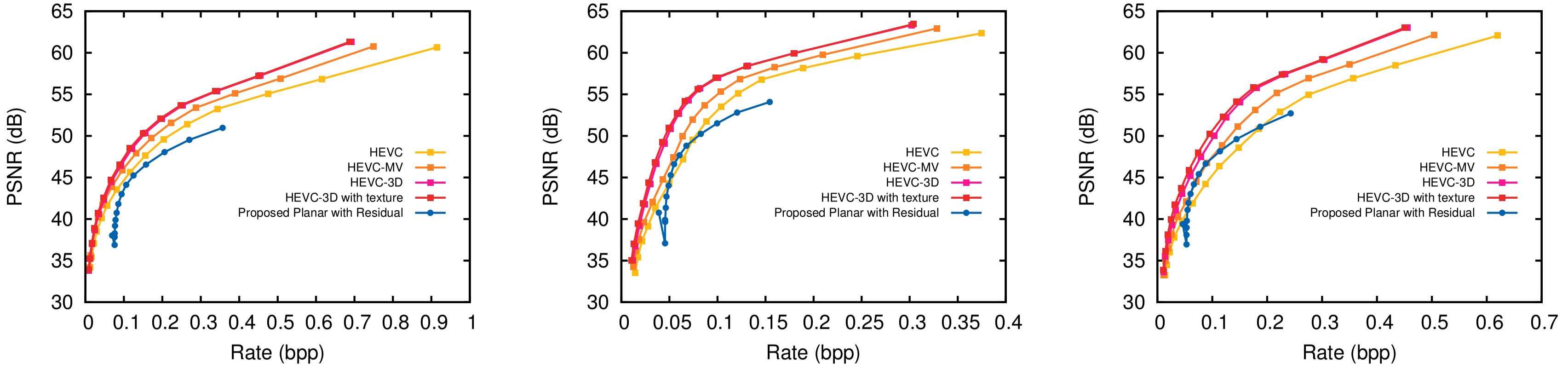


(a) Aloe (b) Art (c) Baby



(d) Books (e) Cloth (f) Cones

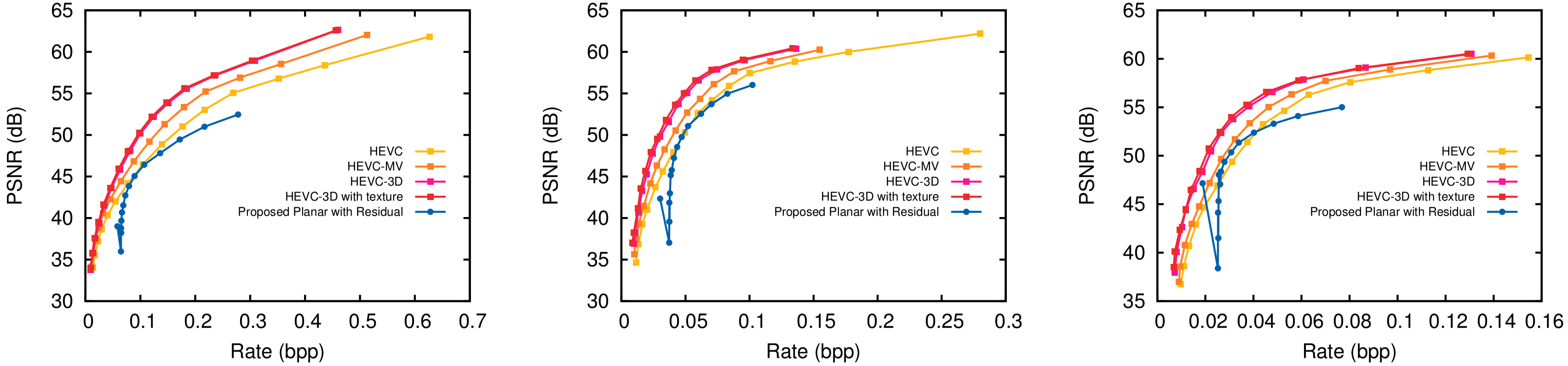
Figure 3.26: Mean PSNR values of the *depth reconstruction* results obtained by HEVC variants and planar prediction with HEVC in residual coding.



(a) Dolls

(b) Lampshade

(c) Laundry

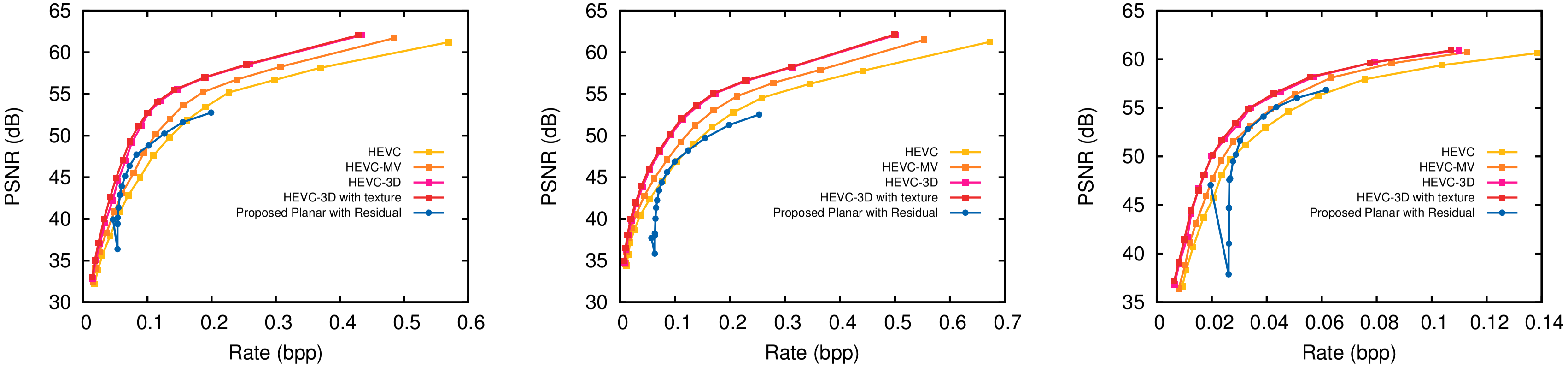


(d) Moebius

(e) Monopoly

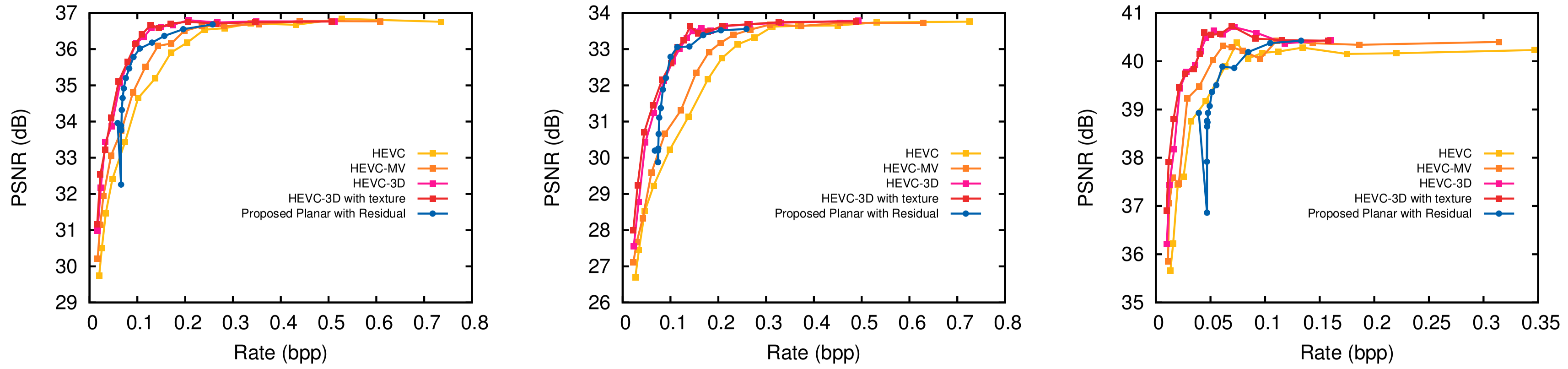
(f) Plastic

Figure 3.27: Mean PSNR values of the *depth reconstruction* results obtained by HEVC variants and planar prediction with HEVC in residual coding.

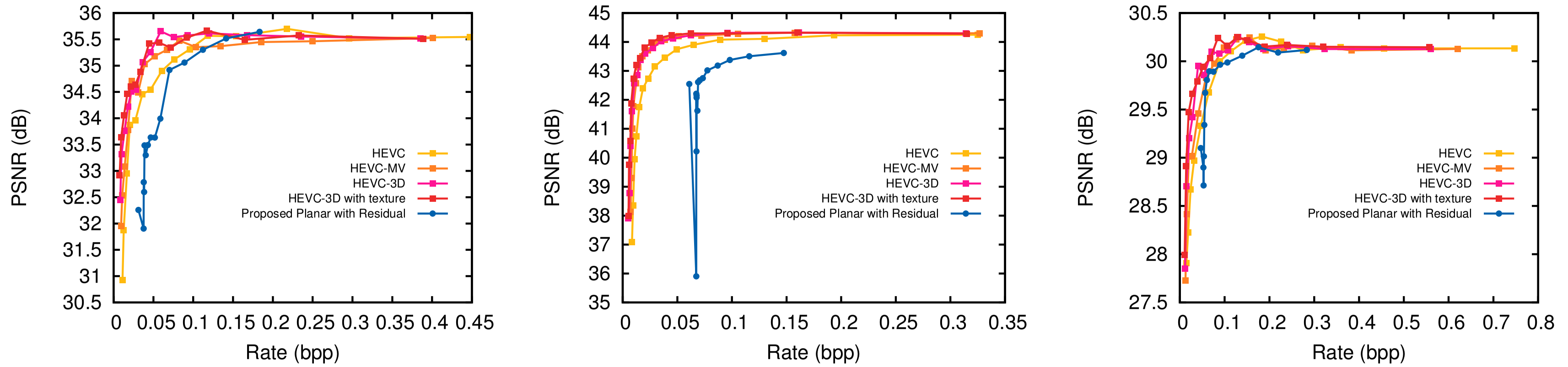


(a) Reindeer (b) Teddy (c) Wood

Figure 3.28: Mean PSNR values of the *depth reconstruction* results obtained by HEVC variants and planar prediction with HEVC in residual coding.

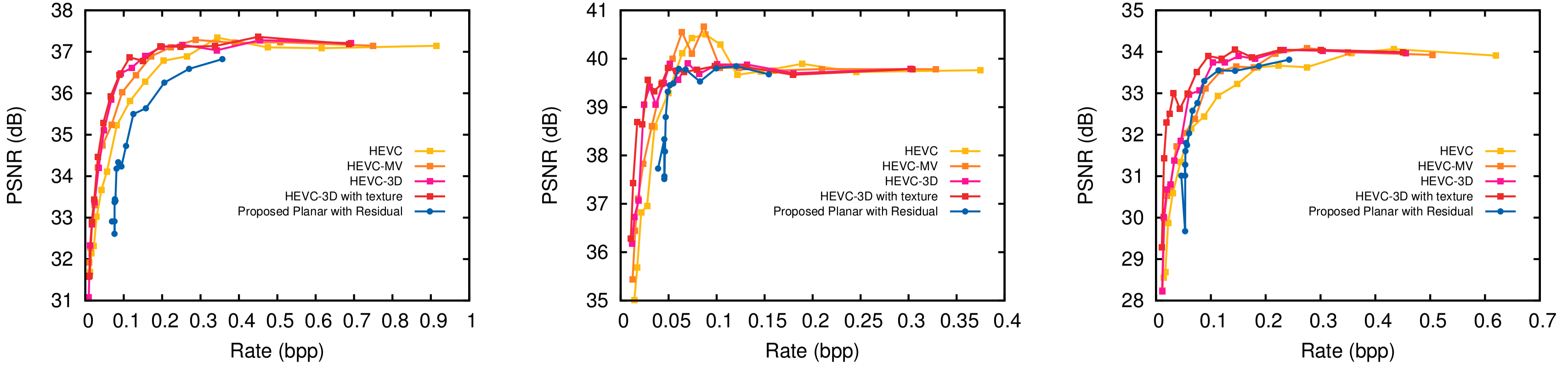


(a) Aloe (b) Art (c) Baby

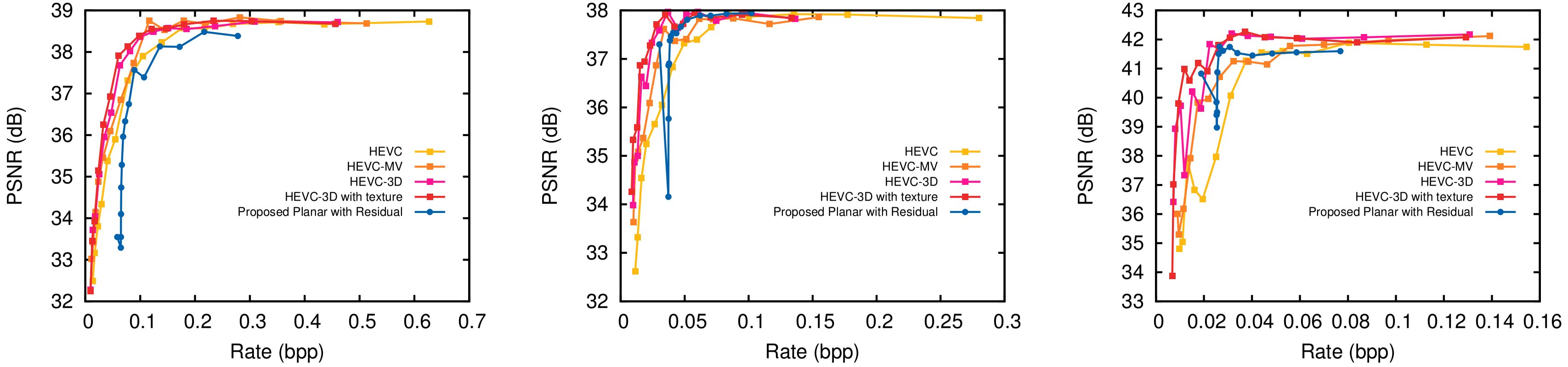


(d) Books (e) Cloth (f) Cones

Figure 3.29: PSNR values of the *novel view rendering* results obtained by the depth reconstructions of HEVC variants and planar prediction with HEVC in residual coding.

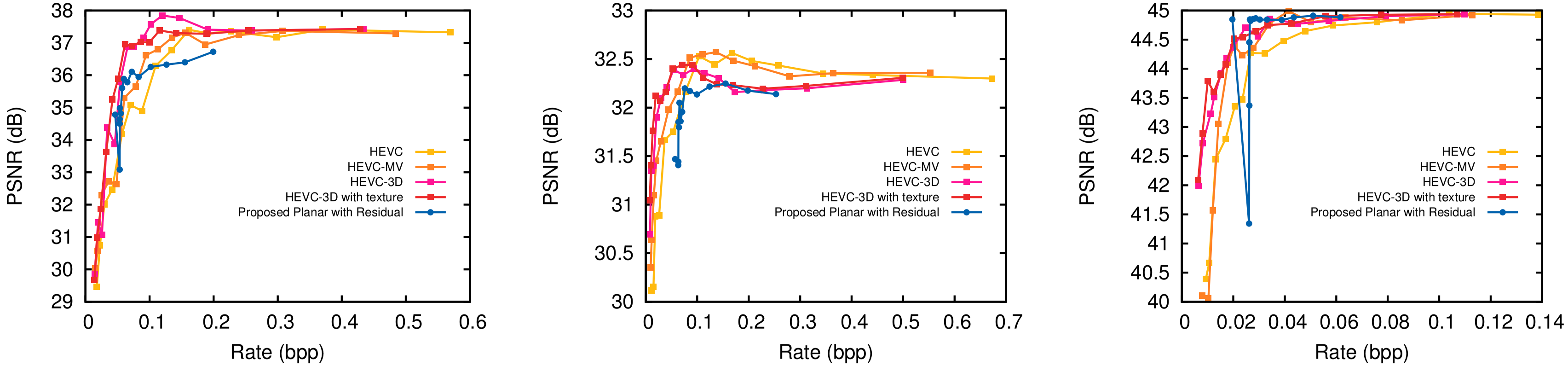


(a) Dolls (b) Lampshade (c) Laundry



(d) Moebius (e) Monopoly (f) Plastic

Figure 3.30: PSNR values of the *novel view rendering* results obtained by the depth reconstructions of HEVC variants and planar prediction with HEVC in residual coding.



(a) Reindeer (b) Teddy (c) Wood

Figure 3.31: PSNR values of the *novel view rendering* results obtained by the depth reconstructions of HEVC variants and planar prediction with HEVC in residual coding.

CHAPTER 4

PLANAR LAYERED MVD REPRESENTATION

The representation of a scene for a high quality novel view rendering is an ongoing research for the last few decades. During this period, the model based rendering techniques advanced in photorealism such that it is now hard to distinguish an image as real or synthetic. However, the real-world geometry is exceedingly complicated that the state of the art scene geometry acquisition techniques still can not provide sufficiently accurate models to utilize them in novel view rendering applications.

The alternative of the geometry based approach is the image based rendering (IBR) techniques. According to the plenoptic sampling theorem, it is possible to obtain perfect light field renderings from the samples of the light field at the minimum sampling rate which is determined by the minimum and maximum depth values of the scene [97], provided that given constraints hold. However, sampling the light field of a scene at the minimum sampling rate still results in a high number of views for the IBR applications. The number of images required for satisfactory rendering results can be reduced by incorporating the geometry information to the scene representation. This is the well-known tradeoff between the geometry and image based representations for rendering [98].

For feasible 3D applications, the scene representation solutions evolved by considering this tradeoff to MVD data format, which is expected to limit the number of views to 2 or 3 with their depth images [25]. The straightforward handling of the MVD format is to consider each texture and depth image in their native format. However, in order to improve compression and rendering capabilities of MVD data, novel representations are proposed in the literature. In the next section, MVD representations in the literature will be summarized and the following section will introduce a planar layer based MVD representation. Finally, the chapter will conclude with experimental results of the proposed MVD representation.

4.1 MVD Representations in the Litearture

The MVD representations can be classified into multi-reference and single reference based representations. In the raw format of the MVD data, each view is a reference point to represent the scene geometry and texture. The raw format has the advantage of backward compatible system designs by maintaining a gradual progress in the representation. The inclusion of the depth modality into the multiview video format brings new redundancies to be considered for efficient compression. Depth compensated inter-view predictions and inter-component predictions are new redundancy removal topics for the MVD representation [31].

As an example, the residual videos are proposed to replace the original texture videos in [99]. The residual videos are obtained by 3D warping of one of the views selected as the base view. This approach reasons its efficiency on the redundancies still exist between the residual images. The proposed representation also has the capability to fully recover the raw MVD data.

In [100], a ray-space based representation is utilized for the MVD data. The set of texture images and depth images are considered as an epipolar plane image (EPI) and an epipolar plane depth image (EPDI), respectively. The plenoptic constraints on EPDI and EPI are utilized to obtain global depth and texture information. The obtained global depth and texture information is represented in the multi-reference structure of the MVD format in a non-redundant way.

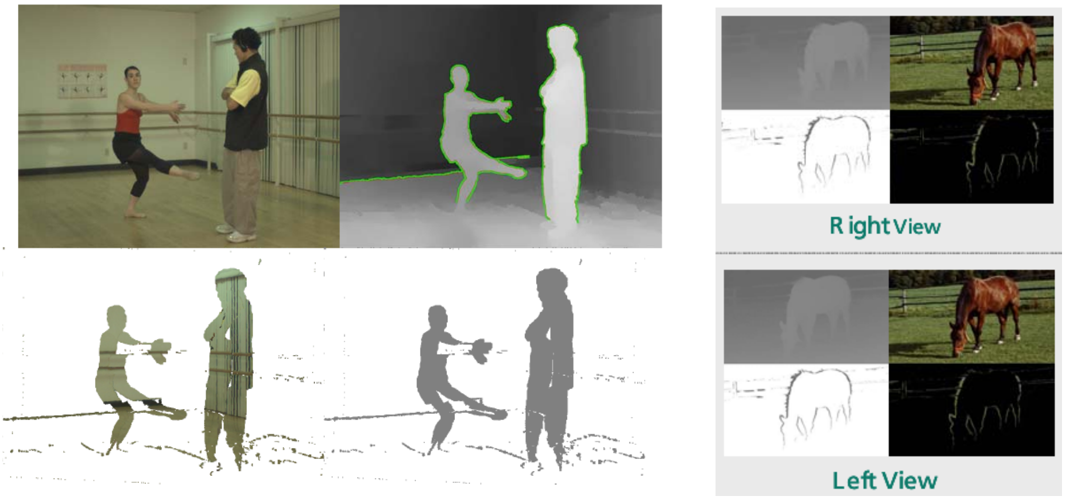
In [29], a non-symmetric multi-reference MVD representation is proposed in order to exploit the inter-view redundancies explicitly. In their study, Domanski et al., considers the mid-view of a multiview set as the base reference view and the side views as the complementary parts of the MVD representation. Only the disoccluded regions with respect to the dominant base view are encoded in the side views. Similar asymmetric approaches are also introduced in [101] as auxiliary information for inpainting occluded regions and in [102] as accumulated occlusion layers.

In [102], the authors also consider accumulating the occluded regions on the reference base view by 3D warping to obtain a single reference representation based on Layered Depth Image (LDI) format [103]. In LDI representation the projection rays of a single reference view are used to sample the visible and occluded surface points of the scene. The depth and color information of each surface point is kept in LDI in order to render novel views of the scene by simple visibility checks. Alternative layer extraction methods are discussed in [104],[105]. A similar LDI based approach is applied by a foreground/background segmentation in [106]. Another LDI based approach defines the depth layers by an object segmentation perspective [107].

In [108], the idea of depth ordered layers is combined with the multi-reference based representation. The so-called Depth Enhanced Stereo (DES) representation consists of two LDIs at the left and right views. The DES format is proposed as a unified, generic and backward compatible format for 3D applications and displays. Another backward compatible representation is proposed in [109] by mixing the conventional stereo and LDI.

In a recent study, [110], a constant depth layer based single reference representation is proposed for multiview images whose depth information is obtained by the method. The number of the constant depth layers is determined according to the plenoptic sampling theory [97] for a given baseline distance between the views. The depth values of the layers are determined by a Lloyd-Max quantization algorithm to minimize the assignment errors on the depth map. Although the main motivation of this study is obtaining high novel view rendering performance, a constant depth layer based texture compression proposed in [111] studies the compression efficiency of a similar layered representation.

Examples for multi-reference and single reference MVD representations are illustrated in Figure 4.1.



(a) LDI (b) DES

Figure 4.1: Examples of MVD representations in the literature. (Reprinted with permission. Copyright IEEE 2009 [108])

4.2 Proposed Planar Layered MVD Representation

As mentioned in the previous section, LDI based representations in the literature defines their layer models according to scene objects, foreground/background relations and/or constant depth approximations. The proposed planar layered MVD representation can be considered as an enhancement of the constant depth layer model to a planar one. While the constant depth approximation is justified by the plenoptic sampling theory in [110], the main motivation of the proposed representation is to define an object-like layers, as in [107], but in a fully automated and more compression friendly manner.

The proposed representation utilizes the previously introduced planar model fitting algorithms to assign a planar layer to each pixel. The number of the layers is regarded as a parameter which tunes the geometric approximation of the representation. Exactly the same planar representations obtained in Chapter 3 for the depth image pairs are packed into a single reference MVD representation.

The single reference view of the proposed representation is a novel view at the middle of the left and right views of the stereo pair. However, the field of view of the reference view is extended in the horizontal direction to cover all the regions visible to left and right views. This means, the image plane of the reference view

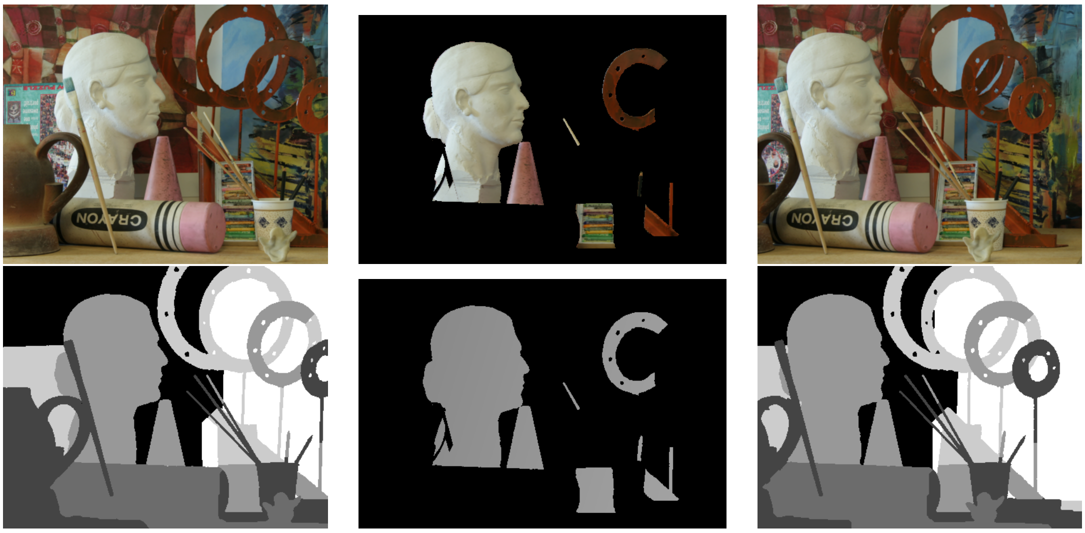


Figure 4.2: A planar layer extraction example for *Art* dataset. Left and right colums are the texture and planar labelings of the stereo pair. Mid column is the texture and the approximated planar depth of the extracted layer.

is extended in the horizontal direction as mentioned in [102].

According to layer definition, the depth maps of the layers are approximated by the corresponding planar model. Hence, the depth/geometry information of the proposed representation can be recovered by the spatial support of the layers and their planar model assignments. The spatial support of each layer is obtained by 3D warping the planar model assignment masks of the left and right views to the reference view in the middle. This operation can be regarded as merging the two reference views to a single reference view.

The texture of the layers is also obtained by merging the texture of side views on the reference view by 3D warping. For each layer, the texture images to be warped are defined by masking the the left and right texture images with the corresponding planar model labelling. At this point it is important to note that the 3D warpings for texture and spatial support are achieved by using the ground truth depth values of the pixels, but not their planar approximations. Otherwise, each view's texture information warped to the reference view may not coincide due to depth approximations. An illustrative example of a planar layer extraction is given in Figure 4.2.

The resulting planar layered MVD representation can be encoded as texture

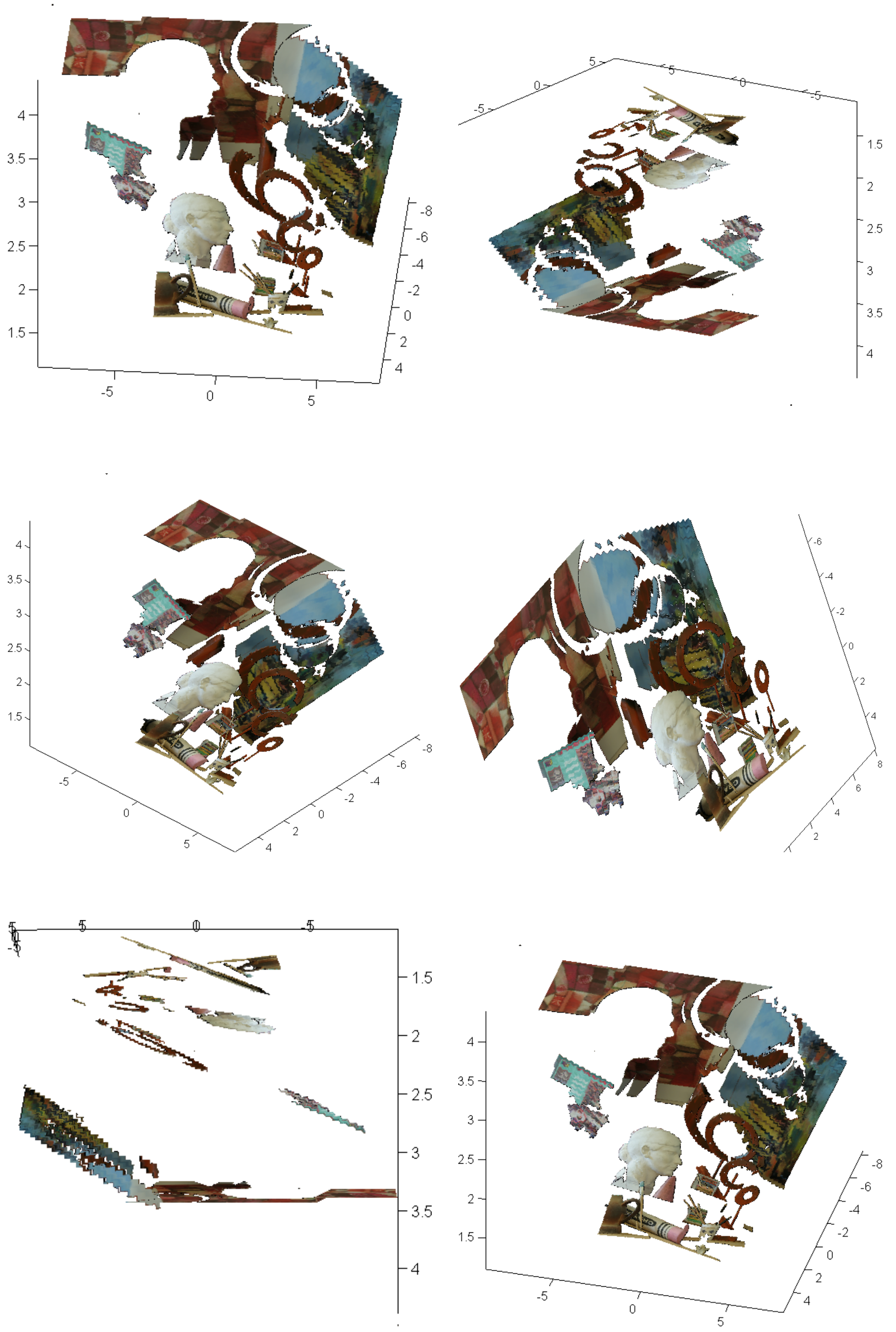


Figure 4.3: A 3D illustration of the planar layered MVD representation of *Art* dataset. Its corresponding layer information in 2D image planes are presented in Figure 4.4. In order to visualize the 3D extent of the planar layers, various camera views are rendered.

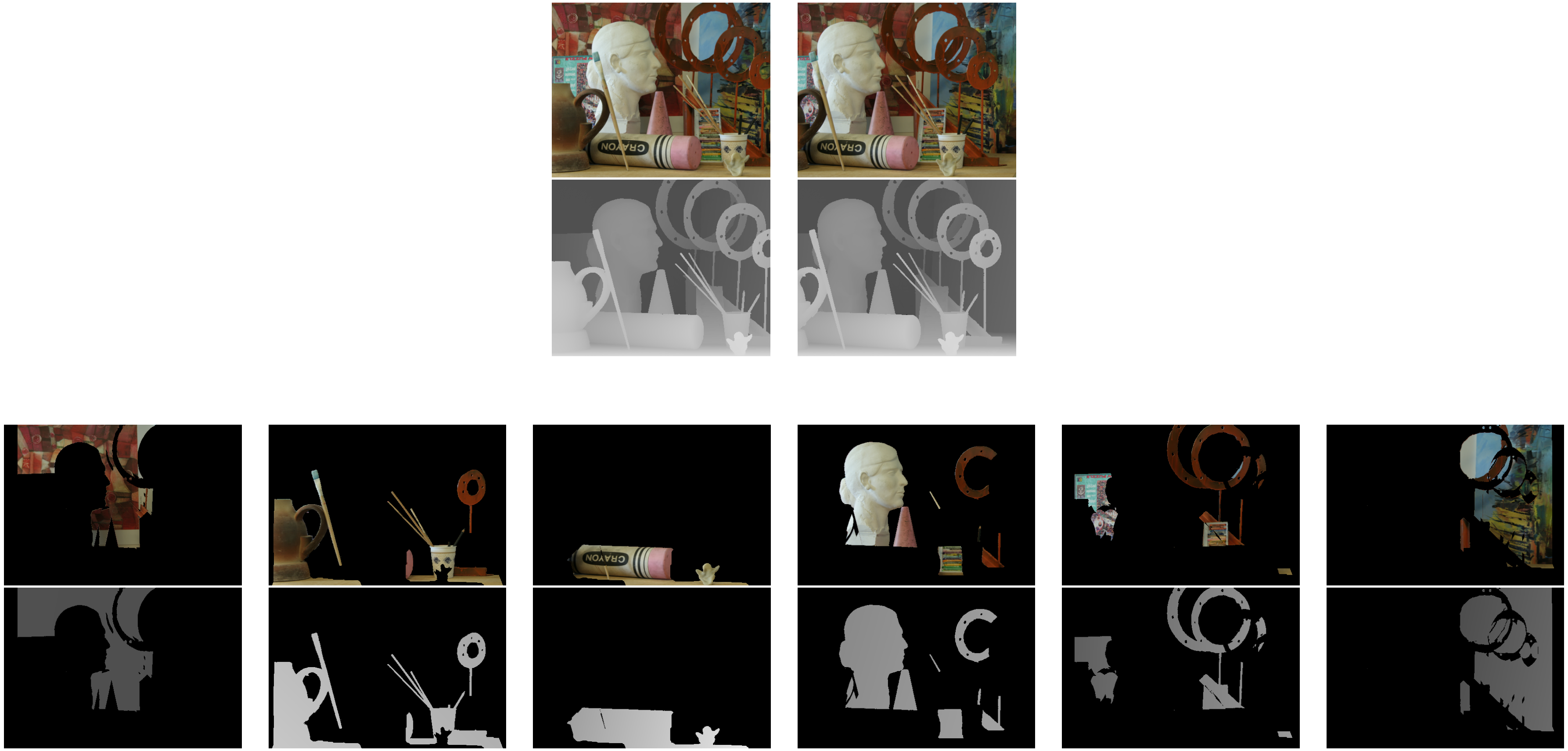


Figure 4.4: A planar layered MVD representation example for *Art* dataset. The top two rows represents the MVD data in its raw format. The bottom two rows are the texture images of the layers and their depth approximations according to the planar model assignments.

layers with their arbitrary boundary shape information and planar model parameters assigned to them. An example of MVD data is represented by 6 planar layers is given in the Figure 4.3 and 4.4 with its planar geometry approximations. The compression of texture layers can be performed efficiently by shape adaptive versions of DCT [112] or DWT [113] based encoders. These shape adaptive approaches also require the shape information of the boundaries explicitly. Hence, the proposed planar layered based MVD representation consists of layers whose shape, texture and planar model should be encoded.

Since the proposed representation merge the pixel information of the two reference view into a single one, the shape masks of the layers overlap at the occluded regions. Hence, the binary shape masks of the layers can be encoded as bit-planes of an image. As an example, the resulting shape image, whose bit-planes are the binary masks of the layers, is given in Figure 4.5 with the labeling images used during its layer extraction.

The novel view rendering for the proposed planar layered representation can be achieved by consecutively 3D warping the texture layers to the desired view. In order to satisfy the visibility constraints of the scene geometry, a z-buffer should be utilized. Since the geometry of the layers is planar, the 3D warping process can be done very efficiently by the well-known texture mapping techniques in computer graphics [114]. Such computer graphics based novel view rendering implementation also handles the z-buffer inherently.

The byte-stream package of the proposed planar layered representation's encoder is similar to the one defined in Section 3.2. The stream package starts with the number of planar models and their parameters, and they are followed by the size of the payloads of the shape and the texture information of the layers. The package ends with the streams of the shape and texture information of the layers consecutively. The byte-package defined for the planar layered based MVD representation is illustrated in Figure 4.6.

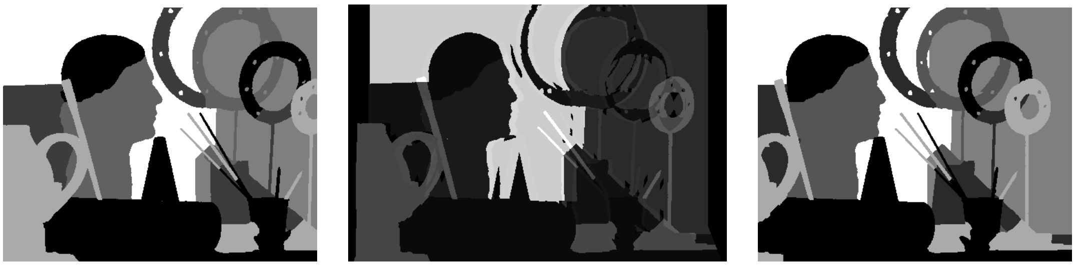


Figure 4.5: The left and right images are the labelings obtained by a planar representation of the stereo depth images. The shape of the extracted layers are bit-plane coded in the middle image. The number of planar models is 8 and histogram equalization is applied to labeling images for better visualization.

4.3 Experiments

The experiments studied the compression and novel view rendering capabilities of the proposed planar layered MVD representation in comparison to state of the art video coding standard, HEVC and its 3D extensions for the MVC and MVD formats and JPEG 2000. Without loss of generality the proposed layered MVD representation is analyzed for the two view case of MVD data.

Since the main problem definition of the thesis is an efficient depth representation for 3D applications, efficient handling and compression schemes for the texture component of the proposed layered representation are not considered and left as a future research topic. However, for the sake of completeness of the representation, lossless compression of the texture information is utilized in order to be able to obtain the streams of the proposed approach.

The layer texture images are handled as ordinary images and a common lossless image compression file format, PNG [115], is utilized to encode them. The concatenated PNG streams of the layered textures form the texture payload of the proposed MVD representation. For a fair comparison, the stereo views of the raw MVD data are also encoded with PNG to obtain its texture payload.

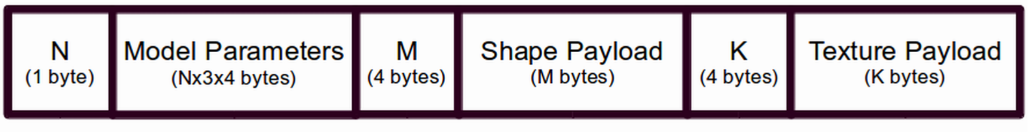
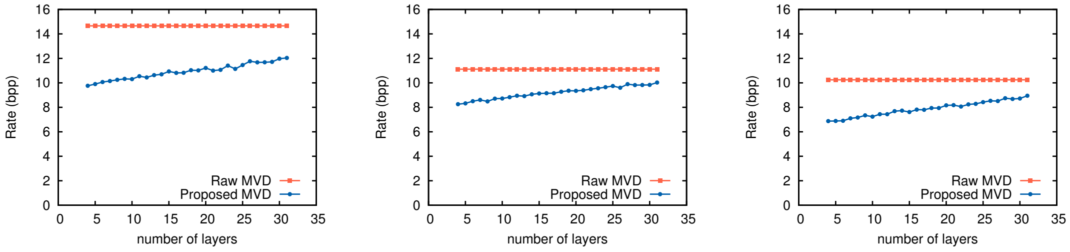


Figure 4.6: The layered planar MVD encoder's byte-stream package definition.

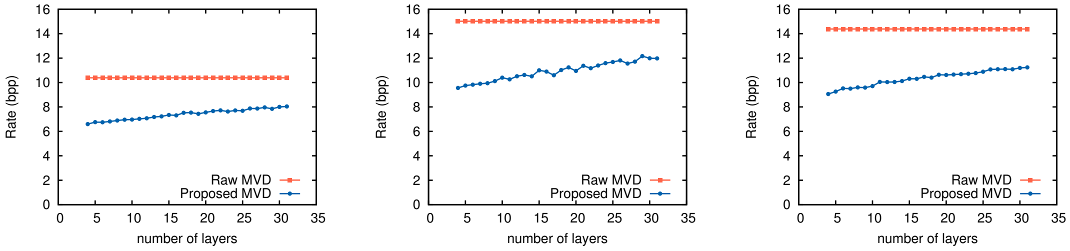
In contrast to a fixed number of views for the raw case, the number of encoded texture image changes by the number of layers utilized in the proposed representation. However, the layer texture images are sparse as shown in Figure 4.4, and their average sparsity increases by the increasing number of layers. In fact, the total texture information decreases by the merging operation of the two reference representation into a single reference representation.

In order to concentrate on the depth representation/compression efficiency of the proposal, the payloads of the texture and the other representation units for the scene geometry are studied separately. In Figure 4.7, the texture payload of the raw MVD data is compared against the proposed planar layered representation for varying number of layers. As an extreme case of single layer planar MVD representation, the layer texture extraction method would obtain a single texture image full of all visible points warped to reference camera in the middle. Hence, the size of the texture information should be almost halved with this rough sketch. By increasing the number of layers, the layered representation's capability to model the occluded regions will increase and the total texture information of the representation should increase. Hence, the dominant trend of increasing texture payload for the increasing number of layers can be found congruent with the expected. Another reason of the increase in the bit budget for the increasing number of layers can be the inherent shape encoding in the texture compression experiments due to utilization of conventional lossless image compression tools for the layer textures. Despite the sparseness of the layer textures is handled with generic tools, the results show that the proposed planar layered representation can enjoy the texture compression efficiency of the single reference based representations by handling the texture in an appropriate form, such as LDI [102].

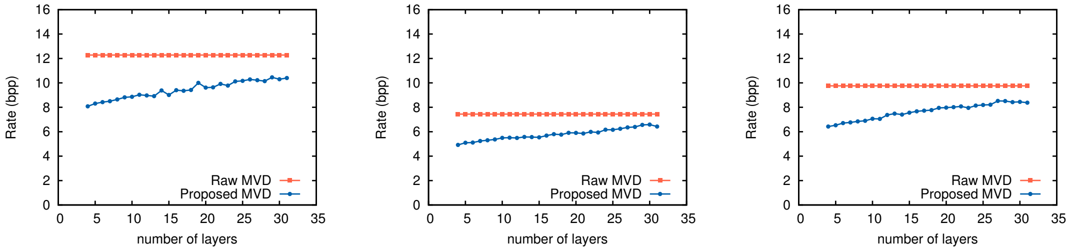
The shape masks of the extracted planar layers are encoded in a lossless manner by the PAQ8 compression tool. Since the shape mask of each layer is bit-plane coded, the layers are grouped into set of maximum 8 layers to define 8-bit per pixel images. The resulting set of images is compressed by PAQ8 and the obtained stream is set as the layer shape information payload of the the proposed representation.



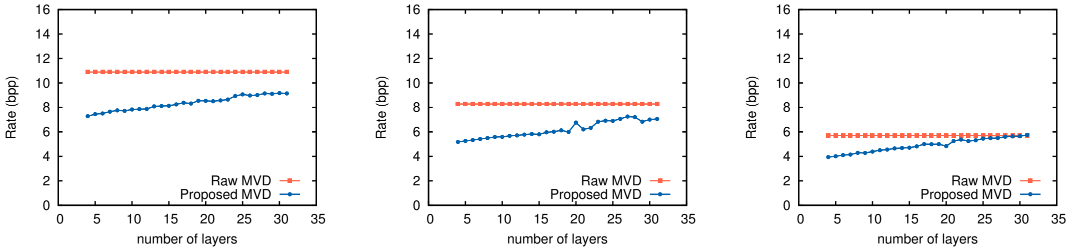
(a) Aloe (b) Art (c) Baby



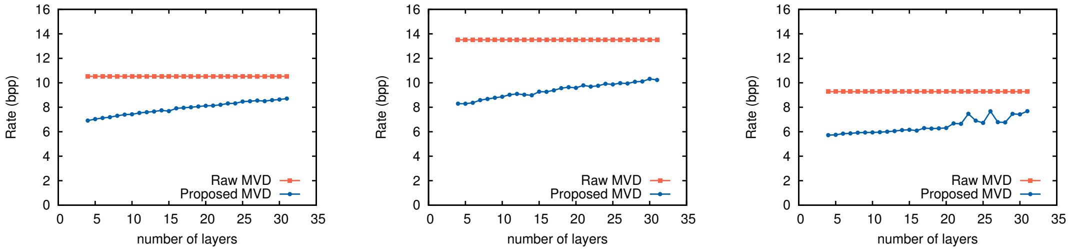
(d) Books (e) Cloth (f) Cones



(g) Dolls (h) Lampshade (i) Laundry



(j) Moebius (k) Monopoly (l) Plastic



(m) Reindeer (n) Teddy (o) Wood

Figure 4.7: The payload of the texture information in bit per pixel for the MVD representations in the raw and the proposed layered format. The payloads are obtained by lossless compression of the texture images by PNG encoder.

The novel view renderings for the raw MVD format is obtained by the reference rendering software provided by the standardization activities on the 3D extension of the HEVC [93]. However, the same software is not utilized for the proposed representation, since the software is designed for a two reference based approach. Accordingly for the proposed representation, the novel view renderings are obtained by a simple 3D warping algorithm with z-buffer. The possible holes left after warping all the layers are filled by extrapolating the texture information horizontally. The extrapolation side of the holes is selected as the deeper one by checking the z-buffer.

The novel view rendering results are analyzed for three equally spaced views between the left and the right views of the given MVD. The PSNR and SSIM metrics are utilized for measuring the novel view rendering quality with respect to ground truth captured images provided by *Middlebury* dataset.

In Figures 4.8-4.15 and 4.16-4.23, the novel view rendering scores in PSNR and SSIM are given, respectively, in three columns for these camera positions. At the mid-view, the proposed planar layered MVD representation provides the best rendering results, especially for the SSIM metric at lower bit rates. However, it performs the worst in general for the novel view renderings located on the left and right quarters of the baseline. The comparative performance change according to the position of the novel view is due to differences between the single and two reference based MVD representations.

In DIBR, the rendering results get better by decreasing the distance of the novel view to a reference view. In fact for the raw MVD representation, the novel views located at the reference views are already given by definition. Due to this fact, the JCT-VC included a 3-view plus 3-depth MVD scenario in their MVD coding standardization activities [25] to avoid the rendering performance drop in the middle of the viewing range of the 3D content. This DIBR related fact is observed as a novel view rendering performance drop at the side views for the proposed planar layered representation, since its single reference is defined to be at the mid point of the stereo pair.

Mid novel view rendering scores around the rendering quality for the ground truth depth case owe to the layer texture extraction method of the proposed approach. The texture information of the layers is extracted according to given depth maps of the views; hence, the texture information of the proposed planar layers does not contain any rendering artifacts due to depth distortions. However, when a novel view is not located at the reference view of the proposed representation, the planar approximations on depth information introduces rendering artifacts.

The sources of the rendering distortions for the proposed representation can be examined in two topics as texture- and geometry-based distortions. The texture-based distortions are introduced by the irreversible nature of the texture accumulation of multiple views into a single reference system. The texture of the regions visible by multiple views are blended by the DIBR process during layer extraction of the proposed approach. Hence, the novel view rendering of a given view in the raw MVD set cannot be recovered in general as long as no supplementary information is added to the representation. Although there are factors, such as lighting and the reflectance properties of the scene geometry that violates the Lambertian assumptions in blending the texture information, similar single reference based representations are proposed in the literature in order to explicitly remove textural redundancies [102],[29].

The geometric distortions introduced by the planar approximations of the scene also cause texture distortions in novel view rendering. Such geometric distortions might result in gradual horizontal shifts in the positions of the scene objects. In case the layer assignments of the proposed representation are congruent with the object boundaries, the planar layers with their texture information become an object representation with some approximations both in geometry and texture. Hence, object-wise planar layer assignments can be considered as manipulating the scene object in 3D space. However, the rate distortion similar energy formulation for the planar layer extraction of the proposed representation might violate the object boundaries. Such layer assignments break the visual integrity of the object in novel view rendering. This characteristic novel view rendering artifact of the proposed representation is shown in Figure 4.24 for *Aloe* dataset.

The texture based rendering distortions cannot be avoided since they are structural distortions of the proposed planar layered representation by definition. However, the geometry based novel view rendering distortions can be dissipated by decreasing the planar fitting error; i.e. by increasing the number of planar layers properly. The expected increase in PSNR and SSIM scores of the novel view renderings is shown on the left and right columns of the Figures 4.8 to 4.23.

Another interesting observation for the proposed method is slightly decreasing trend of the novel view rendering quality at the reference view for increasing bit rates, i.e. increasing number of layers. This phenomena can be explained by the alpha mating related artifacts at the layer boundaries. To increase the number of layers will result in more texture boundaries which are prone to artifacts in DIBR based texture extraction and novel view rendering steps of the proposed planar representation.

4.3.1 Visual Assessment

The objective evaluation of the proposed planar MVD representation by the PSNR and SSIM scores at different view points show that the performance of the proposed representation might vary according to the content of the MVD data. In general, the novel view rendering performance for mid-view is satisfactory due to its single reference based representation. However, the performance of the planar layered representation might drop severely as the view is located further from the reference mid-view. The basic 3D warping algorithm utilized in texture extraction and novel view rendering processes of the proposed representation can be accounted for a meaningful part of the performance drop against a state-of-the-art renderer [93] utilized for the other two reference based experiments. However, the visual characteristics of the proposed representation on novel view rendering and geometric distortions are worth to be analyzed in detail.

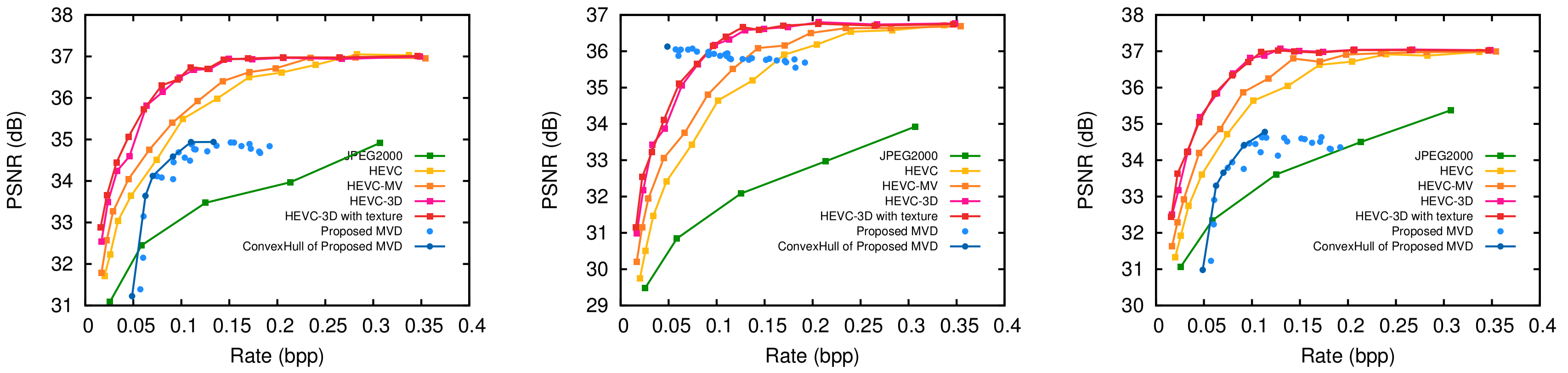
In Figures 4.25 to 4.27, the visual novel view rendering results in the middle and at the left or right quarter of the baseline are presented for three of the datasets in *Middlebury* set. The ground truth camera views of the novel views are also included for visual comparison. The compression instances for JPEG

2000, HEVC and proposed planar layered representation are selected visually according to the PSNR plots. For *Art* and *Moebius* datasets, all the instances provide similar PSNR scores at the shown quarter baseline view. The instances of *Books* dataset are selected to be obtained at the similar bit-rates for each MVD compression approach. The numeric details in PSNR, SSIM and bits per pixel (bpp) of the instances are also given in Tables 4.2 to 4.4.

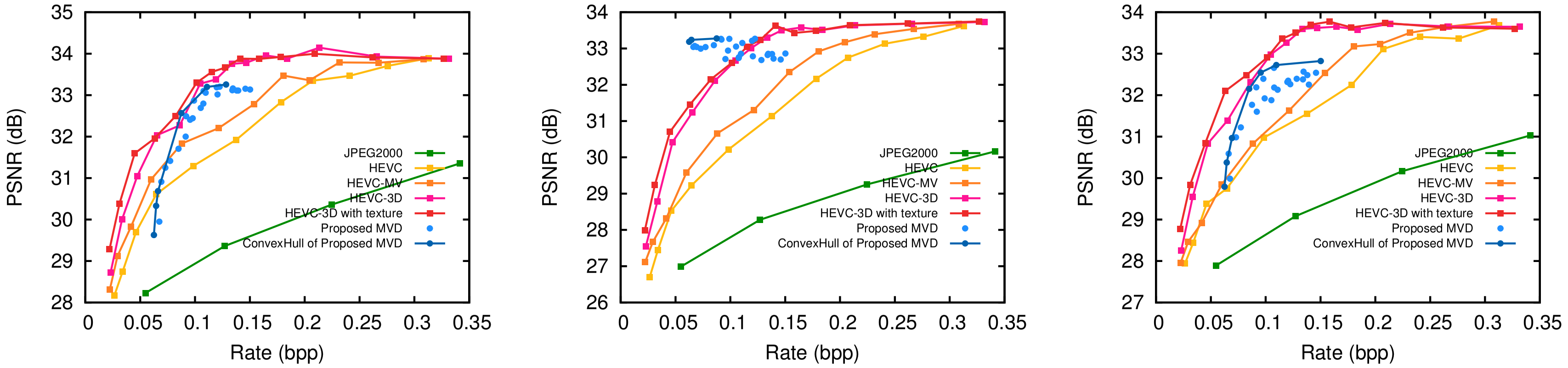
The novel view rendering characteristics of the JPEG 2000 and HEVC are quite similar, but HEVC is more effective in compression due to utilization of various state-of-the-art spatial predictive mechanisms of the standard. Either compression standard might not preserve the sharp depth discontinuities which results in rendering artifacts at the object boundaries. The brushes and pencils in *Art* dataset are rendered as a fuzzy point cloud, whithout any object integrity. The other objects with greater spatial support also show rendering artifacts at the boundaries.

In contrast to the conventional compression standards, the planar layered representation based novel view renderings draws those brushes and pencils clearly. In general the object boundaries are well preserved in the planar case which is an expected positive outcome of the explicit boundary definition of the planar layers. However, the object integrities might still be lost in the planar layered case, if the layer boundaries do not coincide with the object boundaries. Such an example is visible at the bottom of the brush on the left side of the Figure 4.25e. However, in comparison to fuzzy characteristics of the rendering artifacts observed in JPEG 2000 and HEVC, this kind of artifacts are very concrete by the well-defined geometry on the texture layers of the proposed MVD representation.

The similar novel view rendering characteristics can be observed for *Moebius* dataset in Figure 4.26. The game cards and the geometric object at the right bottom of the scene suffer with the fuzzy object boundaries in JPEG 2000 and HEVC experiments. On the other hand there is no striking rendering artifact for the proposed planar case. However, the measured PSNR and SSIM scores of the rendered views located at the right quarter of the baseline are very similar

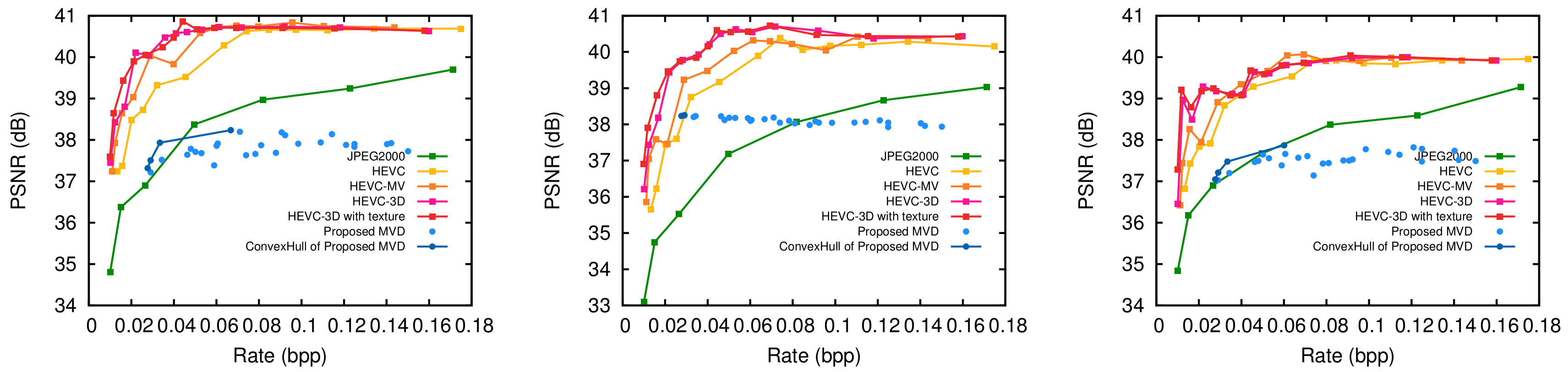


(a) Aloe @0.25 (b) Aloe @0.5 (c) Aloe @0.75

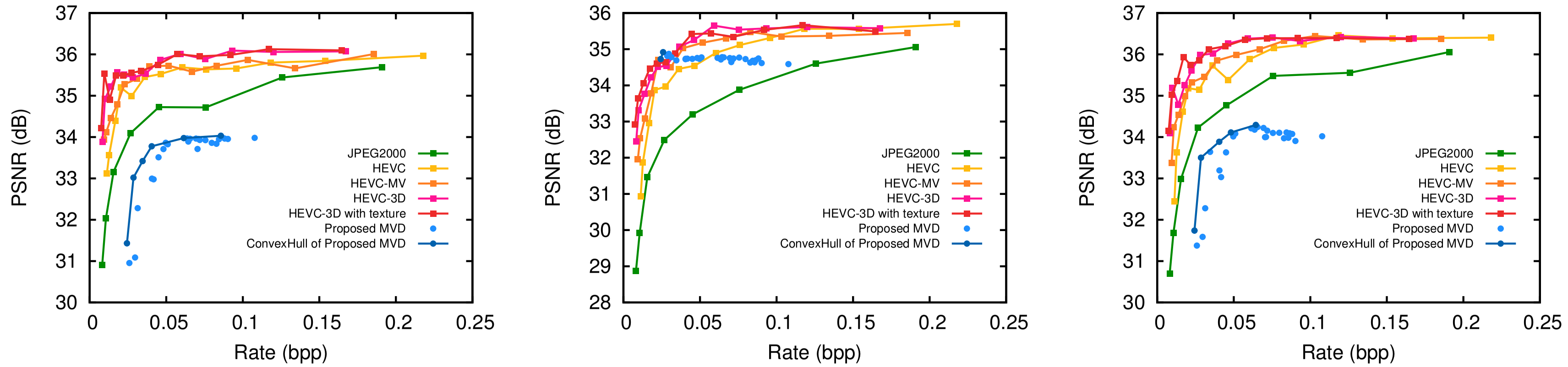


(d) Art @0.25 (e) Art @0.5 (f) Art @0.75

Figure 4.8: The results of experiment, comparing the novel view rendering PSNR values obtained by the proposed and the conventional MVD compression schemes.

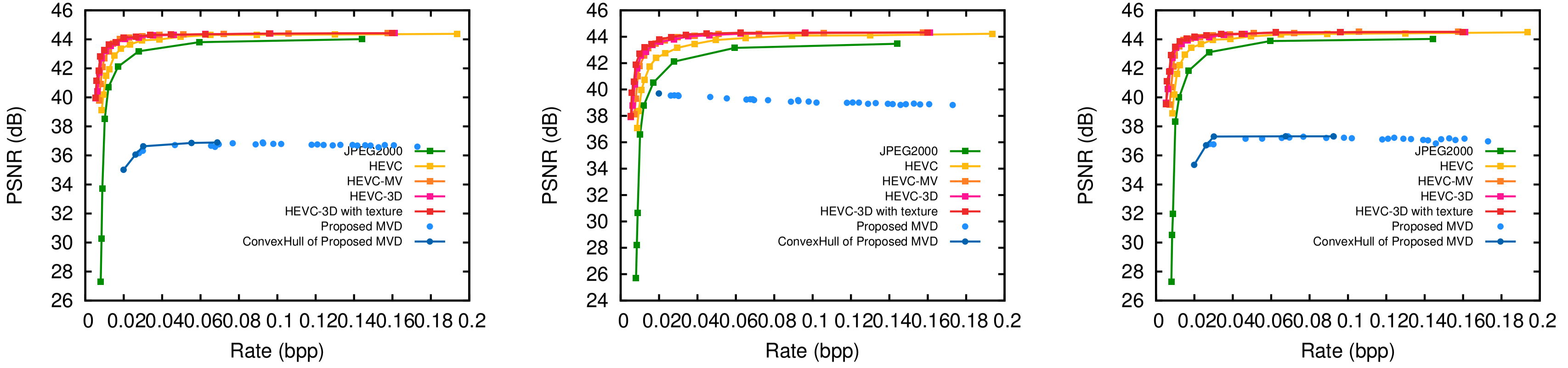


(a) Baby @0.25 (b) Baby @0.5 (c) Baby @0.75

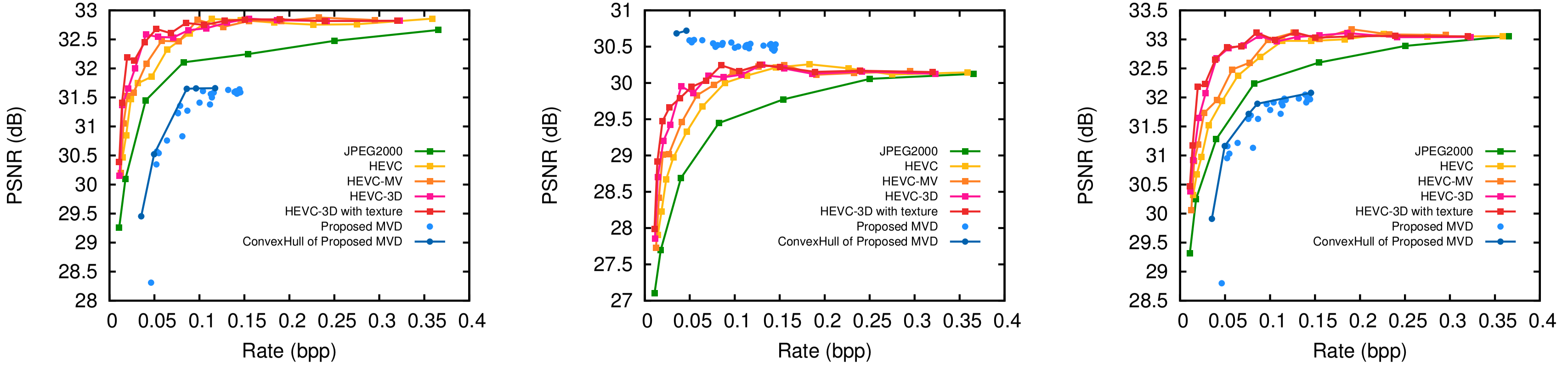


(d) Books @0.25 (e) Books @0.5 (f) Books @0.75

Figure 4.9: The results of experiment, comparing the novel view rendering PSNR values obtained by the proposed and the conventional MVD compression schemes.

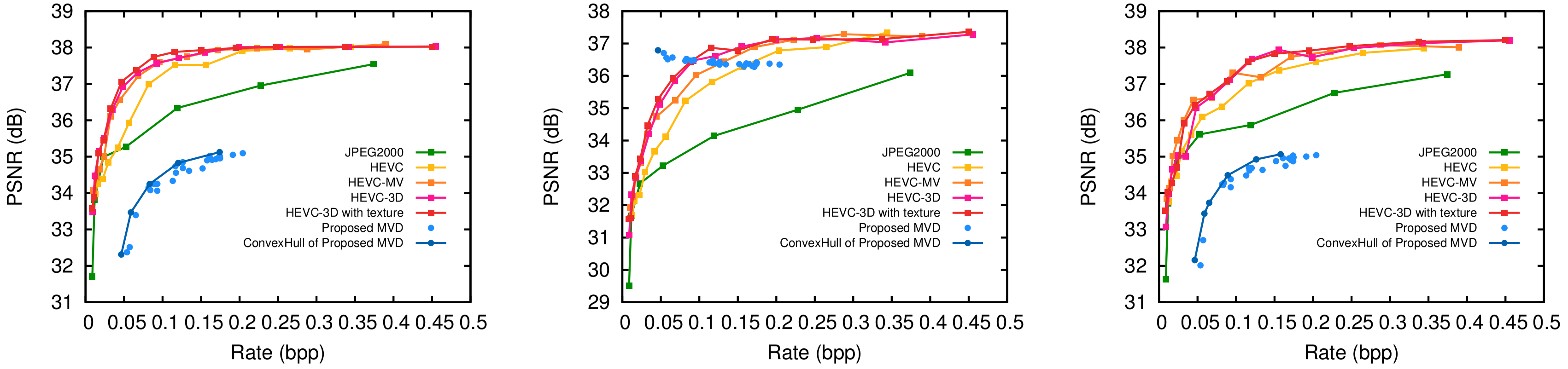


(a) Cloth @0.25 (b) Cloth @0.5 (c) Cloth @0.75



(d) Cones @0.25 (e) Cones @0.5 (f) Cones @0.75

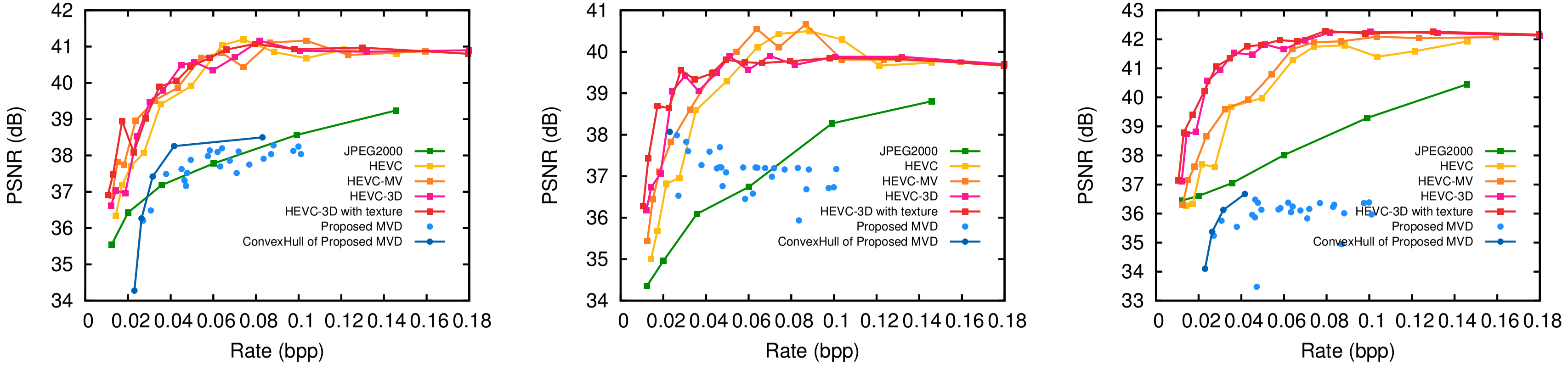
Figure 4.10: The results of experiment, comparing the novel view rendering PSNR values obtained by the proposed and the conventional MVD compression schemes.



(a) Dolls @0.25

(b) Dolls @0.5

(c) Dolls @0.75

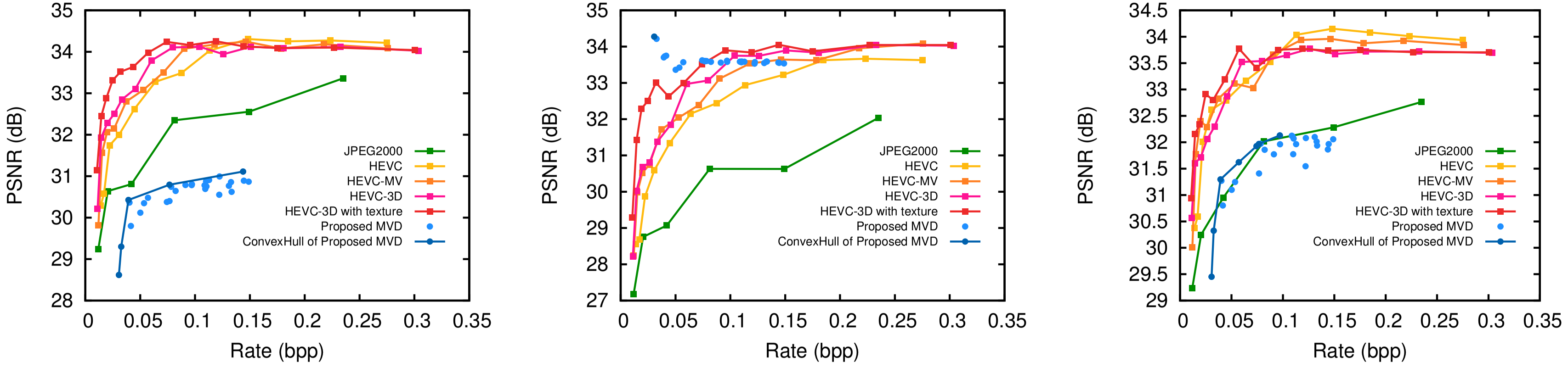


(d) Lampshade @0.25

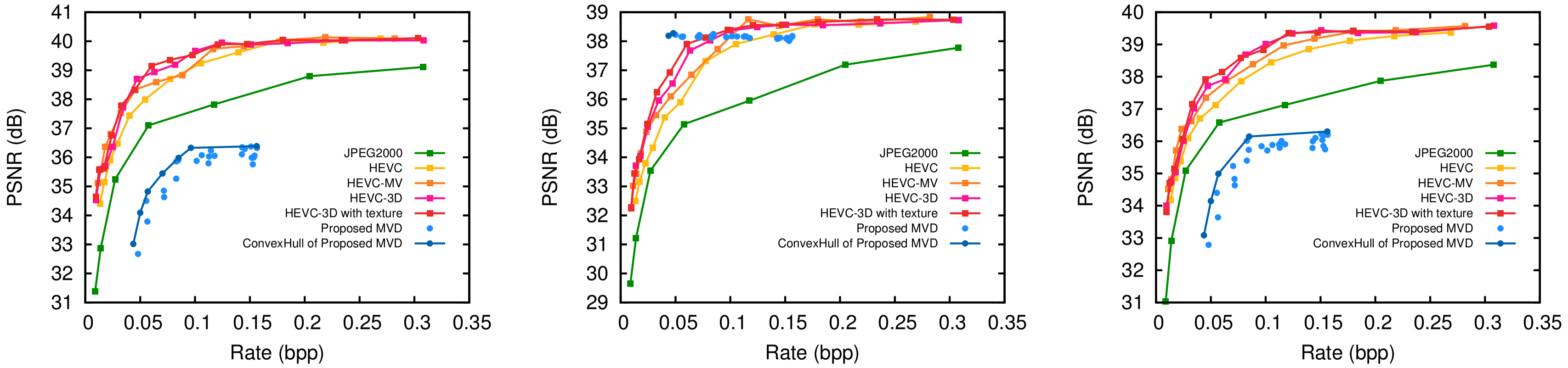
(e) Lampshade @0.5

(f) Lampshade @0.75

Figure 4.11: The results of experiment, comparing the novel view rendering PSNR values obtained by the proposed and the conventional MVD compression schemes.

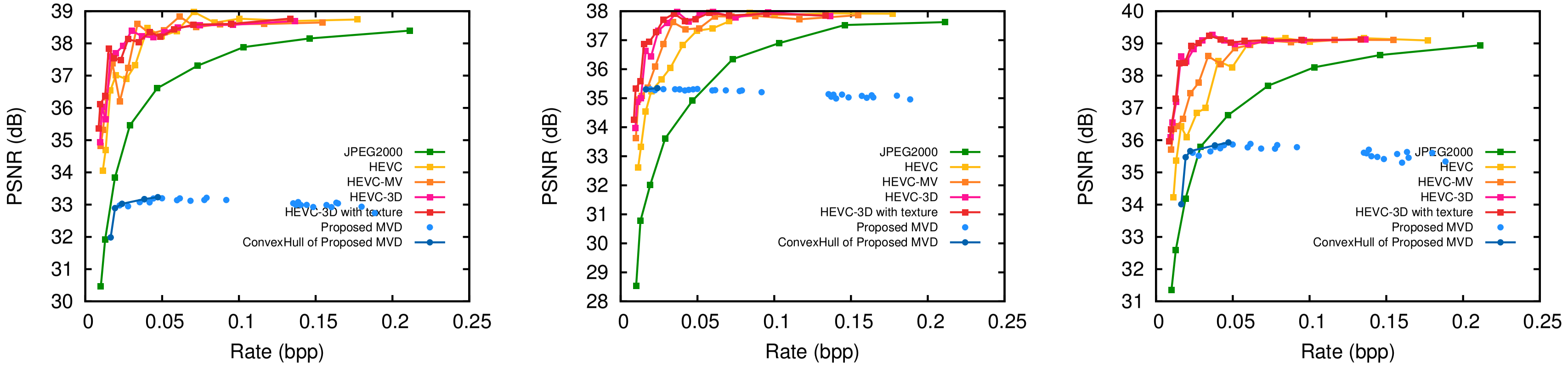


(a) Laundry @0.25 (b) Laundry @0.5 (c) Laundry @0.75

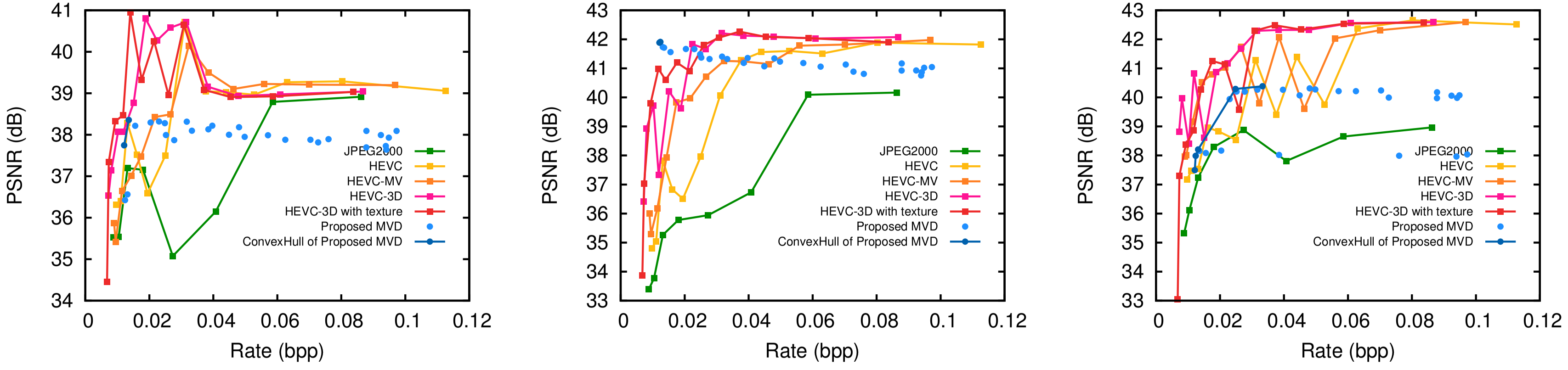


(d) Moebius @0.25 (e) Moebius @0.5 (f) Moebius @0.75

Figure 4.12: The results of experiment, comparing the novel view rendering PSNR values obtained by the proposed and the conventional MVD compression schemes.

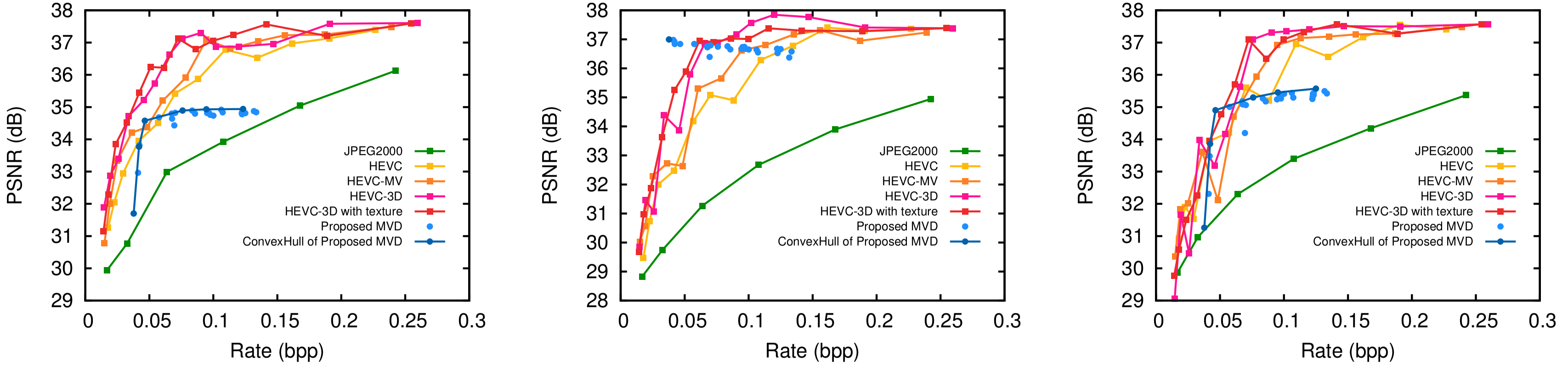


(a) Monopoly @0.25 (b) Monopoly @0.5 (c) Monopoly @0.75

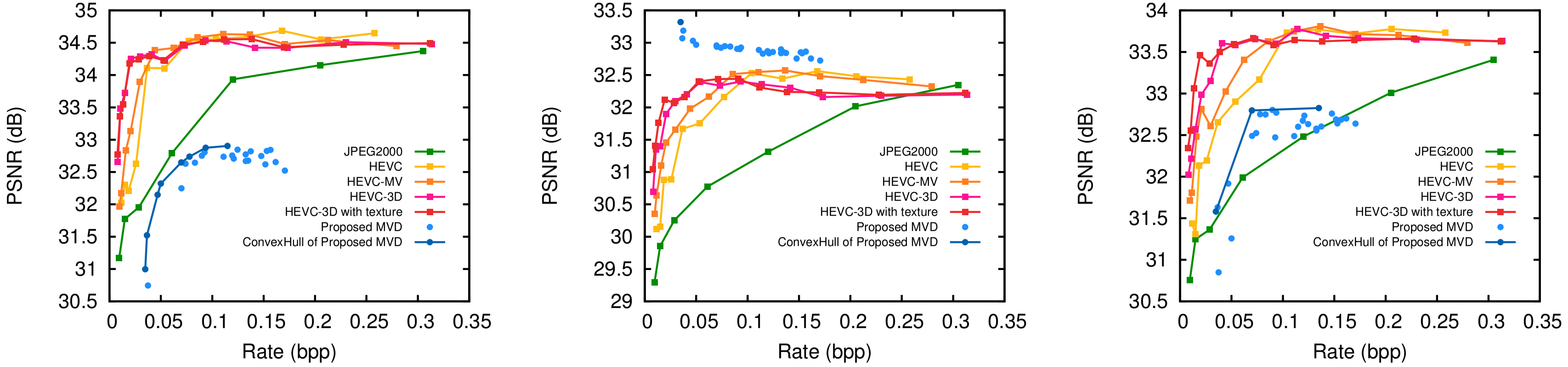


(d) Plastic @0.25 (e) Plastic @0.5 (f) Plastic @0.75

Figure 4.13: The results of experiment, comparing the novel view rendering PSNR values obtained by the proposed and the conventional MVD compression schemes.

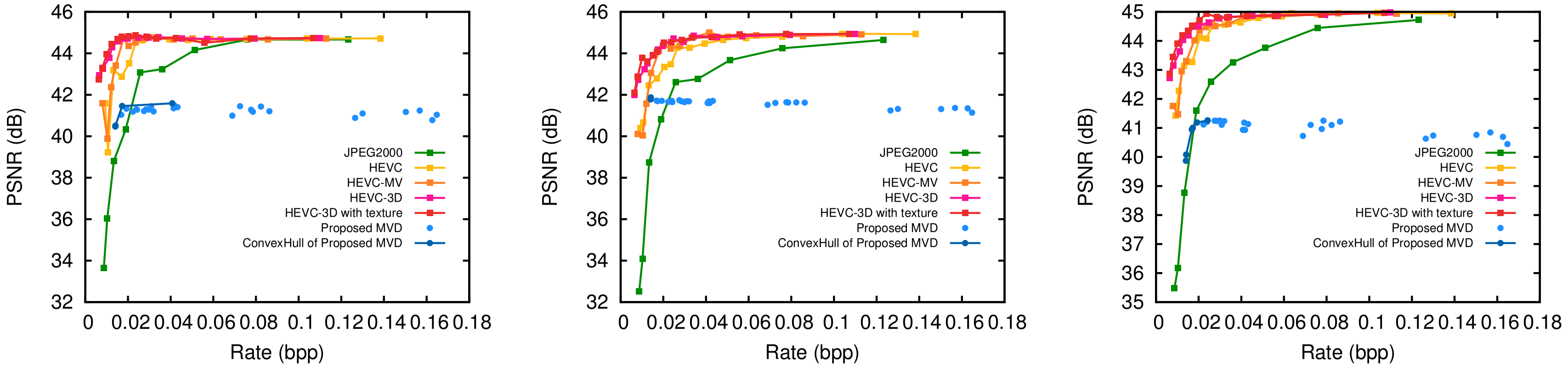


(a) Reindeer @0.25 (b) Reindeer @0.5 (c) Reindeer @0.75



(d) Teddy @0.25 (e) Teddy @0.5 (f) Teddy @0.75

Figure 4.14: The results of experiment, comparing the novel view rendering PSNR values obtained by the proposed and the conventional MVD compression schemes.

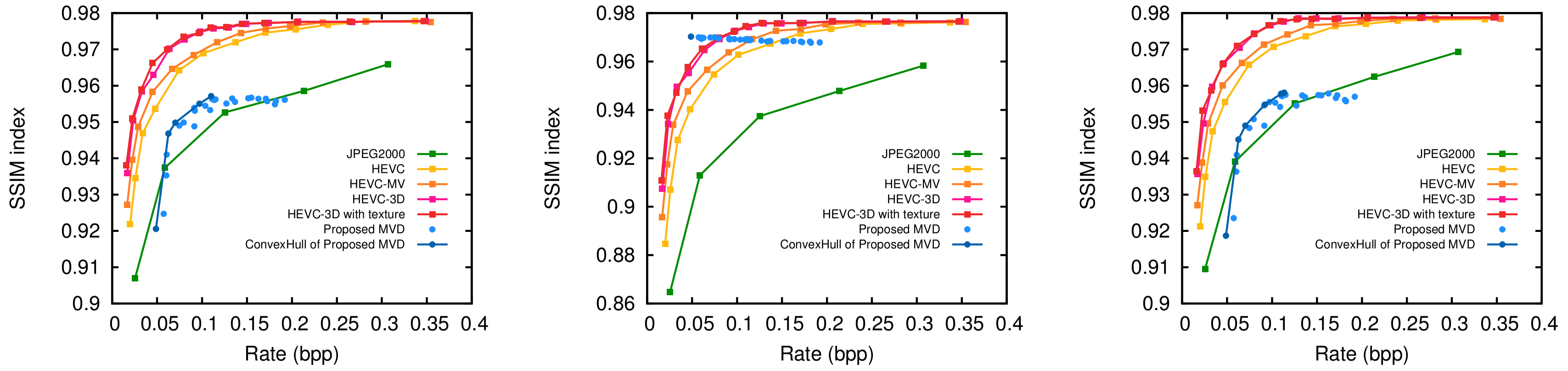


(a) Wood @0.25

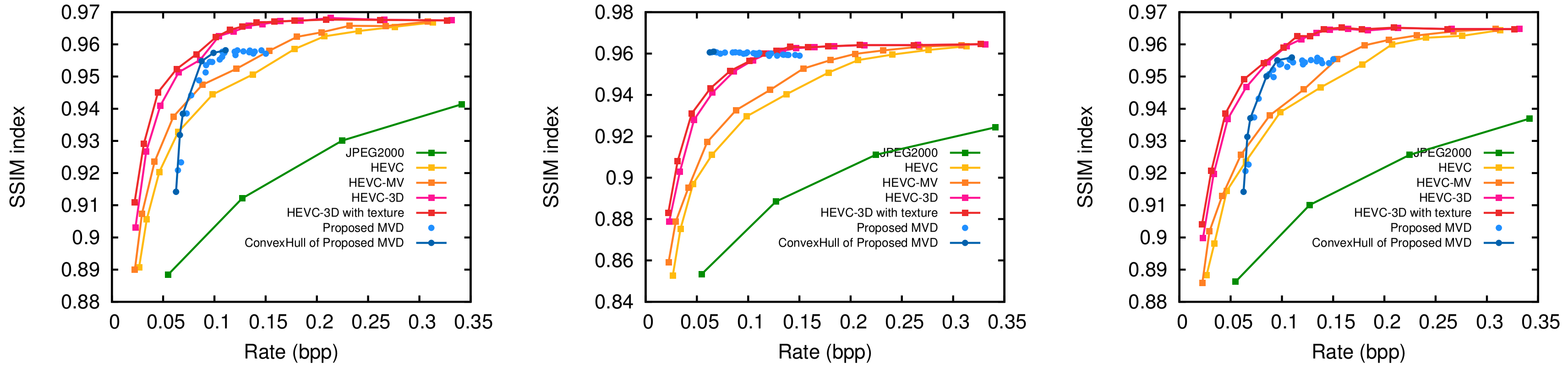
(b) Wood @0.5

(c) Wood @0.75

Figure 4.15: The results of experiment, comparing the novel view rendering PSNR values obtained by the proposed and the conventional MVD compression schemes.

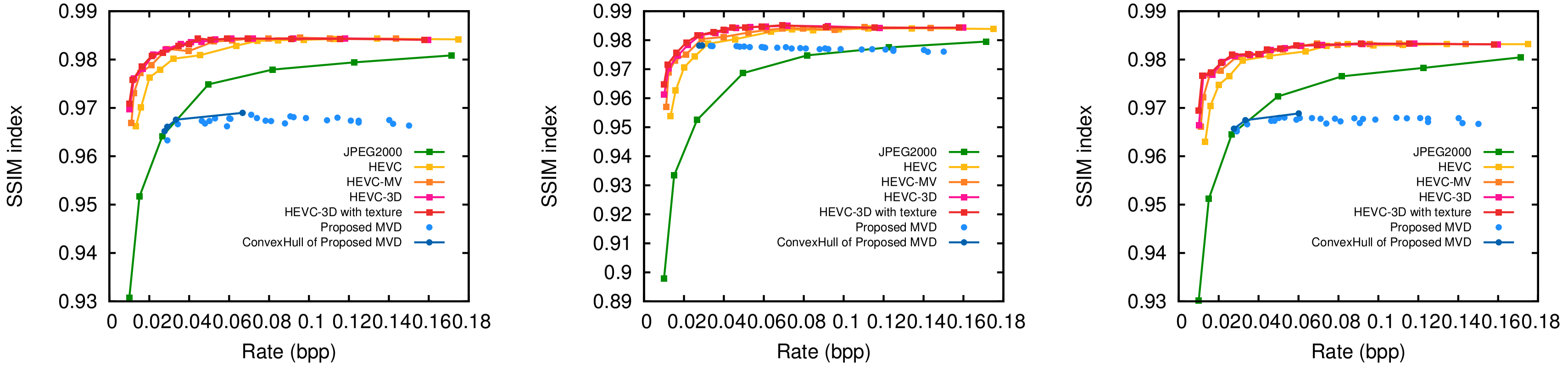


(a) Aloe @0.25 (b) Aloe @0.5 (c) Aloe @0.75



(d) Art @0.25 (e) Art @0.5 (f) Art @0.75

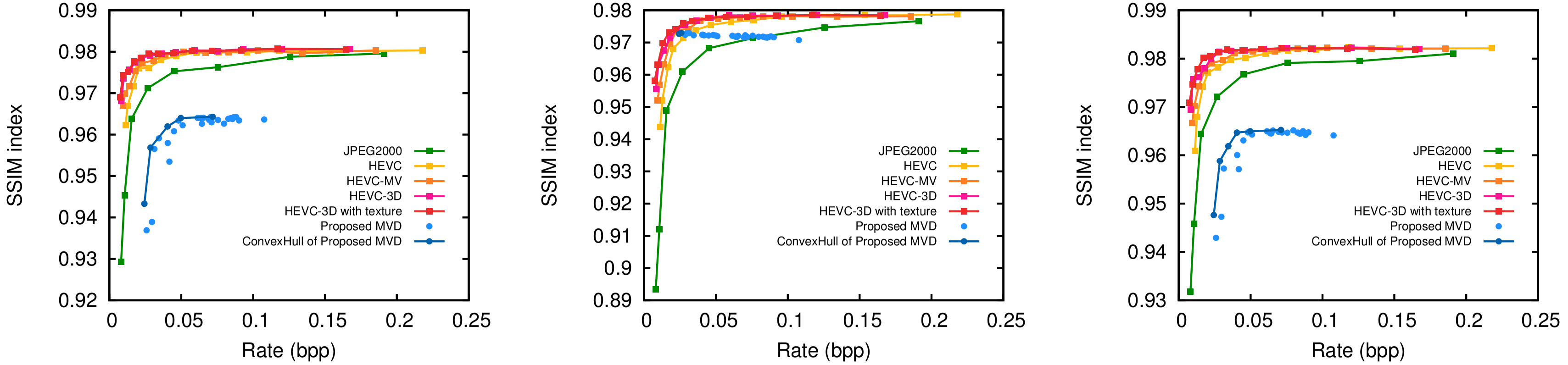
Figure 4.16: The results of experiment, comparing the novel view rendering SSIM scores obtained by the proposed and the conventional MVD compression schemes.



(a) Baby @0.25

(b) Baby @0.5

(c) Baby @0.75

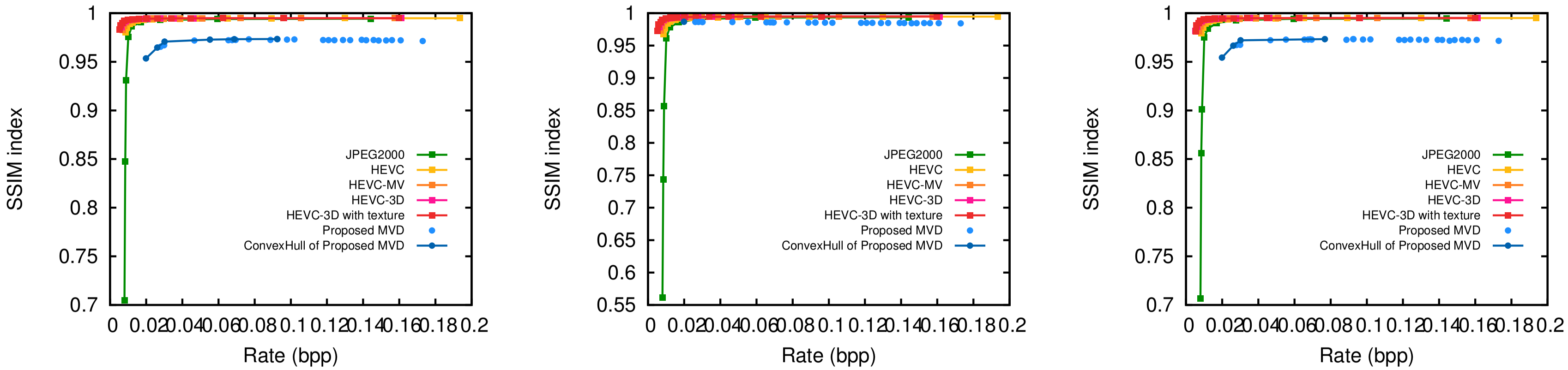


(d) Books @0.25

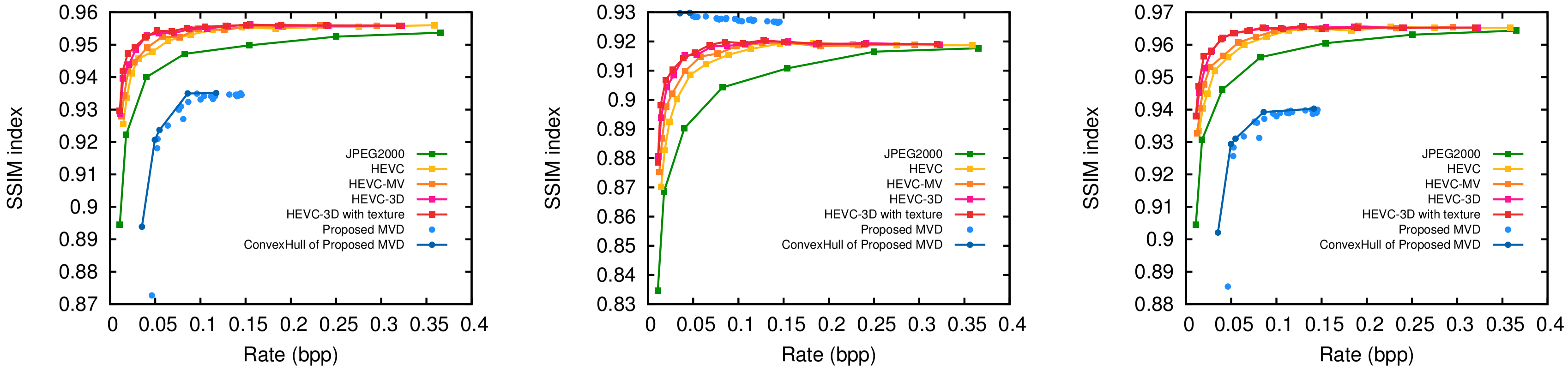
(e) Books @0.5

(f) Books @0.75

Figure 4.17: The results of experiment, comparing the novel view rendering SSIM scores obtained by the proposed and the conventional MVD compression schemes.

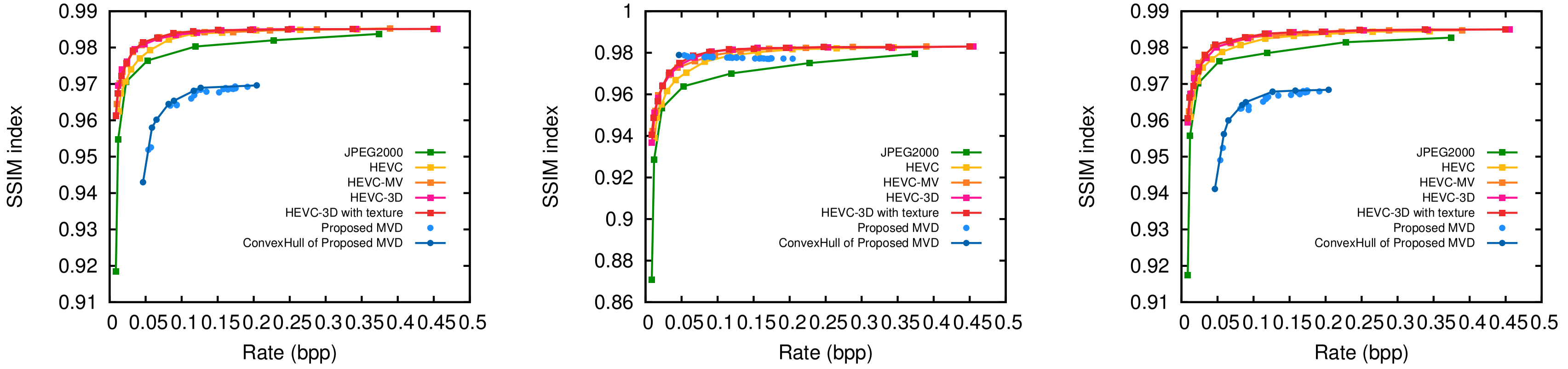


(a) Cloth @0.25 (b) Cloth @0.5 (c) Cloth @0.75

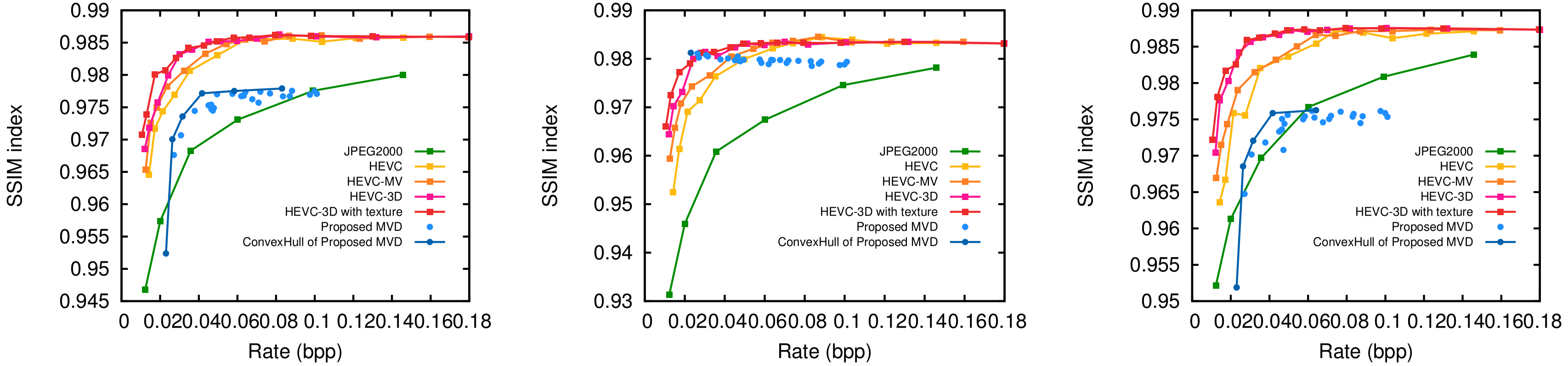


(d) Cones @0.25 (e) Cones @0.5 (f) Cones @0.75

Figure 4.18: The results of experiment, comparing the novel view rendering SSIM scores obtained by the proposed and the conventional MVD compression schemes.

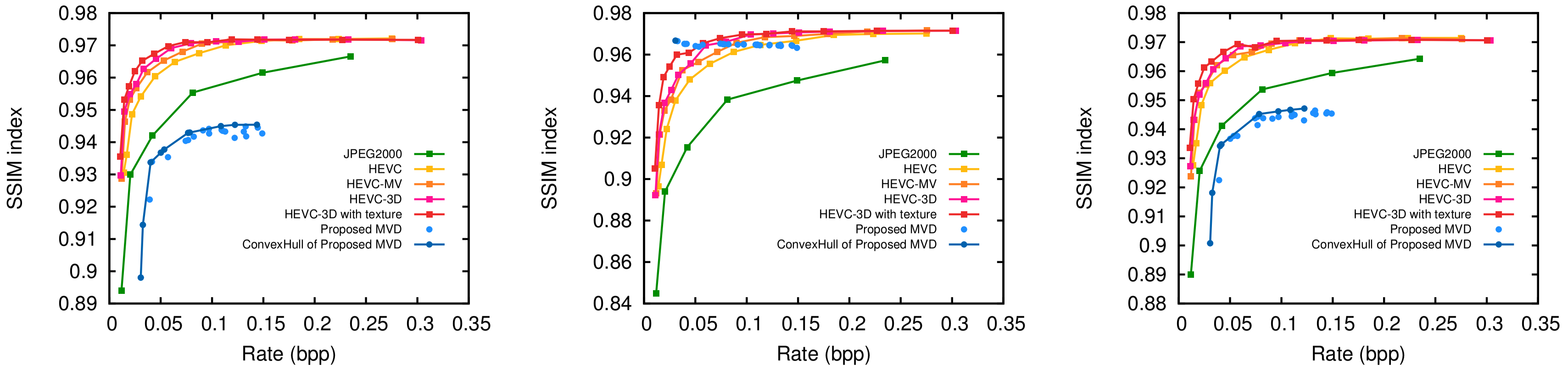


(a) Dolls @0.25 (b) Dolls @0.5 (c) Dolls @0.75

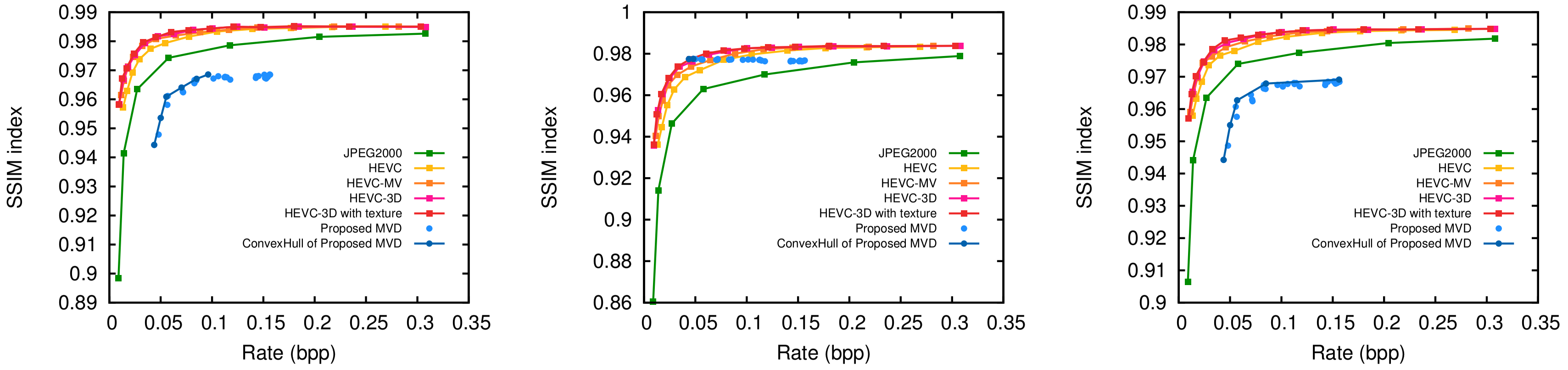


(d) Lampshade @0.25 (e) Lampshade @0.5 (f) Lampshade @0.75

Figure 4.19: The results of experiment, comparing the novel view rendering SSIM scores obtained by the proposed and the conventional MVD compression schemes.

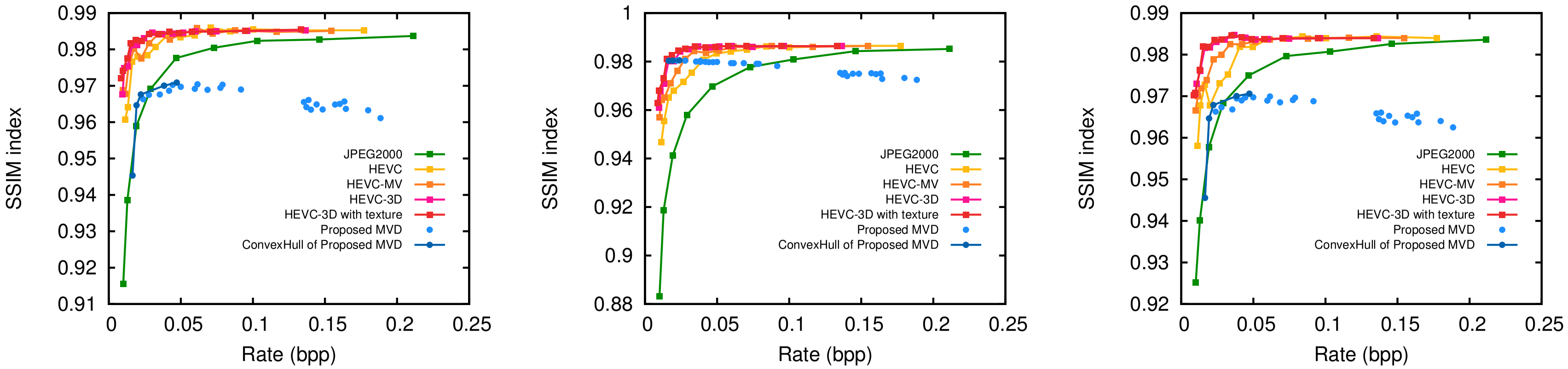


(a) Laundry @0.25 (b) Laundry @0.5 (c) Laundry @0.75

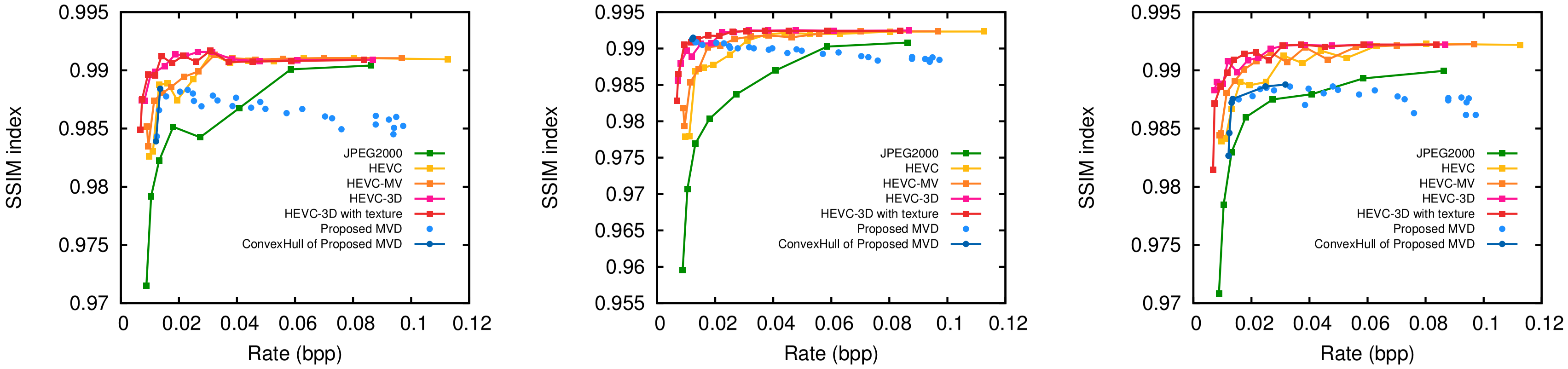


(d) Moebius @0.25 (e) Moebius @0.5 (f) Moebius @0.75

Figure 4.20: The results of experiment, comparing the novel view rendering SSIM scores obtained by the proposed and the conventional MVD compression schemes.

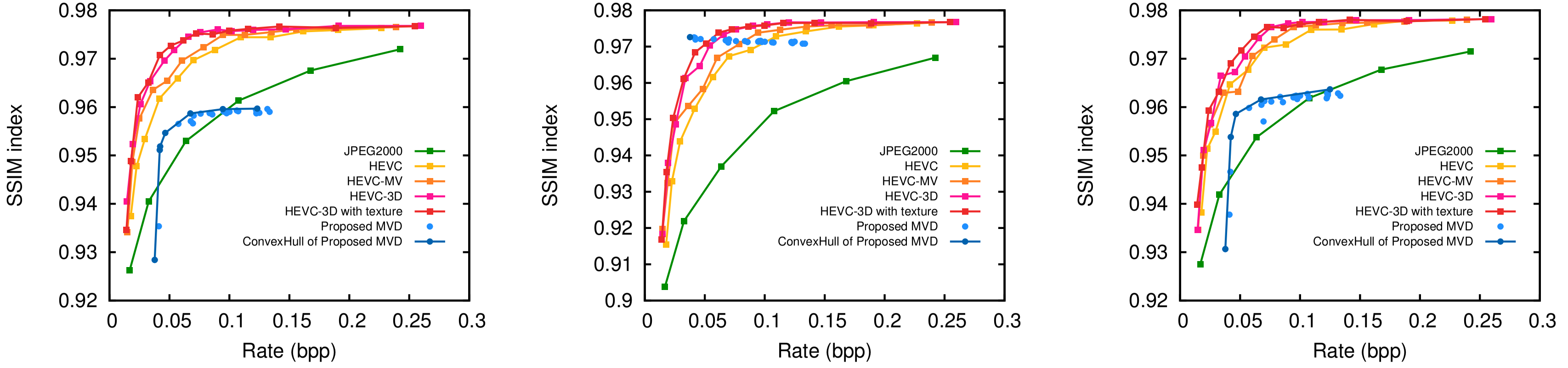


(a) Monopoly @0.25 (b) Monopoly @0.5 (c) Monopoly @0.75

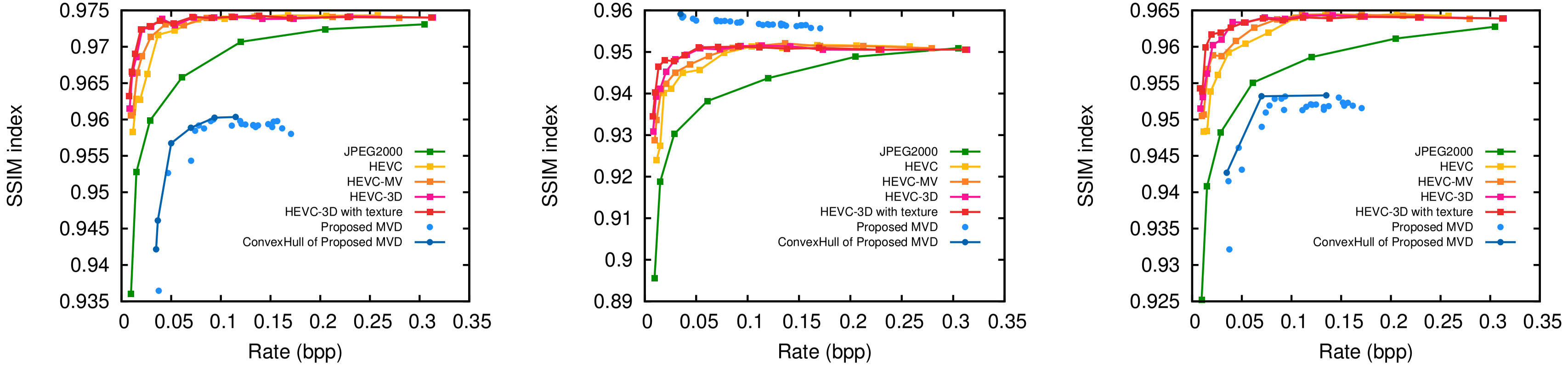


(d) Plastic @0.25 (e) Plastic @0.5 (f) Plastic @0.75

Figure 4.21: The results of experiment, comparing the novel view rendering SSIM scores obtained by the proposed and the conventional MVD compression schemes.

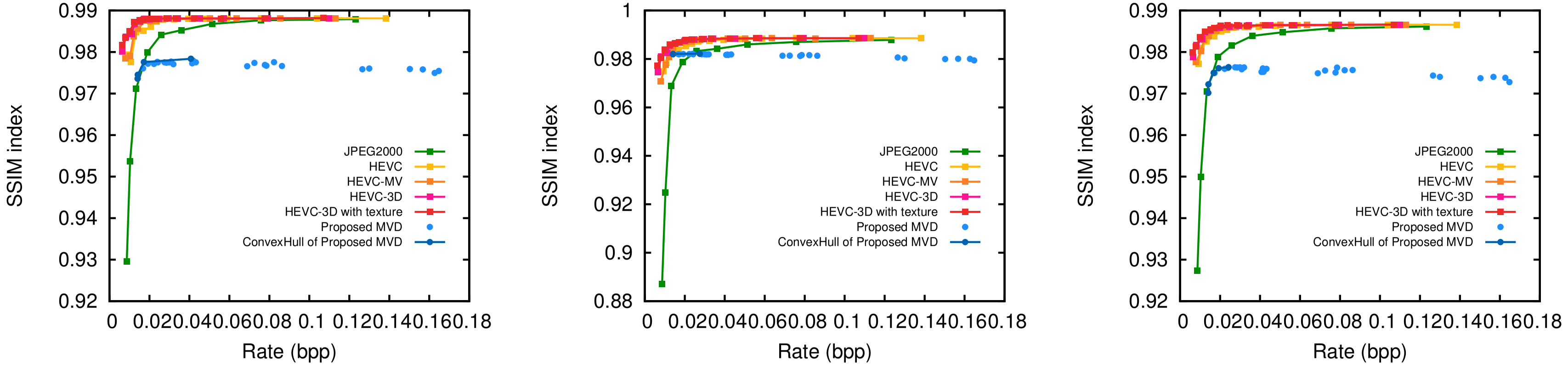


(a) Reindeer @0.25 (b) Reindeer @0.5 (c) Reindeer @0.75



(d) Teddy @0.25 (e) Teddy @0.5 (f) Teddy @0.75

Figure 4.22: The results of experiment, comparing the novel view rendering SSIM scores obtained by the proposed and the conventional MVD compression schemes.



(a) Wood @0.25

(b) Wood @0.5

(c) Wood @0.75

Figure 4.23: The results of experiment, comparing the novel view rendering SSIM scores obtained by the proposed and the conventional MVD compression schemes.

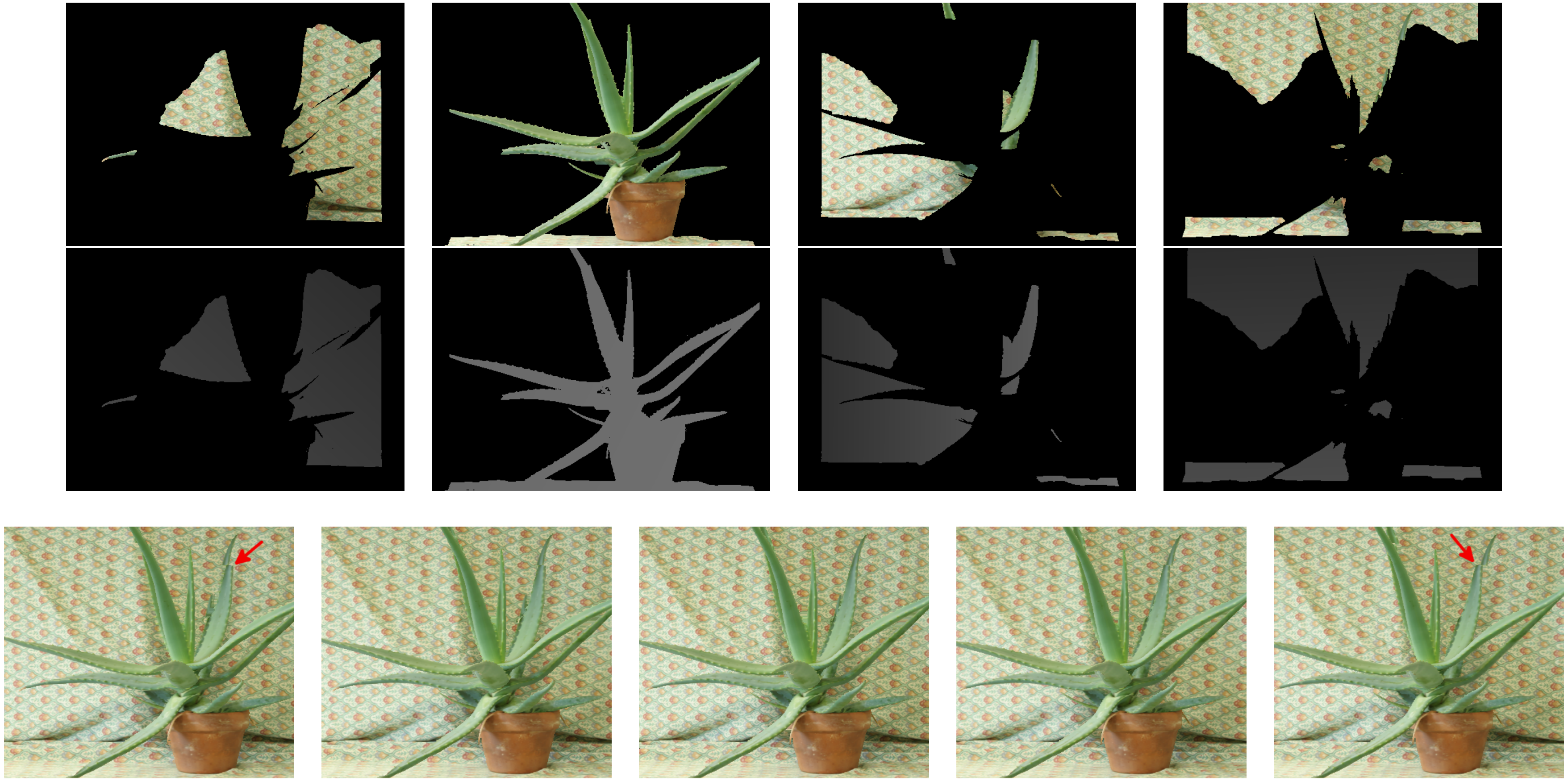


Figure 4.24: Top to bottom; texture and planar approximation of the 4-layered representation of *Aloe* dataset and its novel view rendering results. The distances between the novel views are the quarters of the baseline and starts from the left view of the raw MVD data. The red arrows highlight the artifacts due to layer boundary errors.

for all experiments. The pleasant rendering results obtained for the proposed approach also differs from the others by doubling the bit-rate in order to encode ten planar layers of the representation.

The bit-rate costs of the novel view renderings given in Figure 4.27 is approximately the same for *Books* dataset. Visually the results are also comparable but the PSNR scores of the view at the left quarter of the baseline presents a 2 dB difference between HEVC and proposed planar layered representation (see Table 4.4). The difference between SSIM scores are also congruent with the PSNR scores. The rendering quality difference, visually hard to notice but presented with the objective quality metrics are due to geometric approximation of the scene by the planar models.

Such geometric differences cause the object boundaries to shift away from its ground truth position for any novel view different than the reference view in the middle of the baseline. The magnitude of the shift increases as the position of the novel view gets further from the reference view. The visual example given in Figure 4.24 for the object boundary violation of layer assignments also illustrates the gradually increasing boundary shift towards the side views.

Such mismatches at the object boundaries can cause severe degradations in PSNR and SSIM scores. In order to demonstrate the extent of such geometric distortion over the objective metrics, the PSNR and SSIM scores between two novel view renderings located at the $1/100$ and $2/100$ of the baseline is calculated. The ground truth texture and depth information is utilized during DIBR. 1% baseline distance between the two novel views typically limits the maximum disparity to one pixel for *Middlebury* datasets. The objective results given in Table 4.1 show a wide variance related to texture and geometry of the content.

The visual comparison of the novel view rendering results of the MVD compression approaches shows that the PSNR and SSIM metrics for the visual quality are far from predicting the visual performance of the proposed planar layered representation. The DIBR friendly geometry information of the proposed MVD representation provides much concrete and coherent rendering results with the same PSNR or SSIM scores at the bit-rates, whereas JPEG 2000 and HEVC can

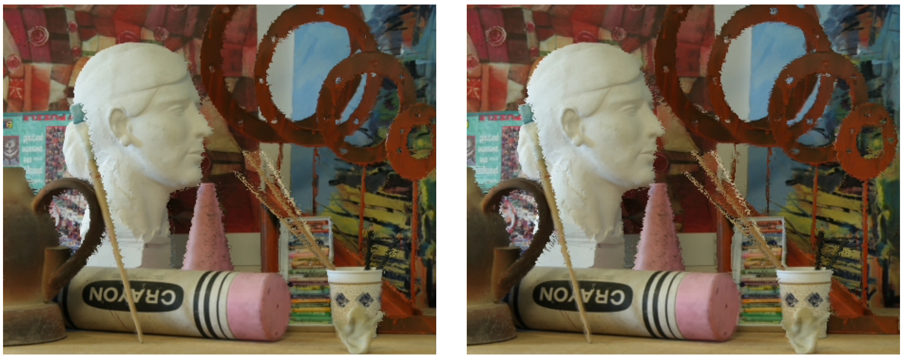
Table 4.1: Objective evaluation of novel view renderings by PSNR and SSIM measures for geometric distortions resulting in horizontal shifts of one pixel at most.

Name	PSNR (dB)	SSIM index
Aloe	39.84	0.985
Art	33.09	0.958
Baby	43.26	0.987
Books	32.24	0.949
Cloth	34.66	0.951
Cones	33.26	0.947
Dolls	34.32	0.961
Lampshade	44.85	0.993
Laundry	32.14	0.949
Moebius	35.19	0.965
Monopoly	33.53	0.962
Plastic	40.54	0.993
Reindeer	32.66	0.937
Teddy	35.57	0.974
Wood	40.19	0.971

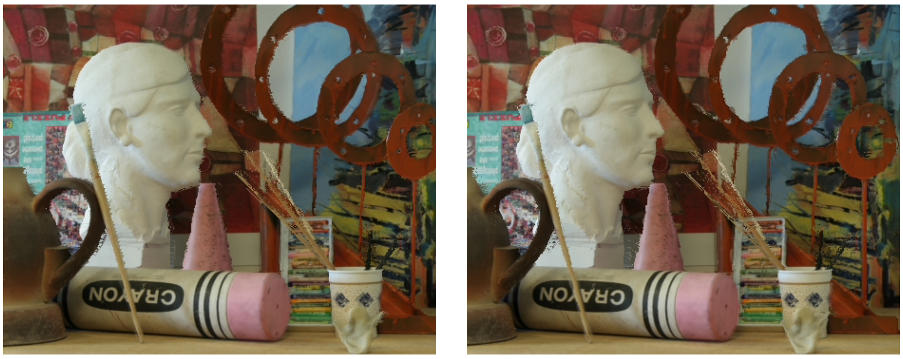
not preserve the sharp depth boundaries. PSNR and SSIM measurements given in Table 4.1 also provide evidence for possible severe drops of these objective metrics due to geometric distortions introduced by the planar approximations. Although 2D visual inspection of the rendering results of the planar approximations shows unnoticeable differences, their effects on depth perception should be evaluated. The boundary shifts directly change the disparity between the stereo pairs and alter the perceived depth. Since the ultimate goal of a 3D application is a satisfactory 3D experience in general, the depth perception of the compressed MVD data will be discussed in the next subsection.

4.3.2 Depth Perception Comparison

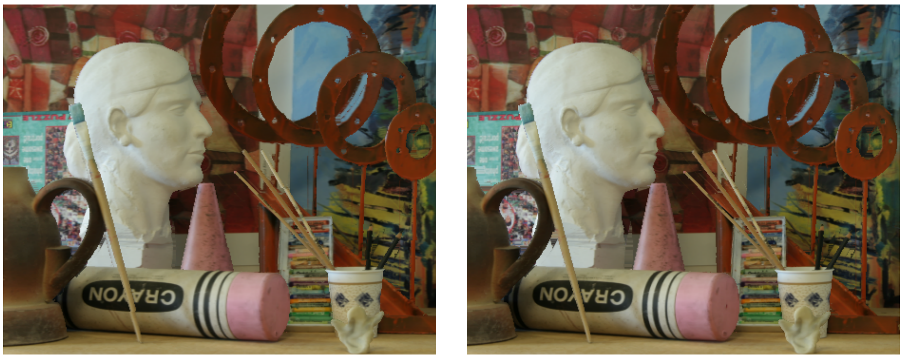
In the 3D research community, quality assessment is an important problem to guide the research activities into efficient solutions for the 3D applications of interest [116]. In general visual communications, the depth perception is the main added value of the 3D applications. HVS has many depth perception mechanisms but the stereopsis is the mainly exploited one in 3D displays [117].



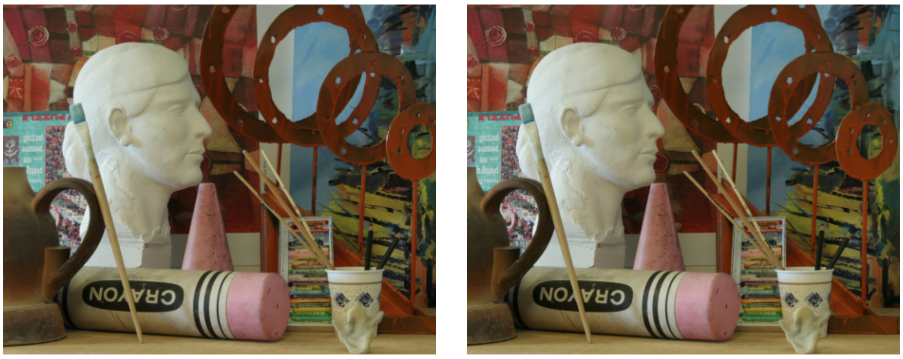
(a) JPEG 2000 rendering @0.25 (b) JPEG 2000 rendering @0.5



(c) HEVC rendering @0.25 (d) HEVC rendering @0.5

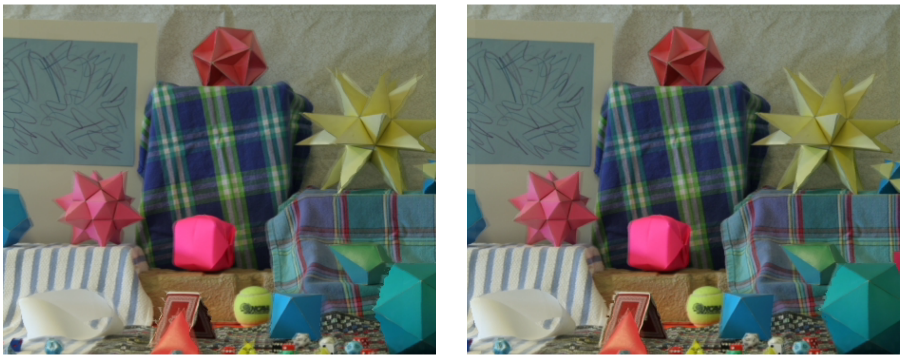


(e) Planar MVD rendering @0.25 (f) Planar MVD rendering @0.5

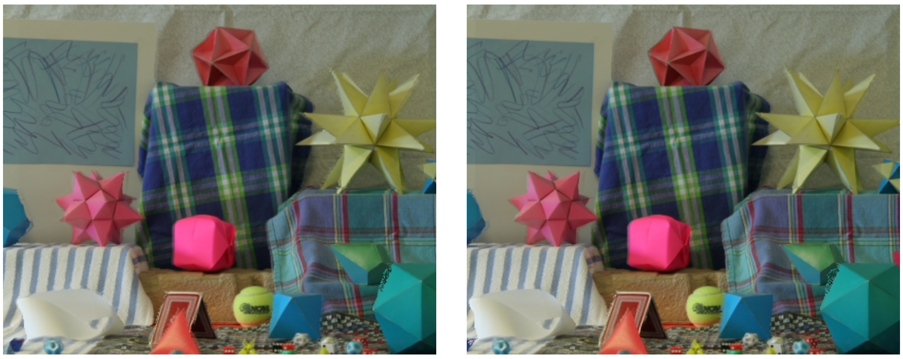


(g) Camera view @0.25 (h) Camera view @0.5

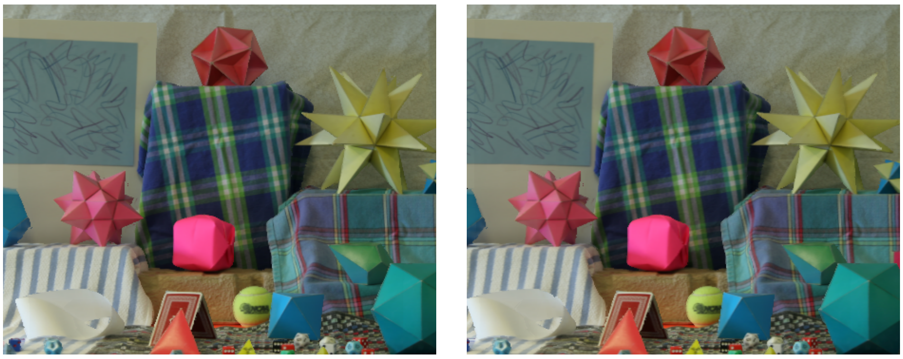
Figure 4.25: Visual comparison of novel view rendering results obtained for the *Art* dataset by different MVD compression schemes and the corresponding ground truth camera views.



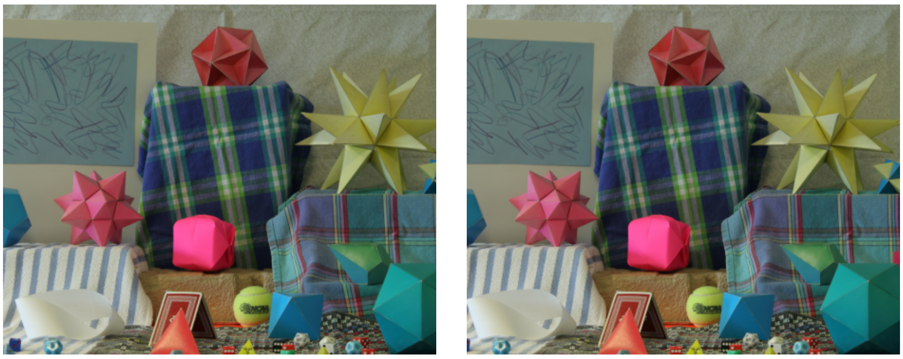
(a) JPEG 2000 rendering @0.5 (b) JPEG 2000 rendering @0.75



(c) HEVC rendering @0.5 (d) HEVC rendering @0.75

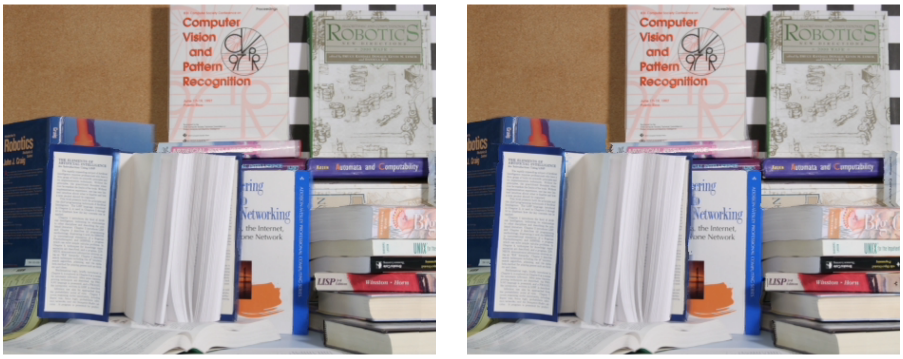


(e) Planar MVD rendering @0.5 (f) Planar MVD rendering @0.75

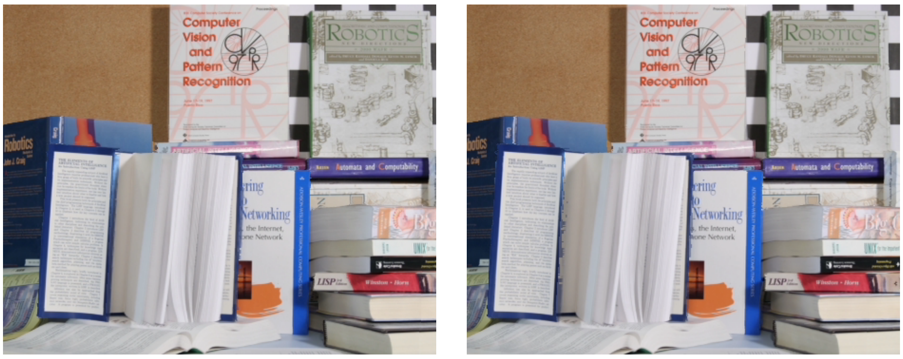


(g) Camera view @0.5 (h) Camera view @0.75

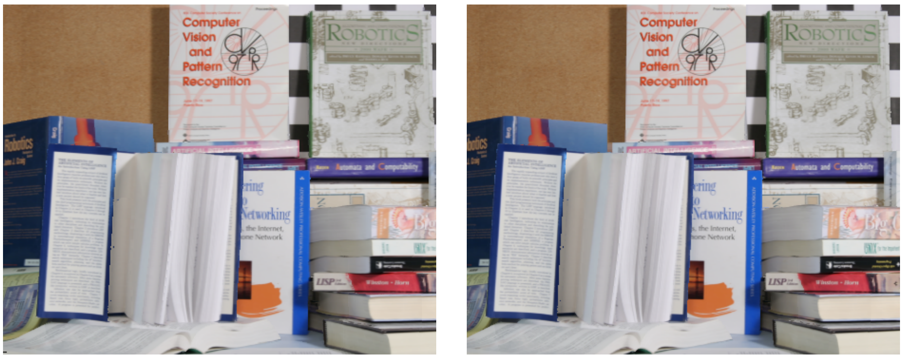
Figure 4.26: Visual comparison of novel view rendering results obtained for the *Moebius* dataset by different MVD compression schemes and the corresponding ground truth camera views.



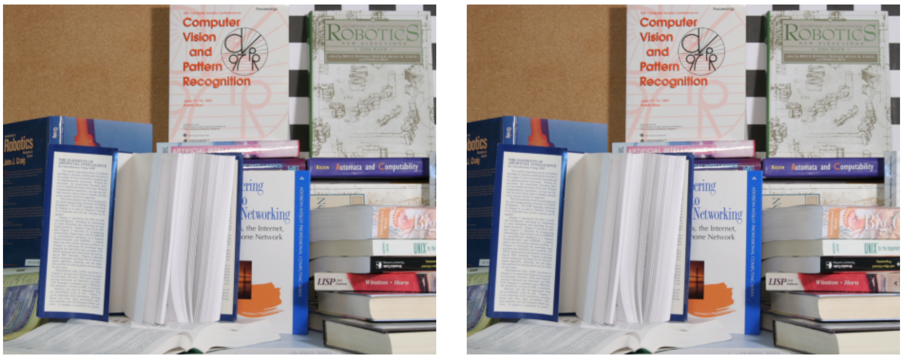
(a) JPEG 2000 rendering @0.25 (b) JPEG 2000 rendering @0.5



(c) HEVC rendering @0.25 (d) HEVC rendering @0.5



(e) Planar MVD rendering @0.25 (f) Planar MVD rendering @0.50



(g) Camera view @0.5 (h) Camera view @0.75

Figure 4.27: Visual comparison of novel view rendering results obtained for the *Books* dataset by different MVD compression schemes and the corresponding ground truth camera views.

Table 4.2: PSNR and SSIM values of the novel view rendering results obtained for *Art* dataset are tabulated.

		PSNR (dB)					SSIM index				
Compression	bpp	@0	@0.25	@0.5	@0.75	@1	@0	@0.25	@0.5	@0.75	@1
JPEG 2000	0.22466	∞	30.36	29.25	30.17	∞	1.0	0.930	0.911	0.925	1.0
HEVC	0.06469	∞	30.61	29.23	29.75	∞	1.0	0.933	0.911	0.924	1.0
Planar	0.06629	28.08	30.69	33.04	30.58	28.90	0.894	0.932	0.961	0.931	0.908

Table 4.3: PSNR and SSIM values of the novel view rendering results obtained for *Moebius* dataset are tabulated.

		PSNR (dB)					SSIM index				
Compression	bpp	@0	@0.25	@0.5	@0.75	@1	@0	@0.25	@0.5	@0.75	@1
JPEG 2000	0.027102	∞	35.24	33.54	35.09	∞	1.0	0.963	0.946	0.964	1.0
HEVC	0.022597	∞	35.90	33.80	35.39	∞	1.0	0.969	0.955	0.968	1.0
Planar	0.056946	32.38	34.82	38.16	34.99	32.80	0.942	0.961	0.977	0.963	0.943

Table 4.4: PSNR and SSIM values of the novel view rendering results obtained for *Books* dataset are tabulated.

		PSNR (dB)					SSIM index				
Compression	bpp	@0	@0.25	@0.5	@0.75	@1	@0	@0.25	@0.5	@0.75	@1
JPEG 2000	0.026721	∞	34.09	32.49	34.23	∞	1.0	0.980	0.961	0.972	1.0
HEVC	0.027412	∞	34.99	33.96	35.14	∞	1.0	0.976	0.971	0.978	1.0
Planar	0.028628	31.30	33.02	34.81	33.50	30.57	0.941	0.957	0.972	0.959	0.933

Measuring the depth perception quality and visual comfort based on stereopsis and other depth clues is among the challenging goals of 3D quality assessment [116]. For MVD-based applications, the 3D quality assessment is also complicated by the DIBR based artifacts [118].

The common approach for evaluating the depth perception is to conduct subjective tests and to obtain mean opinion scores (MOS). By a different perspective, a monoscopic subjective test protocol is also proposed recently in order to analyze the novel view rendering quality under depth compression [119]. However, subjective tests are time consuming and hard to compare objectively. Hence, there is an active research on obtaining an objective metric for assessing the 3D quality of a content. In the literature, there are full reference [120], restricted reference [121],[122], and no reference [123] quality metric proposals for DIBR based applications. Most of these proposals pay attention to measure the reconstruction quality of depth boundaries in their assessment. The correlations between the MOS and the conventional 2D objective metrics are also studied in [124] and it is reported that distinct items can have the same objective scores but very different subjective scores and vice versa.

The various approaches in the literature show the difficulty of the objective evaluation of the depth perception for conventional and DIBR based stereo view. The limited assessment and even contradicting cases are reported for the conventional 2D objective metrics. The results given in the previous subsections for objective scores and 2D visualization are in accordance with these facts of the unsolved 3D assessment problem. In order to analyze the results thoroughly, the depth perception is simulated by the anaglyph method [125] in the Figures 4.28 to 4.30.

Anaglyph stereo is a primitive visualization technique which is prone to ghosting artifacts severely. By the motivation of comparing the depth perception artifacts, the chroma channels of the rendered views are discarded in the anaglyph images to limit the ghosting artifacts. The baseline distance between the rendered stereo pairs for anaglyph images is set to 1/10 of the baseline distance between the stereo pair of the MVD. For example, the anaglyph stereo picture located at the midpoint; i.e. 0.5 of baseline, is composed of rendered views at 0.45 and 0.55 of the baseline. Since the *Middlebury* dataset does not provide sufficiently dense views to create the anaglyph stereo pairs from captured views, the ground truth cases are obtained by DIBR techniques utilizing the original depth and texture information.

The perceived depth for the example given for *Art* dataset is almost the same for all cases. However, the DIBR artifacts at the depth boundaries for the JPEG 2000 and HEVC are disturbing the visual comfort. The proposed planar case is almost as good as the ground truth novel view renderings, except for the object breakdowns at the rendered side view. Similar comments can be stated for the example given for *Moebius* dataset. The example given for *Books* have limited boundary artifacts in the results of JPEG 2000 and HEVC. The visual quality of all compression results is comparable to ground truth rendering result. However, the depth differences between the books stacked on the right cannot be perceived for the proposed planar representation case. This nuance is based on the geometric distortion induced by the planar approximation of that region with a single planar model.

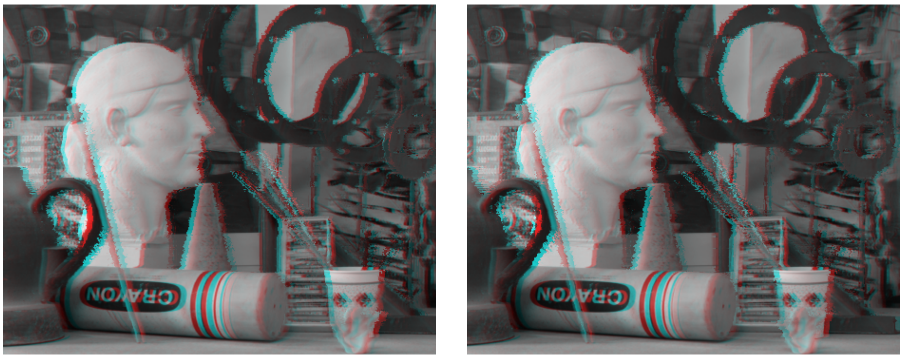
The proposed layered planar MVD representation has a quite different geometric distortion characteristic in comparison to conventional multi reference based MVD representations. In multi reference based MVD cases, the depth distortion on side views of the view to be rendered might result in horizontal texture shifts. In the extreme case, these texture shifts might cause double view or blur on rendered the novel views. Hence, the geometric distortion in conventional two reference based representations is prone to degrade the texture quality of the novel view renderings. However the single reference based representations, such as LDI and the proposed representation, remove the texture incoherencies due to geometric distortions by handling the texture information with a single reference.

An extreme case of geometric distortion is given for *Aloe* dataset in the left column of the Figure 4.31 in anaglyph technique. The planar layered representation of the example was given in Figure 4.24. In order to compare the effects of geometric distortion, the ground truth renderings are also presented in the right column of the figure. The plant's branch towards the camera is perceived very differently in the proposed planar case and the ground truth case. The planar approximation decreased the depth perception of objects coming out of the 2D displaying plane. However, the texture quality and depth perception consistency across the views are quite stable and coherent for the proposed layered planar representation under extreme geometric distortion.

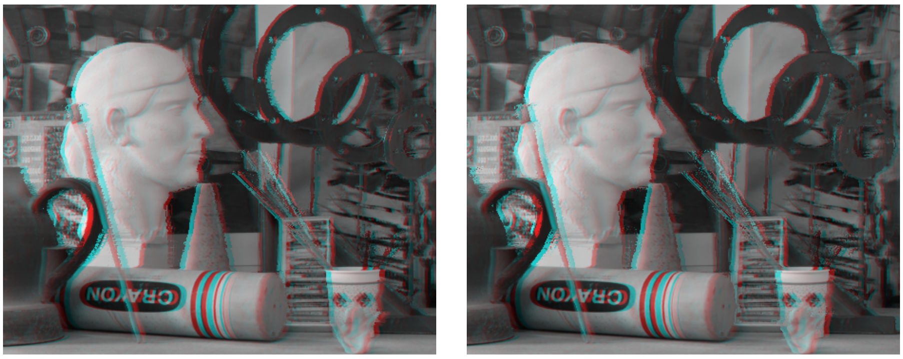
The potential of proposed layered planar MVD representation to tolerate these extreme geometric distortions can be explained by the single reference based representation of the texture information of the layers. While extracting the texture information, the ground truth depth images are utilized in the 3D warping operations. As long as the spatial support of the corresponding planar model does not change, the texture to be obtained for that layer should be the same. Hence, the texture information of the layers is defined according to the labeling images of the planar assignments; however, it is independent of the assigned planar model geometries.

This phenomena can be considered as decoupling of the textural and geometric distortions of the proposed planar MVD representation. To illustrate the textural concreteness of the proposed MVD representation in novel view rendering, exactly the same 4 planar assignment based novel view renderings with the proposed and the raw MVD formats are presented in Figure 4.32 for *Teddy* dataset. While the extreme geometric distortions result in doubled boundaries of the objects, especially in the mid view for the two reference based raw MVD representation, such textural distortions do not exist in the renderings of proposed MVD representation.

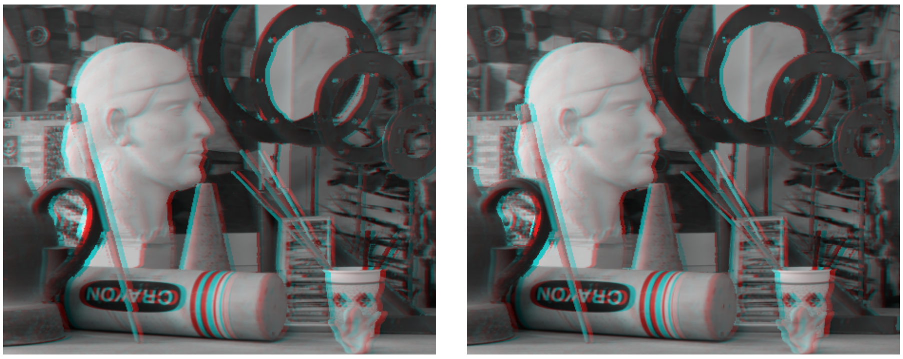
The coherent texture and geometric properties of the proposed layered planar representation are important novelties of the proposal and it might lead to new optimality formulations in depth compression. In case of objective assessment of depth perception quality is available, the geometric distortions might be optimized for a pleasant depth perception without any degradations on the texture coherency of the novel view renderings. Hence, assuming the texture quality is unchanged, the proposed MVD representation should provide encoded MVD data at different depth perception qualities. The concept of *just noticeable depth difference* studied in [126] should play a fundamental role in assessing the reconstruction quality of the depth perception in such a framework.



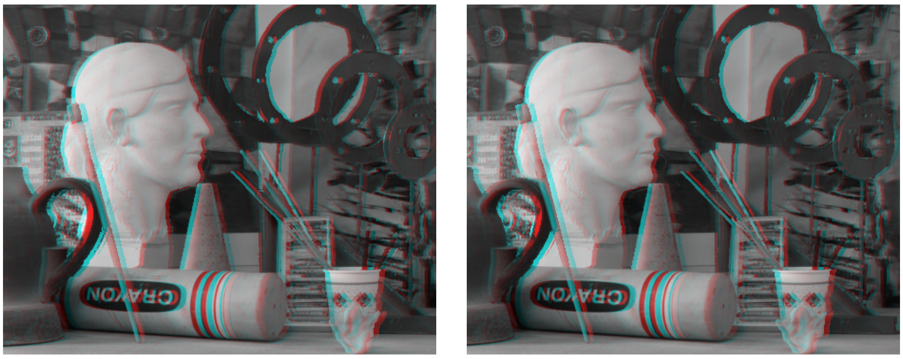
(a) JPEG 2000 rendering @0.25 (b) JPEG 2000 rendering @0.5



(c) HEVC rendering @0.25 (d) HEVC rendering @0.5

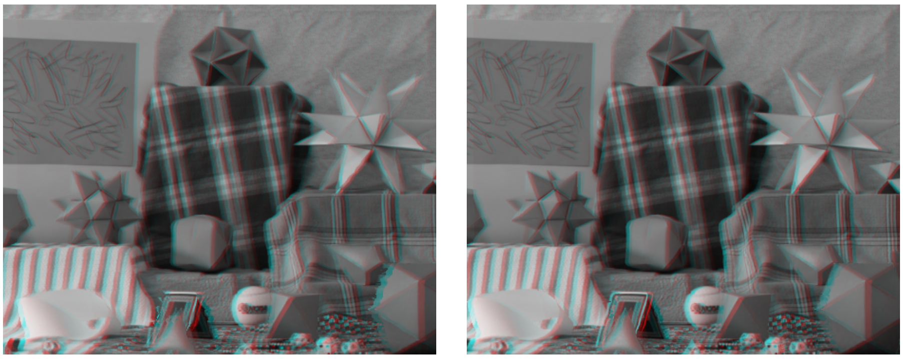


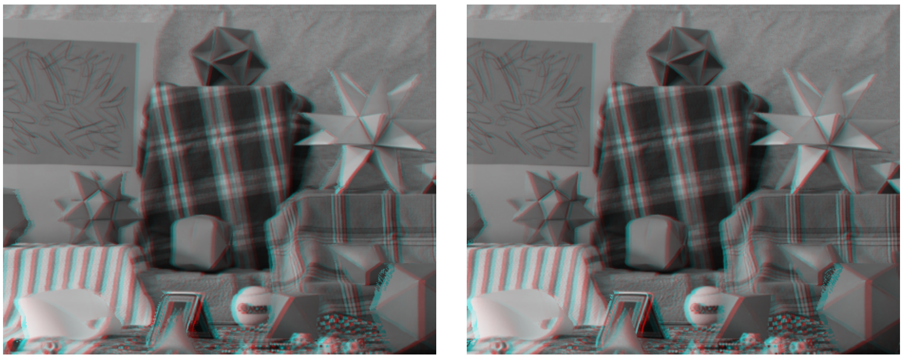
(e) Planar MVD rendering @0.25 (f) Planar MVD rendering @0.5

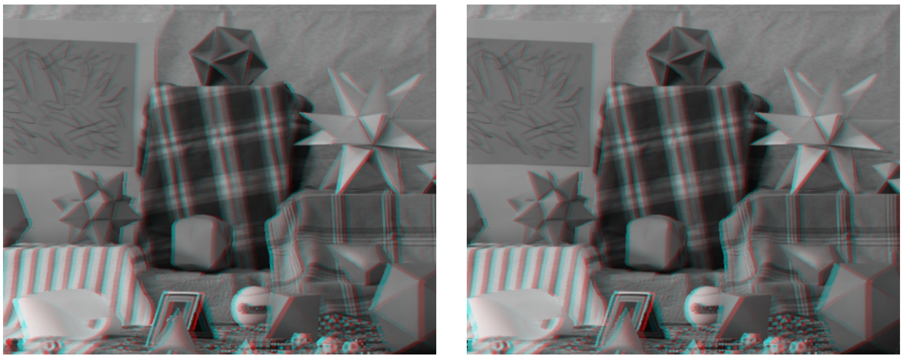


(g) Original MVD rendering @0.25 (h) Original MVD rendering @0.5

Figure 4.28: The comparison of the depth perception of the *Art* dataset for different MVD compression schemes. The views from left to right are located at the left quarter and in the middle of the baseline. The anaglyph technique is applied by red-cyan encoding. Best viewed in digital copy with a proper zoom in.


(a) JPEG 2000 rendering @0.5 (b) JPEG 2000 rendering @0.75


(c) HEVC rendering @0.5 (d) HEVC rendering @0.75


(e) Planar MVD rendering @0.5 (f) Planar MVD rendering @0.75

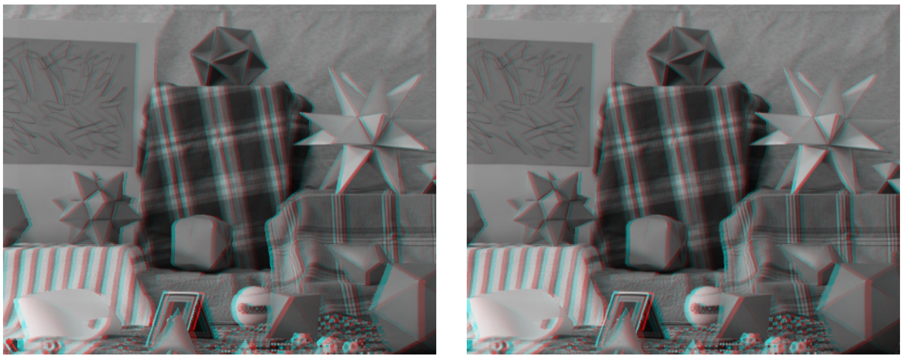
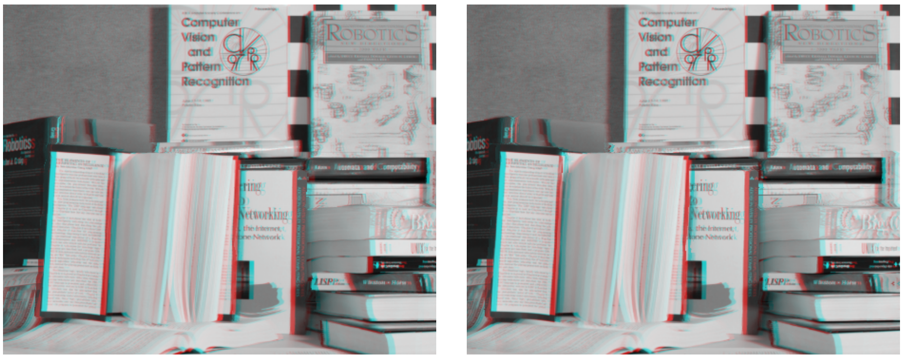
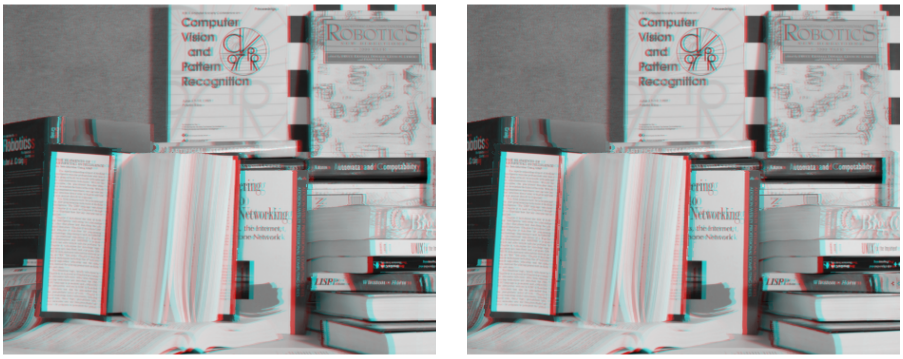

(g) Original MVD rendering @0.5 (h) Original MVD rendering @0.75

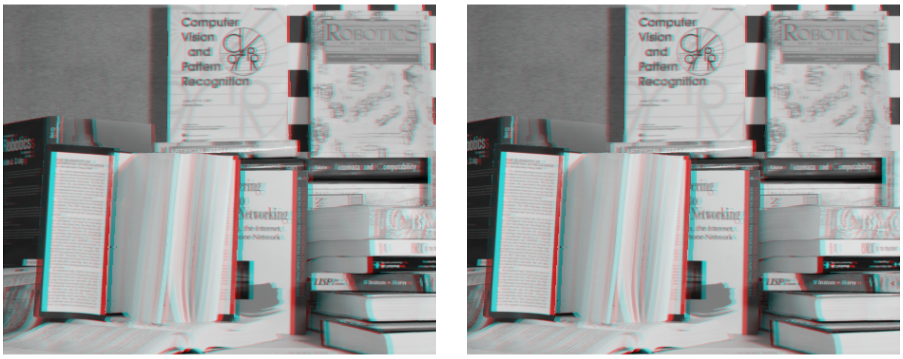
Figure 4.29: The comparison of the depth perception of the *Moebius* dataset for different MVD compression schemes. The views from left to right are located in the middle and at the right quarter of the baseline. The anaglyph technique is applied by red-cyan encoding. Best viewed in digital copy with a proper zoom in.



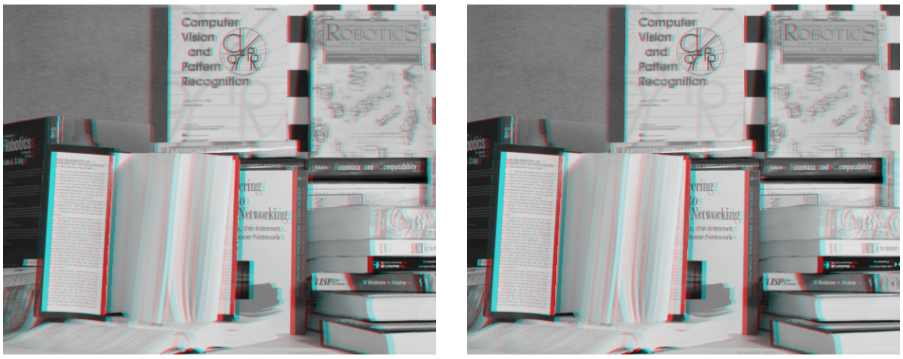
(a) JPEG 2000 rendering @0.25 (b) JPEG 2000 rendering @0.5



(c) HEVC rendering @0.25 (d) HEVC rendering @0.5

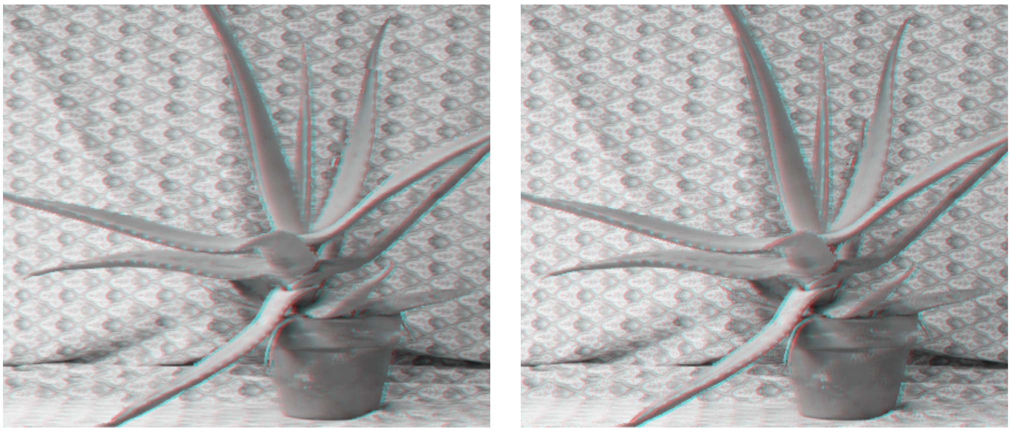


(e) Planar MVD rendering @0.25 (f) Planar MVD rendering @0.5

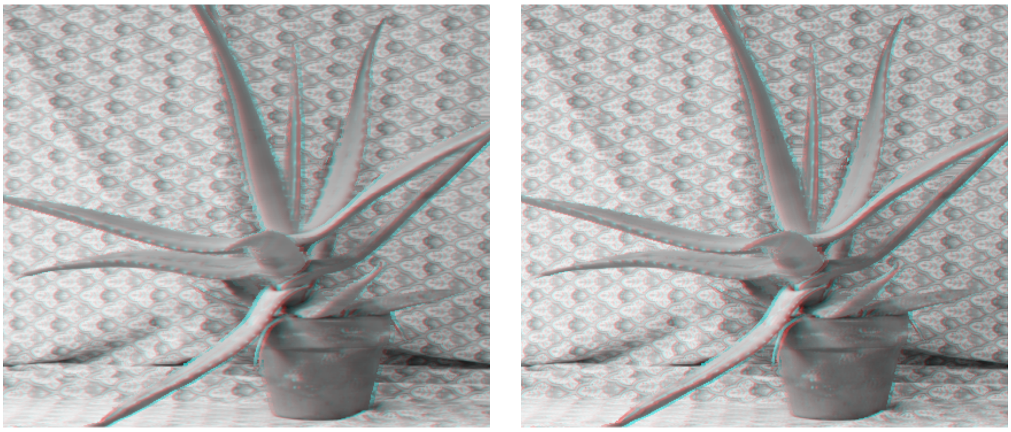


(g) Original MVD rendering @0.25 (h) Original MVD rendering @0.5

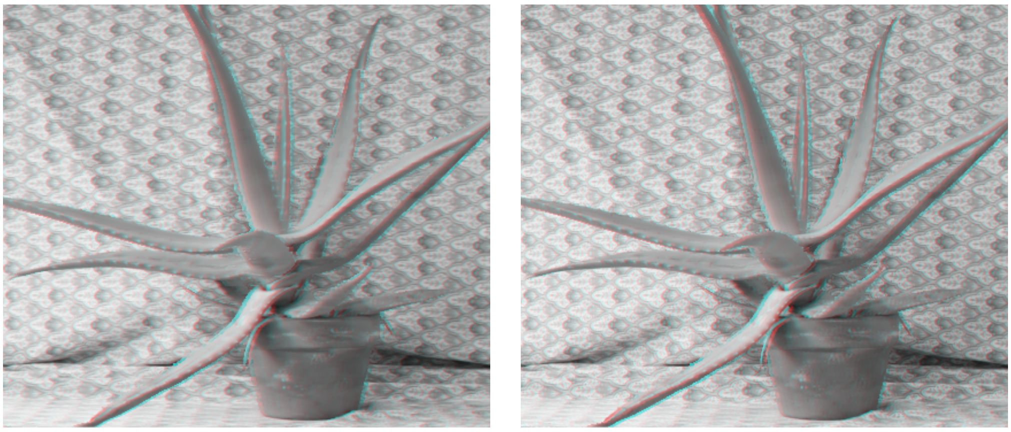
Figure 4.30: The comparison of the depth perception of the *Books* dataset for different MVD compression schemes. The views from left to right are located at the left quarter and in the middle of the baseline. The anaglyph technique is applied by red-cyan encoding. Best viewed in digital copy with a proper zoom in.



(a) Renderings @0.25



(b) Renderings @0.5



(c) Renderings @0.75

Figure 4.31: The depth perception comparison of the layered planar MVD under extreme geometric distortion (4 planar model) with the uncompressed raw MVD format for the *Aloe* dataset. The left column shows the rendering results for the proposed layered planar MVD representation. The anaglyph technique is applied by red-cyan encoding. Best viewed in digital copy with a proper zoom in.

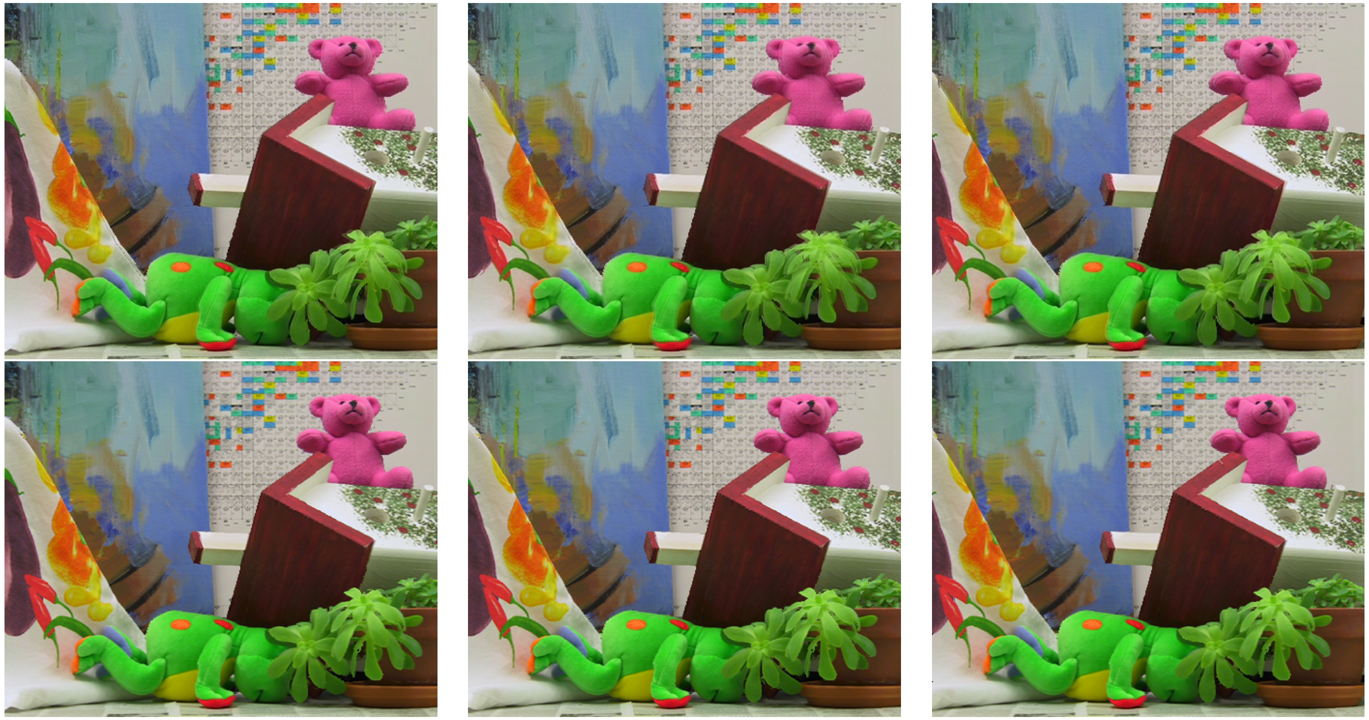


Figure 4.32: Top to bottom, novel view rendering results of raw and proposed MVD representations for the depth images approximated by 4 planar models. Left to right, novel views positioned at the 0.25, 0.5 and 0.75 of the baseline distance.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

5.1 Summary of the Thesis

This thesis proposes, a planar segmentation based approach to handle the two important topics in depth compression for 3DV application: exploitation of the smooth characteristics of the depth modality and the preservation of the clear depth discontinuities for artifact free novel view rendering. Different than the other linear and planar approximations in the literature, the proposed approach defines the planarity in the 3D space of the scene. The multiple depth images of the scene provided by the MVD data are considered as a 3D point cloud to be approximated by 3D planes. In this perspective, the approximation/representation problem of the depth images turned into a co-segmentation problem; i.e. depth estimaton and object segmentation are obtained simultaneously.

The planar co-segmentation of stereo depth images is formulated by an energy minimization framework. A MRF model is utilized in defining the energy costs of the co-segmentation problem. The combination of data and regularization terms, named smoothness and label costs, provided a rate distortion similar tradeoff to the solutions. The optimization problem of the energy terms is approximately solved by a modified PEARL algorithm. The modifications of the PEARL algorithm utilized the smoothness of the depth images to revert the candidate model set to be in an increasing regime. The planar model sampling is achieved by fitting planar models for the connected components of the assignments to obtain object-like representations.

The weights for the energy cost terms are considered in order to obtain planar representations at different reconstruction quality for a practical depth compression framework. A heuristic algorithm is designed to select the favorite solution among the Pareto optimal set of the multiple objectives. The favorite solutions are defined as the ones with the minimum geometric distortion, while satisfying the constraint on the maximum number of planar models to be utilized. By this formulation, the planar reconstructions of the stereo depth images can be obtained for various number of planar models in a systematic way.

The depth compression experiments are performed by encoding the planar models and their spatial support, i.e. by labeling images. The objectives of the co-segmentation energy is designed to be related with the rate distortion optimality of the representations. The proposed segmentation based depth compression approach defines the depth segments with a rate distortion sense different than the other segmentation based approaches in the literature.

The depth compression experiments are conducted in comparison to state-of-the-art DWT and DCT based compression tools, JPEG 2000 and variants of HEVC, respectively. The depth reconstruction and novel view rendering results are evaluated by PSNR and SSIM metrics. Since planar approximations might not be convenient for every scene geometry, the planar representation is also considered as a prediction tool in a residual coding fashion.

Lastly, a novel MVD representation is derived from the planar representations by merging them into a single reference based representation, such as LDI. The layers of the proposed MVD representation are defined as the planar surfaces approximating the stereo depth images. The texture information of the layers is obtained by merging the texture of the pixels assigned to that planar model. Since the texture extraction of the layers can be performed with the ground truth depth images, the proposed layered planar MVD representation decouples the rendering distortions related to geometry and texture in a wide extent.

The different novel view rendering characteristics of the proposed single referenced system are analyzed in comparison to rendering results of the state-of-the-art compression methods. The objective measures are discussed by visual comparison of the results. The effects of the planar approximations on the depth perception are also illustrated for the novel MVD representation.

5.2 Conclusions

The energy based co-segmentation framework is successful in obtaining coherent representation units across the stereo depth images. Visually these segments coincide with the scene object boundaries in general. However, the method is still prone to segmentation errors which can result in broken object renderings at the end.

The heuristic algorithm developed to obtain various planar representations in a systematic way shows the responsiveness of the energy terms on the obtained solution. However, the approximations of the optimization algorithms, GC and PEARL, avoid the obtained solutions as the Pareto optimal instances to create a convex trajectory for the rate distortion plots. The most scattered plots occur at higher rates for the scenes containing limited number of objects or mostly planar surfaces. This observation shows possible breakdowns of the heuristic algorithm in seeking a redundant number of models.

The sampled instances of *Middlebury* dataset shows various textural and geometric variations; hence, the results of the experiments are convenient to make conclusions and generalizations about the compression performance of the proposed approach. As long as the depth images contain descriptive boundaries for the objects in the scene, the planar representations become efficient in depth reconstruction, and especially, in novel view rendering. The analytic representation of the planar regions reconstructs the depth of non-boundary regions almost for free. Since the main compression cost of the planar approach is the boundary encoding of the labeling images, the efficient cases can be explained by this reasoning. For the scenes mostly composed of planar surfaces, the planar compression at low rates gives comparable novel view rendering results with the state of the art MVD compression standard, HEVC-3D.

Although, the planar depth reconstruction might be inferior than the reconstructions of HEVC variants, the sharp discontinuity preserving properties of the proposed planar representation step in to provide better novel view rendering results. This fact shows itself by comparing the relative performance of planar representation in depth reconstruction and novel view rendering. In general, while the planar depth reconstruction results are comparable with the HEVC results, the novel view rendering results become comparable to HEVC-MV or even HEVC-3D. Hence, the proposed planar compression of stereo depth images can be regarded as an efficient approach for rendering applications of 3DV.

The planar representation as a prediction tool provides superior results for the DWT based comparative experiments. However, the same cannot be stated for DCT based experiments. The datasets, which are efficient cases of pure planar representation, show better depth reconstructions and novel view rendering results. Based on this variable performance, the proposed planar approach can be included in the prediction modes of contemporary depth encoders like HEVC-3D.

The proposed layered planar MVD representation's most important aspect is its capabilities in decoupling the distortions on texture and geometry to a wide extent for the novel view renderings. The single reference based definition of the proposed MVD accounts for this property. For representations with multiple references, extreme depth distortions are not tolerable, since they can break down the stereo correspondences between the multiple references. Such stereo correspondence errors result in blurred or doubled views for novel view renderings. However, for the proposed single reference based representation, the texture of layers are extracted according to the ground truth depth images; i.e. correct stereo correspondences. Hence, the texture information of a layer depends only on the labeling images obtained by the planar co-segmentation, but not the utilized planar models. This property constitutes the decoupled texture and geometry of the proposed MVD representation.

The objective comparison in novel view rendering of the proposed MVD representation with raw MVD format encoded by the state-of-the-art encoders is also affected by the decoupling phenomena. The proposed approach has quite high rendering quality even at low rates for the mid-view which is also the reference view of the representation. This performance is the consequence of the ground truth depth utilization in texture extraction. The novel view rendering performance of the proposed approach drops by shifting the view towards the side views. However, the visual evaluation of the rendering results with low scores indicates that the proposed representation should be preferred to renderings obtained from the HEVC encoded raw MVD data. The concrete textural properties of the proposed representation make the difference at this point.

The tolerance of the proposed representation to extreme geometric distortions brings the depth perception quality related questions. For the planar representations using a few number of models, the depth variations in the scene might be lost. Hence, the depth distortions above the just noticeable threshold can degrade the depth perception and 3D experience. Accordingly, while the proposed layered planar representation enjoys the concrete and high quality texture reproduction of the scene, it might suffer with reduced depth perception due to possible rough geometric approximations. The relation between texture quality and depth reproduction in 3D experience is an unsolved problem of 3DV assessment. However, the proposed representation widens the MVD compression alternatives by its tolerance to extreme geometric distortions.

5.3 Future Work

Towards a practical solution for 3DV applications, the first major missing topic of the proposed approach is the texture compression of the MVD data and the second one is the temporal changes in MVD data. Although the scope of the thesis is limited to the depth modality of the MVD data, a complete solution always has the advantage of utilizing all possible redundancy forms like interview, inter modality and temporal. The methods in the literature to exploit these redundancies should be considered and adapted to the proposed approaches and representations.

In order to have a complete idea of the compression efficiency of the proposed layered planar MVD representation, the texture compression of the layers might be the first problem to approach. The lossless texture compression experiment results are promising to obtain a more efficient lossy texture coding for the proposed MVD representation. It is also reasonable to expect the single reference representation to handle the inter-view textural redundancies better by explicitly merging the texture information of multiple visible regions. At this point, the available shape information of the layers should be kept in mind to handle the sparsity of the texture images of the layers.

Exploitation of temporal redundancy in this problem might necessitate various modifications. The temporal smoothness of the depth surfaces can be handled in the energy based formulation of the planar fitting problem. The motion vectors estimated for the temporal predictions of the texture can be considered in enriching the neighborhood definition of the MRF to a spatio-temporal one. More practical approaches might leave the planar model fitting formulation untouched and concentrate on efficient temporal initializations of the MRF solutions. The temporal evolutions of the planar layers may be considered to be encoded or constrained by a global camera motion model. These are just a few possibilities in considering the temporal dimension.

The thesis shows the efficient usage of the planar representation in depth compression for novel view rendering applications by the generic compression tools. The piecewise constant characteristics of the labeling images and the bit-plane coded layer shapes of the proposed MVD can be studied to obtain better compression ratios by tailoring the shape encoder. Context-based entropy coded chain or crack codes based approaches [71], [44] can be considered in improving the shape compression routines of the proposals.

The broken object-like rendering artifacts of proposed MVD representation can be studied to obtain more coherent co-segmentation results with the scene objects. The edge information of the depth and texture images can be introduced in the energy formulation to favor the labeling discontinuities to occur at depth or texture edges. The novel view rendering quality metrics can be considered to advance the model fitting cost terms in the energy formulation. At the extreme case, the planar model based stereo object estimation costs introduced in [48] can be adapted to the rate-distortion like energy formulation of the proposed approaches.

The tolerance of proposed MVD representation to the extreme geometric distortions should be studied to analyze its effects on 3D experience. The limited capabilities of the conventional video metrics in the assessment of 3D experience and depth perception might require subjective test for these analyses. Based on this analysis, possible geometric manipulation applications, such as display adaptation, can be developed in the framework of layered planar MVD representation.

To summarize, the proposed planar representation based approaches are promising for rendering friendly compression applications. Especially, the proposed layered planar MVD representation's properties in decoupling the texture and geometry needs to be analyzed thoroughly. Although there are serious difficulties in the 3D assessment of the results, the proposed MVD representation deserves to be considered as a promising alternative by introducing the extreme geometric manipulations into the 3D compression community.

REFERENCES

[1] D. Minoli, *3DTV Content Capture, Encoding and Transmission: Building the Transport Infrastructure for Commercial Services.* Wiley, 2010.

[2] J.-X. Chai, X. Tong, S.-C. Chan, and H.-Y. Shum, "Plenoptic sampling," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '00, (New York, NY, USA), pp. 307–318, ACM Press/Addison-Wesley Publishing Co., 2000.

[3] H.-Y. Shum and S. B. Kang, "Review of image-based rendering techniques," in *Visual Communications and Image Processing 2000*, pp. 2–13, International Society for Optics and Photonics, 2000.

[4] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vision*, vol. 47, pp. 7–42, Apr. 2002.

[5] E. Stoykova, A. A. Alatan, P. Benzie, N. Grammalidis, S. Malassiotis, J. Ostermann, S. Piekh, V. Sainov, C. Theobalt, T. Thevar, and X. Zabulis, "3-D time-varying scene capture technologies - a survey," vol. 17, pp. 1568–1586, 2007.

[6] P. Benzie, J. Watson, P. Surman, I. Rakkolainen, K. Hopf, H. Urey, V. Sainov, and C. von Kopylow, "A survey of 3DTV displays: techniques and technologies," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 11, pp. 1647–1658, 2007.

[7] A. A. Alatan, Y. Yemez, U. Gudukbay, X. Zabulis, K. Müller, Ç. E. Erdem, C. Weigel, and A. Smolic, "Scene representation technologies for 3DTV—a survey," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 11, pp. 1587–1605, 2007.

[8] A. Smolic, K. Mueller, N. Stefanoski, J. Ostermann, A. Gotchev, G. B. Akar, G. Triantafyllidis, and A. Koz, "Coding algorithms for 3DTV—a survey," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 11, pp. 1606–1621, 2007.

[9] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross, "Nonlinear disparity mapping for stereoscopic 3D," *ACM Transactions on Graphics (TOG)*, vol. 29, no. 4, p. 75, 2010.

[10] Wikipedia, "H.262/MPEG-2 Part 2 — Wikipedia, the free encyclopedia," 2013. [Online; accessed 14-October-2013].

[11] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Mueller, T. Wiegand, *et al.*, "The effects of multiview depth video compression on multiview rendering," *Signal Processing: Image Communication*, vol. 24, no. 1, pp. 73–88, 2009.

[12] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Efficient prediction structures for multiview video coding," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 11, pp. 1461–1473, 2007.

[13] Wikipedia, "H.264/MPEG-4 AVC — Wikipedia, the free encyclopedia," 2013. [Online; accessed 14-October-2013].

[14] S. Shimizu, M. Kitahara, H. Kimata, K. Kamikura, and Y. Yashima, "View scalable multiview video coding using 3-D warping with depth map," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 17, pp. 1485–1495, nov 2007.

[15] E. Martinian, A. Behrens, J. Xin, and A. Vetro, "View synthesis for multiview video compression," in *Picture Coding Symposium*, vol. 37, pp. 38–39, 2006.

[16] B. Özkalaycı, O. S. Gedik, and A. A. Alatan, "3-D structure assisted reference view generation for H. 264 based multi-view video coding," in *Picture Coding Symposium*, vol. 37, pp. 1–4, 2007.

[17] B. Girod, C.-L. Chang, P. Ramanathan, and X. Zhu, "Light field compression using disparity-compensated lifting," in *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, vol. 1, pp. I–373, IEEE, 2003.

[18] 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.

[19] A. J. Woods, T. Docherty, and R. Koch, "Image distortions in stereoscopic video systems," in *IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology*, pp. 36–48, International Society for Optics and Photonics, 1993.

[20] C. Fehn, "A 3D-TV system based on video plus depth information," in *Signals, Systems and Computers, 2003. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2, pp. 1529–1533, IEEE, 2003.

[21] P. Merkle, Y. Wang, K. Müller, A. Smolic, and T. Wiegand, "Video plus depth compression for mobile 3D services," in *3DTV Conference: The True*

Vision-Capture, Transmission and Display of 3D Video, 2009, pp. 1–4, IEEE, 2009.

[22] C. Fehn, "A 3D-TV approach using depth-image-based rendering (dibr)," in *Proc. of VIIP*, vol. 3, 2003.

[23] L. Zhang and W. J. Tam, "Stereoscopic image generation based on depth images for 3D TV," *Broadcasting, IEEE Transactions on*, vol. 51, no. 2, pp. 191–199, 2005.

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

[25] "Call for proposals on 3D video coding technology." ISO/IEC JTC1/SC29/WG11 (MPEG) Doc. N12036, Geneva, Switzerland, March 2011.

[26] P. Merkle, K. Müller, and T. Wiegand, "3D video: acquisition, coding, and display," *Consumer Electronics, IEEE Transactions on*, vol. 56, no. 2, pp. 946–950, 2010.

[27] F. Pereira, "Video compression: Discussing the next steps," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pp. 1582–1583, IEEE, 2009.

[28] I. Daribo, C. Tillier, and B. Pesquet-Popescu, "Motion vector sharing and bitrate allocation for 3D video-plus-depth coding," *EURASIP Journal on Applied Signal Processing*, vol. 2009, p. 3, 2009.

[29] M. Domanski, O. Stankiewicz, K. Wegner, M. Kurc, J. Konieczny, J. Siast, J. Stankowski, R. Ratajczak, and T. Grajek, "High efficiency 3D video coding using new tools based on view synthesis," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3517–3527, 2013.

[30] M. M. Hannuksela, D. Rusanovskyy, W. Su, L. Chen, R. Li, P. Aflaki, D. Lan, M. Joachimiak, H. Li, and M. Gabbouj, "Multiview-video-plus-depth coding based on the advanced video coding standard," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3449–3458, 2013.

[31] K. Müller, H. Schwarz, D. Marpe, C. Bartnik, S. Bosse, H. Brust, T. Hinz, H. Lakshman, P. Merkle, F. Rhee, G. Tech, M. Winken, and T. Wiegand, "3D high-efficiency video coding for multi-view video and depth data," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3366–3378, 2013.

[32] I. T. Union, "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference." ITU-T Recommendation J.144, 2001.

[33] W.-S. Kim, A. Ortega, P. Lai, D. Tian, and C. Gomila, "Depth map coding with distortion estimation of rendered view," *Visual Communication and Information Processing*, vol. 7543–7552, 2010.

[34] G. Tech, H. Schwarz, K. Müller, and T. Wiegand, "3D video coding using the synthesized view distortion change," in *Picture Coding Symposium (PCS), 2012*, pp. 25–28, 2012.

[35] G. Cheung, V. Velisavljevic, and A. Ortega, "On dependent bit allocation for multiview image coding with depth-image-based rendering," *Image Processing, IEEE Transactions on*, vol. 20, no. 11, pp. 3179–3194, 2011.

[36] M. Maitre and M. N. Do, "Joint encoding of the depth image based representation using shape-adaptive wavelets," in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pp. 1768–1771, IEEE, 2008.

[37] A. Sánchez, G. Shen, and A. Ortega, "Edge-preserving depth-map coding using graph-based wavelets," in *Signals, Systems and Computers, 2009 Conference Record of the Forty-Third Asilomar Conference on*, pp. 578–582, IEEE, 2009.

[38] W.-S. Kim, S. K. Narang, and A. Ortega, "Graph based transforms for depth video coding," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 813–816, IEEE, 2012.

[39] R. Mathew, D. Taubman, and P. Zanuttigh, "Scalable coding of depth maps with R-D optimized embedding," *Image Processing, IEEE Transactions on*, vol. 22, no. 5, pp. 1982–1995, 2013.

[40] H. Schwarz, C. Bartnik, S. Bosse, H. Brust, T. Hinz, H. Lakshman, D. Marpe, P. Merkle, K. Müller, H. Rhee, *et al.*, "3D video coding using advanced prediction, depth modeling, and encoder control methods," in *Picture Coding Symposium (PCS), 2012*, pp. 1–4, IEEE, 2012.

[41] P. Lai, A. Ortega, C. C. Dorea, P. Yin, and C. Gomila, "Improving view rendering quality and coding efficiency by suppressing compression artifacts in depth-image coding," in *IS&T/SPIE Electronic Imaging*, pp. 72570O–72570O, International Society for Optics and Photonics, 2009.

[42] P. Zanuttigh and G. M. Cortelazzo, "Compression of depth information for 3D rendering," in *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, 2009*, pp. 1–4, IEEE, 2009.

[43] F. Jager, "Contour-based segmentation and coding for depth map compression," in *Visual Communications and Image Processing (VCIP), 2011 IEEE*, pp. 1–4, IEEE, 2011.

[44] I. Tabus, I. Schiopu, and J. Astola, "Context coding of depth map images under the piecewise-constant image model representation," *Image Processing, IEEE Transactions on*, vol. 22, no. 11, pp. 4195–4210, 2013.

[45] S. Hoffmann, M. Mainberger, J. Weickert, and M. Puhl, "Compression of depth maps with segment-based homogeneous diffusion," in *Scale Space and Variational Methods in Computer Vision*, pp. 319–330, Springer, 2013.

[46] J. Gautier, O. Le Meur, and C. Guillemot, "Efficient depth map compression based on lossless edge coding and diffusion," in *Picture Coding Symposium (PCS), 2012*, pp. 81–84, IEEE, 2012.

[47] T. Sikora, "The MPEG-4 video standard verification model," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 7, no. 1, pp. 19–31, 1997.

[48] M. Bleyer, C. Rother, P. Kohli, D. Scharstein, and S. Sinha, "Object stereo—joint stereo matching and object segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 3081–3088, IEEE, 2011.

[49] S. N. Sinha, D. Steedly, and R. Szeliski, "Piecewise planar stereo for image-based rendering.," in *ICCV*, pp. 1881–1888, Citeseer, 2009.

[50] D. Gallup, J.-M. Frahm, and M. Pollefeys, "Piecewise planar and non-planar stereo for urban scene reconstruction," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 1418–1425, IEEE, 2010.

[51] A.-L. Chauve, P. Labatut, and J.-P. Pons, "Robust piecewise-planar 3D reconstruction and completion from large-scale unstructured point data," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 1261–1268, IEEE, 2010.

[52] E. Imre, A. A. Alatan, and U. Gudukbay, "Rate-distortion efficient piecewise planar 3-D scene representation from 2-D images," *Image Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 483–494, 2009.

[53] P. D. Grünwald, *The minimum description length principle.* MIT press, 2007.

[54] A. A. Alatan and L. Onural, "Estimation of depth fields suitable for video compression based on 3-D structure and motion of objects," *Image Processing, IEEE Transactions on*, vol. 7, no. 6, pp. 904–908, 1998.

[55] A. Blake, P. Kohli, and C. Rother, *Markov Random Fields for Vision and Image Processing.* MIT Press, 2011.

[56] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 259–302, 1986.

[57] S. Brooks and B. Morgan, "Optimization using simulated annealing," *The Statistician*, pp. 241–257, 1995.

[58] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 11, pp. 1222–1239, 2001.

[59] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Understanding belief propagation and its generalizations," *Exploring artificial intelligence in the new millennium*, vol. 8, pp. 236–239, 2003.

[60] J. Zhang, "The mean field theory in EM procedures for markov random fields," *Signal Processing, IEEE Transactions on*, vol. 40, no. 10, pp. 2570–2583, 1992.

[61] H. Isack and Y. Boykov, "Energy-based geometric multi-model fitting," *International journal of computer vision*, vol. 97, no. 2, pp. 123–147, 2012.

[62] P. H. Torr, "Geometric motion segmentation and model selection," *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 356, no. 1740, pp. 1321–1340, 1998.

[63] E. Vincent and R. Laganiére, "Detecting planar homographies in an image pair," in *Image and Signal Processing and Analysis, 2001. ISPA 2001. Proceedings of the 2nd International Symposium on*, pp. 182–187, IEEE, 2001.

[64] S. Birchfield and C. Tomasi, "Multiway cut for stereo and motion with slanted surfaces," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 1, pp. 489–495, IEEE, 1999.

[65] M. Zuliani, C. S. Kenney, and B. Manjunath, "The multiransac algorithm and its application to detect planar homographies," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 3, pp. III–153, IEEE, 2005.

[66] M. S. Datasets, "`http://vision.middlebury.edu/stereo/data`," 2014. [Online; accessed 04-January-2014].

[67] M. Caramia and P. DellOlmo, *Multi-objective management in freight logistics: Increasing capacity, service level and safety with optimization algorithms.* Springer, 2008.

[68] S. Ruzika and M. M. Wiecek, “Approximation methods in multiobjective programming,” *Journal of optimization theory and applications*, vol. 126, no. 3, pp. 473–501, 2005.

[69] M. Ehrgott, *Multicriteria optimization.* Springer, 2005.

[70] G. J. Sullivan and T. Wiegand, “Rate-distortion optimization for video compression,” *Signal Processing Magazine, IEEE*, vol. 15, no. 6, pp. 74–90, 1998.

[71] I. Daribo, G. Cheung, and D. Florencio, “Arithmetic edge coding for arbitrarily shaped sub-block motion prediction in depth video compression,” in *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pp. 1541–1544, IEEE, 2012.

[72] D. Freedman and P. Drineas, “Energy minimization via graph cuts: Settling what is possible,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2, pp. 939–946, IEEE, 2005.

[73] I. H. Witten, R. M. Neal, and J. G. Cleary, “Arithmetic coding for data compression,” *Communications of the ACM*, vol. 30, no. 6, pp. 520–540, 1987.

[74] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov, “Fast approximate energy minimization with label costs,” *International Journal of Computer Vision*, vol. 96, no. 1, pp. 1–27, 2012.

[75] X. Wu and N. Memon, “Context-based, adaptive, lossless image coding,” *Communications, IEEE Transactions on*, vol. 45, no. 4, pp. 437–444, 1997.

[76] D. Marpe, H. Schwarz, and T. Wiegand, “Context-based adaptive binary arithmetic coding in the H. 264/AVC video compression standard,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 7, pp. 620–636, 2003.

[77] M. O. Bici, J. Lainema, K. Ugur, and M. Gabbouj, “Planar representation for intra coding of depth maps,” in *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video , 2011*, pp. 1–4, IEEE, 2011.

[78] J. A. Parker, R. V. Kenyon, and D. E. Troxel, “Comparison of interpolating methods for image resampling,” *Medical Imaging, IEEE Transactions on*, vol. 2, no. 1, pp. 31–39, 1983.

[79] Wikipedia, “Prediction by partial matching — Wikipedia, the free encyclopedia,” 2013. [Online; accessed 24-November-2013].

[80] M. Mahoney, "Adaptive weighing of context models for lossless data compression," *Florida Tech., Melbourne, USA, Tech. Rep*, 2005.

[81] `http://cs.fit.edu/~mmahoney/compression/paq.html`, "The PAQ data compression programs," 2013. [Online; accessed 24-November-2013].

[82] `http://www.imagecompression.info/gralic/`, "Lossless photo compression benchmark," 2013. [Online; accessed 24-November-2013].

[83] `http://www.squeezechart.com/bitmap.html`, "Lossless image compression," 2014. [Online; accessed 23-January-2014].

[84] H. Hampel, R. B. Arps, C. Chamzas, D. Dellert, D. L. Duttweiler, T. Endoh, W. Equitz, F. Ono, R. Pasco, I. Sebestyen, *et al.*, "Technical features of the JBIG standard for progressive bi-level image compression," *Signal Processing: Image Communication*, vol. 4, no. 2, pp. 103–111, 1992.

[85] Wikipedia, "Lempel–Ziv–Markov chain algorithm — Wikipedia, the free encyclopedia," 2013. [Online; accessed 24-November-2013].

[86] Wikipedia, "Portable Network Graphics — Wikipedia, the free encyclopedia," 2013. [Online; accessed 24-November-2013].

[87] Wikipedia, "Zip (file format) — Wikipedia, the free encyclopedia," 2013. [Online; accessed 24-November-2013].

[88] L. Merritt and R. Vanam, "x264: A high performance H. 264/AVC encoder," *online] http://neuron2. net/library/avc/overview_x264_v8_5. pdf*, 2006.

[89] M. J. Weinberger, G. Seroussi, and G. Sapiro, "The LOCO-I lossless image compression algorithm: principles and standardization into JPEG-LS," *Image Processing, IEEE Transactions on*, vol. 9, no. 8, pp. 1309–1324, 2000.

[90] Wikipedia, "Run-length encoding — Wikipedia, the free encyclopedia," 2013. [Online; accessed 24-November-2013].

[91] K. Ugur, K. Andersson, A. Fuldseth, G. Bjontegaard, L. P. Endresen, J. Lainema, A. Hallapuro, J. Ridge, D. Rusanovskyy, C. Zhang, *et al.*, "High performance, low complexity video coding and the emerging HEVC standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 20, no. 12, pp. 1688–1697, 2010.

[92] J. Stankowski, M. Domanski, O. Stankiewicz, J. Konieczny, J. Siast, and K. Wegner, "Extensions of the HEVC technology for efficient multiview video coding," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pp. 225–228, IEEE, 2012.

[93] “Jct3v-b1005_d0.” Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, Shanghai, China, October 2012.

[94] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.

[95] G. Han, J.-R. Ohm, W.-J. Han, and T. Wiegand, “Overview of the high efficiency video coding (hevc) standard,” 2012.

[96] “Dirac specification, `diracvideo.org`,” 2008.

[97] L. McMillan and G. Bishop, “Plenoptic modeling: An image-based rendering system,” in *Proceedings of the 22nd annual conference on computer graphics and interactive techniques*, pp. 39–46, ACM, 1995.

[98] S. B. Kang, R. Szeliski, and P. Anandan, “The geometry-image representation tradeoff for rendering,” in *Image Processing, 2000. Proceedings. 2000 International Conference on*, vol. 2, pp. 13–16, IEEE, 2000.

[99] M. Tanimoto and M. Wildeboer, “Frameworks for FTV coding,” in *Picture Coding Symposium, 2009. PCS 2009*, pp. 1–4, IEEE, 2009.

[100] T. Ishibashi, M. P. Tehrani, T. Fujii, and M. Tanimoto, “FTV format using global view and depth map,” in *Picture Coding Symposium (PCS), 2012*, pp. 29–32, IEEE, 2012.

[101] I. Daribo, G. Cheung, T. Maugey, and P. Frossard, “RD optimized auxiliary information for inpainting-based view synthesis,” in *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2012*, pp. 1–4, IEEE, 2012.

[102] K. Müller, A. Smolic, K. Dix, P. Kauff, and T. Wiegand, “Reliability-based generation and view synthesis in layered depth video,” in *Multimedia Signal Processing, 2008 IEEE 10th Workshop on*, pp. 34–39, IEEE, 2008.

[103] J. Shade, S. Gortler, L.-W. He, and R. Szeliski, “Layered depth images,” in *Proceedings of the 25th annual conference on computer graphics and interactive techniques*, pp. 231–242, ACM, 1998.

[104] V. Jantet, L. Morin, and C. Guillemot, “Incremental-LDI for multi-view coding,” in *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, 2009*, pp. 1–4, IEEE, 2009.

[105] S.-U. Yoon, E.-K. Lee, S.-Y. Kim, Y.-S. Ho, K. Yun, S. Cho, and N. Hur, “Coding of layered depth images representing multiple viewpoint video,” in *Proc. of Picture Coding Symposium (PCS) SS3-2*, pp. 1–6, 2006.

[106] A. Frick, B. Bartczak, and R. Koch, “Real-time preview for layered depth video in 3D-TV,” in *SPIE Photonics Europe*, pp. 77240F–77240F, International Society for Optics and Photonics, 2010.

[107] S.-C. Chan, Z.-F. Gan, K.-T. Ng, K.-L. Ho, and H.-Y. Shum, “An object-based approach to image/video-based synthesis and processing for 3-D and multiview televisions,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 19, no. 6, pp. 821–831, 2009.

[108] A. Smolic, K. Mueller, P. Merkle, P. Kauff, and T. Wiegand, “An overview of available and emerging 3D video formats and depth enhanced stereo as efficient generic solution,” in *Picture Coding Symposium, 2009. PCS 2009*, pp. 1–4, IEEE, 2009.

[109] W. Bruls and R. K. Gunnewiek, “Options for a new efficient, compatible, flexible 3D standard,” in *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pp. 3497–3500, IEEE, 2009.

[110] J. Pearson, M. Brookes, and P. Dragotti, “Plenoptic layer-based modelling for image based rendering.,” *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, 2013.

[111] A. Gelman, P. L. Dragotti, and V. Velisavljevic, “Multiview image coding using depth layers and an optimized bit allocation,” *Image Processing, IEEE Transactions on*, vol. 21, no. 9, pp. 4092–4105, 2012.

[112] T. Sikora and B. Makai, “Shape-adaptive DCT for generic coding of video,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 5, no. 1, pp. 59–62, 1995.

[113] S. Li and W. Li, “Shape-adaptive discrete wavelet transforms for arbitrarily shaped visual object coding,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 10, no. 5, pp. 725–743, 2000.

[114] Wikipedia, “Texture mapping — Wikipedia, the free encyclopedia,” 2013. [Online; accessed 8-December-2013].

[115] W3C, “`http://www.w3.org/TR/PNG/` Portable Network Graphics (PNG) specification (second edition),” 2003. [Online; accessed 9-December-2013].

[116] A. K. Moorthy and A. C. Bovik, “A survey on 3D quality of experience and 3D quality assessment,” in *IS&T/SPIE Electronic Imaging*, pp. 86510M–86510M, International Society for Optics and Photonics, 2013.

[117] S. Reichelt, R. Häussler, G. Fütterer, and N. Leister, “Depth cues in human visual perception and their realization in 3D displays,” in *SPIE Defense, Security, and Sensing*, pp. 76900B–76900B, International Society for Optics and Photonics, 2010.

[118] E. Bosc, P. Le Callet, L. Morin, and M. Pressigout, "Visual quality assessment of synthesized views in the context of 3D-TV," in *3D-TV System with Depth-Image-Based Rendering*, pp. 439–473, Springer, 2013.

[119] E. Bosc, P. Hanhart, P. Le Callet, and T. Ebrahimi, "A quality assessment protocol for free-viewpoint video sequences synthesized from decompressed depth data," in *Fifth International Workshop on Quality of Multimedia Experience*, 2013.

[120] H. Shao, X. Cao, and G. Er, "Objective quality assessment of depth image based rendering in 3DTV system," in *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, 2009*, pp. 1–4, IEEE, 2009.

[121] G. Nur and G. B. Akar, "An abstraction based reduced reference depth perception metric for 3D video," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pp. 625–628, IEEE, 2012.

[122] C. T. Hewage and M. G. Martini, "Reduced-reference quality metric for 3D depth map transmission," in *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2010*, pp. 1–4, IEEE, 2010.

[123] M. Solh and G. AlRegib, "A no-reference quality measure for dibr-based 3D videos," in *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, pp. 1–6, IEEE, 2011.

[124] E. Bosc, R. Pepion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, M. Pressigout, and L. Morin, "Towards a new quality metric for 3-D synthesized view assessment," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 7, pp. 1332–1343, 2011.

[125] Wikipedia, "Anaglyph 3D — Wikipedia, the free encyclopedia," 2013. [Online; accessed 25-December-2013].

[126] D. V. S. De Silva, W. A. C. Fernando, G. Nur, E. Ekmekcioglu, and S. T. Worrall, "3D video assessment with just noticeable difference in depth evaluation," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pp. 4013–4016, IEEE, 2010.

[127] A. T. Delong, *Advances in Graph-Cut Optimization: Multi-Surface Models, Label Costs, and Hierarchical Costs.* PhD thesis, The University of Western Ontario, Ontario, Canada, 2011.

[128] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, vol. 2. Cambridge Univ Press, 2000.

APPENDIX A

GRAPH CUT ALGORITHM

The graph cut algorithm interprets the MRF energy minimization problem as the well-known *maximum flow problem* of the optimization theory. The maximum flow problem for a given network finds the maximum flow capacity from a single source to a single sink. According to max-flow min-cut theorem, the maximum flow is equal to the minimum capacity removed from the network to avoid any flow from source to sink, and it is called *s-t min-cut problem* [55].

In case of MRF, when the energy function is in a submodular quadratic pseudo-boolean form, an equivalent flow network can be constructed and its minimum cut encodes the field configuration for the minimum MRF energy [55]. While the pseudo-boolean form enforces the MRF to be a binary random field, the quadratic form allows pairwise energy terms at most. For the submodularity, all the pairwise energy terms in the MRF energy function should satisfy the condition,

$$\mathcal{V}_{p,q}(0,1) + \mathcal{V}_{p,q}(1,0) \geq \mathcal{V}_{p,q}(0,0) + \mathcal{V}_{p,q}(1,1) \ . \tag{A.1}$$

In graph cut approach, the submodular quadratic pseudo-boolean MRF energy function is represented as flow capacities of directed edges in a graph. The diagram given in Figure A.1 is a simple case of two random variable and it explains the relation between the edge capacities and the unary and pairwise potentials of the MRF model. An s-t cut is a subset of vertices, S, which includes the source node but not the sink node. Hence, the s-t cut encodes a binary assignment to vertices, i.e, the random variables of the MRF, by being or not being an element of set S. The cost of an s-t cut is the sum of the

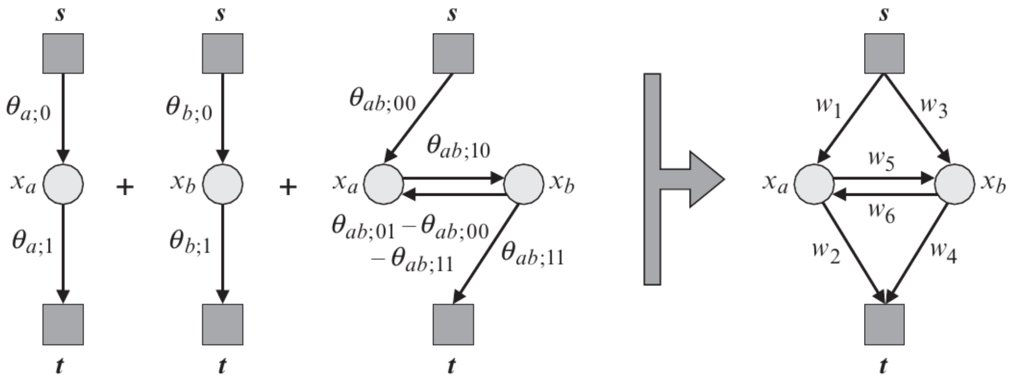


Figure A.1: A graph construction example for two random variables with unary and pairwise energies. (Reprinted with permission. Copyright The MIT Press 2011 [55])

capacities of the directed edges connecting vertices in S to vertices not included in S and it is equivalent to the MRF energy of encoded binary assignment by construction. Possible s-t cuts and the corresponding costs are shown in an example with two random variables in Figure A.2. In optimization theory, there exist algorithms that find the minimum s-t cut, when all the directed edge capacities are nonnegative; this condition is satisfied by the submodularity constraint [55].

Hence, the min s-t cut solution of the graph cut approach solves the second order submodular MRF energies exactly for binary cases. When there are multiple labels for MRF assignments the graph cut approach is utilized to find the efficient updates or moves of a greedy algorithm. Two popular update rules utilizing the graph cut algorithm are the α-expansion and $\alpha - \beta$ swap moves [58]. While in α-expansion the current MRF assignments can be changed to α label or keeps

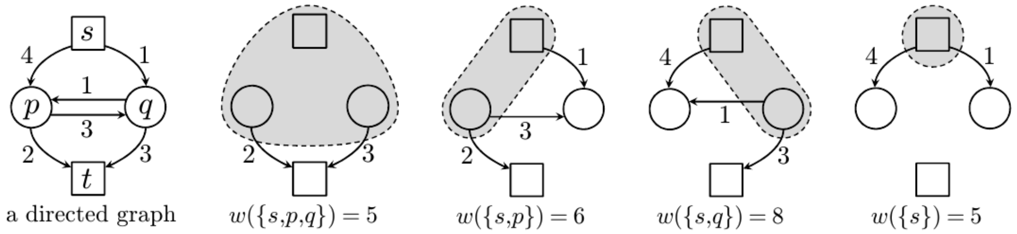


Figure A.2: Possible s-t cuts for two random variable case (Reprinted with permission. Copyright Andrew Thomas Delong 2011 [127]).

its current assignment, in $\alpha - \beta$ swap moves the random variables currently assigned to α or β label can be swapped and the rest keeps its current assignment. Illustrations of the moves are given in Figure A.3. For both assignment update rules, all possible moves can be represented by a binary coding like change or keep its assignment. The best move which makes the maximal decrease in the MRF energy can be obtained by solving the s-t min-cut problem of the corresponding binary representation. The moves are applied for every label or label pair consecutively until no energy decrease is possible with the utilized update rule. The resulting local minimum with the greedy algorithm is a good approximation in general, and for the α-expansion rule it is within a known factor of the global optimum [58].

A.0.1 MRF optimization with label costs

The label cost term potential defined in 2.10 is a higher order potential function of all random variables in the field. It is a global regularization term of the MRF model. However, it violates the quadratic form of the MRF energy; hence, it is not possible to trivially generate a directed graph to solve it as a graph cut problem. In [74] Delong et al. proposed two methods respectively for α-expansion and $\alpha - \beta$ swap moves to minimize the energy terms with label costs.

The method for the $\alpha - \beta$ swap moves proposes to compare the costs of the three possible moves at each assignment updates. The first possible move is the conventional best $\alpha - \beta$ swap move for the energy function without the label

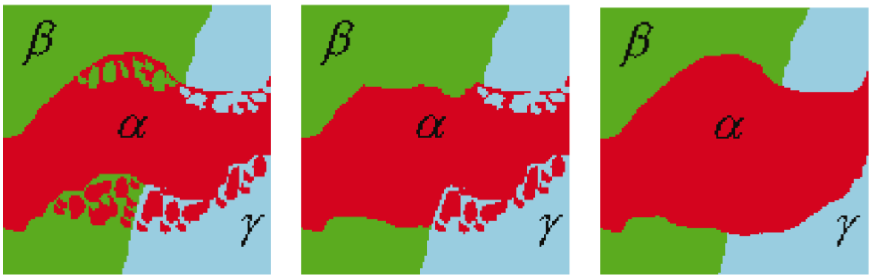


Figure A.3: GC based MRF update rule examples. From left to right; initial labeling map, $\alpha - \beta$ swap move and α-expansion move (Reprinted with permission. Copyright IEEE 1999 [58]).

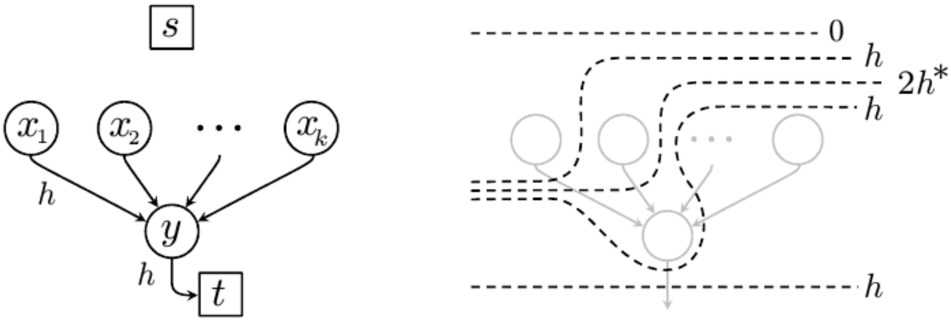


Figure A.4: Graph construction for a label cost (Reprinted with permission. Copyright Andrew Thomas Delong 2011 [127]).

costs. The other two possible moves are the ones which eliminate the α or β assignments by swapping all β values to α or vice versa. The costs of these three assignments are compared with considering the labeling costs and the minimum one is selected as the $\alpha - \beta$ swap move.

The method for the α-expansion move includes higher order label cost terms in the graph construction. According to the general graph construction method proposed in [72], the label costs are converted into quadratic forms with the help of an auxiliary variable, y. The details about quadratic conversion can be found in [127]. The graph constructed according to the label cost in the quadratic form is illustrated in Figure A.4. The right side of this figure shows that any s-t cut that contains more than one variables of type x should also contain the auxiliary variable y to be a minimum s-t cut. With this graph construction trick, the label costs can be encoded in the graph representing the MRF energy.

A simple graph construction example to solve the s-t min-cut problem for an α-expansion move is given in Figure A.5. While the black edges of the graph account for the higher order potential of the label costs, the gray edges account for the combination of unary and pairwise potentials. In the given example, the α-expansion move for the label α considers only the label costs of the β and γ labels since α-move can not.

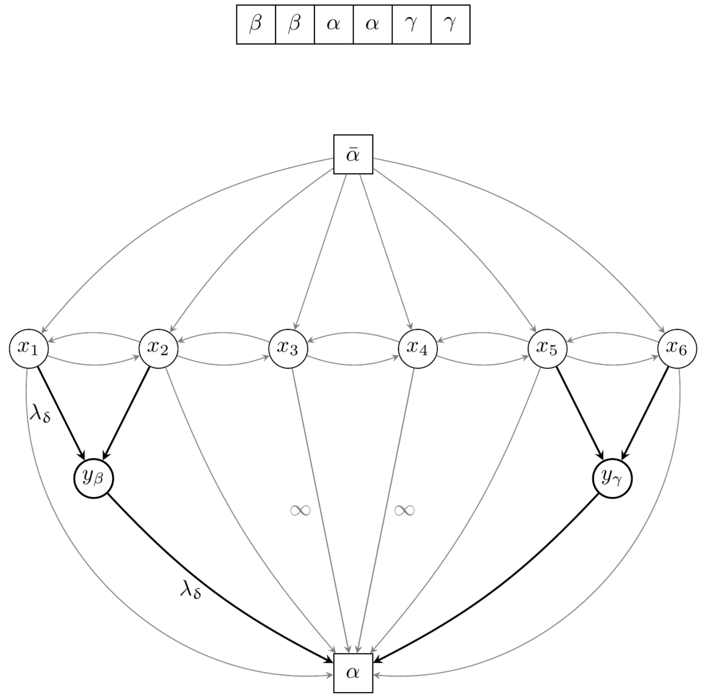


Figure A.5: A graph construction example for an α-expansion move with labeling cost. The current labeling of the 1-D field of 6 variable is given at the top. The auxiliary nodes and the edges for the label cost are highlighted.

APPENDIX B

MIDDLEBURY CAMERA CALIBRATION MATRICES

The projection matrix of a camera maps a point in 3D space to a 2D point on the image plane of the camera. This mapping is a linear function if the points in the domain and the range set are represented in homogenous coordinate systems.

According to pinhole camera model the projection matrix, P, can be decomposed into internal, K, and external calibration matrices [128]. The external calibration matrix is also composed of rotation, R, and translation, T, matrices as,

$$P = K\left[R|T\right] \ . \tag{B.1}$$

The *Middlebury* dataset provides disparity maps of two cameras which have pure horizontal translational shift between them. The disparity maps of the views are given as 8-bit gray scale images which encode the disparity values between 0 and d_{max}, linearly.

In order to obtain a generic camera calibration matrix compatible with the scene geometry, the internal (K)and external (R, T) calibration matrices of the stereo pair are defined as,

$$K = \begin{bmatrix} d_{max} & 0 & w/2 \\ 0 & d_{max} & h/2 \\ 0 & 0 & 1 \end{bmatrix} \ , \tag{B.2}$$

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \ , \tag{B.3}$$

$$T_{left} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad , \quad T_{right} = \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} \quad . \tag{B.4}$$

The w and h are the width and the height of the images respectively.

APPENDIX C

HEVC-3D RENDERER CONFIGURATION

All the novel view rendering experiments in this thesis, utilizing the renderer software given by the HEVC-3D extension, used the following configuration.

Table C.1: Parameter values of HEVC-3D renderer.

RenderDirection	0
RenderMode	0
TemporalDepthFilter	0
SimEnhance	0
ShiftPrecision	2
HoleFillingMode	1
BlendMode	0
BlendZThresPerc	30
BlendUseDistWeight	1
BlendHoleMargin	6
Sweep	0

CURRICULUM VITAE

BURAK OĞUZ ÖZKALAYCI

Birth date	: 04/02/1981	Marital status	: Married
Sex	: Male	Nationality	: Turkish (TC)
Mobile	: +90 533 363 5814	*E-mail*	: bozkalayci@gmail.com

EDUCATION

Degree	**Institution**	**Year of Graduation**
M.S.	METU Electrical and Electronics Eng.	2006
B.S.	METU Mathematics - Double Major	2004
B.S.	METU Electrical and Electronics Eng.	2003
High School	Ankara Atatürk Anatolian High School	1999

PROFESSIONAL EXPERIENCE

Year	**Place**	**Enrollment**
2011-Present	Aselsan Inc., *Ankara, Turkey*	Senior Systems Engineer
2007-2011	Vestek R&D, *Ankara, Turkey*	Senior Design Engineer
2007	Universiteit Gent, *Gent, Belgium*	Research Assistant
2005-2006	METU EEE Dep., *Ankara, Turkey*	Research Assistant
2003-2005	Başkent Uni. EE Dep., *Ankara, Turkey*	Teaching Assistant

PUBLICATIONS

Dissertations

- **Multi-view video coding via dense depth field**, master thesis supervised by A.A.Alatan, 2006.

Journal Publications

- **3D Planar Representation of Stereo De[th Omages for 3DTV Applications**, B.Özkalaycı, and A.A.Alatan, *Submitted to* IEEE Transactions on Image Processing, 2014.
- **Towards a Solution in 3D Reconstruction from Broadcast Video**, S.Knorr, E.İmre, B.Özkalaycı, U.Topay, A.A.Alatan, and T.Sikora, The Journal of Image Communications, Elsevier Signal Processing, 2006.

Conference Publications

- **MRF-Based Planar Co-segmentation for Depth Compression**, B.Özkalaycı, and A.A.Alatan, *Submitted to* International Conference on Image Processing, 2014.
- **A Novel Planar Layered Representation for 3DTV and Freeview TV Applications**, B.Özkalaycı, and A.A.Alatan, International Conference on Multimedia and Expo Hot 3D Workshop, 2013.
- **Occlusion Handling Frame Rate Up-Conversion**, B.Özkalaycı, and A.A.Alatan, International Conference on Image and Signal Processing and Analysis, 2011.
- **Occlusion Adaptive Frame Rate Up-conversion**, B.Özkalaycı, A.A.Alatan, and A.Baştuğ, International Conference on Consumer Electronics-Berlin, 2011 .
- **Sulci Detection for Improving the Accuracy of Cortical Thickness Measurements in Focal Cortical Dysplasia Diagnosis**, B.Özkalaycı, Universiteit Gent FirW PhD Symposium, 2007.
- **3-D Structure Assisted Reference View Generation for H.264 Based Multiview Video Coding**, B.Özkalaycı, S.O.Gedik, A.A.Alatan, Proceedings of Picture Coding Symposium, 2007.
- **Multi-view Video Coding via Dense Depth Estimation**, B.Özkalaycı, and A.A.Alatan, Proceedings of 3DTV-Con Conference, 2007.
- **A Modular Scheme for 2D/3D Conversion of TV Broadcast**, S.Knorr, E.İmre, B.Özkalaycı, U.Topay, A.A.Alatan, and T.Sikora, Proceedings of 3DPVT Conference, 2006.
- **Çoklu Görüntü Kodlama Amaçlı Sık Derinlik Haritası Kestirimi**, B.Özkalaycı, A.A.Alatan, Proceedings of SIU conference, 2006 (National Conference).

Patents

- **Method of, and apparatus for, detecting image boundaries in video data**, B.Özkalaycı, EP2509045, 2012.
- **Motion estimation method for frame rate up conversion applications**, B.Özkalaycı, EP2461565, 2012.
- **Super resolution based N-View + N-Depth multiview video coding**, B.Özkalaycı, E.Taşlı, EP2373046, 2011.
- **Super resolution enhancement for N-View + N-Depth multiview video**, B.Özkalaycı, E.Taşlı, EP2369850, 2011.
- **Motion compensated interpolation**, B.Özkalaycı, EP2360637, 2011.
- **A method for foreground/background discrimination**, B.Özkalaycı, EP2355042, 2011.
- **Motion vector field retiming method**, B.Özkalaycı, EP2334065, 2011.
- **Background motion estimate based halo reduction**, B.Özkalaycı, and A.Baştuğ, EP2237559, 2010.
- **Halo reducing motion-compensated interpolation**, B.Özkalaycı, and A.Baştuğ, EP2237560, 2010.

Misc.

- **A Novel Planar Layered Representation for 3D Content and Its Applications**, B.Özkalaycı, and A.A.Alatan, IEEE COMSOC MMTC E-Letter, 2013.