POZNAŃ UNIVERSITY OF TECHNOLOGY
FACULTY OF ELECTRONICS AND TELECOMMUNICATIONS
CHAIR OF MULTIMEDIA TELECOMMUNICATION AND MICROELECTRONICS

Doctoral Dissertation

Depth Estimation in Free-Viewpoint Television

Dawid Mieloch

Supervisor: prof. dr hab. inż. Marek Domański

Auxiliary supervisor: dr inż. Olgierd Stankiewicz

TABLE OF CONTENTS

ABSTRACT 5

STRESZCZENIE 6

LIST OF SYMBOLS AND ABBREVIATIONS 7

1 INTRODUCTION 10

1.1 Scope of the dissertation 10

1.2 Goals and thesis of the dissertation 13

1.3 Overview of the dissertation 14

1.4 Acknowledgements 15

2 DEPTH ESTIMATION IN FREE-VIEWPOINT TELEVISION 16

2.1 Analysis of depth map quality and processing time in the scope of FTV 16

2.2 Assumptions and requirements for a depth estimation method for FTV 21

3 OVERVIEW OF STATE-OF-THE-ART 22

3.1 Depth estimation fundamentals 22

3.1.1 Local estimation 22

3.1.2 Global estimation 23

3.2 State-of-the-art methods of depth estimation 27

3.2.1 Inter-view consistency of depth maps 27

3.2.2 Usage of segmentation in depth estimation 29

3.2.3 Temporal consistency of depth maps 30

3.2.4 Parallelisation of depth optimisation 31

3.2.5 Other depth estimation methods 31

3.3 Conclusions 33

4 PROPOSED MULTIVIEW DEPTH ESTIMATION METHOD 34

4.1 Overview of the proposed method of depth estimation 34

4.2 Proposed cost function 35

4.2.1 Inter-view matching cost 39

4.2.2 Intra-view discontinuity cost 40

4.3 Proposed temporal consistency enhancement method 42

4.4 Proposed method of estimation parallelisation 45

4.5 Details of the implementation of the proposed depth estimation method 50

5 METHODOLOGY OF EXPERIMENTS 53

5.1 Assessment of the quality of depth maps 53

5.2 Assessment of depth maps temporal consistency 55

5.3 Multiview sequences test set 56

6 EXPERIMENTAL RESULTS 59

6.1 Comparison with the state-of-the-art depth estimation method 59

6.2 Impact of the number of segments on depth maps quality and processing time 64

6.3 Impact of the number of views on depth maps quality and processing time 67

6.4 Impact of temporal consistency enhancement on depth maps quality and processing time 70

6.5 Impact of the parallelisation method on depth maps quality and processing time 75

7 APPLICATION OF PROPOSED DEPTH MAP ESTIMATION METHOD IN FTV SYSTEMS 79

7.1 Acquisition of the new FTV test sequences 79

7.2 Overview of Poznań University of Technology Free-Viewpoint Television system 81

8 SUMMARY OF THE DISSERTATION 83

8.1 Original achievements of the dissertation 83

8.2 Research work done 85

8.3 Conclusions 86

REFERENCES 88

PUBLICATIONS OF THE AUTHOR 101

APPENDIX 107

ABSTRACT

The development of a **novel depth estimation method for free-viewpoint television systems** is the main goal of research presented in this dissertation. In free-viewpoint television the functionalities offered to a viewer are extended by the possibility of controlling the displayed viewpoint of a scene.

The dissertation presents an analysis of the depth estimation process in the scope of free-viewpoint television (FTV) systems. The author of the dissertation defines a set of requirements, which are not met by state-of-the-art methods of depth estimation, and that should be met by a new depth estimation method designed for free-viewpoint television. The main focus is put on the influence of the quality of estimated depth maps on the quality of virtual views and on an analysis of the high complexity of depth estimation. The fundamentals of depth estimation, together with state-of-the-art methods of depth estimation, are also described.

The novel method for depth estimation proposed by the author of the dissertation consists of **three primary achievements of this dissertation: the inter-view consistent segment-based depth estimation method, the temporal consistency enhancement that simultaneously increases temporal consistency of depth maps and reduces the complexity of estimation, and a new method of parallelisation for graph-based depth estimation methods.**

The experimental results contain a comparison of the proposed depth estimation method with the reference method provided by the ISO/IEC MPEG group. The efficiency of the proposed method for variable estimation parameters and the results for the proposed methods of temporal enhancement and parallelisation were also included in the dissertation.

This dissertation also includes a description of new test sequences that are available for the research community and include depth maps estimated using the method proposed by the author. The depth estimation method that is the result of research presented in this dissertation is also a part of the free-viewpoint television system developed by the Chair of Multimedia Telecommunications and Microelectronics of Poznań University of Technology. These achievements prove the high usability of the presented depth estimation method in real FTV systems.

STRESZCZENIE

Stworzenie **nowej metody estymacji map głębi przeznaczonej dla systemów telewizji swobodnego widzenia** jest głównym celem badań, które zostały zaprezentowane w niniejszej rozprawie. W telewizji swobodnego punktu widzenia, możliwości widza są rozszerzone poprzez możliwość kontroli aktualnie oglądanego przez niego punktu widzenia sceny.

W rozprawie zawarto analizę procesu estymacji głębi w zakresie dotyczącym systemów telewizji swobodnego punktu widzenia (ang. free-viewpoint television – FTV). Autor rozprawy definiuje zbiór wymagań, które nie są spełnione przez dotychczas znane metody estymacji głębi, a które powinna spełniać nowa technika estymacji głębi przeznaczona dla telewizji swobodnego punktu widzenia. Główny cel to określenie wpływu jakości map głębi na jakość widoków wirtualnych w systemach telewizji swobodnego punktu widzenia i analiza wysokiej złożoności obliczeniowej procesu estymacji głębi. W rozprawie przedstawione są również podstawy estymacji głębi oraz dotychczas znane metody estymacji.

Nowa metoda estymacji map głębi zaproponowana przez autora składa się na **trzy główne osiągnięcia tej rozprawy: przestrzennie spójną metodę estymacji map głębi opartą na segmentacji widoków, metodę zwiększenia spójności czasowej map głębi zmniejszająca złożoność obliczeniową estymacji oraz nową metodę zrównoleglania procesu optymalizacji opartego na wykorzystaniu grafów.**

Zaprezentowane wyniki eksperymentalne zawierają porównanie prezentowanej metody estymacji z metodą odniesienia rozpowszechnioną przez grupę ISO/IEC MPEG. Wydajność proponowanej metody dla zmiennych parametrów estymacji i wyniki badań proponowanych usprawnień spójności czasowej i zrównoleglania obliczeń również zostały zawarte w rozprawie.

Rozprawa zawiera również opis nowych sekwencji testowych, które zostały udostępnione środowisku badaczy, a które to zawierają mapy głębi wyznaczone za pomocą metody prezentowanej przez autora. Metoda ta jest również częścią systemu telewizji swobodnego punktu widzenia, który został stworzony przez Katedrę Telekomunikacji Multimedialnej i Mikroelektroniki Politechniki Poznańskiej. Te osiągnięcia rozprawy podkreślają dużą przydatność proponowanej metody w rzeczywistych systemach FTV.

LIST OF SYMBOLS AND ABBREVIATIONS

β – smoothing coefficient

β_0 – initial smoothing coefficient

C – set of views

c – view used in depth estimation

c' – view neighbouring to some view

D_p – data term for the point p

$\overline{d}_p$ – currently considered depth of a point p

d_s – currently considered depth of the segment s

D – set of views neighbouring to some view

$M_{s,s'}$ – inter-view matching cost between segments s and s'

$m_{s,s'}$ – core of the inter-view matching cost between segments s and s'

P – set of points of the input view

p – point of the input view

Q – set of points in the neighbourhood of some point

q – point in the neighbourhood of some point

S – set of segments in some view

s – segment in some view

s' – segment in the view c', which corresponds to the segment s in the view c for the currently considered depth d_s

T_P – threshold used to decide if a segment is unchanged in comparison to previous P type depth frame

T_I – threshold used to decide if a segment is unchanged in comparison to previous I type depth frame

μ_s – centre of a segment s

W – set of points in the window of the size specified by the user

w – point in some window W

$V_{p,q}$ – smoothness term for point p and q

T – set of segments neighbouring to some segment

$T[\cdot]$ – 3D transform obtained from intrinsic and extrinsic parameters of cameras

t – segment neighbouring to some segment

$V_{s,t}$ – intra-view discontinuity cost between segments s and t

$[Y\ C_b\ C_r]_p$ – vector of Y, C_b, C_r colour components of the point p

$[\hat{Y}\ \hat{C}_b\ \hat{C}_r]_s$ – vector of average Y, C_b, C_r colour components of the segment s

2D – two-dimensional

3D – three-dimensional

BBB – Big Buck Bunny, a set of multiview test sequences

CPU – central processing unit

DERS – Depth Estimation Reference Software, the state-of-the-art depth estimation method provided by the MPEG community

FTV – free-viewpoint television

GC – graph cut algorithm

GPU – graphics processing unit

HEVC – High Efficiency Video Coding

HM – High Efficiency Video Coding test model

MPEG – Moving Pictures Experts Group (ISO/IEC JTC1/SC29/WG11) of International Standardization Organization (ISO) and International Electrotechnical Commission (IEC)

MVD – multiview video plus depth

PSNR – peak signal-to-noise ratio

QP – quantisation parameter

SAD – sum of absolute differences

SNIC – Simple Non-Iterative Clustering, the superpixel segmentation method

SSD – sum of squared differences

VSRS – View Synthesis Reference Software, the state-of-the-art virtual view synthesis method provided by the MPEG community

WTA – winner-takes-all, a local depth estimation method

1 INTRODUCTION

1.1 Scope of the dissertation

Free-viewpoint television (FTV) [51], [53], [94], [99], [100], [121] provides the ability of free navigation through a natural three-dimensional scene. A user of an FTV system is not limited to watch only the views acquired by cameras – **the idea of free navigation assumes that it should be possible to view a scene from any arbitrary viewpoint and view direction. A scene can be watched from virtual views, placed even between or in front of real cameras of a system** (Fig. 1.1).

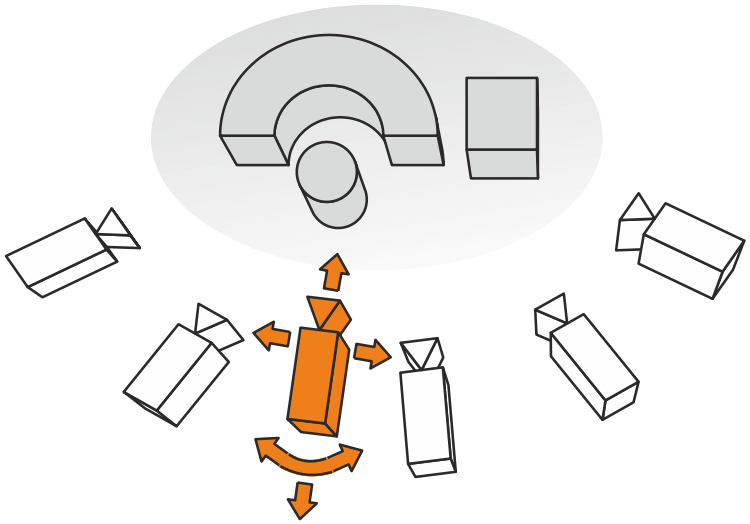


Fig. 1.1. Free navigation through a scene. Fixed positions and directions of real cameras (corresponding to real views) are represented by the white cameras, a virtual view, which can be freely moved by a user, is represented by the orange camera.

First of all, multiple views of a scene are acquired synchronously by cameras of an FTV system. Then, a three-dimensional representation of a scene is estimated and used, together with real views, to synthesise virtual views of a scene [11], [14], [25], [58], [93]. The most common representation of a scene in free-viewpoint television systems is MVD [73] (multiview plus depth), in which a scene is represented as a set of views and their corresponding depth maps. An example of the MVD representation for one of the FTV test sequences is shown in Fig. 1.2. A depth map is a sequence of matrices of depth samples for each point of

the corresponding view. For easy visualisation, these matrices are usually represented as grey-scale images. The depth of a point is represented in this case by grey levels – points that are close to the camera are whiter in a depth map image, while more distant points in the background are darker.

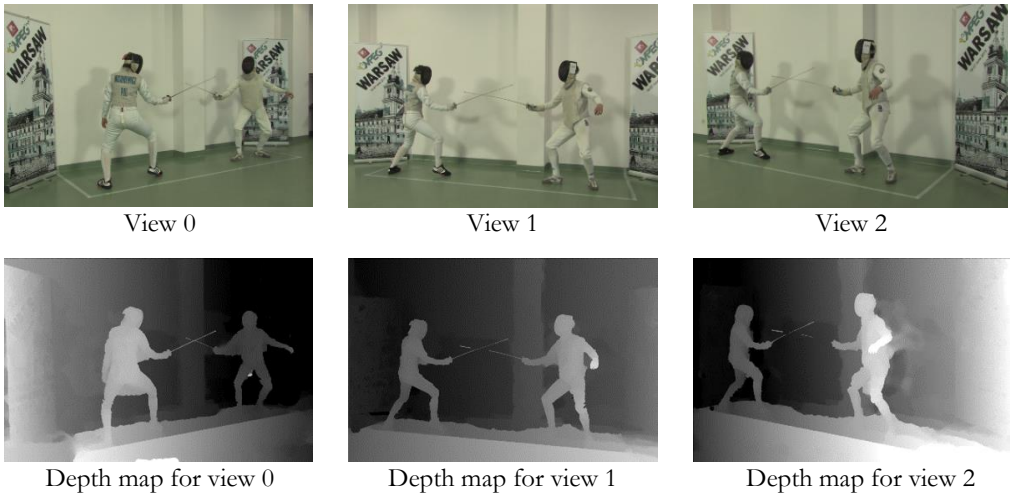


Fig. 1.2. An example of the MVD representation for the "Poznań Fencing2" test sequence [21]. On the top, there are three views of the sequence, while on the bottom, the three depth maps that correspond to the presented views.

Depth maps can be estimated algorithmically using multiple acquired views of a scene. The depth of a point is calculated on the basis of a search for a corresponding point, i.e. a point that represents the same part of a scene in another view. **The estimation of depth maps from the views acquired by cameras of a free-viewpoint television system is the main problem considered by the author in this dissertation.**

Depth maps can be also acquired with depth cameras that measure the distance to an object using, e.g. infrared waves [10]. Nevertheless, the possible applications of depth cameras in free-viewpoint television systems are limited, for instance, because of the low resolution of depth cameras. Moreover, the use of depth cameras in outdoor scenes is even more problematic due to interferences from other infrared illumination sources and a limited measurement range of these cameras.

In a typical structure of an FTV system that provides a functionality of free navigation, the main steps of multiview video acquisition and processing are as follows (Fig. 1.3): synchronous acquisition of multiview video using cameras of the system, camera parameters estimation (in order to calculate parameters and positions of cameras), pre-processing of

a multiview video (e.g. colour correction and correction of lens distortions), the estimation of depth maps for all views, and synthesis of the virtual view selected by a viewer [22].

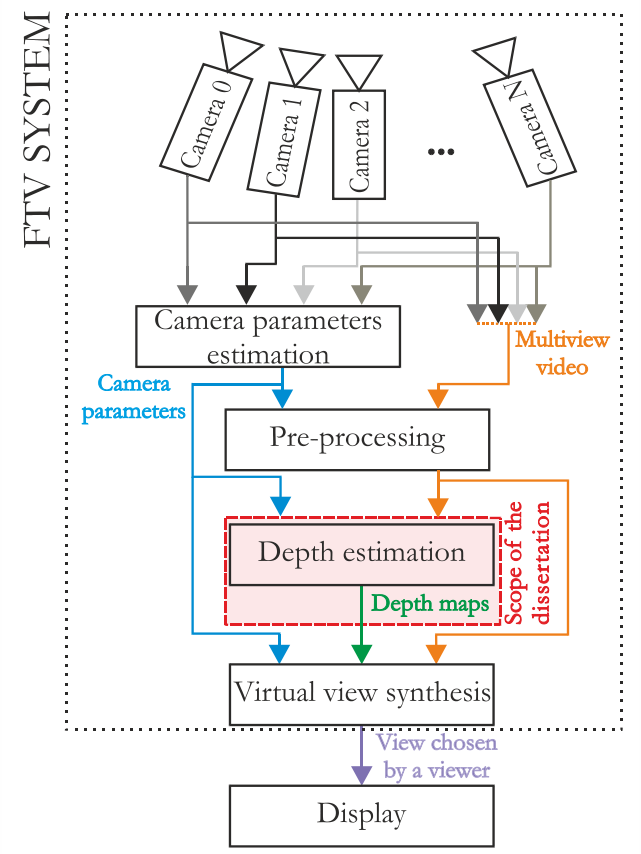


Fig. 1.3. A typical structure of an FTV system. The scope of the dissertation is highlighted with the red rectangle.

In the presented structure of a free-viewpoint television system, the quality of depth maps has a direct impact on the quality of synthesised virtual views [74], and thus on the quality of the experience of navigating through a scene. **The development of a novel depth estimation method that provides the high quality of virtual views in free-viewpoint television systems is the main goal of research presented in this dissertation.**

The topic of depth estimation is currently of high importance, what can be seen in variety and multitude of the noteworthy depth estimation methods presented during last two years [28], [41], [59], [63], [84], [86], [120]. A depth estimation method that can be successfully used for free-viewpoint television has to match very strict requirements that concern the quality of estimated depth maps (described in detail in Chapter 2) Nevertheless, these requirements are not fully ensured by current state-of-the-art methods (as presented in Chapter 3).

Moreover, the variety of hitherto presented free-viewpoint television systems [53] demonstrates that FTV systems vary significantly in the positioning of cameras and their number (from a few to hundreds). Therefore, depth estimation for free-viewpoint television requires also the reduction of the high complexity of estimation (resulting, among other reasons, from the high number of views), which makes it the most complex step of multiview video processing in the FTV.

The necessity of fulfilling all the presented requirements makes it very difficult to develop a versatile depth map estimation method that could be successfully utilised in all FTV systems. Therefore, **research performed for this dissertation significantly influences the further evolution of FTV systems.**

1.2 Goals and thesis of the dissertation

The main goal of this dissertation is to improve the quality of depth maps estimated for the purpose of virtual view synthesis in free-viewpoint television systems in comparison to state-of-the-art depth estimation methods. The focus is also put on the reduction of the complexity of depth estimation.

The thesis of the dissertation is formulated as follows:

It is possible to reduce the processing time of depth estimation and improve the quality of virtual views in free-viewpoint television systems in comparison to the state-of-the-art depth estimation methods by means of image segmentation and temporal consistency enhancement.

In order to prove the thesis stated above, a novel segmentation-based method of depth estimation is proposed by the author. The proposed method is tested and compared with the state-of-the-art depth estimation method DERS developed by the MPEG community [91], through virtual view synthesis performed using estimated depth maps.

Some of the ideas used in the proposed depth estimation method were already described by the author in [22], [68], [70], while the simplified version of the proposed method of depth maps temporal consistency enhancement was described by the author in [69] and [72].

1.3 Overview of the dissertation

Here, the organisation of this dissertation is presented. In Chapter 1, the author describes the scope of the dissertation, together with an introduction to the subject of free-viewpoint television and the estimation of depth maps.

In Chapter 2, the author presents the depth estimation process in the scope of free-viewpoint television systems. The author considers the fundamentals of virtual view synthesis necessary to define a set of requirements that have to be met by a new method of depth estimation. The main focus is put on the influence of the quality of estimated depth maps on the quality of the virtual views and on the analysis of the high complexity of depth estimation.

In Chapter 3, the fundamentals of depth estimation, together with the current state-of-the-art methods of depth estimation, are presented. The descriptions of the presented methods mainly refer to their usability in free-viewpoint television systems.

In Chapter 4, the author presents the novel method for depth estimation proposed by the author of this dissertation. The proposal presented in this chapter consists of 3 primary achievements of this dissertation: the inter-view consistent segment-based depth estimation method, temporal consistency enhancement that simultaneously reduces the complexity of estimation, and the new method of the parallelisation for graph-based depth estimation methods.

In Chapter 5, the author describes the methodology of the experimental verification of depth estimation methods. The author describes and explains the choice of methods of the assessment of the quality and the temporal consistency of depth maps that focus on the usability of depth maps in free-viewpoint television. The set of multiview test sequences that were used in all experiments is also described.

The results of the experiments are presented in Chapter 6. The results include the comparison of the presented depth estimation method and the state-of-the-art method provided by the ISO/IEC MPEG group, as well as the performance of the proposed method for variable parameters of performed estimation: the number of cameras used in estimation and the number of segments in each view. The results also include the performance of the presented temporal consistency enhancement and the proposed parallelisation method.

In Chapter 7, the author presents the applications of the proposed depth estimation method for the purposes of preparing of the new test sequences available for the research community and in the practical free-viewpoint television system that was developed by the Chair of Multimedia Telecommunication and Microelectronics of the Poznań University of

Technology. A brief description of this system, co-created by the author of this dissertation, is also presented.

Finally, in Chapter 8, the author summarises the presented dissertation. In this chapter, the author presents all the original achievements of this dissertation and concludes the performed research on depth estimation in free-viewpoint television systems.

1.4 Acknowledgements

Research presented in this dissertation was supported by the National Centre for Research and Development, Poland under Project no. TANGO1/266710/NCBR/2015 and by the Polish Ministry of Science and Higher Education for the status activity consisting of research and development and associated tasks supporting development of young scientists and doctoral students.

2 DEPTH ESTIMATION IN FREE-VIEWPOINT TELEVISION

Depth maps used in virtual view synthesis have to match a specific set of requirements that result from properties and capabilities of free-viewpoint television systems. This chapter presents a review of major problems of depth estimation in the scope of free-viewpoint television. An analysis of the depth maps properties that are required to perform virtual view synthesis of high quality, together with a discussion on the complexity of the depth maps estimation, are described in Section 2.1. The presented analysis is used in Section 2.2 to state requirements for a versatile depth estimation method, which can be successfully used in any free-viewpoint television system.

In research presented in this dissertation, MVD is used as the representation of a scene. There are several different representations of a 3D scene in FTV systems, e.g. a scene can be represented as a set of 3D points in 3D point cloud representation [109] or as an epipolar plane image [41], [115] that is based on epipolar line geometry [33].

Nevertheless, the importance and spread of MVD representation in new video technologies were also emphasised by the Moving Picture Experts Group (MPEG) of the International Organization for Standardization (ISO). The group issued a Call for Evidence (CfE) on a coding technology for MVD representation [15], [123]. A CfE is a form of an invitation for researchers to show the current state of their compression technologies. Technologies presented by research laboratories that answered that CfE proved that MVD representation can be compressed very efficiently [17], [96], [102], with about 50% reduction of the required bitrate in comparison to simulcast coding. High coding efficiency makes MVD representation very beneficial for multiview systems such as FTV systems. Besides in FTV, MVD representation is used also in 3D scene modelling [10] and some machine vision applications [34], [66], [89], [97].

2.1 Analysis of depth map quality and processing time in the scope of FTV

A. Depth estimation complexity

Depth estimation is the most complex step of the processing of a multiview video in FTV systems. This complexity is the result of a large number of views (even up to hundreds of cameras [53]) and a high resolution of cameras. FTV systems can be classified according to

distances between cameras used to acquire a scene. Systems that use many cameras that are close to each other are described as dense FTV systems, while systems with more widely spaced cameras are described as sparse systems.

Even in a sparse FTV system with, e.g. 10 high-resolution cameras, the complexity of depth estimation is high, because depth has to be estimated for almost 20 million points for each frame. The depth of a point is calculated on the basis of a search for the corresponding point in another view that represents the same part of a scene. Therefore, because the distance between neighbouring cameras in sparse systems is increased in comparison to dense systems, the number of possible depths that have to be checked in order to find the corresponding points in neighbouring views is increased as well. It further increases the complexity of depth estimation.

In the methods of depth estimation that provide the quality of depth maps sufficient for the virtual view purpose, the processing time is usually not lower than few minutes [59], [91]. In proposed structures of FTV systems, depth estimation is assumed to be done by a provider of free navigation service [53], [94], therefore, the functionality of free navigation through a scene is available after some period of time from the moment of an acquisition of a multiview video. Therefore, a minimisation of the complexity of depth estimation has a direct impact on the usefulness of FTV systems.

In order to estimate depth maps of the sufficient quality for free-viewpoint television, it is required not only to reduce the high complexity of estimation but also to ensure the high quality of estimated depth maps, with particular emphasis on the inter-view and temporal consistencies of depth maps, described in the following section.

B. Inter-view and temporal consistency of depth maps

Depth maps are **inter-view consistent** if they represent the same 3D scene. It means that in inter-view consistent depth maps two points that represent the same part of a scene in different views have such values of depth that after three-dimensional projection of these points they represent the same 3D point of a scene. The depth of a point is defined here as the distance from the plane of a camera that acquired this point to the 3D position of this point, while the plane of a camera is understood as the plane that contains the sensor of a camera.

Lack of the inter-view consistency of depth maps has a direct impact on the quality of generated virtual views. In virtual view synthesis, points from a view are projected to a virtual

view using the corresponding depth map and camera parameters (intrinsic camera parameters and the position of the camera represented by extrinsic camera parameters).

A simplified virtual view synthesis is shown in Fig. 2.1. In order to determine the positions of objects that were occluded in some views, it is required to use more than one view in synthesis. Therefore, virtual view synthesis is performed when at least 2 views of a scene and their corresponding depth maps are available. If an object was not occluded, then the colour of this object is usually averaged in the final virtual view, using colours of the object from both real views.

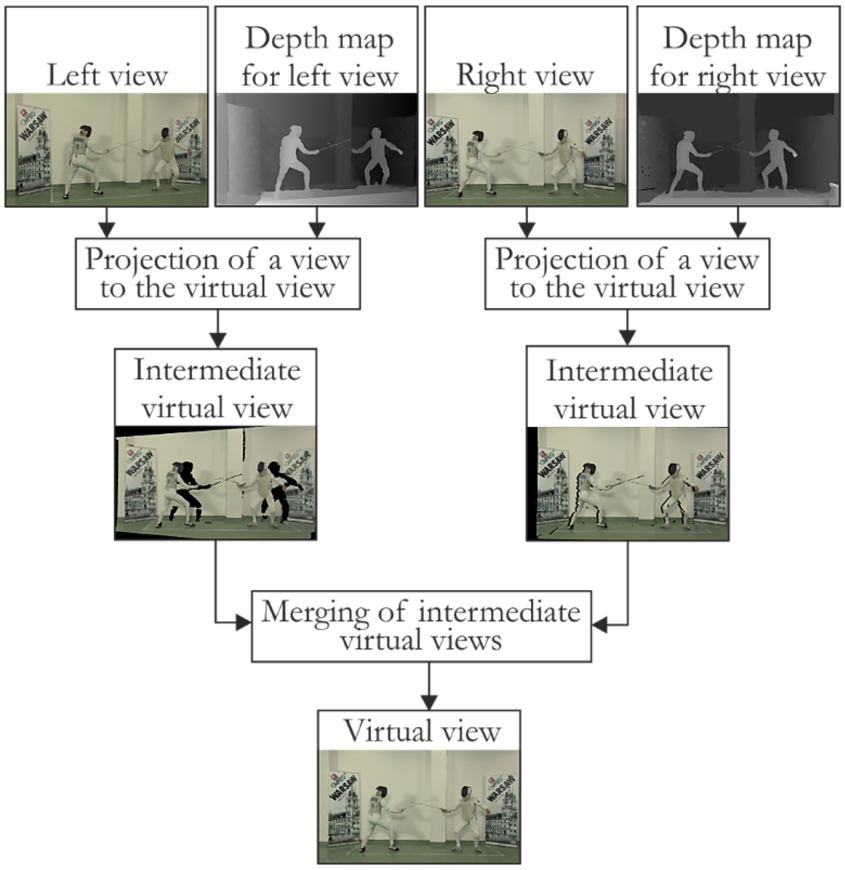
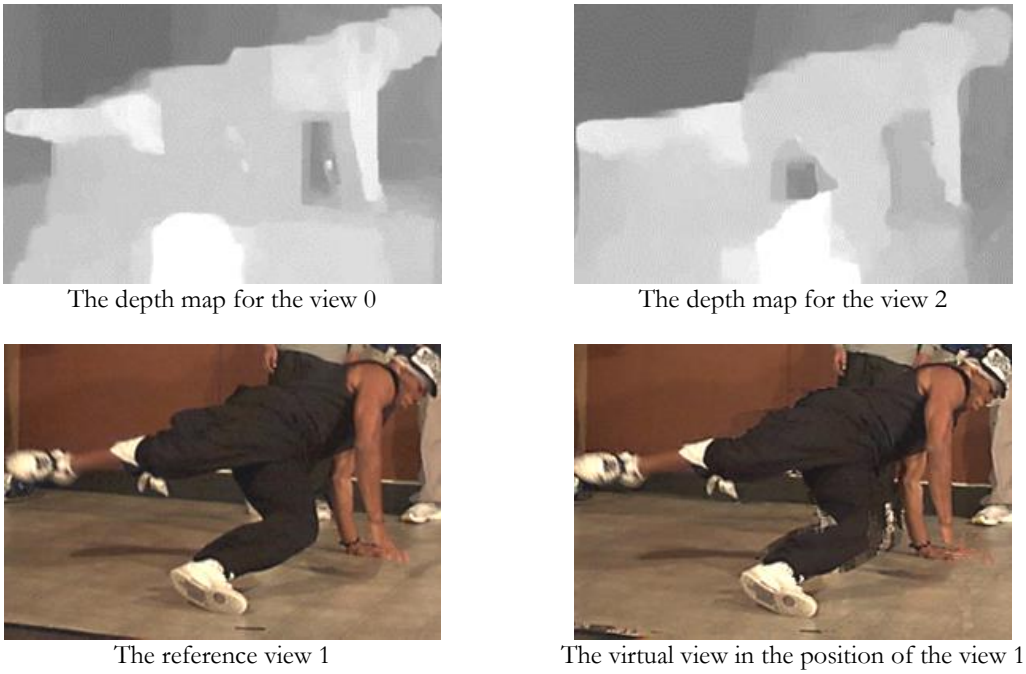


Fig. 2.1. Virtual view synthesis. The black regions in the intermediate virtual views are occluded in the corresponding input views.

Unfortunately, a point of a scene sometimes has such values of depth, that after the 3D projection of this point, the point does not represent the same part of a 3D scene as corresponding points in neighbouring views. In such case, an ambiguity of the position of this point in a virtual view can be seen. Fig. 2.2 presents fragments of depth maps that were estimated independently for two neighbouring views. Independent estimation caused that the depth of the dancer is not inter-view consistent in some parts of the views (e.g. the left leg of

the dancer). It results in visible errors in the virtual view. Lack of inter-view consistency significantly reduces both the objective and the subjective quality of synthesised views [29], [30].



The depth map for the view 0 | The depth map for the view 2

The reference view 1 | The virtual view in the position of the view 1

Fig. 2.2. An example of the result of no inter-view consistency of depth maps on a virtual view synthesis.

The estimation of inter-view consistent depth maps does not only increase the quality of virtual views but also helps to estimate depth in occluded parts of a scene. Even if some parts of a scene are occluded in neighbouring views, then information from further views can be used to estimate depth in these areas, what is not possible when depth maps for input views are estimated independently. In the case of independent estimation it is necessary to perform additional depth inpainting [39], [67], [81], or inter-view consistency refinement [49], [106], what increases the overall processing time of depth maps computation.

Another important aspect of depth maps, which influences the quality of synthesised views, is their temporal consistency. **The temporal consistency** of a depth map means that points of a scene have values of depth assigned consistently in consecutive frames, i.e. a point representing a still object has the same depth value in consecutive frames, or the depth value changes in accordance with the motion of an object [95], [106].

An example of two consecutive frames of depth maps that are not temporally consistent and the virtual views synthesised using these depth maps is presented in Fig. 2.3. It can be seen that the still background of the scene is estimated non-consistently in two consecutive

frames. The visible flickering of the virtual view significantly reduces the perceived quality of free navigation.



The depth map for the frame 0



The depth map for the frame 1



The frame 0 of the virtual view



The frame 1 of the virtual view

Fig. 2.3. An example of the result of no temporal consistency of depth maps on virtual view synthesis.

C. Positioning of cameras

Research on an optimal camera arrangement in multiview systems shows that the positioning of cameras has a significant influence on the quality of the 3D reconstruction of a scene [75], [78], [79]. The author of this dissertation also co-authored the analysis of the optimal camera placement in FTV systems [23], [94]. The results show that if there are many areas in a scene that are occluded by some objects, then it is recommended to arrange cameras around a scene as a set of camera pairs. For scenes with a high number of occlusions, such positioning of cameras increases the quality of synthesised virtual views. Therefore, a depth estimation method developed for free-viewpoint television systems has to be independent of the camera positioning and should estimate depth maps for any number of arbitrarily placed cameras.

When the positioning of cameras is arbitrary, it is necessary to perform the 3D projection of a point using the camera parameters. On the other hand, if optical axes of cameras are parallel, it is only required to search for corresponding points in one dimension (horizontal), because there is no vertical displacement of objects in neighbouring views in such case. The

process of 3D projection is much more complex than for 1D search, therefore, for arbitrarily positioned cameras the process of the correspondence search can be longer.

2.2 Assumptions and requirements for a depth estimation method for FTV

As it is presented in previous sections of this chapter, in order to avoid the virtual view synthesis errors that may worsen the viewing experience of virtual navigation in free-viewpoint television, a method of depth estimation has to meet a specific set of requirements that are the result of the structure of the FTV systems and how the virtual view synthesis process is performed. To conclude, **a method of depth estimation for the FTV systems should be characterised by:**

- the high quality of estimated depth maps, with particular emphasis on inter-view and temporal consistencies and a good representation of edges of objects,

- the possibility of the estimation of multiple depth maps for all views,

- high versatility, i.e. no assumptions about the number and the positioning of cameras can be stated and, moreover, a method can be used for different scenes without any modifications,

- short processing time comparing to the state-of-the-art methods of depth estimation that meet the abovementioned requirements (e.g. a method has to provide the possibility of parallelisation).

The presented requirements were utilised by the author of the dissertation as the basis of a novel method of depth estimation for free-viewpoint television systems (described in Chapter 4). These requirements are also used to assess the usefulness of available state-of-the-art depth estimation methods for FTV systems (presented in Chapter 3).

3 OVERVIEW OF STATE-OF-THE-ART

The topic of depth estimation is very widely discussed, especially in the scope of new immersive visual media. This chapter focuses on a description of depth estimation fundamentals required for the understanding of the dissertation achievements. Moreover, hitherto presented depth estimation methods are presented. Particular emphasis is put on the usability of these methods in free-viewpoint television.

3.1 Depth estimation fundamentals

In this section, the fundamentals of algorithmic depth estimation from a multiview video are presented. The described fundamentals are the basis of the depth map estimation methods described in Sections 3.2.1 – 3.2.4 and the method proposed by the author (Chapter 4).

3.1.1 Local estimation

The process of depth estimation can be performed independently for each point of an input view. In such case, estimation is performed on the basis of the local characteristics of the image, hence this type of depth estimation is called the **local estimation**.

In order to calculate the depth of a point, a point has to be simultaneously visible by at least two cameras. For each point, a search for the most similar corresponding point in another neighbouring view (one or many) is performed. The point in the neighbouring image with the highest similarity to the actually processed point is chosen as the most probable corresponding point and used for depth calculation. The most common name for this approach is "Winner-Takes-All" (WTA) because only the point that provided the highest similarity is the only one that is considered for depth estimation. The use of WTA is the simplest approach in depth local estimation. Because of the low complexity of such methods, very often their real-time applications are available [103].

The similarity of fragments of images is the basis of the correspondence search in depth estimation and is measured through the use of various similarity metrics. The most common

is a sum of absolute/squared difference (SAD/SSD) calculated in a small window. Nevertheless, other similarity metrics like gradient, rank, and census transforms were shown to outperform SAD and SSD [114]. For a wide survey and an analysis of similarity metrics see [48].

The similarity of possibly corresponding points is calculated on the basis of the colour difference between these points, or in small windows that include neighbouring points in order to decrease the influence of noise present in input views. The size and the shape of used windows can be controlled adaptively to the content of input views [22].

Unfortunately, choosing correspondent points only on the basis of the highest similarity, independently from the neighbourhood of points, may lead to the incorrect estimation of depth. Local estimation methods can estimate depth properly only on heavily textured areas of a scene. For one-coloured objects, that usually should have a smooth gradient of the depth (e.g. walls of buildings, floors), an estimated depth map can be highly noised when calculated using local estimation methods. A similarity of points that represent such areas can be high for many different points in neighbouring views. Moreover, the estimation of depth in occluded areas of a scene is impossible with the use of the correspondence search only.

The presented drawbacks of local estimation methods led to research on so-called global estimation methods, described in the following section.

3.1.2 Global estimation

A. Introduction

The depth of a point can be estimated not only on the basis of the correspondence search but also with the use of the depth calculated for the other points during the estimation process. This approach allows the estimation of a smooth depth on surfaces present in a scene. Such type of depth estimation is known as **global estimation**.

The problem of the depth map estimation in global estimation methods can be presented as a cost (goal) function minimisation [44], [45]. In its mostly used form, the cost function is defined as:

$$E(\bar{d}_p) = \sum_{p \in \mathrm{P}} D_p(\bar{d}_p) + \sum_{p \in \mathrm{P}} \sum_{q \in \mathrm{Q}} V_{p,q}(\bar{d}_p, \bar{d}_q), \tag{3.1}$$

where:

P – set of points of the input view,

p – point of the input view,

$\bar{d}_p$ – currently considered depth of a point p,

D_p – data term that represents a cost of assigning the depth d_p to the point p,

Q – set of points in the neighbourhood of the point p,

q – point in the neighbourhood of the point p

$\bar{d}_q$ – currently considered depth of a point q,

$V_{p,q}$ – smoothness term that represents the intra-view discontinuity cost of assignment of the depth d_p to the point p and depth d_q to the point q.

The data term D_p is responsible for the correspondence between points in neighbouring views. D_p is usually calculated as a sum of absolute differences (SAD) between colour values of a point p and a point in a neighbouring view that corresponds to the point p for the considered depth of a point ($\bar{d}_p$). Like in local estimation methods, presented earlier, the SAD metric can be replaced with other similarity metric and can be calculated in a small window that contains neighbouring points in order to minimise the influence of noise on performed depth estimation.

The smoothness term $V_{p,q}$ is introduced to the cost function in order to estimate depth on surfaces of a scene that lack a texture and details (e.g. one-coloured walls), where it is hard to determine the correspondence between points of the neighbouring views. The smoothness term usually utilises a linear discontinuity model that is based on the similarity of depths of neighbouring points:

$$V_{p,q}(\bar{d}_p, \bar{d}_q) = \beta_0 \cdot |\bar{d}_p - \bar{d}_q|, \tag{3.2}$$

where:

β_0 – fixed smoothing coefficient provided by the user in the beginning of estimation,

$\bar{d}_p$ – currently considered depth of a point p,

$\bar{d}_q$ – currently considered depth of a point q.

The presented formulation of depth estimation problem allows estimating a smooth depth on objects of a scene. The cost function (3.1) can be further expanded with another terms, for example, to ensure also inter-view [44] and temporal [56] consistencies.

B. Target function minimisation

A solution for the presented formulation of the cost function (3.1) can be estimated using an appropriate optimisation method. One of the most common of such methods is the graph cut method [6]. Below, a short description of this optimisation algorithm is presented.

In general, the graph cut algorithm is a method for the optimisation of the binary problems [6]. In image processing, this algorithm is used to assign one of two proposed labels to points of an image, e.g. in order to segment a foreground and a background of an image. When the graph cut algorithm is used for the image processing purposes, usually each point of an image is represented as a node in a graph. The cost function is represented as a set of edges between nodes of the graph. In order to find the solution of the labelling problem, the maximum flow problem is solved [5].

An example of a graph constructed for depth estimation is presented in Fig. 3.1. Nodes s and t represent two labels (two different values of depth) that have to be assigned each point of an image. If the presented graph would be used to solve the cost function (3.1), then edges between nodes s and t would represent the data term, while edges between nodes that represent points of an image would represent the smoothness term.

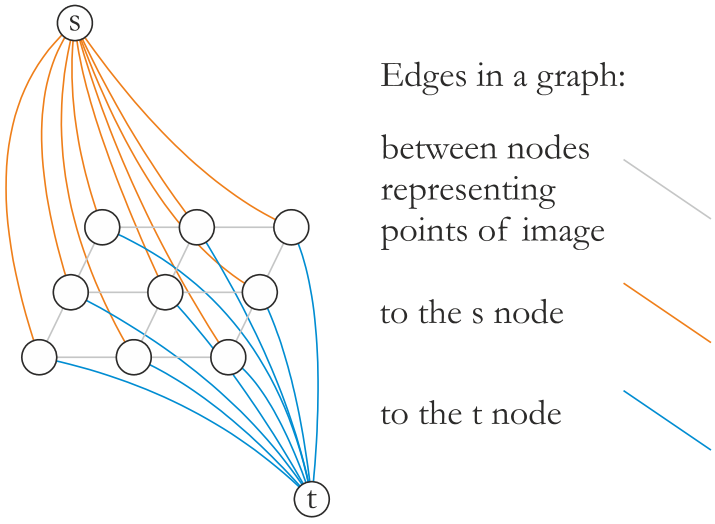


Fig. 3.1. An example of a graph used in the graph cut optimisation.

The graph cut algorithm finds the optimal cut of a graph that assigns nodes of a graph either to the s node or to the t node (Fig. 3.2). The optimal cut represents a minimum of a cost function that is solved by the graph cut algorithm.

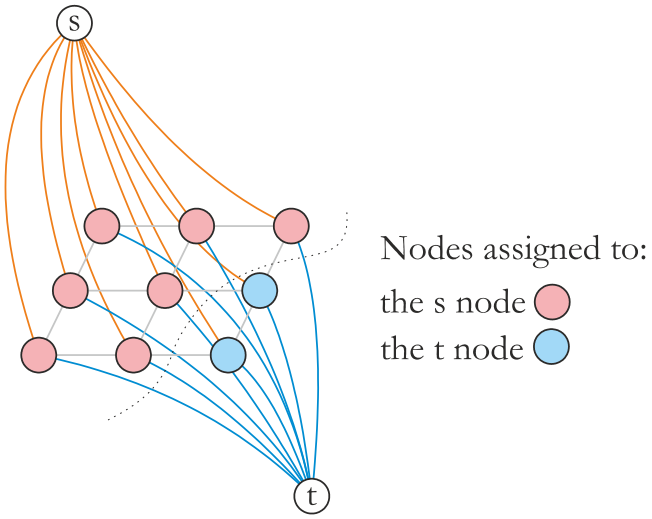


Fig. 3.2. An example of a cut of a graph after the graph cut optimisation.

Solutions of more general problems (i.e. the assignment of one of more than two labels to points of an image – multi-label segmentation) can be estimated by performing a series of graph cut optimisations. The problem of depth estimation expressed as in eq. (3.1) is a multi-label optimisation problem, in which each considered depth is represented by another label. Two most common methods of the minimisation for multi-label problems are the α-β swap and the α-expansion [6].

In the α-β swap, graph cut optimisation is performed for all pairs of n labels α and β, therefore, a large number of optimisations is performed (n^2) in order to achieve the final labelling of an image. Moreover, the α-β swap does not guarantee that the minimum of the optimised function will be found.

In the α-expansion method of the multi-label problem optimisation, in each iteration the graph cut algorithm assigns to each point a label α or leaves the current label (referred as a non-α label). For example, when points of an image are initially labelled with the label a, in the first iteration the graph cut algorithm assigns each point either to the label a, or to another label b. In the latter iteration, the graph cut algorithm assigns each point to the further label c, or leaves the previous labelling (i.e. a or b). Therefore, the α-expansion method requires to perform only n iterations of the graph cut algorithm. The α-expansion, in contrary to the α-β swap, guarantees that the found solution is within a known factor of the global minimum [6].

The α-expansion can be used only if between nodes that represent neighbouring points p and q there is specified a term $U_{p,q}(\alpha,\beta)$ that fulfils the following criteria:

$$U_{p,q}(\alpha,\beta) = 0 \Leftrightarrow \alpha = \beta \,, \tag{3.3}$$

$$U_{p,q}(\alpha,\beta) = U_{p,q}(\beta,\alpha) \geq 0 \,, \tag{3.4}$$

$$U_{p,q}(\alpha,\beta) \leq U_{p,q}(\alpha,\gamma) + U_{p,q}(\gamma,\beta) \,. \tag{3.5}$$

The abovementioned criteria are fulfilled by the smoothness term $V_{p,q}$ (3.2), therefore, it confirms that the α-expansion can be used for the minimisation of the presented cost function (3.1).

The graph cut method is used in the proposed method of depth estimation (Chapter 4), therefore, this algorithm was chosen as an example of the optimisation method. The detailed comprehensive description of the graph cut method, also in other applications than the estimation of depth maps, can be also found in [6], [44], [45]. Another example of the optimisation method, which can be used in depth estimation, is the belief propagation algorithm [98].

3.2 State-of-the-art methods of depth estimation

In this section, the current state-of-the-art methods of depth estimation are presented. Descriptions of the presented methods mainly refer to their usability in free-viewpoint television systems.

The methods are divided into groups that refer to the main achievements of the dissertation described in Chapter 4, i.e. inter-view consistent depth map estimation, the use of the image segmentation in the estimation process, the temporal consistency of depth maps, and the parallelisation of depth optimisation.

3.2.1 Inter-view consistency of depth maps

As considerations presented in Chapter 2 show, the inter-view consistency of depth maps is essential for the depth maps used in virtual view synthesis. Inter-view consistency can be achieved with depth refinement methods, e.g. by the filtering of depth maps projected from

all views to the centre view and another projection of such merged and filtered depth map back to the original positions of views [27], [49], [108].

Depth refinement can be also a part of virtual view synthesis. In [38], depth maps are projected to the centre view, as in other abovementioned methods, but the additional optimisation of a complex cost function calculated on the basis of input views, depth maps and motion vectors is performed. The resulting quality of virtual views is significantly improved, but the complexity of this method makes it impossible to perform real-time virtual view synthesis, desirable in FTV systems.

Depth estimation described in [121] consists of multiple post-processing steps that include the consistency check between neighbouring cameras, the merging of depth maps from neighbouring cameras, and bilateral median filtering of resulting depth maps. Despite complex post-processing, this method allows estimating depth maps in real-time. Unfortunately, the method is adjusted to be used with the multi-camera rig that contains 4 closely spaced cameras with parallel optical axes. It simplifies the depth estimation process because the search for the corresponding points is performed only in one dimension and objects of a scene are visible from the same angle by all cameras. Nevertheless, this method cannot be used in the versatile free-viewpoint television systems that state no assumptions about the positioning and number of cameras.

Other methods are often designed for multi-view systems of different characteristics than FTV systems, e.g. for lightfields [120], or multi camera arrays [52]. Yet another type of methods can be used only for sequences acquired using a moving camera rig, e.g. [88] and [116]. Both mentioned methods ensure temporal and inter-view consistencies: in [88] by the introduction of the new spatio-temporal primitive used in the inter-view matching, while in [116] by performing of optimisation that uses simultaneously all views and dozens of consecutive frames of the sequence.

On the contrary to the abovementioned methods, the formulation of depth estimation problem proposed by the author (presented in Section 4.2) enables depth map estimation for any positioning of cameras and provides a better inter-view consistency not by additional refinement but during the depth estimation process itself. Therefore, in the proposed method, ensuring inter-view consistency does not additionally increase the overall complexity of the depth estimation process.

3.2.2 Usage of segmentation in depth estimation

In order to shorten the processing time of estimation, depth optimisation can be based on segments of an image, instead of on individual points, like in [37]. In this method, the smoothness term is proportional to the length of the boundary between neighbouring segments, while the data term is based on the matching of segments in the neighbouring views. The experimental results show that the use of image segmentation helps to reduce the complexity of depth estimation and decreases errors of estimation that are the result of the poor representation of edges of objects in point-level estimation. Nevertheless, the matching of segments in neighbouring views is effective only when cameras are close to each other. In multiview video acquired with sparse FTV systems, the segmentation of the same object may be significantly different in neighbouring views because of large distances between cameras [8]. Moreover, the proposed cost function does not ensure inter-view consistency of estimated depth maps.

Another method that utilises image segmentation is PMSC (PatchMatch-Based Superpixel Cut [59]). This method uses the smoothing cost that is calculated between neighbouring points of an image and the data cost calculated both for points and segments. Depth estimation is repeated for different numbers of segments and resulting depth maps are merged into one. The method was shown to achieve very good results in terms of the quality of depth maps. Nevertheless, the processing time for high-resolution stereo-pair images is longer than 10 minutes, and because estimation is done for a depth map for one view only, inter-view consistency is not ensured. A similar method, which also is based on a set of depth estimations for different numbers of segments that are merged into one, can be found in [84].

In [86], image segmentation is used only in the correspondence search. The size of the matching window is large but is limited by the boundaries of segments. It merges the advantages of large matching windows (the limitation of the influence of noise) and small windows (the possibility of the correct depth estimation for small objects). The complexity of depth estimation can be also lowered by using segmentation to reduce the number of considered depth candidates. In [12], the number of possible depths for a segment of an image is reduced to only two candidates. Unfortunately, the assumptions made by the authors are not true when the optical axes of the cameras are not parallel.

The use of segmentation in the depth map estimation process is widespread. What distinguishes the depth estimation method proposed by the author in Chapter 4 from the aforementioned methods, is that depth optimisation in the proposed method is based only on

segments of an image. In the presented state-of-the-art methods, optimisation is also sometimes performed on segments (e.g. [59], [84]), but at some step of estimation, point-level optimisation is still required.

3.2.3 Temporal consistency of depth maps

The quality of depth maps estimated using local estimation methods is too low for virtual view synthesis because the inter-view and temporal consistencies of depth maps are usually not assured (the importance of the inter-view and temporal consistencies in FTV is described in Section 2.1). However, some local estimation methods try to utilise the temporal information, e.g. [47] presents a depth estimation method for stereo-pairs, in which the similarity calculated in the previous frames is a part of the matching cost for the subsequent frames. It decreases the influence of noise in the input views on the quality of resulting depth maps, what in some degree increases the temporal consistency of depth maps.

Estimated depth may be post-processed in order to increase its quality using methods of depth map refinement. Therefore, depth refinement lets to initially estimate depth maps using a simpler depth estimation method (e.g. one that does not ensure the temporal consistency of depth maps), what can result in an overall decrease of the complexity. Many methods of the depth quality enhancement utilise image segmentation [26], [81], [104], colour guided filtering [63] or joint temporal and spatial filtering [106]. These methods can be used to increase the resolution of estimated depth maps, reduce noise present in depth maps or improve the representation of edges of objects.

The temporal consistency of depth maps also can be achieved through the use of additional refinement [54], [112]. Such refinement methods are usually based on the segmentation of the background of a scene. Unfortunately, the temporal consistency of objects in the foreground is not increased.

The temporal consistency of depth maps can be also increased with denoising of input views used in depth estimation [91], [95]. The main advantage of such approach is that denoising can be performed independently from depth estimation, therefore, can be used with all depth estimation and refinement methods. On the other hand, an additional step of estimation increases the overall processing time, reducing the main advantage of local estimation methods, i.e. their low computational complexity.

On the contrary to the methods presented above, the new method of temporal consistency enhancement of depth maps, presented by the author in Section 4.3, simultaneously decreases the complexity of the depth estimation process.

3.2.4 Parallelisation of depth optimisation

The graph cut algorithm is characterised by high complexity, which is dependent on the number of nodes and edges used in a graph. There are solutions that use parallelisation in order to speed up the optimisation process. For example, [105] describes the implementation of the graph cut algorithm on the GPU together with the α-expansion. The parallelisation of operations performed on a graph allowed 5-8 times faster computation than for one core of the CPU. The JF-cut [77] provides the even larger reduction of the computation time (up to 40-fold), however, GPU implementations are limited by the relatively small size of GPU memory, making multi-view depth estimation for an FTV system very difficult.

Other implementations utilise distributed computing systems [113] with a very high number of the computing nodes (even few hundred). Nevertheless, despite using such computational power, the speed-up of optimisation in the depth estimation process is around tenfold. In [61], a graph is cut into smaller sub-graphs. The graph cut algorithm is performed on each of the sub-graphs and results of these cuts are merged in the additional step. The speed-up is dependent on the number of available cores in a CPU (using 4 cores decreases computational time 3.5 times), however, this method can be used only for graphs that have no edges between nodes that represent different views.

The proposed parallelisation method, presented in Section 4.4, does not focus on the speeding up of the graph cut process itself, but on the parallelisation of the α-expansion method. Therefore, no assumptions about the construction of a graph have to be stated, thus making proposed inter-view consistent multi-view depth estimation possible to be calculated with the use of parallelisation.

3.2.5 Other depth estimation methods

The methods presented in this section are not directly related to depth estimation in versatile free-viewpoint television systems but are presented to show other interesting directions of research on depth map estimation.

For example, in dense multi-view systems depth maps can be estimated using an epipolar plane image [41], [115] that is a three-dimensional structure in which views are placed sequentially one after another. The process of depth estimation is not based on the search of corresponding points in neighbouring views but on the search of lines that connect corresponding points in all images (i.e. epipolar lines). Estimated epipolar lines are used to calculate the depth of points. These methods enforce the depth to be consistent in all views and are characterised by lower complexity than global estimation methods, however, epipolar lines can be found effectively only in dense multi-view systems.

More recently, a new interesting type of depth estimation methods was introduced, which uses convolutional neural networks to support depth estimation on the basis of a previously prepared database of reference depth maps. The data-driven estimation, although can represent the direction of future research in depth estimation, is still limited to specific applications (e.g. for soccer stadiums footage [9]), stereo pairs [111], or multi-view systems with a very narrow base [28].

The acquisition of the depth of a scene can be also performed using depth sensors. Available solutions use an infra-red illumination of a defined pattern of points which is used to define the 3D position of all object in a scene (e.g. in Microsoft Kinect device [10]) or the calculation of the distance to objects by measuring the time of flight of an emitted infra-red light (in time-of-flight cameras [31]).

The estimation of depth using depth sensors is usually performed in real-time, what enables the use of these sensors in machine vision applications [3]. Moreover, depth sensors can work in very low-light conditions and can easily estimate depth on homogeneous surfaces because these sensors do not use the correspondence search based on the texture of objects, contrary to image-based depth estimation methods.

Unfortunately, the use of depth sensors is very limited in FTV systems. Using only one sensor can be insufficient for modelling of a whole scene [90]. Obviously, using more than one sensor is possible [42], nevertheless, using more devices can be problematic due to possible interferences between depth sensors that acquire the same scene [110]. Interferences from other infrared illumination sources, especially in outdoor scenes, together with a limited resolution of depth cameras [55], [76], lack of temporal consistency [119], and a high level of noise and holes in depth maps [7] further limit possible applications of depth sensors in free viewpoint television systems. Therefore, the main advantage of depth sensors, which is the real-time estimation, is diminished by the necessity of using depth maps refinement methods to meet all requirements for depth maps that can be used for FTV.

3.3 Conclusions

The presented state-of-the-art methods of depth estimation only partially meet the requirements, presented in Section 2.2, that concern their possible use in free-viewpoint television. Especially the simultaneous assurance of inter-view consistency and the temporal consistency of estimated depth maps, together with the relatively low complexity, can be seen as very demanding. The versatility of methods is often very low and only few methods allow the depth maps estimation for any positioning of cameras. It emphasises the need for new depth estimation method that can help in further development of free-viewpoint television systems.

The global estimation methods that utilise segmentation obtain the best results in terms of the quality of the estimated depth maps and decrease the complexity of the process. Nevertheless, none of the presented methods meets all the requirements to be suitable for the practical FTV systems, hence the proposal of a new depth estimation method presented by the author in the following chapter.

4 PROPOSED MULTIVIEW DEPTH ESTIMATION METHOD

In this chapter, the new method of depth estimation for free-viewpoint television systems, developed by the author of the dissertation, is presented.

Section 4.1 presents an overview of the proposed depth estimation method. The method consists of 3 main parts, described in the consecutive sections: inter-view consistent segment-based depth estimation (Section 4.2), temporal consistency enhancement that reduces the complexity of estimation (Section 4.3), and the method of parallelisation of depth estimation that decreases the processing time of estimation (Section 4.4). Details of depth estimation method implementation created by the author of the dissertation are presented in Section 4.5.

The experimental verification of the performance of the respective proposals is presented in Chapter 6.

4.1 Overview of the proposed method of depth estimation

A new method of depth estimation that can be used for view synthesis should ensure the inter-view and temporal consistencies of depth maps and reduce the complexity of the depth estimation process (as it was concluded in the discussion in Section 2.2). The novelty of the proposed method of depth estimation, and its particular usefulness for free-viewpoint television systems, is a result of the joint application of the following ideas:

1) **Depth estimation is performed for segments instead for individual points of input views**, thus the size of segments can be used for controlling the trade-off between the quality of depth maps and the processing time of estimation, without reducing the resolution of estimated depth maps.

2) The utilisation of the new formulation of the cost function, dedicated for the segment-based estimation of depth maps – estimation can be performed for all views simultaneously and **produces depth maps that are inter-view consistent with no assumptions about the positioning of views**: any number of arbitrarily positioned cameras can be used in the estimation process.

3) Estimated depth maps are calculated on a per-pixel basis, although the segmentation of input views is used, because the correspondence search is not limited to segment centres. Therefore, **segmentation does not have to be consistent in all views** and is performed independently for each view, what leads to the reduction of the complexity of depth estimation.

4) The proposed temporal consistency enhancement method utilises depth maps estimated in previous frames in the estimation of depth for a current frame. Proposed enhancement **increases the temporal consistency of depth maps and, simultaneously, decreases the processing time of estimation.**

5) The proposed depth estimation method uses **the novel parallelisation** of the α-expansion method for graph-based depth estimation that significantly reduces the processing time of depth estimation.

The main idea of the proposed algorithm of depth estimation (i.e. inter-view consistent segment-based depth estimation, described in Section 4.2) was already described by the author in [22], [68], [70] and [71].

The simplified method of proposed depth maps temporal consistency enhancement, which in this dissertation was extended with the introduction of "I type" and "B type" depth frames and with excluding of unchanged segments from the estimation process (described in details in Section 4.3), was described by the author in [69] and [72].

4.2 Proposed cost function

The estimation of depth in the proposed method is based on a cost function minimisation, therefore, the proposed method is a global estimation method (the description and the characteristics of global estimation methods are described in Section 3.1.2). The cost function (3.1) described in Section 3.1.2 is not suitable to be used with segments, thus a novel cost function is proposed. The cost function used in the proposed method is based on two segment-based costs: the intra-view discontinuity cost $V_{s,t}$ (the smoothing term) and the inter-view matching cost $M_{s,s'}$ responsible for the inter-view consistency of depth maps, used instead of the point-based data term D_p described in Section 3.1.2:

$$E(d_s) = \sum_{c \in \mathrm{C}} \left\{ \sum_{c' \in \mathrm{D}} \sum_{s \in \mathrm{S}} M_{s,s'}(d_s) + \sum_{s \in \mathrm{S}} \sum_{t \in \mathrm{T}} V_{s,t}(d_s, d_t) \right\}, \tag{4.1}$$

where:

C – set of views,

c – view used in estimation,

D – set of views neighbouring to the view c,

c' – view neighbouring to the view c,

S – set of segments of the view c,

s – segment in the view c,

d_s – currently considered depth of the segment s,

s' – segment in the view c', which corresponds to the segment s in the view c for the currently considered depth d_s ,

$M_{s,s'}$ – inter-view matching cost between segments s and s',

T – set of segments neighbouring (adjacent) to the segment s,

t – segment neighbouring to the segment s,

$V_{s,t}$ – intra-view discontinuity cost between segments s and t,

d_t – currently considered depth of the segment t.

The neighbouring views c' are two views of a scene: the nearest left view and the nearest right view of the view c. If the view c is the leftmost or the rightmost view, then the number of neighbouring views has to be limited to one: the nearest left or the nearest right neighbour of the view c, depending which one is available.

The limitation of c' to only two views does not affect the inter-view consistency of estimated depth maps. The segment s in the view c is not directly connected to the corresponding segments in all views, however, segments in two neighbouring views are also connected with further views. It means that the depth of the segment s affects estimation in all views. Moreover, reducing the number of connections between segments decreases the complexity of depth estimation.

In the proposed method depth maps can be estimated for arbitrarily placed cameras, therefore, optical axes of cameras cannot be assumed to be parallel. When an arrangement of cameras is linear (camera optical axes are parallel), depth can be defined as the distance from the plane of a camera, as in Fig. 4.1a. For the sake of comprehension: the plane of a camera

is a plane that contains the sensor of a camera, and depth levels are planes that are parallel to the plane of a camera and represent depth values possible to estimate. If cameras are placed non-linearly, then the use of such definition results in a different depth of the same 3D point *P* in each camera (Fig. 4.1b).

In the proposed method, the depth of a point is defined as a distance from the plane of the central camera of the system, as it is presented in Fig. 4.1c. This definition of depth uses so-called global depth levels [22]. Use of the global depth levels allows ensuring the inter-view consistency (defined in Section 2.1B) of the depth of the point *P* in all views.

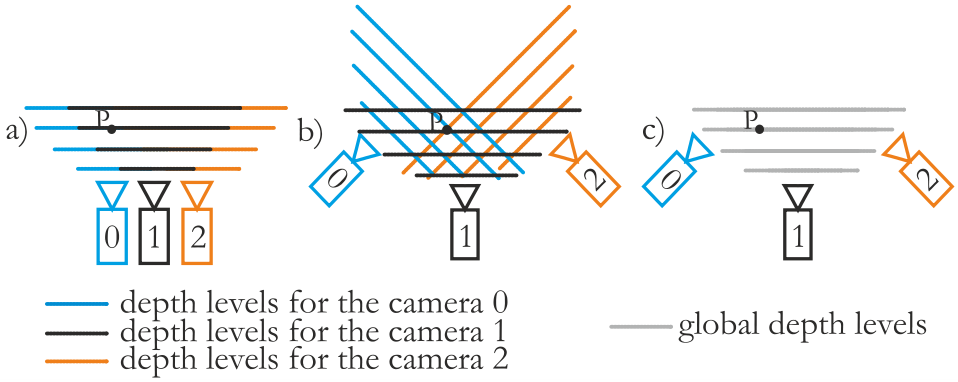


Fig. 4.1. The depth of a point *P* in a scene for: a) a linear arrangement of cameras, b) a non-linear arrangement, c) a non-linear arrangement with global depth levels.

A local minimum of the proposed cost function (4.1) can be estimated using the graph cut algorithm [45], described in Section 3.1.2. The proposed method is based on multi-label problem optimisation, therefore, utilises the α-expansion method [6], also described in Section 3.1.2.

Unlike in typical formulations of graph-based depth estimation, in which each node in a constructed graph represents one point of an input view, in the proposed method each node of a graph corresponds to one segment (Fig. 4.2). Nodes are connected with each other by two types of links that represent the intra-view discontinuity and inter-view matching costs.

Proposed segment-based estimation reduces the number of nodes in a graph in comparison with point-based estimation. The complexity of optimisation is affected by the number of nodes and links present in a graph. Therefore, the use of segments makes the process of optimisation significantly faster. Moreover, depth maps are still estimated in the same resolution as the resolution of the input views, and because of the use of segments, edges of objects in depth maps correspond to the edges in input views. The number of segments, and thus

their size, becomes one of the estimation parameters and can be adjusted. Therefore, the number of segments can be used for controlling the trade-off between the quality of estimated depth maps and the processing time of the depth estimation process.

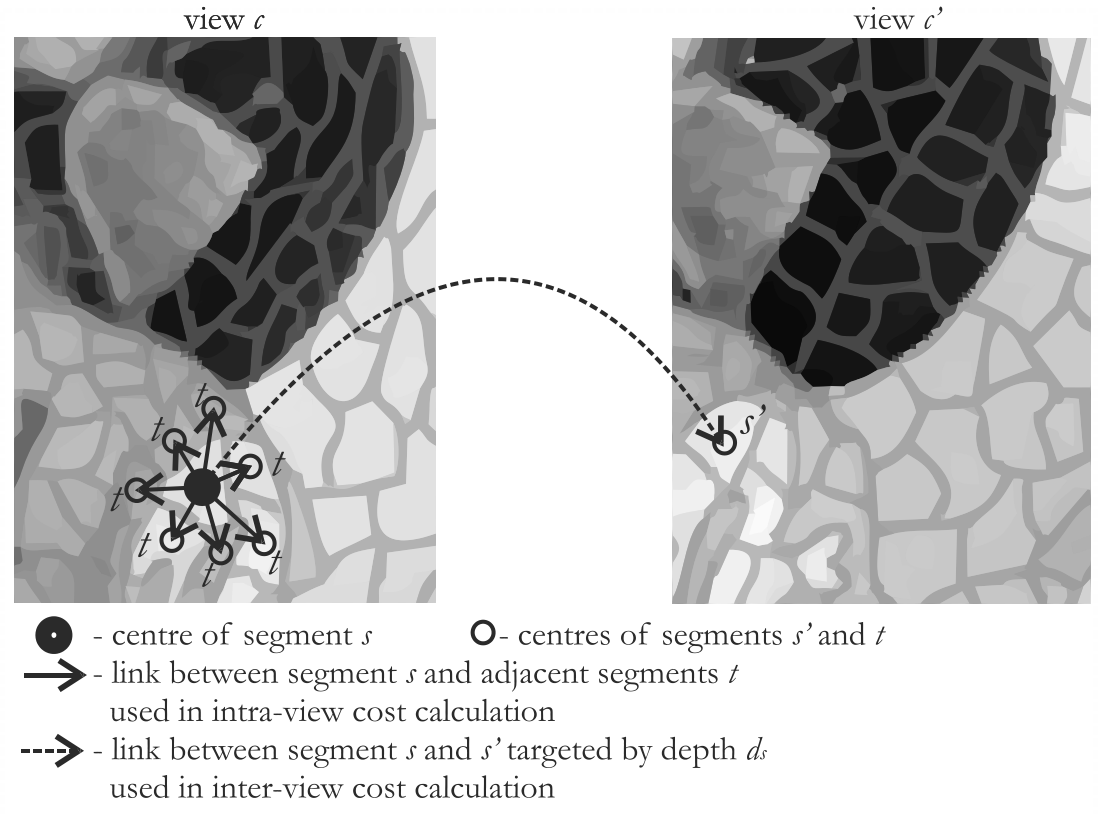


Fig. 4.2. The intra-view discontinuity cost and the inter-view matching cost for a segment *s* for depth estimation performed for 2 views.

Even when segments of a relatively small size are used during estimation (i.e. of the size of 20 points or less), it allows estimating depth maps of high quality significantly faster than when a conventional pixel-based estimation is used. On the other hand, the use of larger segments ensures additional significant reduction of the processing time of estimation, at the expense of a minor loss of quality. The influence of the number of segments on the performance of the presented depth estimation method was tested in the performed experiments described in Section 6.2.

4.2.1 Inter-view matching cost

In order to achieve inter-view consistency of estimated depth maps, the matching cost is not calculated independently for each single view. Instead, the conventional matching cost is replaced with the inter-view matching cost $M_{s,s'}(d_s)$ that is defined between a pair of segments s and s' corresponding to each other for currently considered depth d_s.

The proper matching of whole segments from different views is a difficult operation. Moreover, in the proposed method no assumptions about the positioning of views are made. Therefore, the segmentation of the same object in neighbouring views may significantly vary. It usually results in different shapes and sizes of segments that represent the same parts of a scene. The differences are especially visible if optical axes of cameras are not parallel because objects are visible from different angles in neighbouring cameras. In order to perform inter-view consistent segmentation (i.e. segmentation where objects are segmented in the same way in all views) it is required to use correct depth information as the input of segmentation, which is obviously not available at the beginning of depth estimation.

In order to avoid the abovementioned difficulties, the inter-view matching cost is proposed to be calculated in the pixel-domain in a user-defined window around the centre of a segment and the corresponding point in a neighbouring view. The core of the inter-view matching cost (used later to calculate a final value of the cost), denoted as $m_{s,s'}(d_s)$, is:

$$m_{s,s'}(d_s) = \frac{1}{size(W)} \sum_{w \in \mathsf{W}} \left\| [Y C_b C_r]_{\mu_s + w} - [Y C_b C_r]_{T[\mu_s]+w} \right\|_1 , \tag{4.2}$$

where:

W – set of points in the window of the size specified by the user,

w – point in the window W,

$\|\cdot\|_1$ – L1 distance,

μ_s – centre of a segment s,

$T[\cdot]$ – 3D transform obtained from intrinsic and extrinsic parameters of cameras,

$[Y\ C_b\ C_r]_{\mu_s+w}$ – vector of Y, C_b, C_r colour components of the centre μ_s of the segment s,

$[Y\ C_b\ C_r]_{T[\mu_s]+w}$ – vector of Y, C_b, C_r colour components of the point in a view c' corresponding to the centre μ_s of the segment s in a view c.

In order to achieve the inter-view consistency of depth maps, the value of the inter-view matching cost $M_{s,s'}(d_s)$ has to be calculated as [44]:

$$M_{s,s'}(d_s) = \begin{cases} \min\{0, m_{s,s'}(d_s) - K\} & if\ d_s = d_{s'} \\ 0 & if\ d_s \neq d_{s'} \end{cases}, \tag{4.3}$$

where:

s – segment in the view c,

d_s – currently considered depth of the segment s,

s' – segment in the view c', which corresponds to the segment s in the view c for the currently considered depth d_s ,

$d_{s'}$ – currently considered depth of the segment s',

$M_{s,s'}$ – inter-view matching cost between segments s and s',

$m_{s,s'}$ – core of the inter-view matching cost between segments s and s' [see eq. (4.2)],

K – a positive constant [44] (see Section 4.5 for an discussion on the proper value of K).

The presented definition of the inter-view matching cost does not require segmentation to be inter-view consistent in neighbouring views, therefore, segmentation can be performed independently for each view, reducing the overall complexity of presented depth estimation.

In the proposed method, the centre of a segment can correspond in a neighbouring view to any point, not necessarily to the centre of another segment. Therefore, the presented pixel-domain matching can be used to estimate depth with a high precision, simultaneously reducing the processing time of estimation, as the matching is not performed for all points (the number of matching operations is dependent on the number of segments, not on the number of points in the view).

4.2.2 Intra-view discontinuity cost

In the presented depth estimation method, as mentioned before, estimation is based on segments instead on points of input views. Therefore, the intra-view discontinuity cost cannot be calculated between the neighbouring points within some view (as it was presented in Section 3.1) but is calculated between all adjacent segments within the same view. The cost is calculated as follows:

$$V_{s,t}(d_s, d_t) = \beta \cdot |d_s - d_t| \,, \tag{4.4}$$

where:

β – smoothing coefficient,

d_s – currently considered depth of the segment s,

s – segment in the view c,

t – segment neighbouring to the segment s,

$V_{s,t}$ – intra-view discontinuity cost between segments s and t,

d_t – currently considered depth of the segment t.

Not only the definition of the cost was changed in comparison to the usual definition of the smoothing coefficient but in the proposed method the smoothing coefficient β is not fixed for all segments. Instead, the smoothing coefficient is calculated adaptively using a similarity of two neighbouring segments s and t and β_0 that is an initial smoothing coefficient provided by the user:

$$\beta = \beta_0 / \left\| [\hat{Y}\ \hat{C}_b\ \hat{C}_r]_s - [\hat{Y}\ \hat{C}_b\ \hat{C}_r]_t \right\|_1 \,, \tag{4.5}$$

where:

β – smoothing coefficient,

β_0 – initial smoothing coefficient provided by the user,

$\|\cdot\|_1$ – L1 distance,

s – segment in the view c,

t – segment neighbouring to the segment s,

$[\hat{Y}\ \hat{C}_b\ \hat{C}_r]_s$ – vector of average Y, C_b, C_r colour components of the segment s,

$[\hat{Y}\ \hat{C}_b\ \hat{C}_r]_t$ – vector of average Y, C_b, C_r colour components of the segment t.

When the smoothing coefficient is fixed during depth estimation, it is difficult to determine its best value. It results in depth maps that have edges of objects not properly represented in depth maps (because the smoothing is too high) or, on the other hand, depth maps that are noisy on some surfaces (when the smoothing coefficient is too low).

The proposed definition of the smoothing coefficient that is calculated adaptively to the content of a sequence ensures that when the similarity of adjacent segments is small (e.g. on

edges of an object), the smoothing coefficient also becomes small, thus depths of these segments are not penalised for being discontinuous. Simultaneously, for areas of a scene that have a similar colour (e.g. walls) the smoothing coefficient is high.

As it was presented by the author of this dissertation in [71], the adaptive smoothing coefficient simultaneously increases the quality of estimated depth maps and reduces the processing time of the depth estimation process.

4.3 Proposed temporal consistency enhancement method

In practical free-viewpoint television systems, positions of cameras do not change during the acquisition of a video sequence. Therefore, in such sequence, only a small part of a scene considerably changes in consecutive frames because a background is mostly still through the whole sequence. As it can be seen in Fig. 4.3, for two frames from one of the test sequences, the area that was considerably changed in two consecutive frames is small when compared to the whole scene, even if there is a fast-moving object in the scene (the attacking fencer).

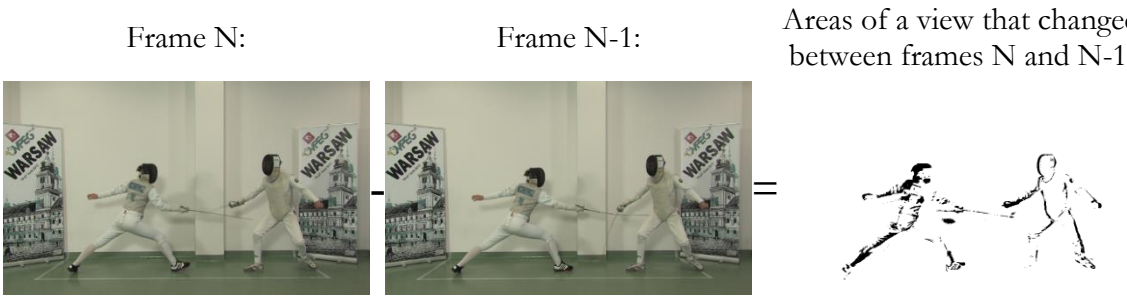


Fig. 4.3. Areas of a view that changed between the two consecutive frames of the "Poznań Fencing2" sequence.

In order to analyse this phenomenon quantitatively, the amount of movement was measured experimentally. The experiment was performed on a set of test sequences used in this dissertation (see Section 5.3). Each frame of sequences was compared with the previous frame and the first frame of sequences. Percentages of points that were changed more by ε for each of $Y\ C_b\ C_r$ colour components were calculated. In the experiment $\varepsilon = 3$ because such small change of a colour of a point (less than 3 for each of $Y\ C_b\ C_r$ components) indicates that with a high probability that point represents the same object, therefore the depth of this point should stay the same in the consecutive frames.

The results for all test sequences are presented in Table 4.1. As it can be seen, on average less than 30% of points change noticeably their colour in comparison with the previous frame. The similarity of points with the first frame of the sequence is also high – on average only 34% of points are noticeably changed in comparison with the first frame of a sequence. For computer-generated sequences (BBB Butterfly and BBB Rabbit) the percentage of points that changed their colour is significantly lower because there is no influence of noise of real camera sensors.

TABLE 4.1. THE PERCENTAGE OF POINTS THAT CHANGED THEIR COLOUR IN COMPARISON TO THE PREVIOUS OR THE FIRST FRAME OF THE SEQUENCE

Test sequence	The percentage of points that changed their value of Y, C_b and C_r components more by 3 with a comparison to:	
	previous frame	first frame
Ballet	43%	42%
Breakdancers	44%	42%
BBB Butterfly	10%	16%
BBB Rabbit	4%	17%
Poznań Blocks	22%	33%
Poznań Blocks2	35%	34%
Poznań Fencing2	36%	45%
Poznań Service2	35%	40%
***Average*:**	**29%**	**34%**

The idea of proposed temporal consistency enhancement is to calculate a new value of depth only for segments that represent fragments of a scene that considerably changed (in terms of their colour) in comparison with the previous or the first frame of the acquired sequence.

The concept of the use of depth maps calculated for previous frames is not novel (see Section 3.2.3) although in other methods is applied in the pixel-based estimation. In the proposed method, depth estimation strictly based on segmentation which is not consistent in subsequent frames. It increases the difficulty of choosing of the proper candidate of depth in previous frames.

The proposed temporal consistency enhancement method allows marking segments as unchanged. Such segments are still used in the calculation of the intra-view discontinuity and

the inter-view matching costs for other segments but are not represented by any node in the structure of an optimised graph. It reduces the number of nodes in a graph, making the optimisation process significantly faster and, moreover, increases the temporal consistency of estimated depth maps.

In the first frame of a depth map, denoted as an "I type" depth frame (through the analogy to the video compression terminologies, in which I frame is compressed without the use of the temporal information), estimation is performed for all segments, as described in previous sections. The following frames (denoted as "P type" depth frames) can utilise depth information from the preceding P type depth frame and the I type depth frame.

The segment p can be marked as unchanged in two cases: if all components of the vector $[\hat{Y}\ \hat{C}_b\ \hat{C}_r]_s$ of average Y, C_b and C_r colour components of segment s changed less than the set threshold T_b in comparison with the segment s_B, which is the collocated segment in the previous P type depth frame, or, if all components of the abovementioned vector $[\hat{Y}\ \hat{C}_b\ \hat{C}_r]_s$ changed less than the threshold T_I in comparison with the segment s_I – the collocated segment in the previous I type depth frame. If any of these two conditions are met, then the segment s adopts depth from the segment s_B or s_I (depending which condition was fulfilled). Therefore, the estimation of depth is performed only for segments that do not meet these conditions.

Thresholds used in the presented enhancement were estimated during the preliminary tests of the proposed method and set as $T_P = 3$ and $T_I = 1$. An example of segments that were marked as unchanged in comparison to the previous I type depth frame and in comparison to the previous P type depth frame is presented in Fig. 4.4.

Introduction of two reference depth frames has a beneficial impact on the visual quality of virtual navigation in free-viewpoint television. Firstly, the adoption of depth from a P type depth frame allows to use earlier estimated depth of objects that changed their position over time (e.g. for objects that were not present in a scene from the beginning of the acquisition, but after their appearing, they remained still). On the other hand, the adoption of depth from the I type depth frame minimises the flickering of depth in the background.

As the results presented in Table 4.1 show, depth estimation may be performed just for one third of points (on the average), because only such part of a view usually represents moving objects. In the presented temporal consistency enhancement the average colour of the whole segment is compared, therefore the influence of a noise is decreased, what further decreases an area of a view where depth has to be estimated.

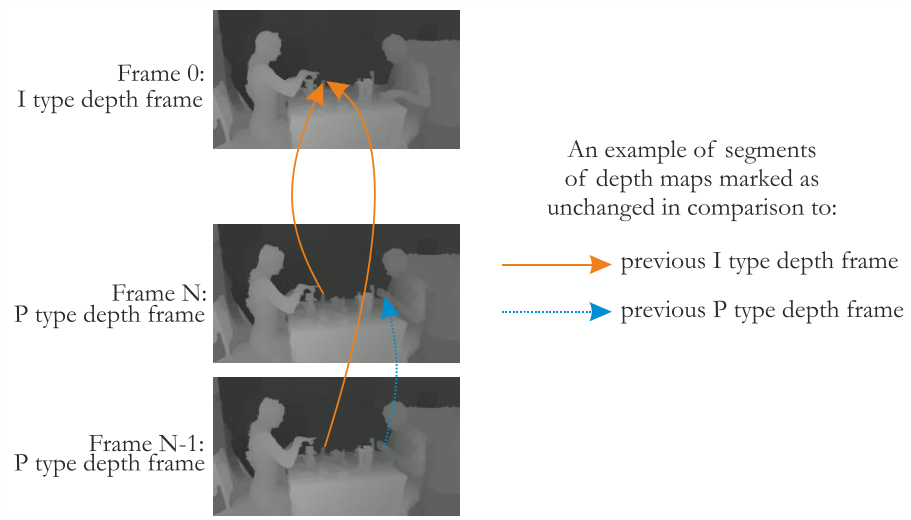


Fig. 4.4. An example of segments marked as unchanged in comparison to I and P type depth frames.

4.4 Proposed method of estimation parallelisation

In order to decrease the overall processing time of depth estimation in the proposed method, depth estimation is performed in parallel. Differently from the methods presented in Section 3.2, in which the parallelisation of the graph cut process is proposed, in the proposed method the whole depth estimation process is parallelised.

First of all, because of the formulation of the inter-view matching cost presented in Section 4.2.1, segmentation can be performed independently for each view. Therefore, the process of segmentation in the proposed method is performed by different threads in order to decrease the overall processing time of depth estimation. In this dissertation, a thread is understood as an independent process that shares no memory with other processes. Different threads can utilise different cores of the computer.

In the proposed method of parallelisation, each of n threads estimates a depth map with the n-times smaller number of depth levels (depth levels are planes that are parallel to the plane of a camera and represent depth values possible to estimate – Section 4.2).

In other depth estimation methods that use the graph cut optimisation (e.g. in [91]), optimisation is performed sequentially, using the α-expansion (described in Section 3.1.2). Depth estimation is started with all points (or superpixels, as in the proposed method) assigned to, e.g. the farthest depth level (denoted as the depth level number 0). In the first

iteration, the graph cut algorithm performs the assignment of points to the closer depth level number 1. Each next iteration of the graph cut optimisation assigns one of closer depth levels. Therefore, one thread performs as many graph cut optimisations as there are levels of depth (Fig. 4.5).

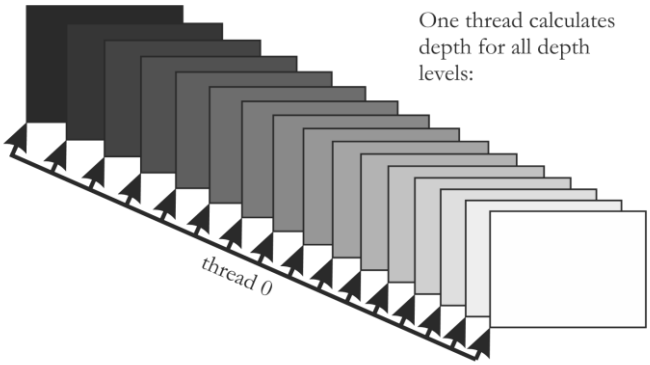


Fig. 4.5. All depth levels are calculated by one thread.
Each rectangle represents a different level of the depth of a scene.

In the proposed method, several simultaneous depth estimation processes are performed. Each depth estimation process performs graph cut optimisations for different sets of depth levels. Therefore, if the computation of depth maps is performed on a CPU that has more than one core, the overall processing time of the depth map estimation can be decreased.

Obviously, even without the use of parallelisation, all cores of a CPU also can be used for depth estimation, e.g. each core can perform the estimation of the depth for different sets of input views (e.g. for each set of 5 cameras of the system), or for different frames of the sequence. Unfortunately, when many standalone depth estimation processes are performed, it results in the loss of the inter-view consistency or the temporal consistency of estimated depth maps. **When the proposed novel parallelisation method is used, both the inter-view and the temporal consistency of depth maps, that are fundamental for the quality of virtual view synthesis, are preserved.**

Depth estimation performed in parallel for different levels of depth can be seen as a straightforward solution to the high complexity of estimation. Nevertheless, the α-expansion method provides the best results if depth estimation is performed sequentially for consecutive levels of depth. The change of the order of the depth levels influences the quality of resulting depth maps. In order to test how the depth levels should be distributed onto threads to minimise the loss of the estimated depth maps quality, the author tests two different ways

of depth levels distribution: depth levels divided into blocks (Fig. 4.6) or interleaved over threads (Fig. 4.7).

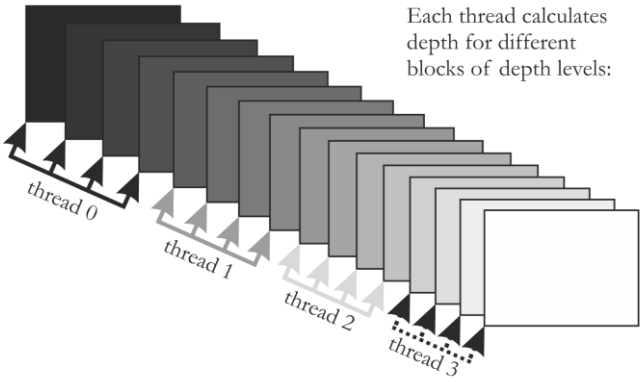


Fig. 4.6. Depth levels are distributed over 4 threads as blocks of depth levels.
Each rectangle represents a different level of the depth of a scene.

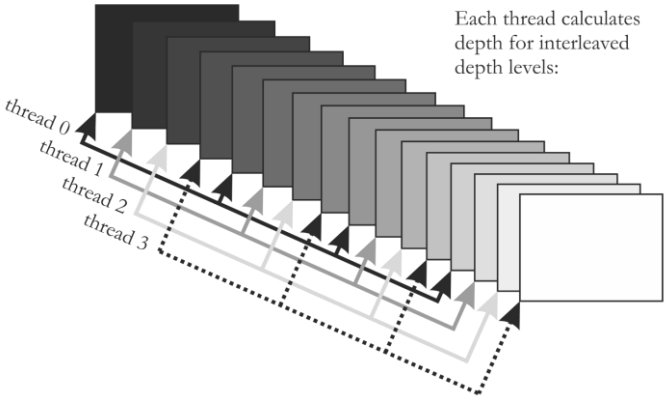


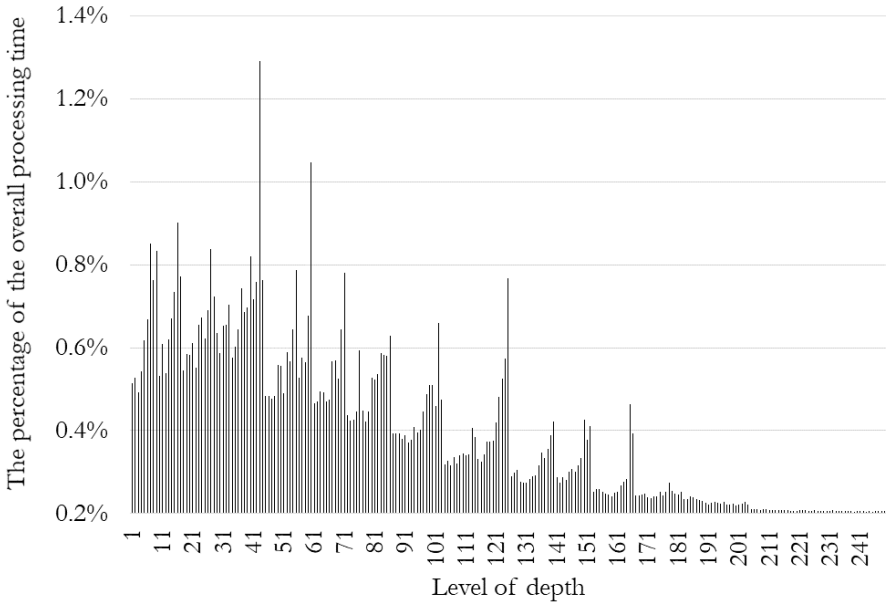
Fig. 4.7. Depth levels are interleaved over 4 threads.
Each rectangle represents a different level of the depth of a scene.

The distribution of depth levels has an influence on the time and the quality of the estimated depth maps. If objects of an acquired scene are placed more densely in some range of depths, then estimation for these depth levels is longer.

Fig. 4.8 presents the relative processing time of depth estimation for each level of depth for the first frame of the "Poznań Blocks2" sequence together with a number of segments that changed their labelling to the currently considered level of depth. Estimation was performed for 250 levels of depth, therefore, each level of depth on average should be calculated in 1/250 (0.4%) of the overall processing time of estimation. Results of the experiment (presented in Fig. 4.8) show that the processing time for different levels of depth varies. For some

levels of depth, estimation was significantly longer than for other levels. The longer processing time of estimation (Fig. 4.8a) corresponds to the increased number of segments that changed that changed their level of depth for the currently considered (Fig. 4.8b). In this sequence most of the objects are in the back of the scene and, except the floor, there are no objects close to cameras. Fig. 4.8a. shows that the processing time needed for the estimation of depth levels that are close to cameras (of high numbers) was significantly shorter.

a)



b)

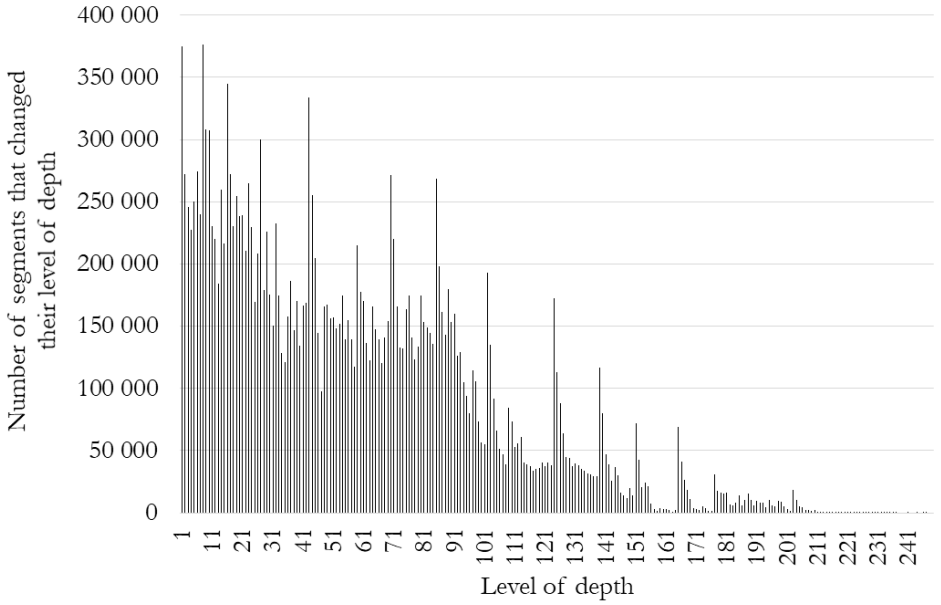


Fig. 4.8. A comparison of: a) a number of segments that changed their level of depth during optimisation performed for a particular level of depth and
b) the processing time of the estimation for a particular level of depth.

Therefore, if depth levels are divided into blocks, the estimation for some threads can be longer, increasing the overall processing time of depth estimation. On the other hand, if depth levels are interleaved, then the processing time of estimation for all threads is nearly equal, but estimated depths tend to be less smooth.

The dependency between the type of the parallelisation and the performance of the depth estimation method was tested experimentally. The results of this experiment are presented in Section 6.5.

Depth maps with the reduced number of depth levels that were calculated by different threads have to be merged into one depth map. The merging process is performed in a similar way as depth estimation, using the optimisation of the cost function (4.1), but only two levels of depth are considered for each segment in one cycle of merging – the depth of a segment from the thread t or the depth from the thread $t + 1$ (Fig. 4.9).

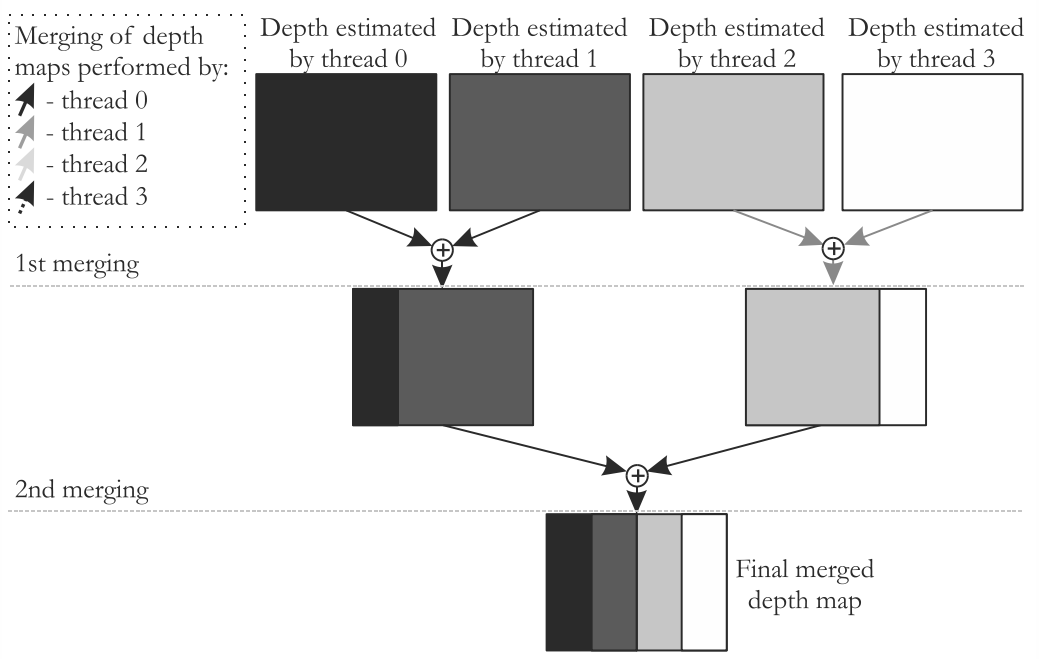


Fig. 4.9. A process of merging of depth maps for the case of the 4-thread parallelisation.

Only two depth maps can be merged into one by one thread during a cycle of merging because the minimisation of the function is performed using the graph cut algorithm (which can be used to optimise binary problems only). Therefore, for n threads $\lceil \log_2(n) \rceil$ of additional cycles are needed to estimate the final depth map that contains all depth levels. E.g. if four threads are used during depth estimation, then two additional merging cycles are sufficient to estimate the final depth map. Usually, the number of depth levels is much larger than

100, therefore, the process of merging increases the processing time of computations in a negligible degree.

4.5 Details of the implementation of the proposed depth estimation method

The proposed method was implemented by the author in the C++ language and includes all the necessary functions required to estimate depth maps for an input sequence. In this section are presented the details of the implementation which concern the use of solutions presented by other researchers and the selection of the constants used in the proposed method.

Only two solutions, i.e. the method of the minimisation of the goal function and the segmentation of input views, were not written by the author of this dissertation. Use of other methods of optimisation and segmentation than described below is possible, however, the characteristics of used methods meet requirements that concern their usability in the presented depth estimation method, therefore, tests of other optimisation and segmentation methods were left beyond the scope of this dissertation.

A. Optimisation method

The presented depth estimation method uses the maxflow-v3.01 library [126], which is an implementation of the algorithm of the function minimisation using the graph cut method described in [5].

As it was demonstrated in [101], the improvement of a problem formulation has a significantly larger influence on depth estimation performance than the selection of the optimisation method. Additionally, the graph cut method, in comparison with the belief propagation, the competitive method of the function minimisation, handles better with penalties between nodes of the graph [101]. Therefore, in the proposed method of depth estimation, where graph construction is based only on dependencies between segments, using the graph cut method is advisable and favourable.

B. Segmentation method

The proposed method of depth estimation uses the superpixel segmentation SNIC (Simple Non-Iterative Clustering [1]). The SNIC method meets the requirements stated by the proposed method of depth estimation. First of all, segments calculated using SNIC represent small, meaningful regions of an image, not whole objects. If each segment would represent a different object of a scene, then an entire object would have the same depth. Moreover, the number of segments can be freely changed. It allows to control the trade-off between the quality of depth maps and the processing time of estimation (Section 4.2).

The SNIC method is characterised by low complexity (what reduces the overall processing time of depth estimation) and achieves one of the lowest segmentation errors when compared to state-of-the-art methods [1], what positively influences the representation of edges of objects in depth maps. For other state-of-the-art superpixel segmentation methods see [2], [13], [60], [62], [85], [107], [118].

In the proposed method of depth estimation, the implementation of the SNIC method provided by the authors of [130] is used. In the comparison with the original algorithm, a one change was made by the author of the presented dissertation – instead in the *CIELAB* space, in order to avoid the recalculation of a colour space, the segments are calculated using the *YUV* colour space.

The SNIC method requires to choose two parameters of segmentation: in the proposed framework the compactness factor, which determines how compacted the superpixels are [130], is set to $m = 5$ and 8-connected segments are used.

C. Selection of the constant *K*

Here, a procedure of the selection of the constant K value used in the of the inter-view matching cost $M_{s,s'}$ (4.3) is presented.

The value of constant K should be chosen with accordance to noise present in an input sequence and to differences in colour characteristics of cameras used. Simultaneously, K has to be selected so that the inter-view matching cost $M_{s,s'}$ is not dominated by the intra-view discontinuity cost $V_{s,t}$, as a sum of these two costs constitutes the cost function (4.1) of the depth optimisation.

The inter-view matching cost is calculated as a sum of differences between Y, C_b, and C_r colour components of the centre of a segment and the corresponding point in another view (4.2). For the preliminary experiments, the proposed method was tested for different values

of K, from $K = 1$ to 765, as this range is limited by the maximum possible difference between two points in Y C_b C_r colour space ($255 \cdot 3$). The chosen final value of K is 30, as it provides the high quality of estimated depth maps for all tested sequences.

5 METHODOLOGY OF EXPERIMENTS

In this chapter, the methods of the experimental verification of the proposed depth estimation method are presented. The used methods of the assessment of the quality and the temporal consistency of depth maps focus on the usability of depth maps in free-viewpoint television.

5.1 Assessment of the quality of depth maps

The problem of the assessment of depth maps quality is widely known to the research community. Nevertheless, available quality metrics often focus on the measurement of the distortion of depth maps caused by encoding [35], [36], [43], [87] or transmission using error-prone networks [65].

Usually, the quality of depth maps is understood as the accuracy of depth maps, i.e. the quality is based on the measurement of the absolute difference between a ground-truth depth map and an estimated depth map. Unfortunately, lack of ground truth depth maps for natural test sequences, especially ones that represent dynamic scenes, leads to difficulties in determining the quality of the estimated depth maps that is understood as their accuracy.

The available databases with ground-truth depth maps do not correspond to the characteristics of free-viewpoint television. The newest Middlebury database [80] is widely used by the research community and allows to easily evaluate the performance of a depth estimation method and compare it with other methods. Unfortunately, the comparison of depth estimation methods in this database is performed for a set of rectified stereo-pair images acquired using two cameras with parallel optical axes, while in free-viewpoint television systems any number of arbitrarily positioned cameras can be used. Moreover, the dataset includes only one frame for each scene, therefore, the temporal consistency of depth maps, which is a significant part of research presented in this dissertation, also cannot be measured using this database.

Other databases of ground-truth depth maps (e.g. one of the newest databases – the ETH3D Benchmark [83]) also focus on the use of multi-camera systems of different properties than FTV, e.g. on moving camera rigs, or on the 3D reconstruction of static scenes.

Because of the abovementioned limitations of available depth map databases, **the quality of depth maps in research presented in this dissertation is measured indirectly**

through the quality of virtual views synthesised using estimated depth maps. For an FTV viewer, the quality of virtual views determines the overall quality of a free-viewpoint television system. Therefore, virtual views can be used as a good determinant of the performance of a depth estimation method designed for FTV systems. Moreover, the temporal consistency of depth maps can be also measured using this methodology (as presented in Section 5.2). The presented methodology is used mostly in the evaluation of depth estimation algorithms for the free navigation purposes (e.g. in [49], [54], [55], [108]). This methodology was proposed also as a part of the 3D framework of the ISO/IEC MPEG group [128].

The ground truth depth maps are not required in the presented approach. Therefore, it is easier to produce test video sequences that can be used for the experimental evaluation of the depth estimation method. The presented measurement of the quality of depth maps was also used by the author in the previous works that concerned the estimation of depth, e.g. in [68], [69], and [70].

The process of the depth maps quality measurement by means of virtual view synthesis is presented in Fig. 5.1. The estimation of depth maps used for the assessment of depth estimation method performance is performed for 5 views of each video test sequence (used test sequences are presented in Section 5.3), with exception for the experiments that test the impact of the number of views on the depth maps quality and the processing time of estimation (presented in Section 6.3).

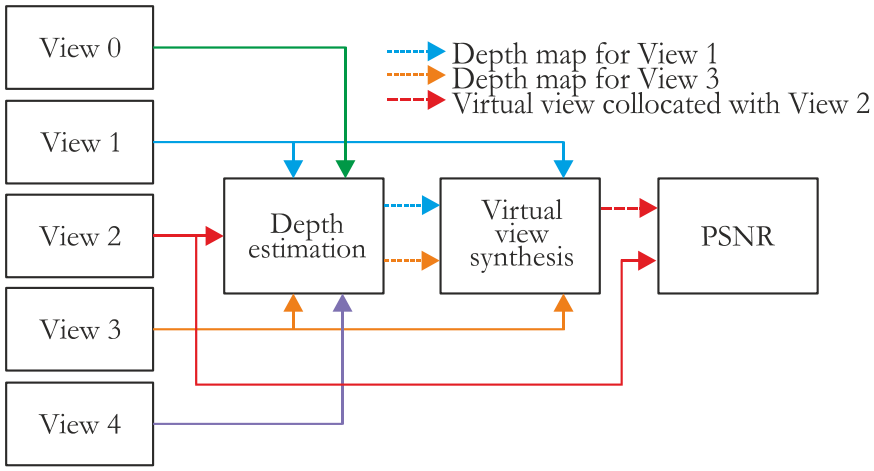


Fig. 5.1. A measurement of the quality of depth maps using virtual view synthesis.

After depth estimation, the Views 1 and 3 and their corresponding depth maps are used to synthesise the virtual view collocated with the real View 2. In the end, the virtual view and

the collocated View 2 are used in the calculation of the PSNR (peak signal-to-noise ratio) between them. The presented process that includes depth estimation and virtual view synthesis is repeated for 50 frames of each sequence, and obtained values of the PSNR are averaged.

PSNR is used as a metric of the virtual views quality, because as it is presented in [40], PSNR of a virtual view is a measure that is statistically the most corresponding to the subjective evaluation of depth maps quality. For the virtual view synthesis purposes, the VSRS (View Synthesis Reference Software [93]) was used. The VSRS was developed and is constantly improved by the ISO/IEC MPEG community.

5.2 Assessment of depth maps temporal consistency

The method of the quality assessment for depth maps presented in Section 5.1 does not take into account the temporal consistency of depth maps (the definition of temporal consistency can be found in Section 2.1). Due to abovementioned lack of ground-truth depth maps for FTV videos, the direct measurement of temporal consistency is difficult.

The size of depth maps after encoding is one of the objective measures of the temporal consistency of depth maps [95], [117]. The efficiency of video compression is highly dependent on the use of information from the other frames of a video. A video that contains a high amount of movement or quickly changing objects (e.g. a depth map that is not temporally consistent) cannot be compressed as efficiently as a video that contains a mainly static scene.

Nevertheless, the scope of this dissertation is to achieve improved quality of the virtual video produced for a user of FTV system. Therefore, in order to measure the temporal consistency of depth maps, instead of depth maps encoding, the encoder is used on the synthesised virtual views. Lack of the temporal consistency of depth maps results in visible flickering of the virtual view (Section 2.1). Therefore, the lower is the temporal consistency of depth maps, the lower is the efficiency of the encoding of virtual views.

The results are expressed as an average luma bitrate difference in comparison to the virtual view synthesised with depth maps that were not temporally enhanced, as it is presented in Fig. 5.2. The difference in the bitrate is calculated using Bjøntegaard metric [4], which measures the difference using at least four bitrate and quality pairs. These pairs are acquired from the encoding of a video for different quantisation parameters.

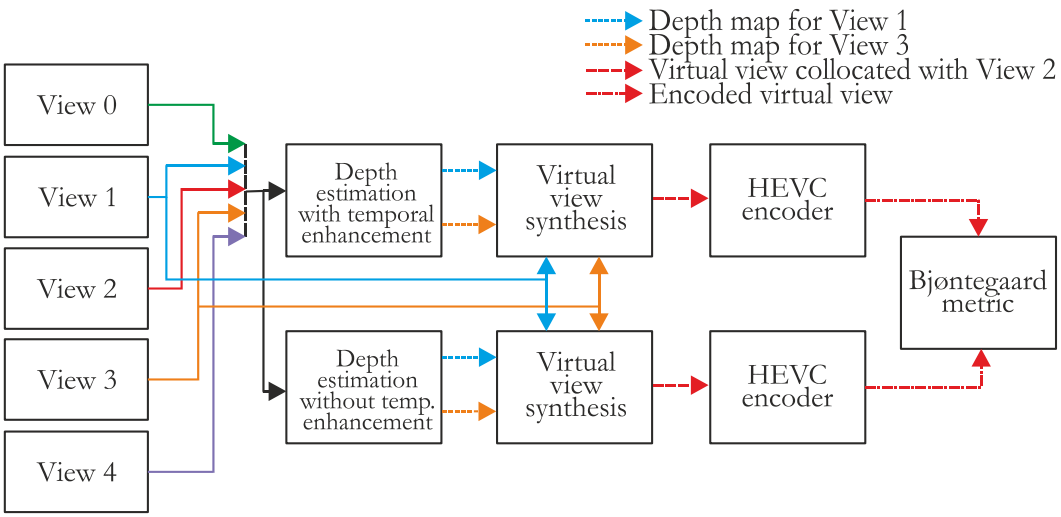


Fig. 5.2. A measurement of the temporal consistency of depth maps through the encoding of a virtual view.

For the encoding process, the HEVC (High Efficiency Video Coding) encoder is used [125]. HEVC is the state-of-the-art encoder and provides a very high efficiency of the encoding process. There are new compression technologies in development (e.g. VVC [131]) or on the market (e.g. recently presented AV1 [32]), nevertheless, HEVC is already widely implemented and used.

The encoder is set in the low-delay mode, so only the first frame of a virtual view is encoded as an intra frame. Such settings of the encoder increase the influence of the temporal consistency of an encoded sequence on the final bitrate and the quality of the encoded sequence. In the performed experiments the HM 16.15 [125] reference software is used, using the MPEG common test conditions and the software reference configurations (both available in [125]).

The described method of measurement of the temporal consistency of depth maps was previously used by the author of the dissertation in [72].

5.3 Multiview sequences test set

In all experiments, a set of 8 multiview test sequences is used (Table 5.1). These sequences differ not only in their content – the sequences have different arrangements of cameras and the resolution of acquired views. Examples of frames for one view of each sequence are

presented in Fig. 5.3. Used test sequences are the part of MPEG-I visual test materials [127] and are used in free-viewpoint television related research (e.g. in [11], [57], [82]).

TABLE 5.1. TEST SEQUENCES USED IN DISSERTATION

Name of the test sequence	Resolution	Used views	Sequence source
Ballet Breakdancers	1024×768	0 to 7	Microsoft Research [122]
BBB Butterfly BBB Rabbit	1280×768	6, 12, 19, 26, 32, 38, 45, 52	Holografika [46]
Poznań Blocks Poznań Blocks2 Poznań Fencing2 Poznań Service2	1920×1080	0 to 7	Poznań University of Technology [20], [21]

Test sequences, due to their various properties, differ also in the complexity of depth estimation. Fast motion present in the "Breakdancers" and "Poznań Fencing2" sequences hinders the achievement of the temporal consistency of estimated depth maps. Moreover, the "Poznań Fencing2" sequence shows two fencers in similar white uniforms in front of the white wall, therefore, the process of the correspondence search in neighbouring views is much more difficult than in sequences where objects are more colourful, e.g. the "Poznań Blocks" sequence. On the other hand, in the "Poznań Blocks" sequence the cameras of the system are placed uniformly on an arc, not as a camera pairs (as in "Poznań Blocks2", "Poznań Fencing2" and "Poznań Service2" sequences). A uniform distribution of cameras, in combination with the high number of occluded areas in the scene, is shown in [23] and [94] to be not optimal for the free-viewpoint television system, making the "Poznań Blocks" sequence very challenging for depth estimation.

Two test sequences are synthetic, i.e. the "BBB Butterfly" and "BBB Rabbit" sequences. While for these sequences depth estimation is not affected by noise present in naturally acquired sequences, these sequences include heavily detailed areas, e.g. blades of grass in "BBB Rabbit" and a fur in "BBB Butterfly".



"Ballet" [122]



"Breakdancers" [122]



"BBB Butterfly" [46]



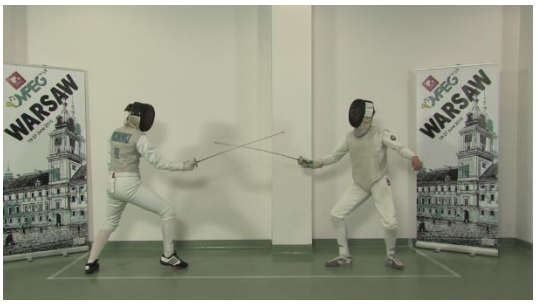
"BBB Rabbit" [46]



"Poznań Blocks" [23]



"Poznań Blocks2" [21]



"Poznań Fencing2" [21]



"Poznań Service2" [21]

Fig. 5.3. Examples of views from used test sequences.

6 EXPERIMENTAL RESULTS

In this chapter, results of performed experiments that tested the performance of the proposed depth estimation method are presented. The experiments were conducted in order to compare the proposed depth estimation method with the state-of-the-art method DERS (Section 6.1), test the impact of the number of segments (Section 6.2) and the number of views used during depth estimation on the quality of depth maps and the processing time of their estimation (Section 6.3). The presented enhancement that increases the temporal consistency of depth maps was also tested (Section 6.4). Finally, in Section 6.5, the performance of the proposed parallelisation method is described.

All experiments were performed on the Intel Core i7-5820K CPU (3.3 GHz clock) machines equipped with 64 GB of the operational memory. The calculations were performed using one core of the CPU, with the exception of the test of the parallelisation method, where the number of used cores varied from 1 to 6.

6.1 Comparison with the state-of-the-art depth estimation method

The proposed method was compared with the state-of-the-art depth estimation method implemented in Depth Estimation Reference Software developed by the MPEG community [50], [91]. DERS uses a graph-based method that states no assumptions about the positioning of cameras and is available for the research community in its entirety. The algorithm used in DERS, like the proposed depth estimation method, is based on the graph cut optimisation. Therefore, taking on the account the capabilities of DERS, it is the reasonable reference depth estimation method for the depth estimation method proposed by the author.

In the proposed method, the number of segments per one view used in the depth estimation process was 100 000 for HD sequences. For such number of segments their average size is 20 points, what ensures good representation of details of a scene and, on the other hand, significantly reduces the complexity of depth estimation. In order to achieve the similar size of segments for sequences of lower resolution, the number of segments for these sequences was set to 50 000. Other parameters of the depth estimation process were the same for the proposed method and DERS: the smoothing coefficient was set to 1, the size of the window

in the inter-view matching was 3×3, estimation for 250 levels of depth, 5 views used for depth estimation.

The comparison of the depth maps quality for the proposed method and for DERS is presented in Table 6.1, together with the comparison of the processing time of depth estimation per one view. For all test sequences, the proposed method estimates depth maps of higher quality than the reference method. The quality of virtual views (which expresses the quality of depth maps) synthesised using depth maps estimated by the proposed method is on average more than 2.5 dB higher than for depth maps estimated using DERS. The highest increase of PSNR is larger than 5 dB for the "Poznań Blocks" sequence. The lowest PSNR of a virtual view for the DERS is below 22 dB, while for the proposed method the lowest PSNR is 25.5 dB. For the proposed method only for one sequence the PSNR is below 27 dB. For DERS there are five such sequences.

TABLE 6.1. COMPARISON OF THE QUALITY OF VIRTUAL VIEWS SYNTHESISED USING DEPTH MAPS ESTIMATED USING THE PROPOSED METHOD AND DERS

Test sequence	PSNR of virtual view [dB]			Processing time of depth estimation per one view [s]		
	DERS	Proposal	Gain	DERS	Proposal	Ratio
Ballet	27.93	**28.69**	0.76	882	**499**	57%
Breakdancers	31.13	**32.19**	1.06	949	**255**	27%
BBB Butterfly	29.97	**33.20**	3.23	593	**279**	47%
BBB Rabbit	22.59	**27.21**	4.62	744	**92**	12%
Poznań Blocks	21.97	**27.20**	5.23	1445	**313**	22%
Poznań Blocks2	25.67	**28.12**	2.45	1060	**210**	20%
Poznań Fencing2	26.74	**28.60**	1.86	2254	**391**	17%
Poznań Service2	23.69	**25.51**	1.82	2780	**305**	11%
		Average:	**2.63**		***Average:***	**27%**

Simultaneously, the processing time of estimation for the proposed method is always shorter than for DERS – on average almost 4 times. What is important, the reduction of the processing time of estimation is the greatest for the Full-HD sequences, therefore, the proposed method is ready to be used with cameras of a high resolution. It is the effect of the use of segmentation in depth estimation – the complexity of estimation in the proposed method is dependent on the number of segments, not on the resolution of input views.

What has to be stressed out, the results shown in this section show only the performance of the core of the proposed depth estimation method – the proposed temporal consistency enhancement and the parallelisation methods (tested in Sections 6.4 and 6.5, respectively), that significantly reduce the processing time of depth estimation, were not used during depth estimation for the experiment presented in this section.

Fig. 6.1 shows an example of the visual comparison of depth maps calculated using the proposed method and DERS for the "Breakdancers" sequence. The edges of objects present in the scene are represented much better in depth maps produced by the proposed method, i.e. the edges of objects in the depth map correspond to the edges visible in the view. It is the result of the proposed intra-view discontinuity cost that is not the same for the whole image but is calculated adaptively to the content of a view. It increases the possibility of the proper reproduction of the discontinuities in depth maps, without any blurred edges. As it is described in Section 2.1, the proper representation of edges in depth maps has a significant impact on the quality of virtual view synthesis.

The next feature of depth maps that should be present in order for depth maps to be used for free-viewpoint television systems is their inter-view consistency. DERS method does not allow to estimate depth maps for all input views simultaneously. The depth map for the View 2 is estimated using Views 0, 1 and 2, the depth map for View 4 – in the second, independent estimation, using Views 3, 4 and 5. Therefore, as it can be seen in Fig. 6.1, the depth of objects was estimated differently for Views 2 and 4 (especially in the background and on the floor). In the proposed method depth maps are inter-view consistent in all views because, despite using the same 5 views for depth estimation as in DERS, only one joint depth estimation is performed for all views.

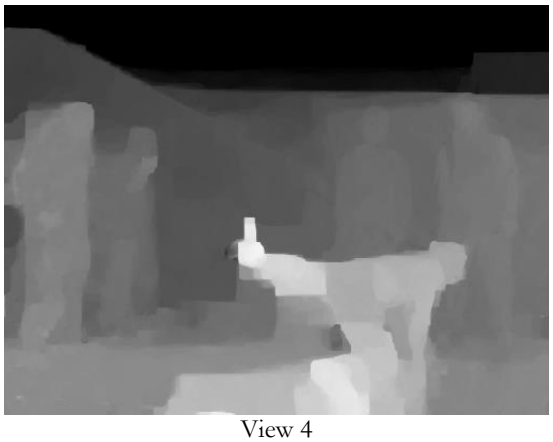
Fig. 6.2 shows a comparison of virtual views that were synthesised using depth maps estimated with the proposed method and DERS, together with the reference view. When the proposed method of depth estimation is used, the virtual view is much more similar to the reference view. The number of the errors visible in the virtual view was decreased, especially on the objects in the foreground.

As the presented results show, the proposed method can be successfully used for free-viewpoint television. **The proposed method allows to estimate depth maps of higher quality and simultaneously reduces the complexity of the whole estimation process in comparison to the state-of-the-art method implemented in DERS.** It is a result of numerous developed improvements that are linked by the use of segmentation.

Depth maps calculated using DERS:



View 2



View 4

Depth maps calculated using the proposed method:



View 2



View 4

Input views:



View 2



View 4

Fig. 6.1. A comparison of depth maps calculated using the proposed method and DERS for the "Breakdancers" sequence.



Virtual view synthesised using
depth maps calculated with DERS



Enlarged virtual view:



Virtual view synthesised using
depth maps calculated with the proposed method



Enlarged virtual view



Reference view



Enlarged reference view

Fig. 6.2. A comparison of virtual views synthesised using depth maps calculated using the proposed method and DERS for the "Breakdancers" sequence.

6.2 Impact of the number of segments on depth maps quality and processing time

This section includes results of the experiment that tested the impact of the number of segments used in the proposed method on estimated depth maps. In order to determine this impact, the process of depth estimation was repeated 7 times: for 1 000, 5 000, 10 000, 25 000, 50 000, 100 000, and 150 000 segments per one view. All other parameters of estimation were the same (5 views, the smoothing coefficient equal to 1, the size of the window in the inter-view matching was 3×3, I type depth frame every 10 frames, estimation performed for 250 levels of depth).

Fig. 6.3 shows the PSNRs of virtual views synthesised using depth maps estimated for the different number of segments, together with the processing time of depth estimation, averaged for all test sequences. Results of the experiment for individual sequences are presented in Table 6.2.

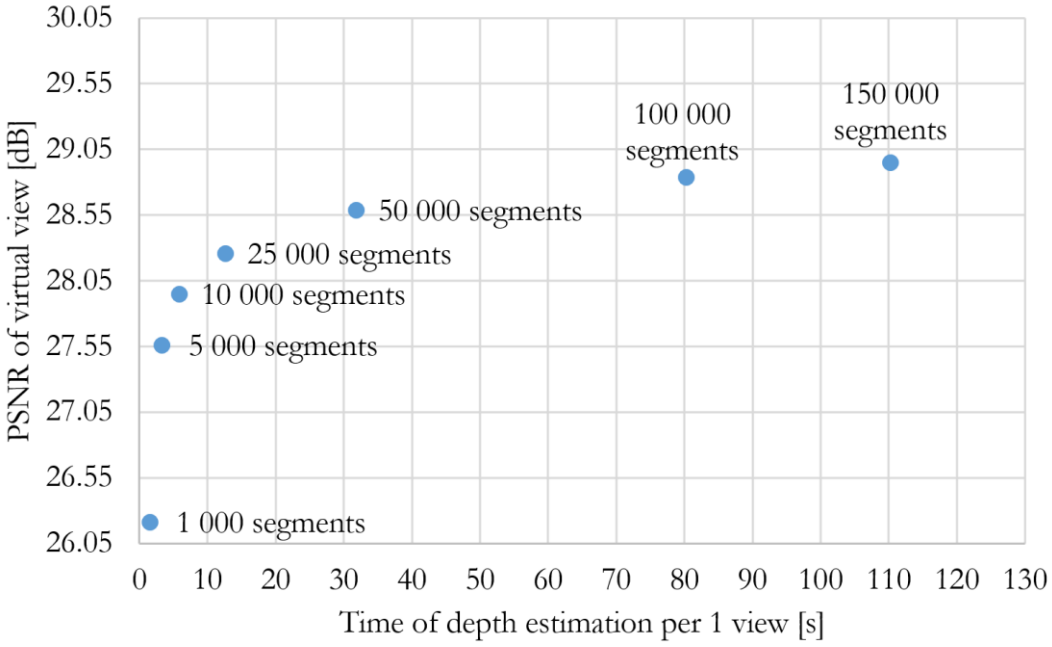


Fig. 6.3. The average quality of a virtual view synthesised using depth maps estimated for a different number of segments per one view and a processing time of depth estimation.

The results show that the depth maps estimated for the higher number of segments per view provide the higher quality of virtual view synthesis. This way, the quality of depth maps becomes changeable – **the trade-off between the quality of depth maps and the processing time of estimation can be controlled.** When only 1 000 of segments are used per

one view in the estimation process, the quality of depth maps is equal to the average quality of depth maps estimated using DERS (which is compared to the proposed method in Section 6.1), but the processing time of the estimation process is significantly shorter and equal to only 2 seconds.

TABLE 6.2. THE QUALITY OF VIRTUAL VIEWS SYNTHESISED USING DEPTH MAPS ESTIMATED FOR A DIFFERENT NUMBER OF SEGMENTS USED IN ESTIMATION

Test sequence	Number of segments						
	1 000	5 000	1 000	25 000	50 000	100 000	150 000
	Mean PSNR of the virtual views [dB]						
Ballet	26.55	28.02	28.30	28.53	28.64	28.83	28.85
Breakdancers	29.12	30.85	31.40	31.67	32.00	32.14	32.15
BBB Butterfly	30.26	32.07	32.36	32.79	33.08	33.23	33.25
BBB Rabbit	24.84	26.08	26.64	26.91	27.14	27.43	27.73
Poznań Blocks	22.43	24.55	25.27	25.74	26.60	27.14	27.26
Poznań Blocks2	25.94	26.93	27.16	27.54	27.87	28.10	28.19
Poznań Fencing2	26.17	27.29	27.60	27.92	28.19	28.43	28.61
Poznań Service2	24.35	24.66	24.80	24.93	25.17	25.38	25.51
Average:	**26.21**	**27.55**	**27.94**	**28.25**	**28.59**	**28.84**	**28.94**

The highest increase of the quality of depth maps can be seen between 1 000 and 25 000 segments per view. Despite the number of segments increased 25 times, the average processing time of estimation increases only 6 times. On the other hand, increasing the number of segments above 100 000 does not change the quality of depth maps significantly (only 0.1 dB), but the mean processing time of estimation is noticeably increased. The processing times of estimation for the individual sequences are presented in Appendix in Table A.1.

An example of the visual comparison of depth maps estimated for 1 000, 25 000 and 150 000 segments per one view for the "Poznań Service2" test sequence is presented in Fig. 6.4. As it can be seen in the enlarged fragments of depth maps, the level of details and the correctness of estimated depth maps increases with the number of used segments.



Depth map calculated using 1 000 segments



Enlarged depth map



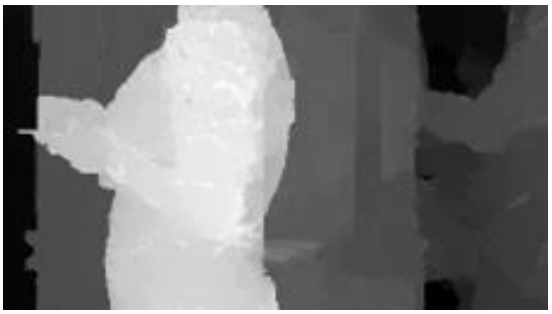
Depth map calculated using 25 000 segments



Enlarged depth map



Depth map calculated using 150 000 segments



Enlarged depth map



Reference view



Enlarged reference view

Fig. 6.4. A comparison of depth maps calculated for a different number of segments for the "Poznań Service2" sequence.

The possibility of having a control on the processing time of depth estimation is very useful for a free-viewpoint television system. When a small number of segments is used, the processing time of estimation is significantly shortened. The quality of the estimated depth is lower but still comparable to the reference depth estimation method. Such fast depth estimation can be used in order to perform a fast verification of the system, e.g. to check if the estimated camera parameters are correct, or cameras of the system are properly synchronised.

6.3 Impact of the number of views on depth maps quality and processing time

In this section, the proposed method is tested for different numbers of views used during depth estimation. For each test sequence, depth estimation was repeated for 6 cases: for 3, 4, 5, 6, 7, and 8 views used in estimation. All other parameters of estimation were the same: 100 000 segments per one view, the smoothing coefficient equal to 1, the size of the window in the inter-view matching was 3×3, I type depth frame every 10 frames, estimation performed for 250 levels of depth.

The quality of virtual views synthesised using depth maps estimated for different numbers of input views, together with the processing time of depth estimation, are presented in Fig. 6.5. Results for individual sequences are presented in Table 6.3. As results show, with the increase of the number of views used during estimation the quality of depth maps is increased. The difference in PSNRs of virtual view is equal to 0.4 dB when 8 views instead of 3 views are used. The processing time of estimation is also slightly increased, nevertheless, the difference becomes negligible when more than 5 views are used (the processing time of estimation is longer by less than 10%). **Taking into account the requirement of ensuring the inter-view consistency of depth maps** (described in Section 2.1), **for free-viewpoint television systems it is recommended to estimate depth maps for all views simultaneously.**

The processing times of estimation for the individual sequences are presented in Appendix in Table A.2.

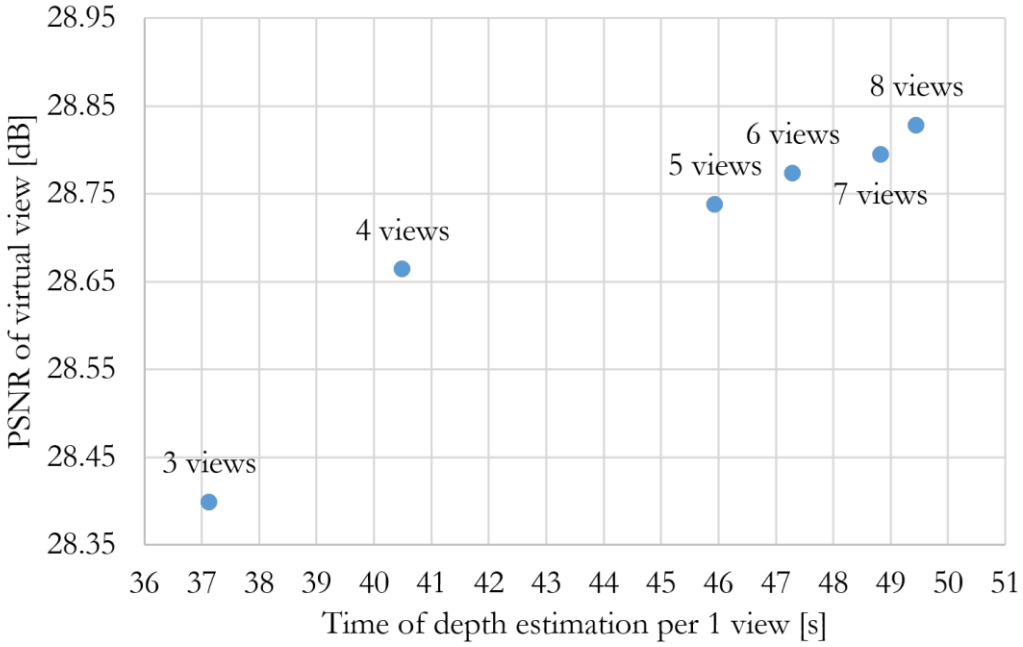


Fig. 6.5. The average quality of a virtual view synthesised using depth maps estimated for a different number of views used in the estimation process and a processing time of depth estimation.

TABLE 6.3. THE QUALITY OF VIRTUAL VIEWS SYNTHESISED USING DEPTH MAPS ESTIMATED FOR A DIFFERENT NUMBER OF VIEWS USED IN ESTIMATION

Test sequence	Number of views					
	3	4	5	6	7	8
	Mean PSNR of the virtual views [dB]					
Ballet	28.68	28.78	28.63	28.93	28.98	29.01
Breakdancers	31.98	32.05	31.99	32.24	32.14	32.28
BBB Butterfly	32.32	33.13	33.08	33.63	33.62	33.55
BBB Rabbit	26.82	27.25	27.13	27.60	27.27	27.44
Poznań Blocks	25.95	26.37	27.13	26.32	26.66	26.66
Poznań Blocks2	28.14	28.27	28.09	27.73	27.94	27.89
Poznań Fencing2	28.12	28.15	28.42	28.64	28.68	28.72
Poznań Service2	25.14	25.27	25.37	25.06	25.03	25.02
Average:	**28.39**	**28.66**	**28.73**	**28.77**	**28.79**	**28.82**

An example of the visual comparison of depth maps estimated using 3 views and 8 views of the "Poznań Blocks" sequence, presented in Fig. 6.6, shows that the depth maps are changed mostly in parts of the scene that are occluded in some views. Enlarged fragments of depth maps show, that in case of the presented sequence**, the increased number of views used in estimation allowed to estimate the depth of the background even in the part surrounded by fingers of the person present in the scene. The proper estimation of the depth of the background is very significant for the subjective quality of virtual views.** If the depth of the background is estimated incorrectly, it causes that some fragments of the background slide in front of objects when a viewer changes a viewpoint of a scene. It significantly reduces the subjective quality of free navigation, therefore, the overall perceived quality of an FTV system.



Depth map calculated using 3 views



Enlarged depth map



Depth map calculated using 8 views



Enlarged depth map



Reference view



Enlarged reference view

Fig. 6.6. The comparison of depth maps calculated for the different number of input views for the "Poznań Blocks" sequence.

6.4 Impact of temporal consistency enhancement on depth maps quality and processing time

This section describes the impact of the proposed temporal consistency enhancement on the quality of estimated depth maps, their temporal consistency, and the overall processing time of depth estimation. The estimation of depth maps for test sequences was repeated for different numbers of P type depth frames between I type depth frames: no P depth frames between I depth frames (only I depth frames – no use of temporal information), 4 P frames (i.e. I type depth frame every 5 frames), 9 P frames, 24 P frames, and finally 49 P frames (only first frame of estimated depth maps was I type depth frame). Other parameters of estimation were the same for all cases: 5 views, 100 000 segments per one view, the smoothing coefficient equal to 1, the size of the window in the inter-view matching was 3×3, estimation performed for 250 levels of depth.

Firstly, as in previous experiments, the quality of estimated depth maps was measured using virtual view synthesis. The results (the virtual views quality and the processing time of depth estimation) averaged for all sequences are presented in Fig. 6.7, while individual results for each sequence are presented in Table 6.4.

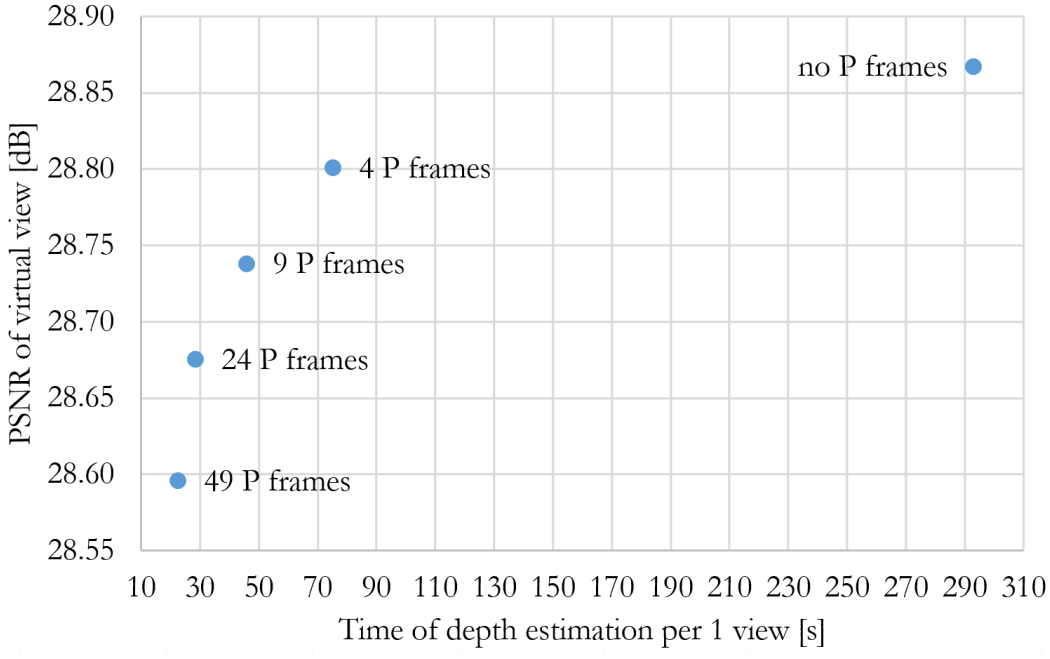


Fig. 6.7. The average quality of a virtual view synthesised using depth maps estimated for a different number of P type depth frames between I type depth frames and the processing time of depth estimation.

TABLE 6.4. THE QUALITY OF VIRTUAL VIEWS SYNTHESISED USING DEPTH MAPS ESTIMATED FOR A DIFFERENT NUMBER OF P TYPE DEPTH FRAMES USED IN ESTIMATION

Test sequence	Number of P type depth frames between I type depth frames				
	0	4	9	24	49
	Mean PSNR of the virtual views [dB]				
Ballet	28.69	28.68	28.63	28.74	28.75
Breakdancers	32.19	32.13	31.99	31.95	31.75
BBB Butterfly	33.20	33.14	33.08	32.97	32.93
BBB Rabbit	27.21	27.16	27.13	27.12	27.08
Poznań Blocks	27.20	27.19	27.13	26.98	26.79
Poznań Blocks2	28.12	28.11	28.09	28.06	28.03
Poznań Fencing2	28.60	28.50	28.43	28.36	28.35
Poznań Service2	25.51	25.45	25.37	25.19	25.04
Average:	28.84	28.80	28.73	28.67	28.59

Use of P type depth frames in an insignificant degree decreases the quality of depth maps (measured as PSNR of the virtual view synthesised using such depth maps). The PSNR of a virtual view for depth maps with only one I type depth frame is only 0.25 dB lower than if only I type depth frames are used. On the other hand, in the same comparison, **the processing time of depth estimation is significantly decreased – depth estimation is almost 15 times faster**. When only I type depth frames are used (therefore, the proposed temporal consistency enhancement is not used), the average processing time of depth estimation per 1 view is almost 5 minutes. With the proposed enhancement the processing time is reduced to around 20 seconds when I type depth frame is used every 50 frames.

The processing times of estimation for the individual sequences are presented in Appendix in Table A.3. The quality for individual frames of each sequence for estimation using I type depth frames and for 49 P type depth frame is presented in Appendix in Fig. A.1 – Fig. A.16.

Fig. 6.8 shows an example of the comparison of 3 consecutive frames of depth maps for the "Ballet" sequence, estimated without temporal consistency enhancement (no P type depth frames) and estimated with the proposed temporal consistency enhancement. Use of P type depth frames during depth estimation increases the temporal consistency of depth maps – the depth of the background is the same in the consecutive frames. It significantly

reduces the flickering which can be seen in virtual views that were synthesised using the depth maps without temporal consistency enhancement.

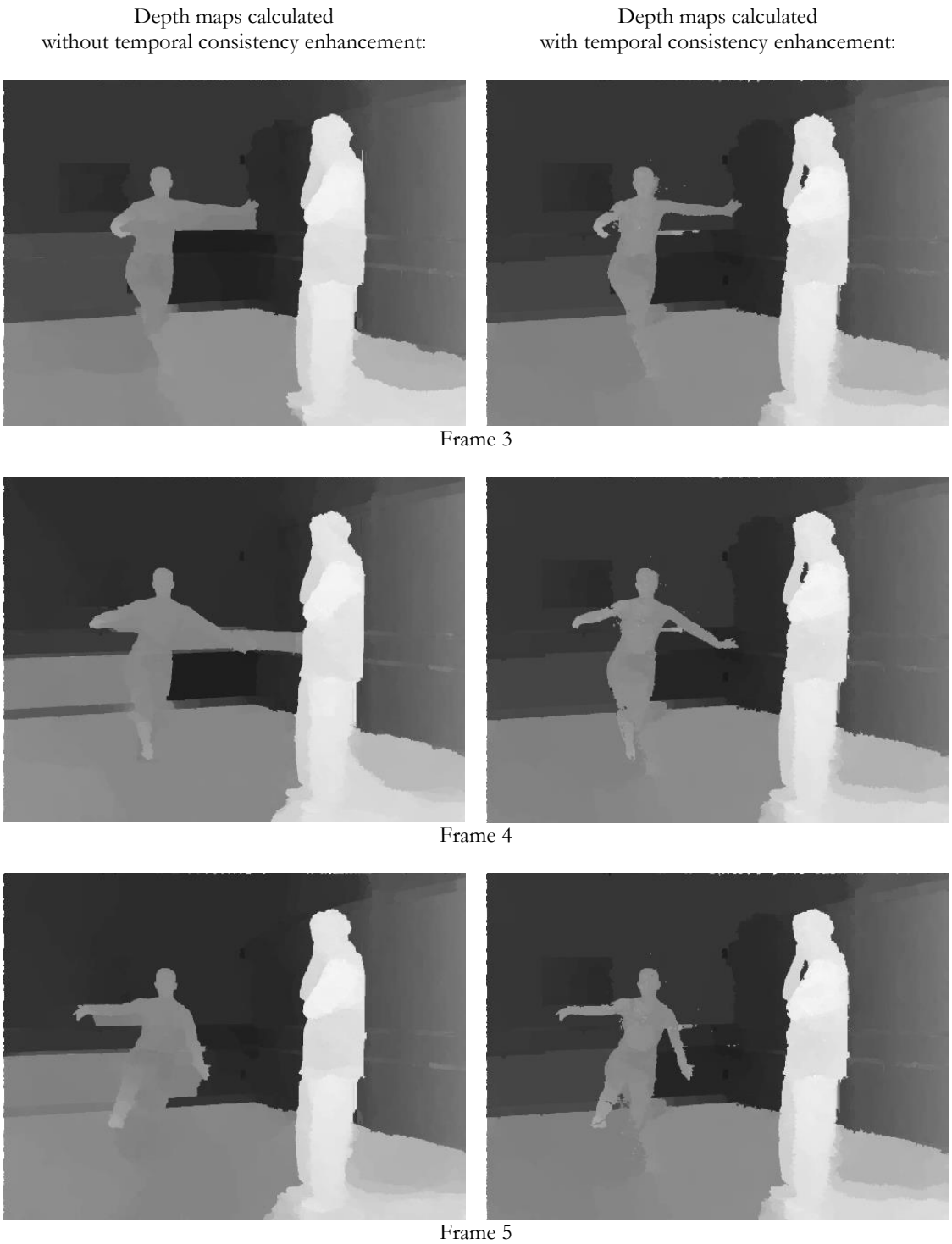


Fig. 6.8. A comparison of depth maps calculated with and without temporal consistency enhancement for the "Ballet" sequence.

In order to evaluate the increase of temporal consistency, the synthesised virtual views were encoded with HEVC encoder using the methodology described in details in Section 5.2. The increase of temporal consistency is measured by the reduction of the bitrate of the encoded virtual view synthesised using the depth maps estimated with proposed temporal consistency enhancement, in comparison with the encoded virtual view synthesised using depth maps without temporal consistency enhancement. The average reduction of the bitrate is calculated using the Bjøntegaard metric. The bitrate reductions are presented in Table 6.5, the PSNRs and bitrates for all used quantisation parameters are presented in Table 6.6.

On average, the use of the proposed temporal consistency enhancement of the depth estimation method causes very large bitrate reduction for the encoded virtual views – up to 32%. Such reduction of the bitrate indicates that the performance of the prediction of the encoder is significantly increased, therefore, it indicates that the consecutive frames of the virtual views are much more similar. The change of the quality of depth maps is not considerably changed (Fig. 6.7), what together with the abovementioned increase of the encoder performance, confirms the **significant increase of the temporal consistency of depth maps for the proposed temporal consistency enhancement.**

TABLE 6.5. AVERAGE LUMA BITRATE REDUCTIONS OF ENCODED VIRTUAL VIEWS SYNTHESISED USING DEPTH MAPS ESTIMATED FOR A DIFFERENT NUMBER OF P TYPE DEPTH FRAMES BETWEEN I TYPE DEPTH FRAMES

Test sequence	Number of P type depth frames			
	4	9	24	49
	Encoded virtual views bitrate reduction compared to virtual views synthesised using depth maps with no P type depth frames			
Ballet	-14.4%	-19.0%	-20.5%	-20.8%
Breakdancers	-24.9%	-33.1%	-34.4%	-34.1%
BBB Butterfly	-18.8%	-29.2%	-34.6%	-38.3%
BBB Rabbit	-7.9%	-8.0%	-11.3%	-19.7%
Poznań Blocks	-4.5%	-5.7%	-7.3%	-6.1%
Poznań Blocks2	-30.7%	-35.0%	-36.9%	-39.9%
Poznań Fencing2	-30.7%	-53.5%	-68.8%	-75.8%
Poznań Service2	-21.6%	-30.0%	-23.5%	-21.7%
Average:	**-19.2%**	**-26.7%**	**-29.7%**	**-32.0%**

TABLE 6.6. THE BITRATE AND QUALITY OF VIRTUAL VIEWS ENCODED USING HEVC. VIRTUAL VIEWS ARE SYNTHESISED USING DEPTH MAPS ESTIMATED FOR A DIFFERENT NUMBER OF P TYPE DEPTH FRAMES BETWEEN I TYPE DEPTH FRAMES

Test sequence	QP	Number of P type depth frames between I type depth frames									
		0		4		9		24		49	
		Bitrate [Mbps]	PSNR [dB]	Bitrate [Mbps]	PSNR [dB]	Bitrate [Mbps]	PSNR [dB]	Bitrate [Mbps]	PSNR [dB]	Bitrate [Mbps]	PSNR [dB]
Ballet	22	4.6	41.8	3.9	41.8	3.7	41.8	3.6	41.8	3.6	41.8
	27	1.7	39.8	1.5	39.9	1.4	39.9	1.4	40.0	1.4	40.0
	32	0.7	37.8	0.6	37.9	0.6	37.9	0.6	37.9	0.6	38.0
	37	0.3	35.9	0.3	35.9	0.3	35.9	0.3	35.9	0.3	35.9
Breakdancers	22	8.7	40.3	8.4	40.4	8.4	40.4	8.1	40.5	8.0	40.5
	27	2.8	38.2	2.7	38.3	2.7	38.3	2.7	38.3	2.6	38.5
	32	1.1	36.4	1.1	36.5	1.1	36.5	1.1	36.5	1.1	36.7
	37	0.5	34.7	0.5	34.8	0.5	34.8	0.5	34.9	0.5	35.1
BBB Butterfly	22	5.5	45.4	5.2	45.4	5.1	45.4	5.1	45.4	5.1	45.4
	27	2.5	42.0	2.4	42.0	2.4	42.1	2.3	42.1	2.3	42.1
	32	1.2	39.1	1.1	39.1	1.1	39.1	1.1	39.1	1.1	39.1
	37	0.5	36.5	0.5	36.5	0.5	36.5	0.5	36.6	0.5	36.5
BBB Rabbit	22	14.8	40.4	10.2	40.8	7.6	41.2	5.5	41.5	4.7	42.0
	27	6.0	36.0	5.0	36.6	3.8	37.1	2.7	37.4	2.3	37.8
	32	2.9	32.7	2.1	33.2	1.6	33.6	1.1	33.9	1.0	34.2
	37	0.9	30.2	0.7	30.5	0.6	30.8	0.4	31.0	0.4	31.2
Poznań Blocks	22	25	41.5	19.2	41.6	18.2	41.7	17.8	41.7	18.2	41.8
	27	9.5	38.1	7.5	38.2	7.0	38.4	6.8	38.3	7.0	38.4
	32	3.5	35.5	2.8	35.6	2.6	35.7	2.5	35.7	2.5	35.7
	37	1.3	33.4	1.0	33.5	0.9	33.6	0.9	33.6	0.9	33.6
Poznań Blocks2	22	26.5	40.1	22.2	40.1	20.8	40.2	19.9	40.2	19.6	40.3
	27	7.7	37.6	6.5	37.7	6.0	37.8	5.7	37.9	5.6	37.9
	32	2.4	35.7	2.1	35.8	1.9	35.9	1.8	35.9	1.8	36.0
	37	0.8	34.1	0.7	34.1	0.6	34.2	0.6	34.2	0.6	34.3
Poznań Fencing2	22	32.3	40.5	30.9	40.8	30.1	40.8	29.8	40.8	29.5	40.8
	27	14.5	36.7	13.4	37.6	12.9	37.7	12.6	37.7	12.4	37.7
	32	5.9	33.4	5.3	34.7	5.1	34.8	4.9	34.8	4.7	34.9
	37	2.0	32.0	1.8	32.5	1.7	32.5	1.7	32.5	1.6	32.7
Poznań Service2	22	46.2	39.7	37.0	39.7	34.7	39.8	35.6	39.8	36.1	39.8
	27	18.1	36.3	14.8	36.5	13.8	36.7	14.4	36.6	14.5	36.6
	32	6.7	33.6	5.7	33.8	5.4	33.9	5.7	33.8	5.7	33.7
	37	2.3	31.6	2.0	31.7	1.8	31.7	2.0	31.5	2.1	31.4

6.5 Impact of the parallelisation method on depth maps quality and processing time

In the last experiment, the impact of the proposed parallelisation method on the quality of estimated depth maps and the processing time was tested. Depth estimation was repeated for different scenarios of parallelisation for each test sequence: without the use of the parallelisation (depth estimation using 1 thread), and for the parallelisation using 2, 4, and 6 threads. The highest number of used threads is limited by the number of standalone cores available in used CPU (Intel Core i7-5820K). Moreover, the tests were performed for both proposed schemes of the parallelisation: for interleaved levels of depth and for depth levels divided into blocks of depth levels.

All parameters of the depth map estimations, besides the number of used threads, were the same: 5 views, 100 000 segments per one view, the smoothing coefficient equal to 1, the size of window in the inter-view matching was 3×3, I type depth frame every 10 frames, estimation for 250 levels of depth.

The results of the experiment (the quality of estimated depth maps, measured by the PSNR of a virtual view, and processing time of depth estimation) are presented in Fig. 6.9. The quality of virtual views synthesised using depth maps estimated for different parallelisation schemes for individual sequences is presented in Table 6.7.

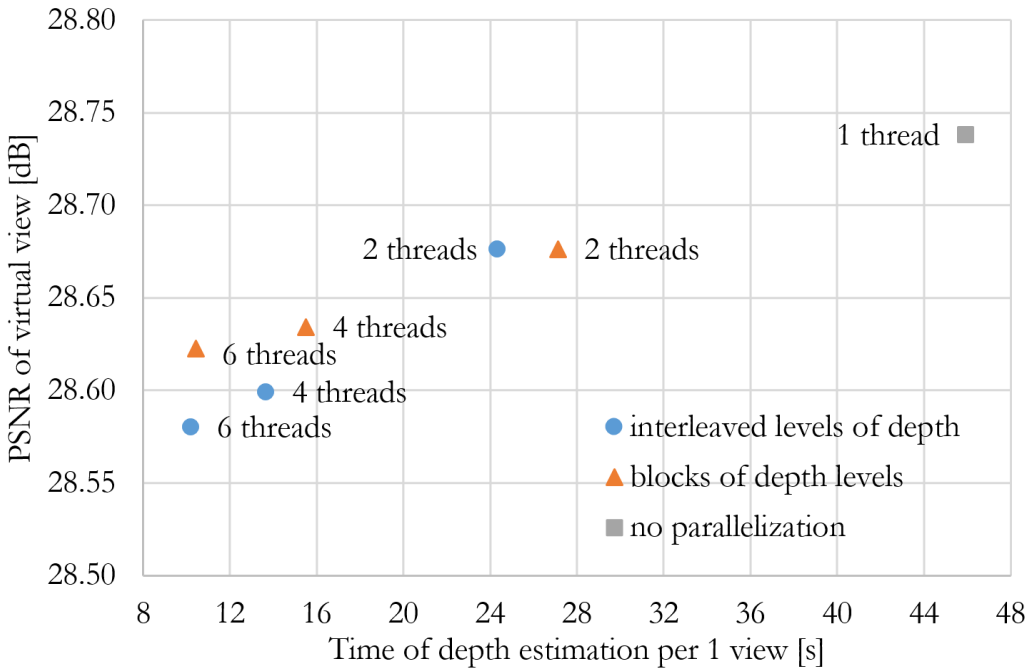


Fig. 6.9. The average quality of a virtual view synthesised using depth maps estimated for different parallelisation schemes and a processing time of depth estimation.

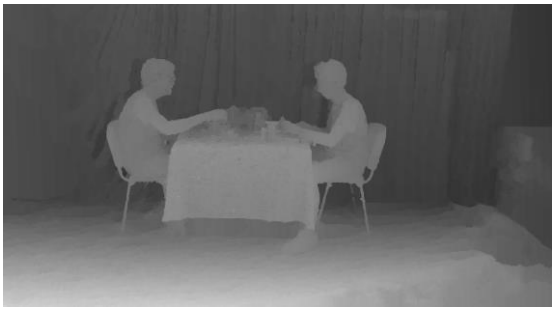
TABLE 6.7. THE QUALITY OF VIRTUAL VIEWS SYNTHESISED USING DEPTH MAPS ESTIMATED FOR A DIFFERENT PARALLELISATION TYPES

Test sequence	Parallelisation type						
	None	Interleaved levels of depth			Blocks of depth levels		
	Number of threads used in depth estimation						
	1	2	4	6	2	4	6
	Mean PSNR of virtual views [dB]						
Ballet	28.64	28.64	28.72	28.71	28.34	28.30	28.18
Breakdancers	32.00	32.05	32.06	31.98	31.93	31.87	31.87
BBB Butterfly	33.08	33.09	32.95	32.85	33.20	33.13	33.19
BBB Rabbit	27.14	27.11	27.10	27.07	27.04	26.97	27.02
Poznań Blocks	27.14	26.67	26.33	26.47	27.12	27.08	27.04
Poznań Blocks2	28.10	28.07	28.01	27.97	28.10	28.09	28.10
Poznań Fencing2	28.43	28.38	28.22	28.22	28.41	28.36	28.36
Poznań Service2	25.38	25.39	25.40	25.36	25.27	25.27	25.22
Average:	**28.74**	**28.68**	**28.60**	**28.58**	**28.68**	**28.63**	**28.62**

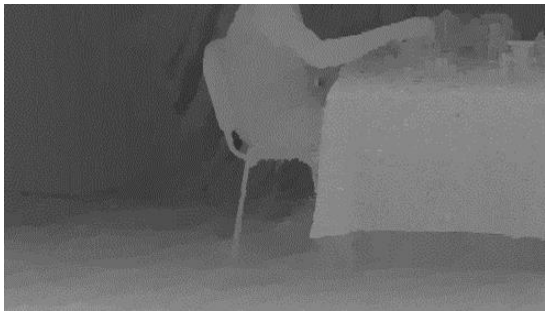
Even for 6 threads the quality decrease in comparison with estimation without the parallelisation is insignificant (only 0.1 dB) but the processing time of estimation decreases more than 4.5 times. The difference in the quality for both proposed parallelisation schemes increases with the number of the threads used, in favour of the depth levels distribution as blocks of depth levels.

An example of the visual comparison of depth maps for sequence "Poznań Blocks2" (Fig. 6.10) confirms the slightly better quality of depth maps estimated using blocks of depth levels. It can be seen that when depth levels are distributed as blocks, the result of depth estimation is more smooth than for interleaved levels of depth.

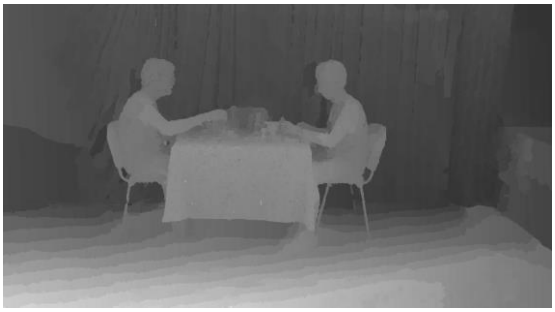
The results of the performed experiment confirm that when the levels of depth are distributed onto threads as blocks of depth levels, the processing time of estimation is slightly longer than for interleaved levels of depth. It is the result of a typical distribution of depth levels in the scene – when depth levels are divided into threads as blocks of depth levels, estimation for some threads is longer (see Section 4.4 for details).



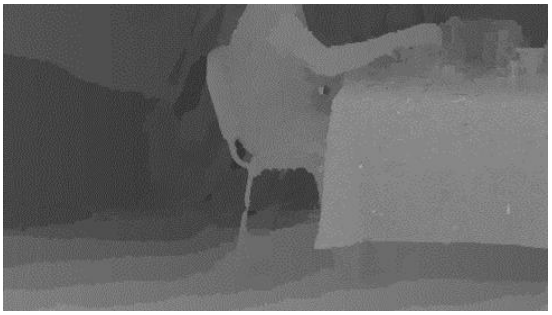
Depth map calculated using one thread



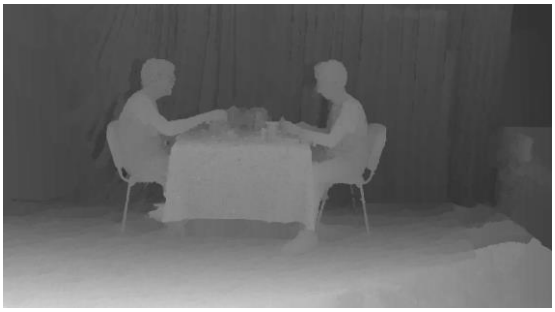
Enlarged depth map



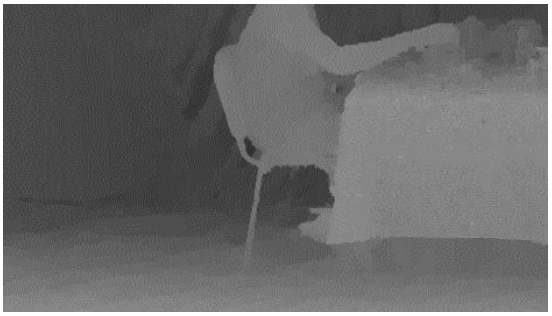
Depth map calculated using 6 threads
with interleaved levels of depth



Enlarged depth map



Depth map calculated using 6 threads
with levels of depth distributed as blocks



Enlarged depth map



Reference view



Enlarged reference view

Fig. 6.10. A comparison of depth maps calculated for different parallelisation types for "Poznań Blocks2" sequence.

On the other hand, the difference in the processing time of estimation becomes negligible when more than 4 threads are used (Fig. 6.11), therefore, because of the slightly higher quality of estimated depth maps, use of blocks of depth levels is recommended.

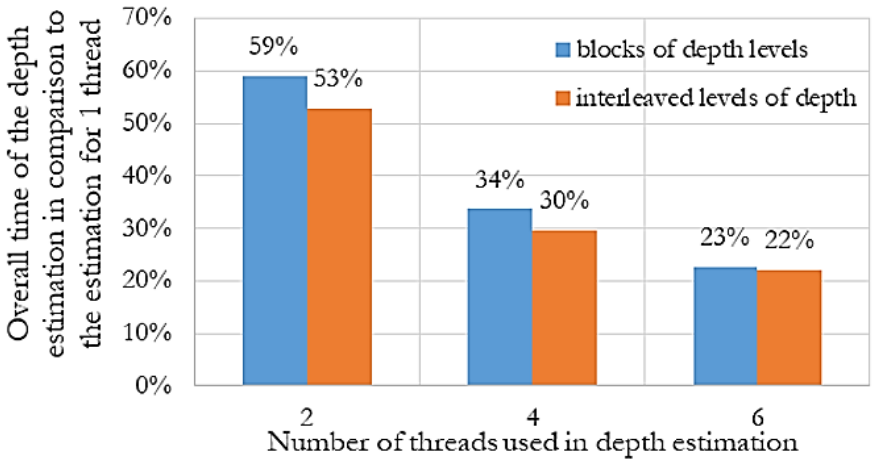


Fig. 6.11. The relative processing time of the parallel depth estimation in comparison to estimation for one thread.

The presented results show that the proposed parallelisation method can be successfully used for the proposed depth estimation method – the processing time of estimation is significantly decreased and the quality of depth maps is simultaneously maintained. Moreover, both the inter-view and the temporal consistency of depth maps, that are fundamental for the quality of virtual view synthesis, are preserved when the proposed parallelisation is used. The method is fully scalable, so the constantly increasing number of cores in modern CPUs can be fully utilised for further reduction of the processing time of depth estimation.

7 APPLICATION OF PROPOSED DEPTH MAP ESTIMATION METHOD IN FTV SYSTEMS

The presented depth estimation method, which is the main achievement presented in this dissertation, was used for the estimation of depth maps for new test sequences that were released for the MPEG community as new test material for a further research on the free navigation. The description of the acquisition process and prepared test sequences is presented in Section 7.1.

The new method of depth estimation was also already implemented in the experimental free-viewpoint television systems developed by the Chair of Multimedia Telecommunications and Microelectronics of the Poznań University of Technology in cooperation with the industry partner. The new free-viewpoint television system is presented in Section 7.2. The presented depth estimation method is a part of the rendering server that prepares the MVD representation required for virtual view synthesis in user terminals.

7.1 Acquisition of the new FTV test sequences

The acquisition of new free-viewpoint television test sequences provided for the research community was performed using a set of 10 wireless camera modules, developed in the Chair of Multimedia Telecommunications and Microelectronics [18], [22]. Each module, which can be mounted on an individual tripod, consists of the Canon XH G1 Full-HD camera, the battery that is sufficient for few hours of recording, and FPGA-controlled devices: the recorder and the wireless synchronisation module.

Each camera module is independent, therefore, cameras of the system can be positioned freely around a scene – the positioning of cameras can be adjusted to the characteristics of the acquired scene. The cameras of the system can be located not only indoors, but the system is also capable of an outdoor acquisition of a multiview video – Fig. 7.1.

The possibility of an arbitrary positioning of cameras led to the use of this FTV system for the research on the optimal positioning of cameras, co-authored by the author of this dissertation [16]. Therefore, the acquired test sequences are not limited to ones that use equally distributed cameras in the front of a scene, e.g. "Poznań Blocks" sequence [20], but also the positioning as a set of pairs may be used – "Poznań Blocks2", "Poznań Fencing2" and "Poznań Service2" sequences [21]. These sequences are used not only in the research

presented in this dissertation but also are available for the MPEG community as a set of new test sequences for further free navigation explorations. The sequences are also provided with camera parameters that were calculated using the camera parameter estimation method provided by the author of this dissertation.



Fig. 7.1. The acquisition of indoor and outdoor multiview test sequences.

The sequences "Poznań Blocks2" and "Poznań Fencing2" were released together with depth maps that were estimated using the depth map estimation method presented in this dissertation. The proposed test sequences with depth maps were very well received by the MPEG community, what can be seen in the resolutions of the 115th meeting of the MPEG that was held in June 2016 in Geneva [129]. **The sequences were described by the researchers that are a part of the MPEG community as ones that enable to create a satisfying free navigation experience for the end user of the FTV system. It confirms the high quality of the presented depth estimation method and its particular usefulness for free-viewpoint television.**

The proposal of Poznań University of Technology, together with free navigation videos for "Poznań Blocks2" and "Poznań Fencing2" sequences, can be found also in the public home page of the MPEG group [124] and as the additional media available for [94] in the IEEE Xplore database.

The proposed multiview test sequences were already used by other international researchers that work on the development of new methods for free-viewpoint television, e.g. to test new virtual view synthesis methods [11], the depth maps and virtual views compression methods [64], [82], and frameworks for the processing of multiview videos [57].

7.2 Overview of Poznań University of Technology Free-Viewpoint Television system

The experience in free navigation obtained during research presented in Section 7.1 was used by the Chair of Multimedia Communication and Microelectronics in the development of the new FTV system with cooperation with the industry partner. The project was supported by the National Centre for Research and Development, Poland under Project no. TANGO1/266710/NCBR/2015.

The new system uses GoPro HERO 4 cameras, of which 38 are mounted in the sports hall of the Poznań University of Technology, around the basketball and volleyball field, as presented in Fig. 7.2. The synchronisation and the control of cameras are performed using the controller developed for this system.



Fig. 7.2. The FTV system in the PUT sports hall.

Furthermore, the system consists of the scene representation server that provides depth maps for acquired video, using the method of depth estimation presented in this dissertation, and the rendering server that synthesises a new virtual viewpoint to the end user in real time.

The demonstration of the system, that included free navigation for a basketball match that was streamed to the smartphones of the audience, was presented during the 121st meeting of the MPEG group in Gwangju, Korea [24] and on the 2017 IEEE International Conference on Image Processing in Beijing, China [19]. One of the frames of the basketball match acquired using the presented system, together with the depth map estimated using the proposed method and used in demonstrations for virtual view synthesis, are presented in Fig. 7.3.

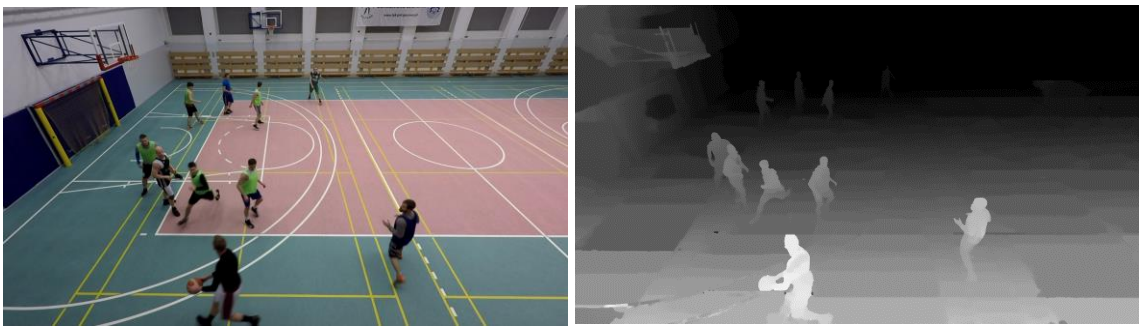


Fig. 7.3. A view and depth map from the acquired basketball match.

The test sequences for the FTV systems usually are shorter than 30 seconds, therefore, do not reflect the real scenario of the TV transmission. The limited number of frames enables to allow a long processing time of depth estimation. On the other hand, with the use of the presented FTV system a whole sports event can be acquired, so the time which can be reserved for depth estimation purposes has to be limited. **The decreased complexity of the proposed depth estimation method and the high quality of estimated depth maps led to the implementation the proposed method in such practical, operational free-viewpoint television system.**

8 SUMMARY OF THE DISSERTATION

The presented dissertation focuses on depth estimation in free-viewpoint television systems. The development of **a novel depth estimation method for free-viewpoint television systems** is the main goal of research presented in this dissertation. The focus is put on increasing the quality of depth maps estimated for the purpose of virtual view synthesis in free-viewpoint television systems and on the reduction of the complexity of depth estimation.

Section 8.1 presents the achievements that are the result of research conducted by the author, Section 8.2 summarises works conducted during the creation of the dissertation, while Section 8.3 presents the general conclusions and shows the future directions of research.

8.1 Original achievements of the dissertation

The primary original achievements of the dissertation concern the development of a novel depth estimation method for free-viewpoint television systems:

- **A novel depth estimation method** which produces depth maps that are inter-view consistent with no assumptions about locations of real cameras that acquire video. Estimation is performed for segments instead for individual pixels, thus the size of segments can be used for controlling the trade-off between the quality of depth maps and the processing time of estimation, without reducing the resolution of the estimated depth maps. The proposed method allows estimating depth maps of higher quality than the state-of-the-art reference method DERS [91] provided by the MPEG community. The experiments show the mean increase of the quality of virtual views estimated with the proposed method by almost 3 dB over the reference method. Moreover, the method provides a shorter processing time than the reference method – the proposed method (without further proposed enhancements described below) is on average about 4 times faster than the reference method.

- **A novel temporal consistency enhancement method**: depth maps estimated in previous frames are utilised in the estimation of the depth for the current frame, increasing the consistency of depth maps and simultaneously reducing the processing time of estimation. This new improvement allows decreasing the processing time of estimation by up

to 15 times in comparison with the unmodified depth estimation method. Moreover, the significant increase of temporal consistency was confirmed by the substantial improvement of the compression of virtual views synthesised using temporally enhanced depth maps. The results show up to 32% of the bitrate reduction in comparison with the use of depth estimation without the proposed temporal enhancement.

- **A novel parallelisation method** of the α-expansion method for graph-based depth estimation. The proposed parallelisation scheme significantly reduces the processing time of depth estimation (up to 4.5 times when 6 cores of CPU are used) with a negligible loss of the quality of estimated depth maps (0.1 dB).

Furthermore, besides the main achievements presented above, two secondary achievements of the dissertation that increase the usefulness of conducted research are presented in this dissertation:

- **Contribution to multiview test sequences production:** new multiview test sequences were released for the research community together with the depth maps estimated using the method presented in this dissertation. The sequences with provided depth maps were described by the researchers of the MPEG community as ones that enable a satisfying free navigation experience for the end user of the FTV system, confirming the high quality of the presented depth estimation method and its particular usefulness for free-viewpoint television.

- **Contribution to the development of an FTV system**, which comprises the proposed depth estimation method as a substantial part of the system. The reduced complexity of the proposed depth estimation method and the high quality of estimated depth maps led to the implementation of the proposed method in such a practical, operational free-viewpoint television system developed by the Chair of Multimedia Telecommunications and Microelectronics of the Poznań University of Technology.

8.2 Research work done

The research work described in the dissertation required substantial amounts of time used for extensive tests and the implementation of the proposed depth estimation method. The problem of the theoretical assessment of the quality of depth maps is widely known to the research community. Therefore, in order to fully assess the performance of the presented depth estimation method in the scope of free-viewpoint television systems, the author of the dissertation conducted a series of complex and comprehensive experiments that were performed for hundreds of different depth estimation configurations.

The author prepared a framework that includes not only all necessary functions used to estimate depth maps for an input sequence but also allows to perform the assessment of the quality (using virtual view synthesis) and temporal consistency of depth maps (using the encoding of virtual views). It required the author of the dissertation to work with more than 60 000 lines of code, almost 15 000 of which were written by the author.

If the experiments had been performed using only one core of the used CPU, then the time of processing would have been longer than 200 days. For this reason, all computations were performed on 6 computing servers that allow a high degree of parallelisation. The distribution of computing tasks between the servers, together with the supervision of the computations, were performed using scripts that were also prepared by the author.

All depth maps and virtual views computed for experimental verification of the presented method constitute collectively more than two hours of video material, though the used test sequences were not longer than 2 seconds each. The size of all data used for the experiments is almost 500 GB.

The presented depth estimation method is a part of the representation server used in an operational free-viewpoint television system [94] (described in Section 7.2). The implementation of the method designed for the representation server was also prepared by the author. The code written by the author constitutes more than 50% of the code lines used in the representation server and about 35% of all code lines in used in the whole FTV system.

8.3 Conclusions

The thesis of the dissertation states that **it is possible to reduce the processing time of depth estimation and improve the quality of virtual views in free-viewpoint television systems in comparison to the state-of-the-art depth estimation methods by means of image segmentation and temporal consistency enhancement.**

In order to prove the thesis, the author presented a new depth map estimation method that is strictly based on the use of segmentation in the estimation process and compared the proposal with the state-of-the-art method, which also uses graph-based optimisation, but does not utilise segmentation. As the performed comprehensive experiments show (Section 6.1), even when a very low number of segments is used during depth estimation, the quality of depth maps estimated using the proposed method is better than in the state-of-the-art method. Simultaneously, the processing time is significantly shorter for the proposed method.

Moreover, the author proposed a novel method of temporal enhancement of depth maps. The method was shown (in Section 6.4) to significantly improve the temporal consistency of estimated depth maps, which is crucial for the quality of virtual views. In addition, the proposed temporal enhancement method decreases the processing time.

Therefore, both the use of segmentation and temporal enhancement can be used to reduce the processing time of depth estimation and improve the quality of virtual views. It confirms the thesis of the presented dissertation.

As it was underlined in the description of the method presented in Section 4.1, the proposed depth estimation method can be used with any number of arbitrarily located cameras and ensures good inter-view consistency and temporal consistency of estimated depth maps. Moreover, the processing time of depth estimation can be further significantly decreased when the proposed parallelisation method is used (Section 6.5). Therefore, it can be concluded that **the proposed method meets all requirements to be successfully used in versatile free-viewpoint television systems** (Section 2.2).

The particular usefulness of the presented depth estimation method was already confirmed by its implementation in an operational free-viewpoint television system developed by the Chair of Multimedia Telecommunications and Microelectronics of the Poznań University of Technology. Depth maps that can be successfully used for virtual view synthesis can be estimated using the proposed method in under 5 seconds per one frame. Therefore, the overall time from the acquisition of an FTV sequence to the presentation of a free navigation becomes acceptable, which can advance the practical use of FTV systems.

The presented algorithm of depth estimation will be further developed by the author. The future work will focus on the further extension of the versatility of the method by extensive tests of the proposed method using other visual systems, e.g. in multiview systems that utilise lightfield or omnidirectional cameras. The reduction of the number of estimation parameters that have to be provided by the user of the prepared software is also planned.

REFERENCES

[1] R. Achanta and S. Süsstrunk, "Superpixels and Polygons Using Simple Non-iterative Clustering," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, pp. 4895-4904.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua and S. Süsstrunk, "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274-2282, Nov. 2012.

[3] M. Allodi, A. Broggi, D. Giaquinto, M. Patander and A. Prioletti, "Machine learning in tracking associations with stereo vision and lidar observations for an autonomous vehicle," in *2016 IEEE Intelligent Vehicles Symposium (IV)*, Gothenburg, 2016, pp. 648-653.

[4] G. Bjøntegaard, "Calculation of average PSNR differences between RD986 curves," ISO/IEC JTC 1/SC 29/WG 11, Doc. MPEG M15378, Austin, TX, 2001.

[5] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124-1137, Sept. 2004.

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

[7] P. Buyssens, M. Daisy, D. Tschumperlé and O. Lézoray, "Superpixel-based depth map inpainting for RGB-D view synthesis," *2015 IEEE International Conference on Image Processing (ICIP)*, Quebec City, QC, 2015, pp. 4332-4336.

[8] J. Cai and C. Jung, "Image-guided depth propagation using superpixel matching and adaptive autoregressive model," in *2015 Visual Communications and Image Processing (VCIP)*, Singapore, 2015, pp. 1-4.

[9] K. Calagari, M. Elgharib, P. Didyk, A. Kaspar, W. Matusik and M. Hefeeda, "Data Driven 2-D-to-3-D Video Conversion for Soccer," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 605-619, March 2018.

[10] M. Camplani, T. Mantecón and L. Salgado, "Depth-Color Fusion Strategy for 3-D Scene Modeling With Kinect," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1560-1571, Dec. 2013.

[11] B. Ceulemans, S. P. Lu, G. Lafruit and A. Munteanu, "Robust Multiview Synthesis For Wide-Baseline Camera Arrays," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2235-2248, Sept. 2018.

[12] W. Chen, M. J. Zhang and Z. X. Xiong, "Fast semi-global stereo matching via extracting disparity candidates from region boundaries," *IET Computer Vision*, vol. 5, no. 2, pp. 143-150, March 2011.

[13] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603-619, May 2002.

[14] L. Do, G. Bravo, S. Zinger and P. H. N. de With, "Real-time free viewpoint DIBR on GPUs for large base-line multi-view 3DTV videos," in *2011 Visual Communications and Image Processing (VCIP),* Tainan, 2011, pp. 1-4.

[15] M. Domański, A. Dziembowski, A. Grzelka, Ł. Kowalski, D. Mieloch, J. Samelak, O. Stankiewicz, J. Stankowski, K. Wegner, "[FTV AHG] Extended results of Poznan University of Technology proposal for Call for Evidence on Free-Viewpoint Television," ISO/IEC JTC1/SC29/WG11, Doc. MPEG M38246, Geneva, Switzerland, 30 May - 03 June 2016.

[16] M. Domański, M. Bartkowiak, A. Dziembowski, T. Grajek, A. Grzelka, A. Łuczak, D. Mieloch, J. Samelak, O. Stankiewicz, J. Stankowski, K. Wegner, "New results in free-viewpoint television systems for horizontal virtual navigation," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*, Seattle, WA, 2016, pp. 1-6.

[17] M. Domański, A. Dziembowski, T. Grajek, A. Grzelka, Ł. Kowalski, M. Kurc, A. Łuczak, D. Mieloch, R. Ratajczak, J. Samelak, O. Stankiewicz, J. Stankowski, K. Wegner, "Methods of high efficiency compression for transmission of spatial representation of motion scenes," in *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, Turin, 2015, pp. 1-4.

[18] M. Domański, A. Dziembowski, A. Kuehn, M. Kurc, A. Łuczak, D. Mieloch, J. Siast, O. Stankiewicz, K. Wegner, "Experiments on acquisition and processing of video for free-viewpoint television," in *2014 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, Budapest, 2014, pp. 1-4.

[19] M. Domański, A. Dziembowski, T. Grajek, A. Grzelka, K. Klimaszewski, D. Mieloch, R. Ratajczak, O. Stankiewicz, J. Siast, J. Stankowski, K. Wegner, "Demonstration of a simple free viewpoint television system," in *2017 IEEE International Conference on Image Processing (ICIP)*, Beijing, 2017, pp. 4589-4591.

[20] M. Domański, A. Dziembowski, M. Kurc, A. Łuczak, D. Mieloch, J. Siast, O. Stankiewicz and K. Wegner, "Poznan University of Technology test multiview video sequences acquired with circular camera arrangement – "Poznan Team" and "Poznan Blocks" sequences", ISO/IEC JTC1/SC29/WG11, Doc. MPEG M35846, Geneva, 2015.

[21] M. Domański, A. Dziembowski, A. Grzelka, D. Mieloch, O. Stankiewicz and K. Wegner, "Multiview test video sequences for free navigation exploration obtained using pairs of cameras", ISO/IEC JTC1/SC29/WG11, Doc. MPEG M38247, Geneva, 2016.

[22] M. Domański, A. Dziembowski, D. Mieloch, A. Łuczak, O. Stankiewicz and K. Wegner, "A practical approach to acquisition and processing of free viewpoint video," in *2015 Picture Coding Symposium (PCS)*, Cairns, QLD, pp. 10-14.

[23] M. Domański, A. Dziembowski, A. Grzelka and D. Mieloch, "Optimization of camera positions for free-navigation applications," in *2016 International Conference on Signals and Electronic Systems (ICSES)*, Krakow, 2016, pp. 118-123.

[24] M. Domański, A. Dziembowski, T. Grajek, A. Grzelka, D. Mieloch, R. Ratajczak, J. Samelak, O. Stankiewicz, J. Stankowski, K. Wegner, "Free-viewpoint television demonstration for sports events," ISO/IEC JTC1/SC29/WG11, Doc. MPEG M41994, Gwangju, Korea, 22-26 January 2018.

[25] A. Dziembowski, A. Grzelka, D. Mieloch, O. Stankiewicz, K. Wegner and M. Domański, "Multiview synthesis — Improved view synthesis for virtual navigation," in *2016 Picture Coding Symposium (PCS)*, Nuremberg, pp. 1-5.

[26] A. Dziembowski, A. Grzelka, D. Mieloch, O. Stankiewicz and M. Domański, "Depth map upsampling and refinement for FTV systems," in *2016 International Conference on Signals and Electronic Systems (ICSES)*, Krakow, 2016, pp. 89-92.

[27] E. Ekmekcioglu, V. Velisavljevic and S. T. Worrall, "Content Adaptive Enhancement of Multi-View Depth Maps for Free Viewpoint Video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 2, pp. 352-361, April 2011.

[28] J. Fácil, A. Concha, L. Montesano and J. Civera, "Single-View and Multi-View Depth Fusion," *IEEE Robotics and Automation Letters*, vol. 2, no. 4, pp. 1994-2001, Oct. 2017.

[29] L. Fang, N. M. Cheung, D. Tian, A. Vetro, H. Sun and O. C. Au, "An Analytical Model for Synthesis Distortion Estimation in 3D Video," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 185-199, Jan. 2014.

[30] L. Fang, Y. Xiang, N. M. Cheung and F. Wu, "Estimation of Virtual View Synthesis Distortion Toward Virtual View Position," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 1961-1976, May 2016.

[31] S. Foix, G. Alenya and C. Torras, "Lock-in Time-of-Flight (ToF) Cameras: A Survey," *IEEE Sensors Journal*, vol. 11, no. 9, pp. 1917-1926, Sept. 2011.

[32] D. Grois, T. Nguyen and D. Marpe, "Coding efficiency comparison of AV1/VP9, H.265/MPEG-HEVC, and H.264/MPEG-AVC encoders," in *2016 Picture Coding Symposium (PCS)*, Nuremberg, 2016, pp. 1-5.

[33] R. Hartley, A. Zisserman, "Multiple view geometry in computer vision" (2nd ed.), Cambridge Univ. Press, 2004.

[34] J. Hernández-Aceituno, R. Arnay, J. Toledo and L. Acosta, "Using Kinect on an Autonomous Vehicle for Outdoors Obstacle Detection," *IEEE Sensors Journal*, vol. 16, no. 10, pp. 3603-3610, 2016.

[35] C. Hewage and M. Martini, "Reduced-reference quality metric for 3D depth map transmission," in *2010 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video*, Tampere, 2010, pp. 1-4.

[36] C. Hewage, S. Worrall, S. Dogan, S. Villette and A. Kondoz, "Quality Evaluation of Color Plus Depth Map-Based Stereoscopic Video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 304-318, April 2009.

[37] L. Hong and G. Chen, "Segment-based stereo matching using graph cuts," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, 2004, pp. I-I.

[38] H. Hsu, C. K. Chiang and S. H. Lai, "Spatio-Temporally Consistent View Synthesis From Video-Plus-Depth Data With Global Optimization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 1, pp. 74-84, Jan. 2014.

[39] W. Jang and Y. S. Ho, "Efficient disparity map estimation using occlusion handling for various 3D multimedia applications," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 4, pp. 1937-1943, November 2011.

[40] W. Jang, T. Y. Chung, J. Y. Sim and C. S. Kim, "FDQM: Fast Quality Metric for Depth Maps Without View Synthesis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 7, pp. 1099-1112, July 2015.

[41] L. Jorissen, P. Goorts, G. Lafruit and P. Bekaert, "Multi-view wide baseline depth estimation robust to sparse input sampling," in *2016 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, Hamburg, pp. 1-4.

[42] Y. Kang and Y. S. Ho, "High-quality multi-view depth generation using multiple color and depth cameras," in *2010 IEEE International Conference on Multimedia and Expo*, Suntec City, pp. 1405-1410.

[43] W. S. Kim, A. Ortega, P. Lai, D. Tian and C. Gomila, "Depth map distortion analysis for view rendering and depth coding," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, Cairo, 2009, pp. 721-724.

[44] V. Kolmogorov and R. Zabih, "Multi-camera Scene Reconstruction via Graph Cuts," in *Proceedings of the 7th European Conference on Computer Vision-Part III (ECCV '02)*, London, UK, pp. 82-96.

[45] V. Kolmogorov and R. Zabin, "What energy functions can be minimized via graph cuts?," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 147-159, Feb. 2004.

[46] P. Kovacs, "[FTV AHG] Big Buck Bunny light-field test sequences". ISO/IEC JTC1/SC29/WG11, Doc. MPEG M35721, Geneva, 2015.

[47] J. Kowalczuk, E. T. Psota and L. C. Pérez, "Real-time temporal stereo matching using iterative adaptive support weights," in *IEEE International Conference on Electro-Information Technology , EIT 2013*, Rapid City, SD, 2013, pp. 1-6.

[48] S. Krig, "Computer Vision Metrics", Springer International Publishing Switzerland, 2016.

[49] M. Kurc, O. Stankiewicz and M. Domański, "Depth map inter-view consistency refinement for multiview video," in *2012 Picture Coding Symposium*, Krakow, 2012, pp. 137-140.

[50] G. Lafruit, K.Wegner, M. Tanimoto, "FTV software framework," ISO/IEC JTC1/SC29/WG11, Doc. MPEG N15349, Warsaw, Poland, June 2015.

[51] G. Lafruit, M. Domański, K. Wegner, T. Grajek, T. Senoh, J. Jung, P. Kovács, P. Goorts, L. Jorissen, A. Munteanu, B. Ceulemans, P. Carballeira, S. García and M. Tanimoto, "New visual coding exploration in MPEG: Super-MultiView and Free Navigation in Free viewpoint TV," in *2016 Proceedings of the Electronic Imaging Conference: Stereoscopic Displays and Application*, San Francisco, 2016, pp. 1-9.

[52] Z. Lee and T. Q. Nguyen, "Multi-Array Camera Disparity Enhancement," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2168-2177, Dec. 2014.

[53] C. Lee, A. Tabatabai, K. Tashiro, "Free viewpoint video (FVV) survey and future research direction," *APSIPA Transactions on Signal and Information Processing,* vol. 4, Oct. 2015.

[54] G. Lee, B. Li and C. Chen, "Content-adaptive depth map enhancement based on motion distribution," in *2014 IEEE Visual Communications and Image Processing Conference*, Valletta, 2014, pp. 482-485.

[55] J. Lei, L. Li, H. Yue, F. Wu, N. Ling and C. Hou, "Depth Map Super-Resolution Considering View Synthesis Quality," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1732-1745, April 2017.

[56] J. Lei, J. Liu, H. Zhang, Z. Gu, N. Ling and C. Hou, "Motion and Structure Information Based Adaptive Weighted Depth Video Estimation," *IEEE Transactions on Broadcasting*, vol. 61, no. 3, pp. 416-424, Sept. 2015.

[57] T. Lenertz, G. Lafruit, "Modular Parallelization Framework for Multi-Stream Video Processing", in Proceedings of the 2016 ACM on Multimedia Conference, Amsterdam, 2016, pp. 1192-1196.

[58] S. Li, C. Zhu and M. T. Sun, "Hole Filling with Multiple Reference Views in DIBR View Synthesis," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 1948-1959, Aug. 2018.

[59] L. Li, S. Zhang, X. Yu and L. Zhang, "PMSC: PatchMatch-Based Superpixel Cut for Accurate Stereo Matching," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 3, pp. 679-692, March 2018.

[60] Y. Liu, M. Yu, B. J. Li and Y. He, "Intrinsic Manifold SLIC: A Simple and Efficient Method for Computing Content-Sensitive Superpixels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 653-666, March 2018.

[61] J. Liu and J. Sun, "Parallel graph-cuts by adaptive bottom-up merging," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, 2010, pp. 2181-2188.

[62] M. Liu, O. Tuzel, S. Ramalingam and R. Chellappa, "Entropy rate superpixel segmentation," *CVPR 2011*, Providence, RI, 2011, pp. 2097-2104.

[63] W. Liu, X. Chen, J. Yang and Q. Wu, "Robust Color Guided Depth Map Restoration," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 315-327, Jan. 2017.

[64] A. Lu, Y. Zhang and L. Yu, "General Synthesized View Distortion Estimation for Depth Map Compression of FTV," *2016 Data Compression Conference (DCC)*, Snowbird, UT, 2016, pp. 496-505.

[65] S. Marcelino, S. Faria, R. Pepion, P. L. Callet, S. Soares and P. Assuncao, "Quality evaluation of depth map error concealment using a perceptually-aware objective metric," in *2015 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, Lisbon, 2015, pp. 1-4.

[66] K. Matusiak, P. Skulimowski and P. Strumiłło, "Improving matching performance of the keypoints in images of 3D scenes by using depth information," in *2017 International Conference on Systems, Signals and Image Processing (IWSSIP)*, Poznan, 2017, pp. 1-5.

[67] D. Miao, J. Fu, Y. Lu, S. Li and C. W. Chen, "Texture-assisted Kinect depth inpainting," in *2012 IEEE International Symposium on Circuits and Systems*, Seoul, 2012, pp. 604-607.

[68] D. Mieloch, A. Dziembowski, A. Grzelka, O. Stankiewicz and M. Domański, "Graph-based multiview depth estimation using segmentation," in *2017 IEEE International Conference on Multimedia and Expo (ICME),* Hong Kong, pp. 217–222.

[69] D. Mieloch, A. Dziembowski, A. Grzelka, O. Stankiewicz and M. Domański, "Temporal enhancement of graph-based depth estimation method," in *2017 International Conference on Systems, Signals and Image Processing (IWSSIP)*, Poznań, 2017, pp. 1-4.

[70] D. Mieloch, A. Dziembowski, A. Grzelka, "Estymacja głębi dla systemów wielowidokowych," Przegląd Telekomunikacyjny, Vol. 86, no. 6, 2017, pp. 479-482,

[71] D. Mieloch, A. Dziembowski, A. Grzelka, "Segmentacja obrazu w estymacji map głębi," *Przegląd Telekomunikacyjny*, vol. 88, no. 6, 2016, pp. 241-244.

[72] D. Mieloch, A. Grzelka, "Segmentation-based method of increasing the depth maps temporal consistency," International Journal of Electronics and Telecommunication, vol. 64, no. 3, pp. 293-298.

[73] K. Muller, P. Merkle and T. Wiegand, "3-D Video Representation Using Depth Maps," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 643–656, April 2011.

[74] G. Nur, S. Dogan, H. K. Arachchi and A. M. Kondoz, "Impact of depth map spatial resolution on 3D video quality and depth perception," in *2010 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video*, Tampere, pp. 1–4.

[75] G. Olague and R. Mohr, "Optimal camera placement for accurate reconstruction," *Pattern Recognit.*, vol. 35, pp. 927–944, 2002.

[76] J. Park, H. Kim, Y. W. Tai, M. S. Brown and I. S. Kweon, "High-Quality Depth Map Upsampling and Completion for RGB-D Cameras," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5559-5572, Dec. 2014.

[77] Y. Peng, L. Chen, F. X. Ou-Yang, W. Chen and J. H. Yong, "JF-Cut: A Parallel Graph Cut Approach for Large-Scale Image and Video," *IEEE Transactions on Image Processing*, vol. 24, no. 2, pp. 655-666, Feb. 2015.

[78] N. Qian, C.-Y. Lo, "Optimizing camera positions for multi-view 3D reconstruction," in *2015 International Conference on 3D Imaging (IC3D),* Liege, 2015, pp. 1-8.

[79] P. Rahimian and J. K. Kearney, "Optimal camera placement for motion capture systems," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 3, pp. 1209-1221, March 1 2017.

[80] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German Conference on Pattern Recognition (GCPR 2014)*, Münster, Germany.

[81] M. Schmeing and X. Jiang, "Superpixel-Based Disocclusion Filling in Depth Image Based Rendering," in *2014 22nd International Conference on Pattern Recognition*, Stockholm, 2014, pp. 1073-1078.

[82] J. Schneider, J. Sauer and M. Wien, "Enhanced view synthesis prediction for coding of non-coplanar 3D video sequences," *2016 Picture Coding Symposium (PCS)*, Nuremberg, 2016, pp. 1-5.

[83] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, A. Geiger, "A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos," in *2017 Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp.2538-2547.

[84] D. Seo and K. H. Jo, "Multi-layer superpixel-based MeshStereo for accurate stereo matching," in *2017 10th International Conference on Human System Interactions (HSI)*, Ulsan, 2017, pp. 242-245.

[85] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888-905, Aug 2000.

[86] H. Shi, H. Zhu, J. Wang, S. Yu and Z. Fu, "Segment-based adaptive window and multi-feature fusion for stereo matching," *Journal of Algorithms & Computational Technology*, vol. 10, no. 1, 2016, pp.3-11.

[87] D. De Silva, W. Fernando, S. Worrall and A. Kondoz, "A novel depth map quality metric and its usage in depth map coding," in *2011 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, Antalya, 2011, pp. 1-4.

[88] M. Sizintsev and R. P. Wildes, "Spatiotemporal Stereo and Scene Flow via Stequel Matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, no. 6, pp. 1206-1219, June 2012.

[89] P. Skulimowski and P. Strumiłło, "Obstacle localization in 3D scenes from stereoscopic sequences," in *2007 15th European Signal Processing Conference*, Poznan, 2007, pp. 2095-2099.

[90] Y. Song and Y. S. Ho, "High-resolution depth map generator for 3D video applications using time-of-flight cameras," *IEEE Transactions on Consumer Electronics*, vol. 63, no. 4, pp. 386-391, November 2017.

[91] O. Stankiewicz, "Stereoscopic depth map estimation and coding techniques for multiview video systems", Doctoral dissertation, Poznań University of Technology, 2013.

[92] O. Stankiewicz, K. Wegner, M. Tanimoto and M. Domański, "Enhanced Depth Estimation Reference Software (DERS) for Free-viewpoint Television", ISO/IEC JTC1/SC29/WG11 Doc. MPEG M31518, Geneva, 2013.

[93] O. Stankiewicz, K. Wegner, M. Tanimoto and M. Domański, "Enhanced view synthesis reference software (VSRS) for Free-viewpoint Television", ISO/IEC JTC 1/SC 29/WG 11, Doc. MPEG M31520, Geneva, 2013.

[94] O. Stankiewicz, M. Domański, A. Dziembowski, A. Grzelka, D. Mieloch and J. Samelak, "A Free-viewpoint Television System for Horizontal Virtual Navigation," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2182-2195, Aug. 2018.

[95] O. Stankiewicz, M. Domański and K. Wegner, "Estimation of temporally-consistent depth maps from video with reduced noise," in *2015 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, Lisbon, pp. 1-4.

[96] J. Stankowski, Ł. Kowalski, J. Samelak, M. Domański, T. Grajek and K. Wegner, "3D-HEVC extension for circular camera arrangements," in *2015 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON),* Lisbon, 2015, pp. 1-4.

[97] X. Suau, J. Ruiz-Hidalgo and J. R. Casas, "Real-Time Head and Hand Tracking Based on 2.5D Data," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 575-585, June 2012.

[98] J. Sun, N.-N. Zheng and H.-Y. Shum, "Stereo matching using belief propagation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 787-800, July 2003.

[99] M. Tanimoto, "FTV standardization in MPEG," in 2014 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), Budapest, pp. 1-4.

[100] M. Tanimoto, M. Panahpour Tehrani, T. Fujii and T. Yendo, "FTV for 3-D Spatial Communication," *Proceedings of the IEEE*, vol. 100, no. 4, pp. 905-917, April 2012.

[101] M. Tappen and W. T. Freeman, "Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters," in *Proceedings Ninth IEEE International Conference on Computer Vision*, Nice, France, 2003, pp. 900-906.

[102] G. Tech, Y. Chen, K. Müller, J. R. Ohm, A. Vetro and Y. K. Wang, "Overview of the Multiview and 3D Extensions of High Efficiency Video Coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 35-49, Jan. 2016.

[103] B. Tippetts, D. Jye Lee, K. Lillywhite and J. Archibald, "Review of stereo vision algorithms and their suitability for resource-limited systems", *Journal of Real-Time Image Processing*, vol. 11, no. 1, pp. 5-25, Jan. 2013.

[104] M. Van den Bergh, D. Carton and L. Van Gool, "Depth SEEDS: Recovering incomplete depth data using superpixels," in *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, Tampa, FL, 2013, pp. 363-368.

[105] V. Vineet and P.J. Narayanan, "Solving multi-label MRFs using incremental alpha-expansion move on the GPUs," in *Proceedings of the 9th Asian conference on Computer Vision*, Xi,An, 2009.

[106] X. Wang, C. Zhu, S. Li, J. Xiao and T. Tillo, "Depth filter design by jointly utilizing spatial-temporal depth and texture information," in *2015 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting,* Ghent, 2015, pp. 1-5.

[107] C. Wang, Y. Guo, J. Zhu, L. Wang and W. Wang, "Video Object Co-Segmentation via Subspace Clustering and Quadratic Pseudo-Boolean Optimization in an MRF Framework," *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 903-916, June 2014.

[108] C. Wei, C. K. Chiang and S. H. Lai, "Iterative depth recovery for multi-view video synthesis from stereo videos," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, Siem Reap, 2014, pp. 1-8.

[109] K. Wei, Y. L. Huang and S. Y. Chien, "Point-based model construction for free-viewpoint TV," in *2013 IEEE Third International Conference on Consumer Electronics Berlin (ICCE-Berlin)*, Berlin, 2013, pp. 220-221.

[110] S. Xiang, L. Yu, Q. Liu and Z. Xiong, "A gradient-based approach for interference cancelation in systems with multiple Kinect cameras," in *2013 IEEE International Symposium on Circuits and Systems (ISCAS2013)*, Beijing, pp. 13-16.

[111] P. Yadati and A. M. Namboodiri, "Multiscale two-view stereo using convolutional neural networks for unrectified images," in *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, Nagoya, pp. 346-349.

[112] C. Yao, T. Tillo, Y. Zhao, J. Xiao, H. Bai and C. Lin, "Depth Map Driven Hole Filling Algorithm Exploiting Temporal Correlation Information," *IEEE Transactions on Broadcasting*, vol. 60, no. 2, pp. 394-404, June 2014.

[113] P. Yim, "Acceleration of the Graph Cut with High Performance Computing", technical report, Virtla Scalpel, Inc., available online: http://virtualscalpel.com/technology.html

[114] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proc. 3rd ECCV*, 1994, pp. 151-158.

[115] Y. Zhang et al., "Light-Field Depth Estimation via Epipolar Plane Image Analysis and Locally Linear Embedding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 739-747, April 2017.

[116] G. Zhang, J. Jia, T. T. Wong and H. Bao, "Consistent Depth Maps Recovery from a Video Sequence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 974-988, June 2009.

[117] R. Zhang, Z. Peng, M. Yu, G. Jiang and W. Bi, "A novel depth spatial-temporal consistency enhancement algorithm for high compression performance," in *2011 4th International Congress on Image and Signal Processing*, Shanghai, 2011, pp. 34-37.

[118] Y. Zhang, R. Hartley, J. Mashford and S. Burn, "Superpixels via pseudo-Boolean optimization," in *2011 International Conference on Computer Vision*, Barcelona, 2011, pp. 1387-1394.

[119] J. Zhu, L. Wang, J. Gao and R. Yang, "Spatial-Temporal Fusion for High Accuracy Depth Maps Using Dynamic MRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 899-909, May 2010.

[120] H. Zhu, Q. Wang and J. Yu, "Occlusion-Model Guided Antiocclusion Depth Estimation in Light Field," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 965-978, Oct. 2017.

[121] F. Zilly, C. Riechert., M. Muller., P. Eisert, T. Sikora and P. Kauff, "Real-time generation of multi-view video plus depth content using mixed narrow and wide baseline," *Journal of Visual Communication and Image Representation*, vol. 25, no. 4, pp. 632-648, 2014.

[122] C. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder and R. Szeliski, "High-quality video view interpolation using a layered representation," *ACM Transactions on Graphics*, vol. 23, pp. 600-608, Aug. 2004.

[123] "Call for Evidence on Free-Viewpoint Television: Super-Multiview and Free Navigation", ISO/IEC JTC1/SC29/WG11, Doc. MPEG N15348, June 2015, Poland, Warsaw.

[124] Experiencing New Points Of View With Free Navigation. [Online]. Available: https://mpeg.chiariglione.org/news/experiencing-new-points-view-free-navigation

[125] HEVC reference codec. [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/

[126] Maxflow-v3.01 library. [Online]. Available: http://vision.csd.uwo.ca/code/

[127] "Overview of MPEG-I Visual Test Materials", ISO/IEC JTC1/SC29/WG11, Doc. MPEG N17718, Ljublana, Slovenia, 2018.

[128] "Overview of 3D video coding", ISO/IEC JTC1/SC29/WG11, Doc. MPEG N9784, Archamps, France, May 2008.

[129] "Resolutions for 115th meeting", ISO/IEC JTC1/SC29/WG11, Doc. MPEG N16152, Geneva, CH, June 2016.

[130] SNIC superpixels. [Online]. Available: https://ivrl.epfl.ch/research/snic_superpixels

[131] "Working Draft 1 of Versatile Video Coding (VTM 1)", ISO/IEC JTC1/SC29/WG11, Doc. MPEG W17669, San Diego, USA, April 2018.

PUBLICATIONS OF THE AUTHOR

International journals:

1. O. Stankiewicz, M. Domański, A. Dziembowski, A. Grzelka, D. Mieloch, J. Samelak, „A Free-viewpoint Television System for Horizontal Virtual Navigation," *IEEE Transactions on Multimedia,* vol. 20, no. 8, pp. 2182-2195, Aug. 2018,

2. D. Mieloch, A. Grzelka, „Segmentation-based method of increasing the depth maps temporal consistency," *International Journal of Electronics and Telecommunication*, vol. 64, no. 3, pp. 293-298, 2018,

3. D. Mieloch, O. Stankiewicz, M. Domański, „Depth map estimation for free-viewpoint television," prepared for *IEEE Transactions on Multimedia*.

Polish journals:

1. D. Mieloch, A. Dziembowski, A. Grzelka, „Estymacja głębi dla systemów wielowidokowych," *Przegląd Telekomunikacyjny,* Vol. 86, No. 6, 2017, pp. 479-482,

2. A. Dziembowski, A. Grzelka, D. Mieloch, „Zwiększanie rozdzielczości obrazu i mapy głębi w celu poprawy jakości syntezy widoków wirtualnych," *Przegląd Telekomunikacyjny*, Vol. 86, No. 6, 2017, pp. 405-408,

3. D. Mieloch, A. Dziembowski, A. Grzelka, „Segmentacja obrazu w estymacji map głębi," *Przegląd Telekomunikacyjny*, Vol. 88, No. 6, 2016, pp. 241-244,

4. A. Dziembowski, A. Grzelka, D. Mieloch, „Wielowidokowa synteza w systemach telewizji swobodnego punktu widzenia," *Przegląd Telekomunikacyjny*, Vol. 88, No. 6, 2016, pp. 233-236,

5. M. Domański, A. Dziembowski, A. Kuehn, D. Mieloch, „Telewizja swobodnego punktu widzenia – nowa usługa czy futurystyczna wizja?," *Przegląd Telekomunikacyjny*, No. 8-9/2014, pp. 734-737,

6. A. Dziembowski, A. Kuehn, A. Łuczak, D. Mieloch, K. Wegner, „Realizacja eksperymentalnego systemu telewizji swobodnego punktu widzenia z łukowym ustawieniem kamer," *Przegląd Telekomunikacyjny*, No. 6/2014, pp. 161-164.

Proceedings of conferences:

a. **The flagship worldwide leading IEEE conferences:**

1. M. Domański, A. Dziembowski, T. Grajek, A. Grzelka, K. Klimaszewski, D. Mieloch, R. Ratajczak, O. Stankiewicz, J. Siast, J. Stankowski, K. Wegner, „Demonstration of a simple free viewpoint television systems," IEEE International Conference on Image Processing, Beijing, China, 17-20 Sept. 2017,

2. D. Mieloch, A. Dziembowski, A. Grzelka, O. Stankiewicz, M. Domański, „Graph based multiview depth estimation using segmentation," *IEEE International Conference on Multimedia and Expo ICME 2017,* Hong Kong, 10-14 July 2017,

3. M. Domański, M. Bartkowiak, A. Dziembowski, T. Grajek, A. Grzelka, A. Łuczak, D. Mieloch, J. Samelak, O. Stankiewicz, J. Stankowski, K. Wegner, „New results in free-viewpoint television systems for horizontal virtual navigation," *2016 IEEE International Conference on Multimedia and Expo ICME,* Seattle, USA, 11-15 July 2016,

4. M. Domański, A. Dziembowski, T. Grajek, A. Grzelka, Ł. Kowalski, M. Kurc, A. Łuczak, D. Mieloch, R. Ratajczak, J. Samelak, O. Stankiewicz, J. Stankowski, K. Wegner, „Methods of High Efficiency Compression for Transmission of Spatial Representation of Motion Scenes," *IEEE International Conference on Multimedia and Expo ICME 2015*, Torino, Italy, June 29-July 3 2015.

b. **Other conferences (Published in IEEE Xplore and indexed in Web of Science):**

1. D. Mieloch, „Intrinsic parameters estimation for a multiview system," *International Conference on Signals and Electronic Systems, ICSES 2018,* Kraków, Poland, September 10-12, 2018 (already published as conference proceedings),

2. J. Stankowski, A. Grzelka, D. Mieloch, K. Wegner, „Processing pipeline for real-time remote delivery of virtual view in FTV systems," *International Conference on Signals and Electronic Systems, ICSES 2018,* Kraków, Poland, September 10-12, 2018 (already published as conference proceedings),

3. M. Domański, A. Dziembowski, T. Grajek, A. Grzelka, D. Mieloch, O. Stankiewicz, J. Stankowski and K. Wegner, „Real-time virtual navigation provision by simple means," *International Conference on Signals and Electronic Systems, ICSES 2018,* Kraków, Poland, September 10-12, 2018 (already published as conference proceedings),

4. D. Mieloch, A. Dziembowski, A. Grzelka, O. Stankiewicz, M. Domański, „Temporal Enhancement of Graph-Based Depth Estimation Method," *IEEE International Conference on Systems, Signals and Image Processing IWSSIP 2017*, Poznań, Poland, 22-24 May 2017,

5. A. Dziembowski, A. Grzelka, D. Mieloch, O. Stankiewicz, M. Domański, „Enhancing View Synthesis with Image and Depth Map Upsampling," *IEEE International Conference on Systems, Signals and Image Processing IWSSIP 2017,* Poznań, Poland, 22-24 May 2017,

6. A. Dziembowski, A. Grzelka, D. Mieloch, O. Stankiewicz, K. Wegner, M. Domański, „Multiview Synthesis – improved view synthesis for virtual navigation," *Picture Coding Symposium 2016,* Nuremberg, Germany, 4-7 December, 2016,

7. M. Domański, A. Dziembowski, A. Grzelka, D. Mieloch, „Optimization of camera positions for free-navigation applications," *International Conference on Signals and Electronic Systems, ICSES 2016,* Kraków, Poland, September 5-7 2016,

8. A. Dziembowski, A. Grzelka, D. Mieloch, O. Stankiewicz, M. Domański, „Depth map upsampling and refinement for FTV systems," *International Conference on Signals and Electronic Systems, ICSES 2016*, Kraków, Poland, September 5-7 2016,

9. A. Dziembowski, M. Domański, A. Grzelka, D. Mieloch, J. Stankowski, K. Wegner, „The influence of a lossy compression on the quality of estimated depth maps," *23rd International Conference on Systems, Signals and Image Processing, IWSSIP 2016,* Bratislava, Slovakia, 23-25 May 2016,

10. M. Domański, A. Dziembowski, D. Mieloch, A. Łuczak, O. Stankiewicz, K. Wegner, „A Practical Approach to Acquisition and Processing of Free Viewpoint Video," *31st Picture Coding Symposium PCS 2015*, Cairns, Australia, 31 May - 3 June 2015,

11. M. Domański, A. Dziembowski, A. Kuehn, M. Kurc, A. Łuczak, D. Mieloch, J. Siast, O. Stankiewicz, K. Wegner, „Experiments on acquisition and processing of video for free-viewpoint television," *3DTV Conference 2014,* Budapest, Hungary, 2-4 July 2014.

c. **ISO/IEC MPEG Documents:**

1. A. Dziembowski, D. Mieloch, K. Wegner, O. Stankiewicz, M. Domański, „Proposal of enhanced version of View Synthesis Reference Software with multiple input views," ISO/IEC JTC1/SC29/WG11 MPEG2018, M42941, Ljublana, Slovenia, 16-20 July 2018,

2. M. Domański, D. Losiewicz, T. Grajek, O. Stankiewicz, K. Wegner, A. Dziembowski, D. Mieloch, „Extended VSRS for 360 degree video," ISO/IEC JTC1/SC29/WG11 MPEG2018, M41990, Gwangju, Korea, 22-26 January 2018,

3. M. Domański, A. Dziembowski, T. Grajek, A. Grzelka, D. Mieloch, R. Ratajczak, J. Samelak, O. Stankiewicz, J. Stankowski, K. Wegner, „Free-viewpoint television demonstration for sports events," M41994, Gwangju, Korea, 22-26 January 2018,

4. K. Wegner, O. Stankiewicz, A. Dziembowski, D. Mieloch, M. Domański, „Exploration Experiments on Omnidirectional 6-DoF/3-DoF+ Rendering," ISO/IEC JTC1/SC29/WG11 MPEG2017, M41807, Macau, China, 23-27 October 2017,

5. K. Wegner, O. Stankiewicz, A. Dziembowski, D. Mieloch, M. Domański, „Evaluation of step-in/step-out capability of state-of-the-art view synthesis technology," ISO/IEC JTC1/SC29/WG11 MPEG2017, M40809, Torino, Italy, 17-21 July 2017,

6. K. Wegner, O. Stankiewicz, A. Dziembowski, D. Mieloch, M. Domański, „Omnidirectional 6-DoF/3-DoF+ rendering," ISO/IEC JTC1/SC29/WG11 MPEG2017, M40806, Torino, Italy, 17-21 July 2017,

7. M. Domański, A. Dziembowski, A. Grzelka, Ł. Kowalski, D. Mieloch, J. Samelak, J. Stankowski, O. Stankiewicz, K. Wegner, „Experimental video coding software for Free Navigation applications," ISO/IEC JTC1/SC29/WG11 MPEG2016, M39527, Chengdu, China, 17-21 October 2016,

8. M. Domański, A. Dziembowski, A. Grzelka, Ł. Kowalski, D. Mieloch, J. Samelak, J. Stankowski, O. Stankiewicz, K. Wegner, „Coding results for Poznan Fencing 2 and Poznan Blocks 2 test sequences in Free Navigation scenario," ISO/IEC JTC1/SC29/WG11 MPEG2016, M39215, Chengdu, China, 17-21 October 2016,

9. M. Domański, A. Dziembowski, A. Grzelka, Ł. Kowalski, D. Mieloch, J. Samelak, O. Stankiewicz, J. Stankowski, K. Wegner, „[FTV AHG] Extended results of Poznan University of Technology proposal for Call for Evidence on Free-Viewpoint Television," ISO/IEC JTC1/SC29/WG11 MPEG2016, M38246, Geneva, Switzerland, 30 May - 03 June 2016,

10. M. Domański, A. Dziembowski, A. Grzelka, Ł. Kowalski, D. Mieloch, J. Samelak, O. Stankiewicz, J. Stankowski, K. Wegner, „ [FTV AHG] Technical Description of Poznan University of Technology proposal for Call for Evidence on Free-Viewpoint Television," ISO/IEC JTC1/SC29/WG11 MPEG2016, M37893, San Diego, US, 22-26 February 2016,

11. M. Domański, A. Dziembowski, T. Grajek, A. Grzelka, M. Kurc, A. Łuczak, D. Mieloch, J. Siast, O. Stankiewicz, K. Wegner, „ [FTV AHG] Video and depth multiview test sequences acquired with circular camera arrangement – "Poznan Service" and "Poznan People"," ISO/IEC JTC1/SC29/WG11 MPEG2015, Doc. M36569, Warsaw, Poland, 20-27 June 2015,

12. M. Domański, K. Klimaszewski, A. Dziembowski, D. Mieloch, A. Łuczak, O. Stankiewicz, K. Wegner, „Freeview Navigation (FN) anchor generation using 3D-HEVC with depth for "Poznan Blocks" sequence," ISO/IEC JTC1/SC29/WG11 MPEG2015, Doc. M36565, Warsaw, Poland, 20-27 June 2015,

13. M. Domański, A. Dziembowski, M. Kurc, A. Łuczak, D. Mieloch, J. Siast, O. Stankiewicz, K. Wegner, „Poznan University of Technology test multiview video sequences acquired with circular camera arrangement – "Poznan Team" and "Poznan

Blocks" sequences," ISO/IEC JTC1/SC29/WG11 MPEG2015, Doc. M35846, Geneva, Switzerland, 14-20 February 2015,

14. M. Domański, A. Dziembowski, K. Klimaszewski, A. Łuczak, D. Mieloch, O. Stankiewicz, K. Wegner, „Comments on further standardization for free-viewpoint television," ISO/IEC JTC1/SC29/WG11 MPEG2015, Doc. M35842, Geneva, Switzerland, 14-20 February 2015,

15. M. Domański, D. Mieloch, A. Dziembowski, O. Stankiewicz, K. Wegner, „Super multiview image compression: results for Bee sequence (FTV EE3)," ISO/IEC JTC 1/SC 29/WG 11, Doc. MPEG2014/m35070, Strasbourg, France, 20-24 October 2014,

16. K. Wegner, O. Stankiewicz, A. Dziembowski, D. Mieloch, M. Domański, „First version of depth maps for Poznan Blocks multiview video test sequence," ISO/IEC JTC1/SC29/WG11 MPEG2015, Doc. M32248, San Jose, USA, 13-17 January 2014,

17. M. Domański, A. Dziembowski, A. Kuehn, M. Kurc, A. Łuczak, D. Mieloch, J. Siast, O. Stankiewicz, K. Wegner, „Poznan Blocks - a multiview video test sequence and camera parameters for Free Viewpoint Television," ISO/IEC JTC1/SC29/WG11 MPEG2015, Doc. M32243, San Jose, USA, 13-17 January 2014.

APPENDIX

TABLE A.1. THE PROCESSING TIMES FOR DEPTH MAPS ESTIMATED FOR A DIFFERENT NUMBER OF SEGMENTS USED IN THE ESTIMATION PROCESS

Test sequence	Number of segments						
	1 000	5 000	1 000	25 000	50 000	100 000	150 000
	Processing time of depth estimation per one view and one frame [s]						
Ballet	0.93	2.86	5.91	16.34	62.87	190.48	198.35
Breakdancers	2.09	4.12	6.34	11.62	24.34	50.44	89.50
BBB Butterfly	2.35	4.12	6.20	10.76	20.25	40.81	71.57
BBB Rabbit	0.96	2.80	5.48	13.57	41.12	98.64	99.61
Poznań Blocks	1.00	2.60	6.38	16.98	41.08	119.36	146.17
Poznań Blocks2	2.27	4.15	6.49	12.32	25.96	60.91	114.37
Poznań Fencing2	0.96	2.30	4.27	8.29	16.34	28.07	64.81
Poznań Service2	2.26	4.16	6.23	11.57	23.02	53.94	97.93
Average:	**1.60**	**3.39**	**5.91**	**12.68**	**31.87**	**80.33**	**110.29**

TABLE A.2. THE PROCESSING TIMES FOR DEPTH MAPS ESTIMATED FOR A DIFFERENT NUMBER OF VIEWS USED IN THE ESTIMATION PROCESS

Test sequence	Number of views					
	3	4	5	6	7	8
	Processing time of depth estimation per one view and one frame [s]					
Ballet	38.48	45.91	62.87	60.77	62.32	65.94
Breakdancers	44.97	50.67	50.44	55.12	54.13	57.68
BBB Butterfly	40.03	38.61	40.81	42.18	43.01	43.52
BBB Rabbit	32.99	37.64	41.12	42.87	45.87	46.67
Poznań Blocks	37.23	36.51	41.08	44.26	42.49	44.06
Poznań Blocks2	43.59	48.91	60.91	55.59	67.09	58.07
Poznań Fencing2	14.13	15.44	16.34	17.04	17.55	17.75
Poznań Service2	45.64	50.32	53.94	60.55	58.14	61.93
Average:	**37.13**	**40.50**	**45.94**	**47.30**	**48.83**	**49.45**

TABLE A.3. THE PROCESSING TIMES FOR DEPTH MAPS ESTIMATED FOR A DIFFERENT NUMBER OF P TYPE DEPTH FRAMES USED IN THE ESTIMATION PROCESS

Test sequence	Number of P type depth frames between I type depth frames				
	0	4	9	24	49
	Processing time of depth estimation per one view and one frame [s]				
Ballet	499.15	116.55	62.87	32.90	21.33
Breakdancers	313.13	82.36	50.44	31.64	25.30
BBB Butterfly	209.79	60.17	40.81	29.44	25.60
BBB Rabbit	254.75	66.36	41.12	25.63	20.55
Poznań Blocks	278.94	69.77	41.08	20.11	14.59
Poznań Blocks2	391.17	96.57	60.91	40.90	33.25
Poznań Fencing2	91.75	25.00	16.34	11.57	9.83
Poznań Service2	305.46	84.67	53.94	35.59	29.44
Average:	**293.02**	**75.18**	**45.94**	**28.47**	**22.49**

TABLE A.4. THE PROCESSING TIMES FOR DEPTH MAPS ESTIMATED FOR A DIFFERENT PARALLELISATION TYPES

Test sequence	Parallelisation type						
	None	Interleaved levels of depth			Blocks of depth levels		
	Number of threads used in depth estimation						
	1	2	4	6	2	4	6
	Processing time of depth estimation per one view and one frame [s]						
Ballet	62.87	32.90	18.88	12.96	62.87	29.51	21.08
Breakdancers	50.44	28.69	16.10	12.58	50.44	31.02	18.76
BBB Butterfly	40.81	21.68	12.28	9.96	40.81	23.27	14.16
BBB Rabbit	41.12	20.17	11.64	8.22	41.12	23.57	12.62
Poznań Blocks	41.08	22.08	13.39	10.78	41.08	29.56	11.45
Poznań Blocks2	60.91	32.19	17.32	12.44	60.91	39.06	21.42
Poznań Fencing2	16.34	8.46	4.56	3.33	16.34	10.11	6.08
Poznań Service2	53.94	28.38	15.01	11.18	53.94	30.93	18.49
Average:	**45.94**	**24.32**	**13.65**	**10.18**	**45.94**	**27.13**	**15.51**

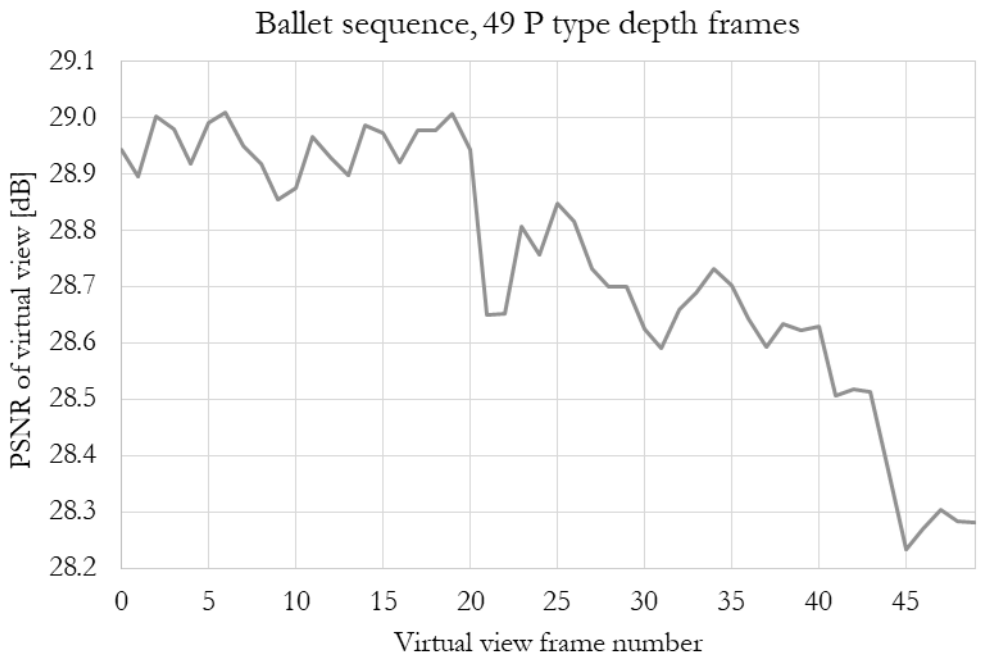


Fig. A.1. The quality of a virtual view for each frame for the "Ballet" sequence.

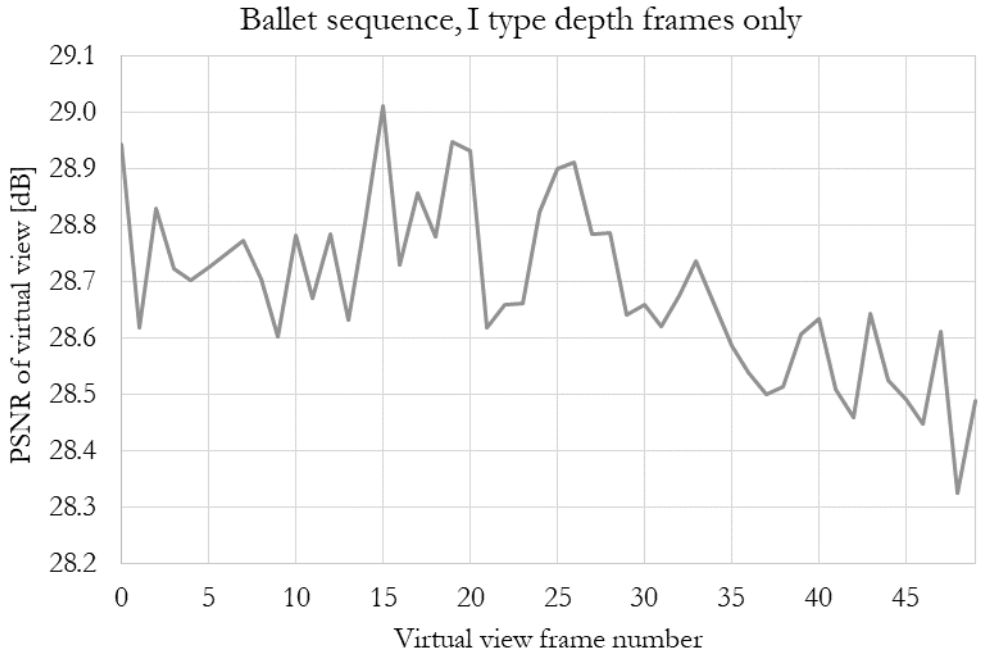


Fig. A.2. The quality of a virtual view for each frame for the "Ballet" sequence.

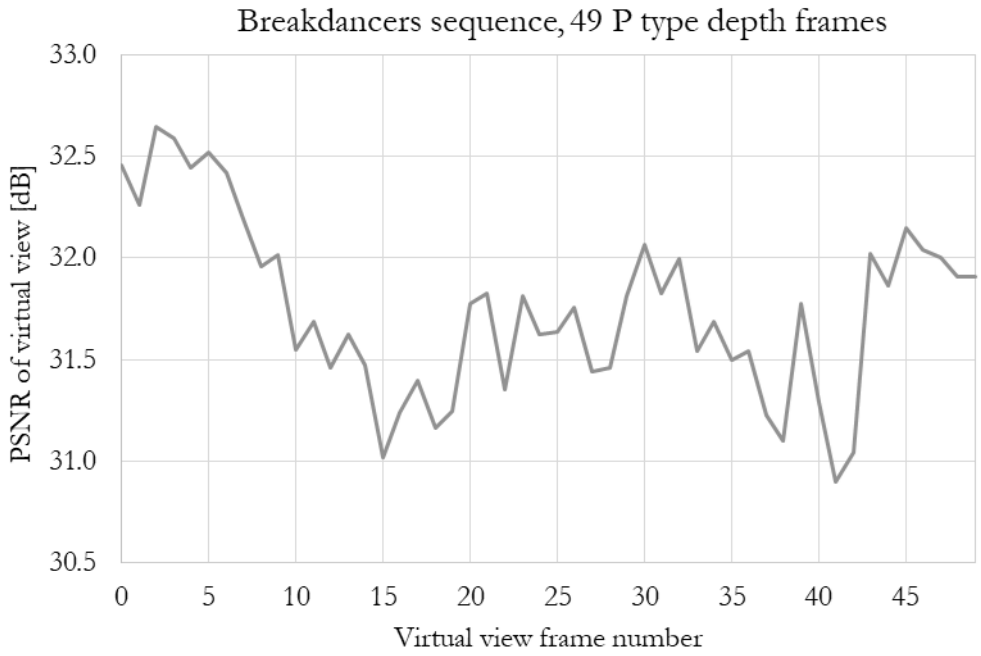


Fig. A.3. The quality of a virtual view for each frame for the "Breakdancers" sequence.

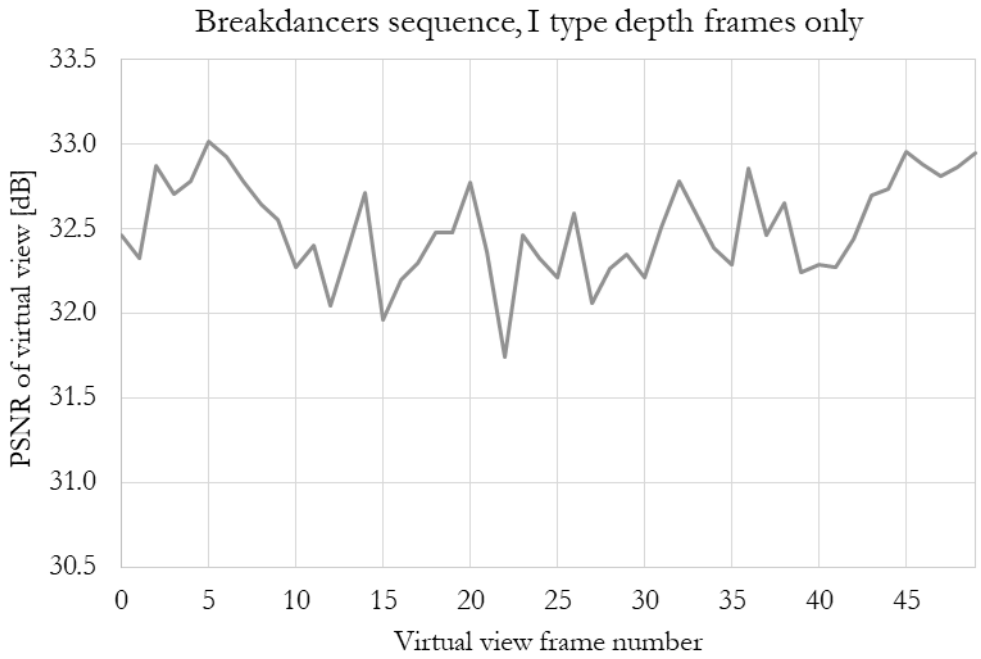


Fig. A.4. The quality of a virtual view for each frame for the "Breakdancers" sequence.

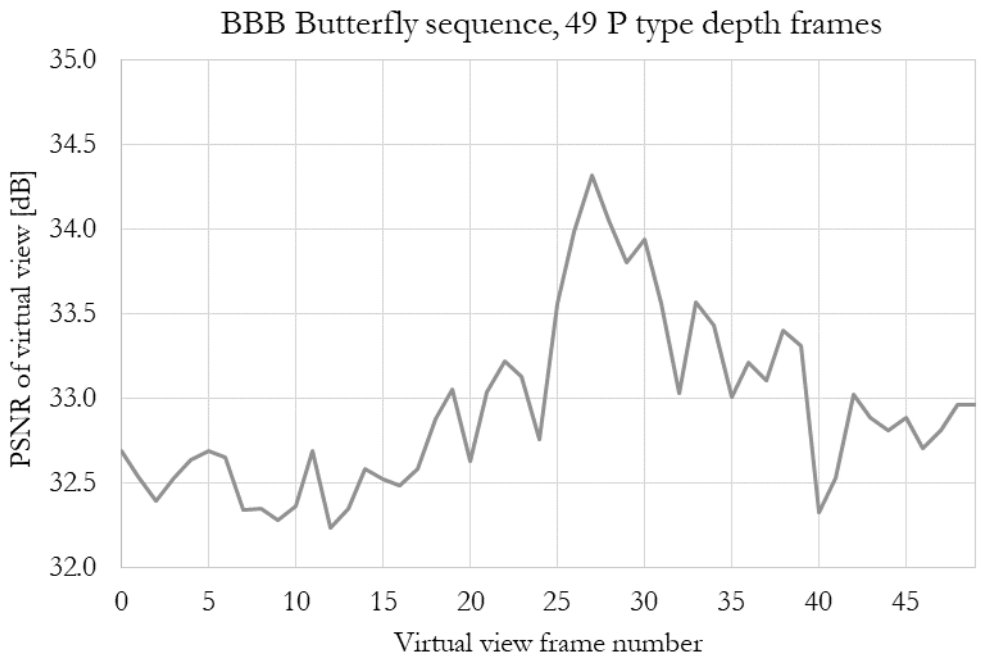


Fig. A.5. The quality of a virtual view for each frame for the "BBB Butterfly" sequence.

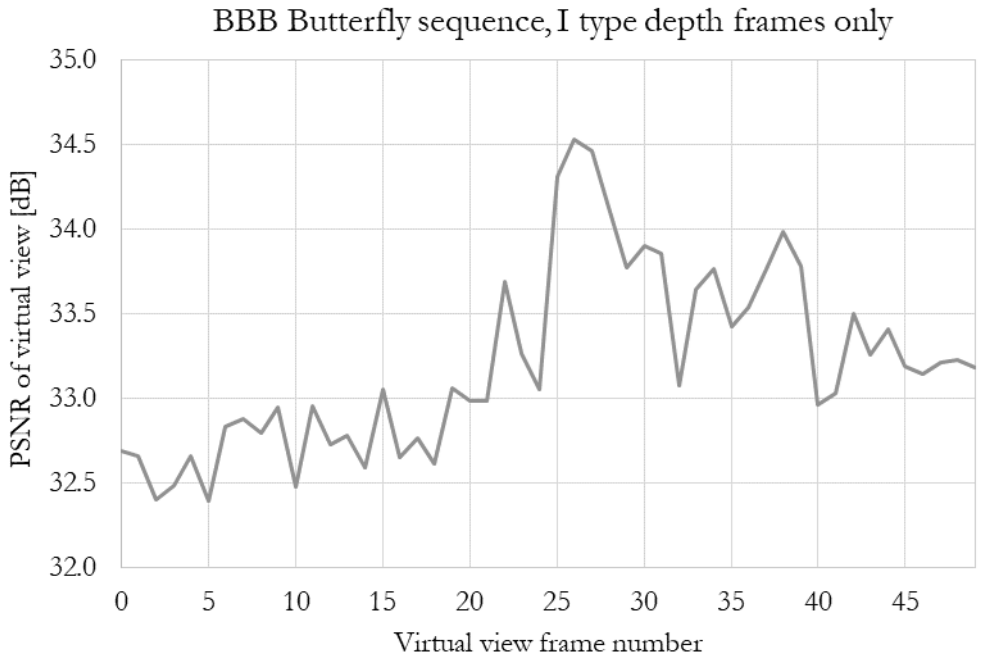


Fig. A.6. The quality of a virtual view for each frame for the "BBB Butterfly" sequence.

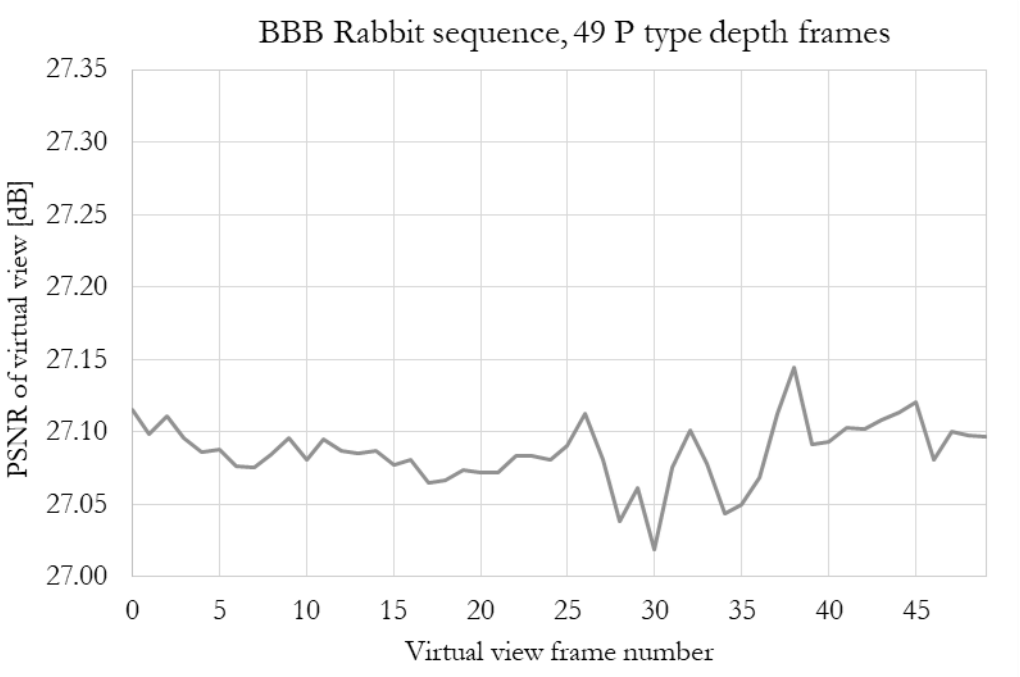


Fig. A.7. The quality of a virtual view for each frame for the "BBB Rabbit" sequence.

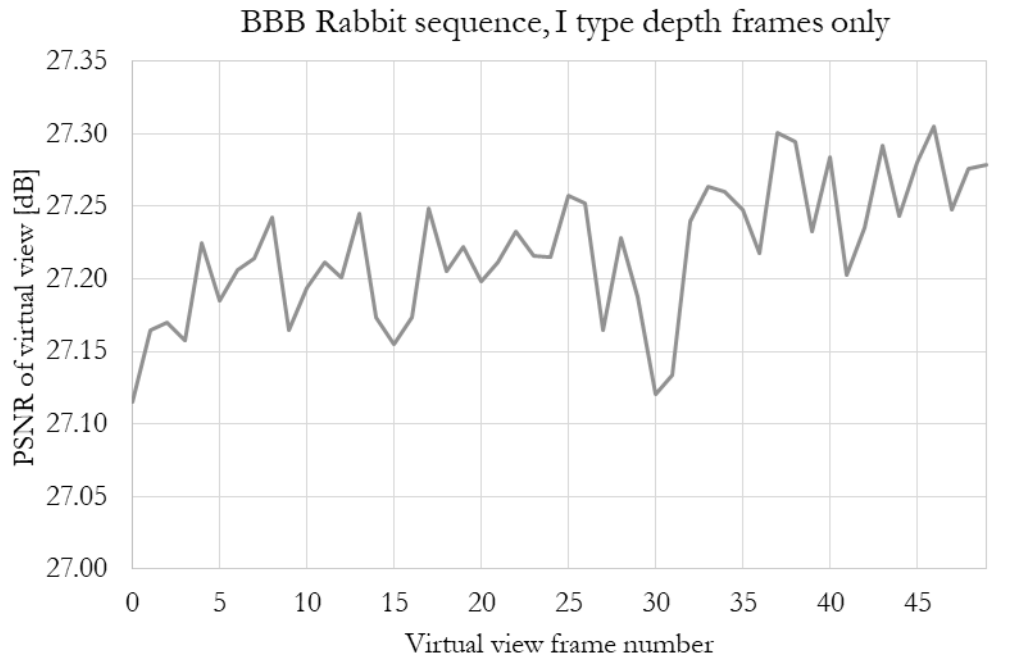


Fig. A.8. The quality of a virtual view for each frame for the "BBB Rabbit" sequence.

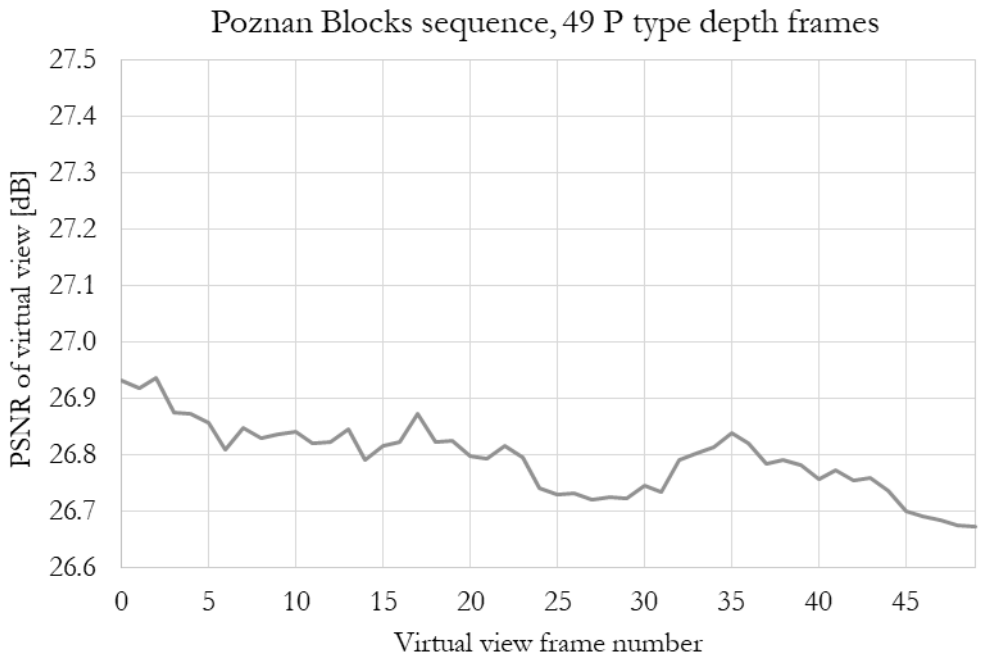


Fig. A.9. The quality of a virtual view for each frame for the "Poznań Blocks" sequence.

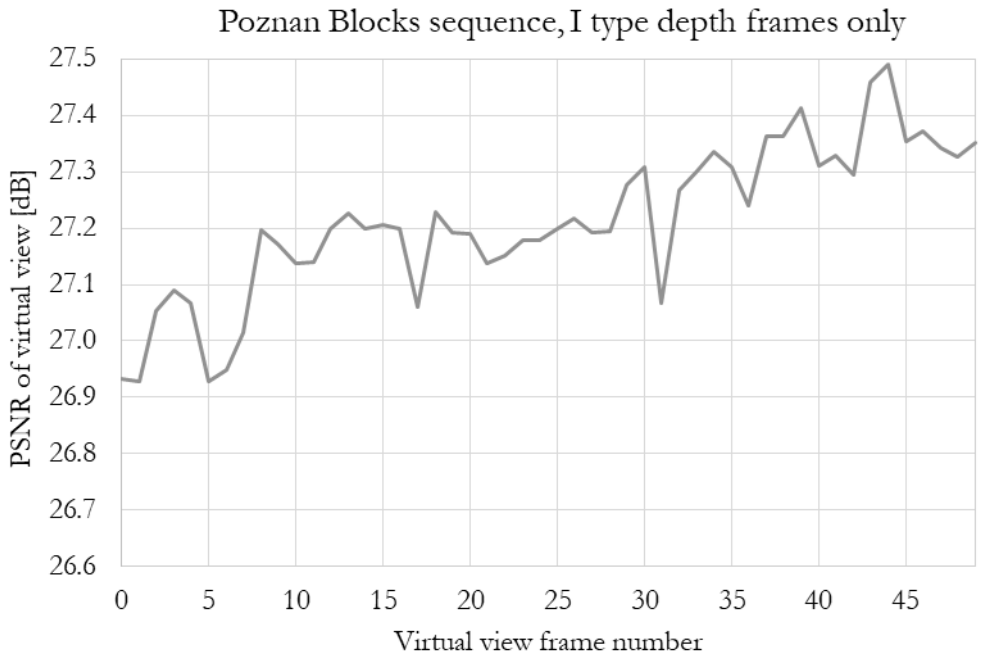


Fig. A.10. The quality of a virtual view for each frame for the "Poznań Blocks" sequence.

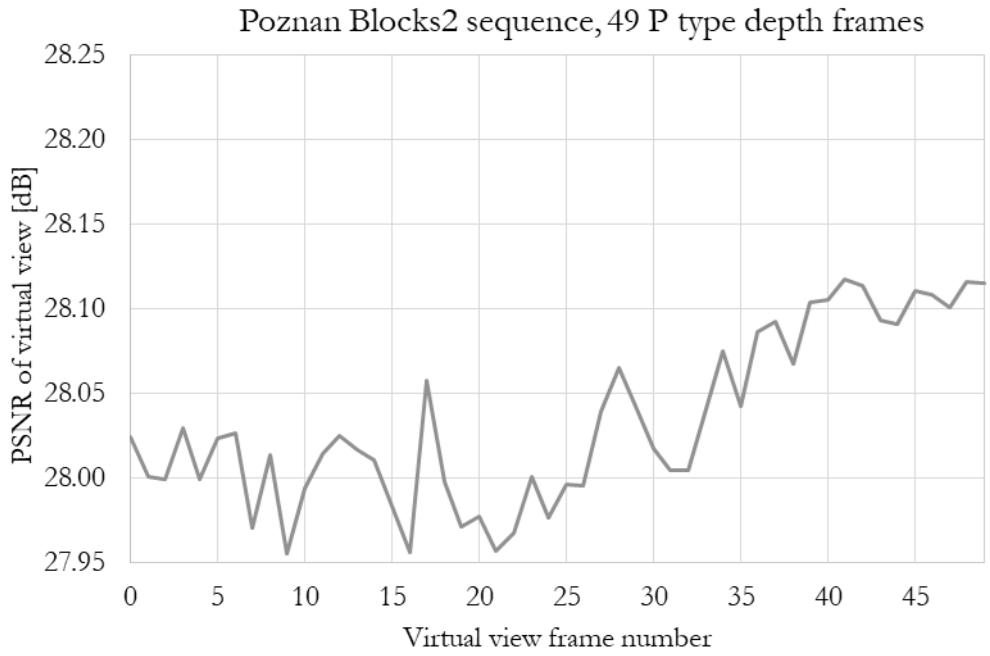


Fig. A.11. The quality of a virtual view for each frame for the "Poznań Blocks2" sequence.

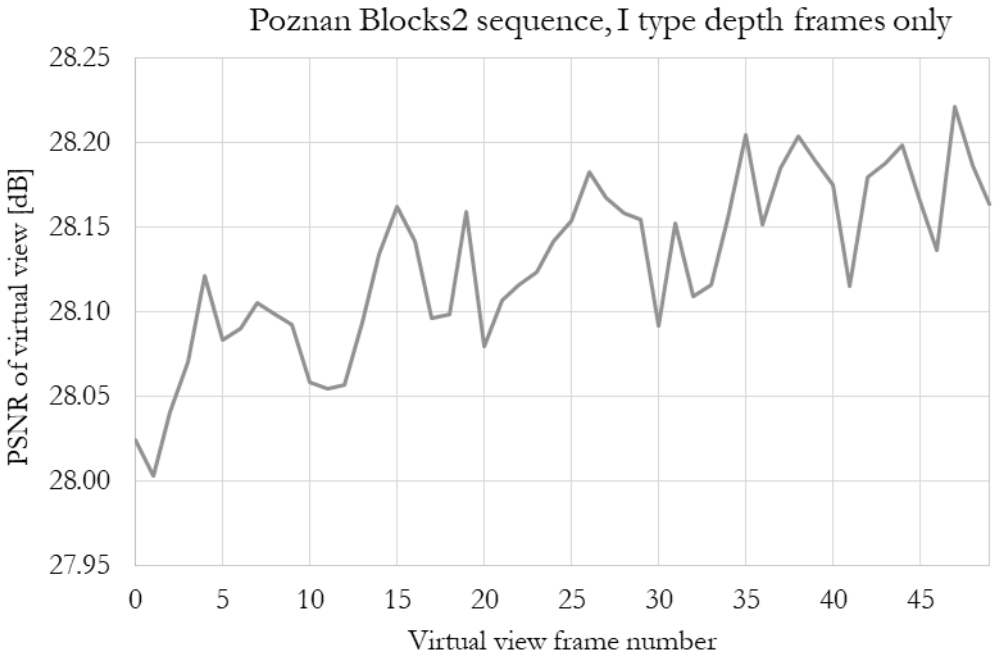


Fig. A.12. The quality of a virtual view for each frame for the "Poznań Blocks2" sequence.

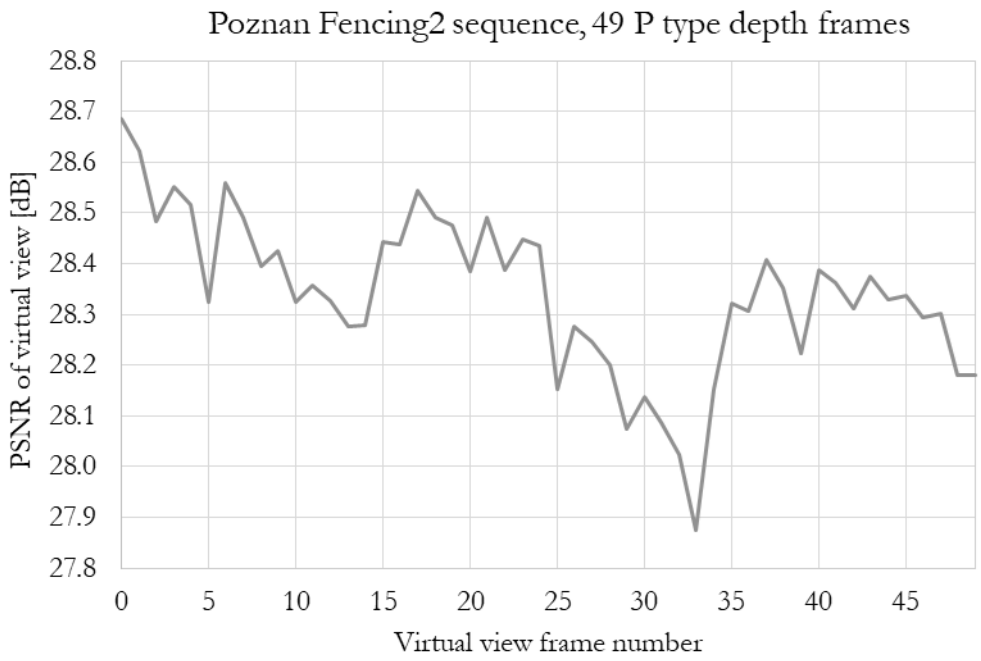


Fig. A.13. The quality of a virtual view for each frame for the "Poznań Fencing2" sequence.

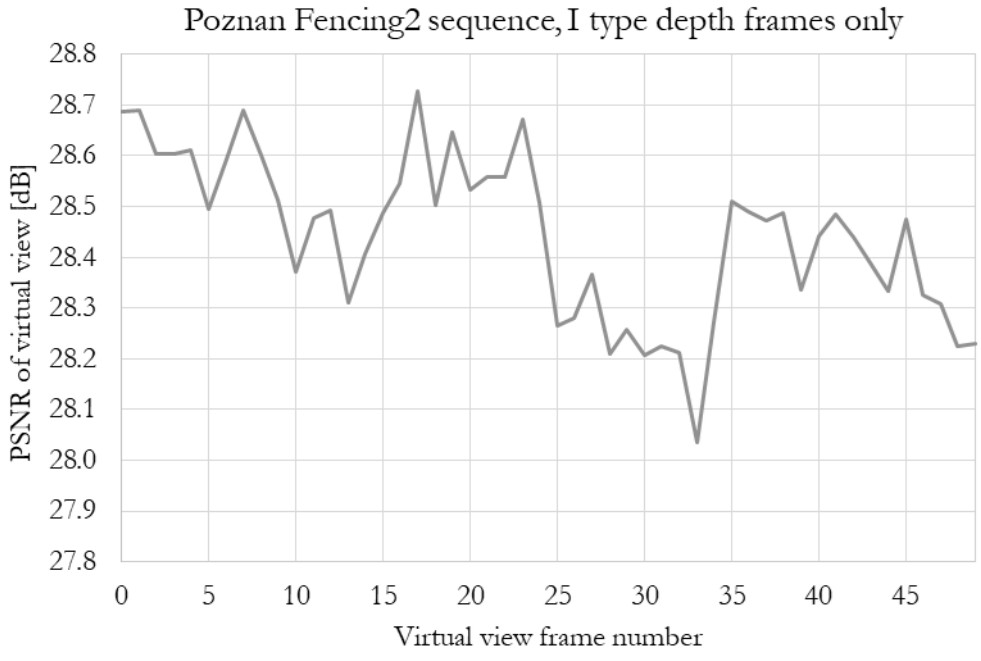


Fig. A.14. The quality of a virtual view for each frame for the "Poznań Fencing2" sequence.

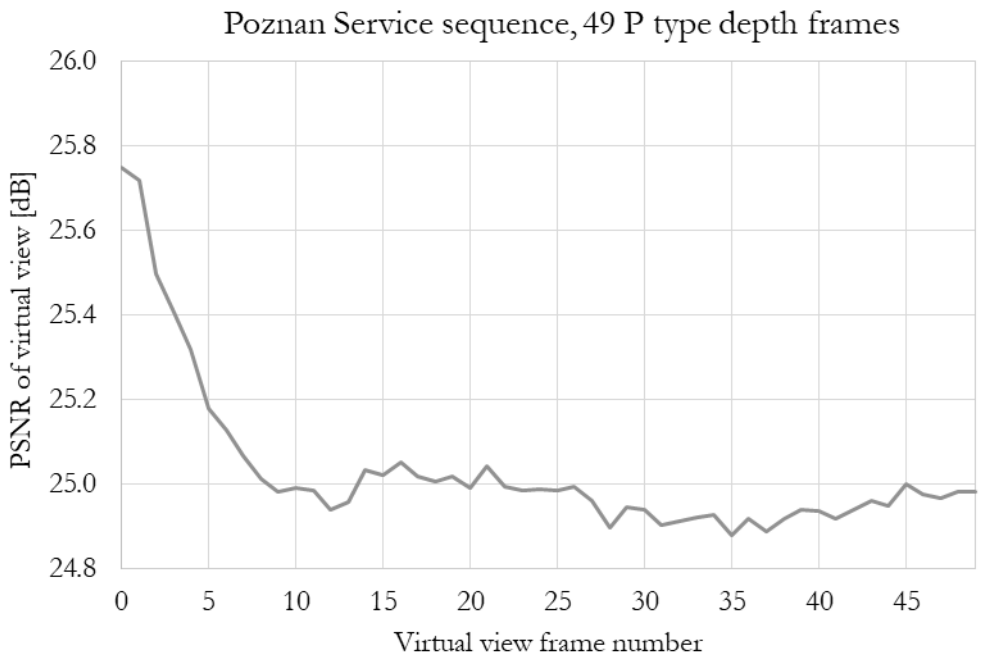


Fig. A.15. The quality of a virtual view for each frame for the "Poznań Service2" sequence.

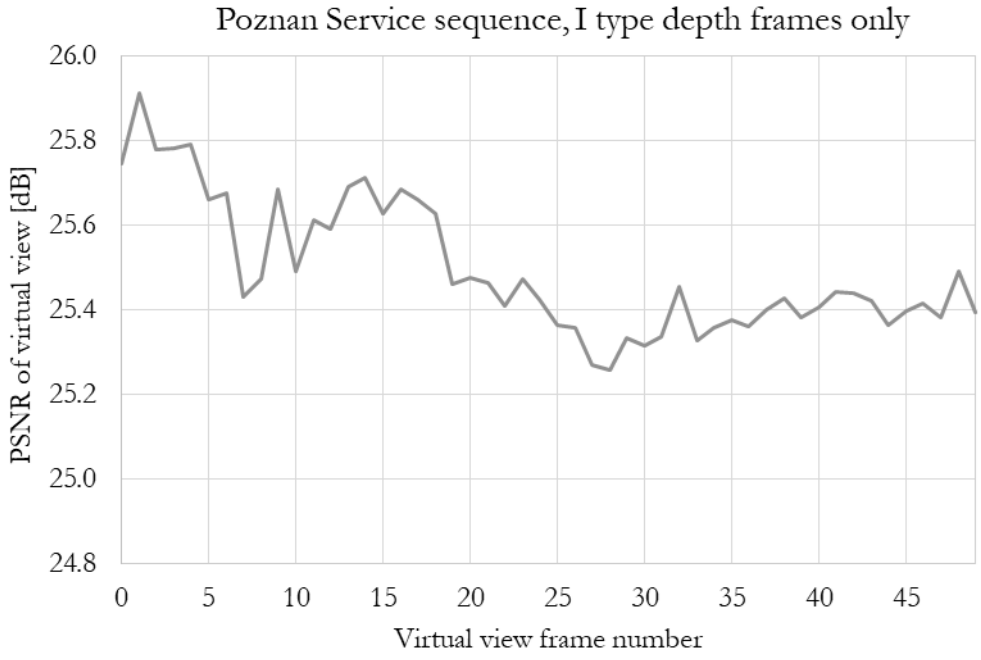


Fig. A.16. The quality of a virtual view for each frame for the "Poznań Service2" sequence.