Poznań University of Technology

Faculty of Electronics and Telecommunications

Chair of Multimedia Telecommunication and Microelectronics

Doctoral Dissertation

Stereoscopic depth map estimation and coding techniques for multiview video systems

Olgierd Stankiewicz

Supervisor: Prof. dr hab. inż. Marek Domański

POZNAN UNIVERSITY OF TECHNOLOGY
Faculty of Electronics and Telecommunications
Chair of Multimedia Telecommunications and Microelectronics

Pl. M. Skłodowskiej-Curie 5
60-965 Poznań

www.multimedia.edu.pl

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

A part of this dissertation related to depth estimation was partially supported by National Science Centre, Poland, according to the decision DEC-2012/07/N/ST6/02267.

A part of this dissertation related to depth coding was partially supported by National Science Centre, Poland, according to the decision DEC-2012/05/B/ST7/01279.

This dissertation has been partially co-financed by European Union funds as a part of European Social Funds.

Online edition 1, 2015
ISBN 978-83-942477-0-6

This dissertation is dedicated to my beloved parents: Zdzisława and Jerzy, who gave me wonderful childhood and all opportunities for development and fulfillment in life.

I would like to thank all important people in my life, who have always been in the right place and supported me in difficult moments, especially during realization of this work.

I would like to express special thanks and appreciation to professor Marek Domański, for his time, help and ideas that have guided me towards completing this dissertation.

Rozprawa ta dedykowana jest moim ukochanym rodzicom: Zdzisławie i Jerzemu, którzy dali mi cudowne dzieciństwo oraz wszelkie możliwości rozwoju i spełnienia życiowego.

Chciałbym podziękować wszystkim ważnym osobom w moim życiu, które zawsze były we właściwym miejscu i wspierały mnie w trudnych chwilach, w szczególności podczas realizacji niniejszej pracy.

Chciałbym również wyrazić szczególne podziękowania oraz wyrazy wdzięczności panu profesorowi Markowi Domańskiemu, za jego czas, pomoc oraz pomysły, które doprowadziły mnie do ukończenia tej rozprawy.

Table of contents

Abstract

The dissertation deals with the problems of stereoscopic depth estimation and coding in multiview video systems, which are vital for development of the next generation three-dimensional television.

The depth estimation algorithms known from literature, along with theoretical foundations are discussed. The problem of estimation of depth maps with high quality, expressed by means of accuracy, precision and temporal consistency, has been stated. Next, original solutions have been proposed.

Author has proposed a novel, theoretically founded approach to depth estimation which employs Maximum A posteriori Probability (MAP) rule for modeling of the cost function used in optimization algorithms. The proposal has been presented along with a method for estimation of parameters of such model. In order to attain that, an analysis of the noise existing in multiview video and a study of inter-view correlation of corresponding samples of pictures have been done.

Also, a novel technique for precision and accuracy enhancement of estimated depth maps is proposed. The technique employs an original Mid-Level Hypothesis algorithm which refines depth map in post-processing.

Yet another, independent achievement of the dissertation is a novel technique for estimation of temporally consistent depth maps with use of noise removal from video prior to the depth estimation itself.

In the dissertation, also, depth coding techniques are discussed. On a background of techniques known from the literature, the problem of depth representation suitable for coding, using legacy compression technology is stated. Author of the dissertation has proposed a novel method of representation of the depth, which employs non-linear transformation, which can be used in order to increase of compression performance in depth map coding.

The proposed non-linear depth representation has been accepted by international group of experts (MPEG) and **adopted to new 3D extensions of ISO/IEC 14496-10 and ITU Rec. H.264 international video coding standards**, describing new generation of 3D video coding technologies, known under names of “MVC+D” and “AVC-3D”.

All of the proposed algorithms have been implemented and their performance has been verified experimentally. The obtained results have been presented in the dissertation.

The following theses have been formulated and proven:

T1) Depth estimation can be improved by usage of modeling of the cost function based on maximization of a posteriori probability.

T2) Precision and accuracy of estimated depth maps can be improved in post-processing with iterative insertion of intermediate values, controlled using view synthesis.

T3) Temporal consistency of estimated depths can be improved using noise removal from input multiview video.

T4) Non-linear representation of depth can be employed in order to improve compression efficiency of depth maps in 3D video systems.

Additional achievement not related directly to the theses yet presented in the dissertation is author’s contribution to production of multiview video sequences that are currently broadly used for test purposes by international research teams, also in research done in the context of standardization in MPEG and JCT-3V expert groups.

In the dissertation, also shown is author’s participation in the development of a 3D video codec, prepared at Chair of Multimedia Telecommunication and Microelectronics of Poznan University of Technology. The codec has been submitted as a proposal for “Call for Proposals for 3D Video Coding Technology” issued by ISO/IEC MPEG group. The excellent results achieved by the proposed codec are shown on the background of exemplary proposals resulting from works of competitive research centers in the world.

Streszczenie

Rozprawa dotyczy problemów wyznaczania i kodowania map głębi stereoskopowej w systemach obrazu wielowidokowego, istotnych dla rozwoju telewizji trójwymiarowej nowej generacji.

W pracy omówiono znane z literatury techniki estymacji głębi wraz z niezbędnymi podstawami teoretycznymi. Wykazano istotę znanego problemu estymacji map głębi charakteryzujących się wysoką jakością, określoną przez dużą dokładność, precyzję i spójność w czasie. Następnie, zaproponowano autorskie rozwiązania problematyki.

Zaprezentowano nowe, bazujące na rozważaniach teoretycznych, podejście do estymacji głębi, opierające się na regule maksymalizacji prawdopodobieństwa a posteriori (Maximum A posteriori Probability) do modelowania funkcji kosztu wykorzystywanej przez algorytmy optymalizacyjne. Przedstawiono również metodę estymacji parametrów takiego modelu. Metoda ta była efektem przebadania szumu występującego w wielowidokowych sekwencjach wizyjnych oraz analizy zagadnienia międzywidokowej korelacji odpowiadających sobie próbek w obrazach.

Zaprezentowano również nową technikę umożliwiającą zwiększanie precyzji i dokładności estymowanych map głębi, poprzez zastosowanie przetwarzania końcowego (postprocessing) z pomocą autorskiego algorytmu hipotezy wartości pośredniej (Mid-Level Hypothesis).

Kolejnym, niezależnym osiągnięciem pracy jest opracowanie i przedstawienie oryginalnej techniki estymacji map głębi, spójnych w dziedzinie czasu, polegającej na wykorzystaniu redukcji szumu w sekwencjach wizyjnych przed dokonaniem samej estymacji.

W rozprawie, rozważono również techniki kodowania głębi. Na tle metod znanych z literatury, wskazano istniejący problem kodowania map głębi stereoskopowej z wykorzystaniem istniejących rozwiązań technologicznych. Następnie, przedstawiono rozwiązanie tego problemu. Autor rozprawy zaprezentował nową metodę reprezentacji głębi, wykorzystującą nieliniową transformację, która umożliwia usprawnienie kompresji map głębi.

Opracowany przez autora sposób nieliniowej reprezentacji głębi (non-linear depth representation) został zaakceptowany przez międzynarodową grupę ekspertów MPEG i **zaadoptowany do specyfikacji nowych rozszerzeń międzynarodowych norm ISO/IEC 14496-10 oraz rekomendacji ITU H.264,** opisujących nową generację technologii

kodowania ruchomego obrazu trójwymiarowego, znanych pod nazwami „MVC+D” i „AVC-3D”.

Wszystkie zaproponowane algorytmy zostały zaimplementowane a ich wydajność sprawdzona eksperymentalnie. Otrzymane wyniki przedstawiono w niniejszej rozprawie.

Sformułowano i udowodniono następujące tezy pracy:

T1) Estymacja głębi może być usprawniona z wykorzystaniem modelowania funkcji kosztu, bazującego na regule maksymalizacji prawdopodobieństwa a posteriori.

T2) Precyzja i dokładność estymowanych map głębi może być zwiększona w kroku przetwarzania kończącego, poprzez iteracyjne wstawianie wartości pośrednich głębi, kontrolowane z pomocą syntezy widoków.

T3) Spójność w dziedzinie czasu estymowanych map głębi może być zwiększona poprzez redukcję szumów w wejściowych sekwencjach wielowidokowych.

T4) Nieliniowa reprezentacja głębi umożliwia zwiększenie sprawności kompresji map głębi w systemach obrazu trójwymiarowego.

Jako dodatkowe osiągnięcie w rozprawie, nie bezpośrednio związane z tezami pracy, przedstawiono udział autora w przygotowaniu wielowidokowych sekwencji wizyjnych, które obecnie są wykorzystywane jako sekwencje testowe przez międzynarodowe zespoły badawcze, także w badaniach prowadzonych przez grupy ekspertów ISO/IEC MPEG i JCT-3V.

W niniejszej rozprawie przedstawiony został również wkład autora w przygotowanie kodeka trójwymiarowych sekwencji wizyjnych, stworzonego w Katedrze Telekomunikacji Multimedialnej i Mikroelektroniki Politechniki Poznańskiej. Kodek ten zgłoszono do udziału w konkursie “Call for Proposals for 3D Video Coding Technology” zorganizowanym przez grupę ISO/IEC MPEG. Doskonałe wyniki uzyskane przez kodek, przedstawiono w rozprawie na tle przykładowych rozwiązań będących wynikiem prac konkurencyjnych ośrodków badawczych z całego świata.

List of terms, symbols and abbreviations

2D	Two-Dimensional
3D	Three-Dimensional
3D-ATM	AVC based 3D Test Model – a reference software developed by MPEG implementing 3D extenstions to AVC (MVC+D and AVC-3D)
3DTV	Three-Dimensional Television
AVC	Advanced Video Coding technology described in ISO/IEC 14496-10:2013 [111] and ITU Rec. H.264 international coding standards
AVC-3D	Common name of "AVC compatible video-plus-depth extension", a 3D video coding technology [117][118][119] that is expected to be described in Annex J of ISO/IEC 14496-10:2012 and ITU Rec. H.264 video coding standard
BJM	Bjøntegaard metric [127] of compression performance
BP	Belief propagation algorithm
CfP	"Call for Proposals on 3D Video Coding Technology" issued by MPEG group [129]
d	Disparity, distance (in pixels) between positions of given point in distinct views
d_{min}	Minimal disparity
d_{max}	Maximal disparity
d_{step}	Quantization step of given disparity representation (minimal step between each of consecutive disparity values), expressed as a multiple of the spatial sampling period in matched images
δ	Normalized disparity, i.e. disparity d scaled to range $0 \ldots \delta_{max}$.
δ_{max}	Maximal normalized disparity value for given representation – e.g. 255 for 8-bit representation.
DERS	Depth Estimation Reference Software [126], the state-of-the-art reference software developed by MPEG
DIBR	Depth-Image-Based Rendering
DSIS	Double Stimulus Impairment Scale, subjective evaluation method [128]
$E[\cdot]$	Expected value operator
EHP	Extended High Profile, a configuration profile of 3D-ATM software [120] reflecting AVC-3D video coding technology
FTV	Free-view Television
GC	Graph cuts algorithm
$H[\cdot]$	Histogram operator
HMM	Hidden Markov Model
HP	High Profile, a configuration profile of 3D-ATM software [120] reflecting MVC+D video coding technology
JCT-3V	ITU-T/ISO/IEC Joint Collaborative Team on 3D Video Coding Extension Development
HEVC	High Efficiency Video Coding technology described in ISO/IEC 23008-2:2013 (MPEG-H Part 2) [121] and ITU Rec. H.265 international coding standards

MAP	Maximum A posteriori Probability
MCNRR	Motion-Compensated Noise Reduction with Refinement technique
MOS	Mean Opinion Score
MPEG	Moving Pictures Experts Group of International Standardization Organization (ISO) and International Electrotechnical Commission (IEC)
MRF	Markov Random Field
MVC	Common name of "Multiview Video Coding" a multiview video coding technology [112][113] described in Annex H of ISO/IEC 14496-10:2012 and ITU Rec. H.264 video coding standard
MVC+D	Common name of "MVC Extension for Inclusion of Depth Maps", a 3D video coding technology [114][115][116] described in Annex I "Multiview and Depth video coding" of ISO/IEC 14496-10:2012 and ITU Rec. H.264 video coding standard
MVD	Multiview Video plus Depth
MV-HEVC	Common name of "Multiview HEVC", a 3D video coding technology [122] currently being under standardization
NDR	Non-linear Depth Representation
Pixel	A fragment of an image, characterized by its coordinates (e.g. x, y), value (e.g. scalar luminance, or vector: red, green and blue) and size, which (in both dimensions) is equal to the sampling period of the image in which given pixel is located
Pixel precision	Level of the detail in which position in image can be expressed, related to full-pixel precision, which corresponds to a single sampling period in the image
PCC	Pearson Correlation Coefficient, also linear correlation coefficient
PSNR	Peak Signal-to-Noise Ratio
QP	Quantization parameter for video
SAD	Sum of Absolute Differences
SBNR	Still Background Noise Reduction technique
SEI	Supplemental Enhancement Information
Smoothing Coefficient	A control parameter of Depth Estimation Reference Software (DERS)
SSD	Sum of Squared Differences
τ	Transformed, coded disparity
τ_{max}	Maximum transformed, coded disparity
VCEG	Video Coding Experts Group
VSRS	View Synthesis Reference Software, a state-of-the-art reference software developed by MPEG [124][125]
WTA	Winner-Takes-All, a brute-force depth estimation technique
z	Distance (z -value) from the view plane of the camera system to given point
z_{near}	The nearest considered distance (z -value) in the camera system
z_{far}	The furthest considered distance (z -value) in the camera system

Chapter 1. Introduction

1.1. The scope of the dissertation

There-dimensional (3D) video gains a lot of attention nowadays. Constantly there is progress in a wide variety of fields related to 3D: from the interest of the customers, through the production of content, the availability of 3D-compatible hardware, to the technology that lays underneath (coding and transmission solutions and standards). Even though there are some skeptic voices about the future of 3D [145], there are strong expectations [144][146][147] that the market of 3D video will extend even further in the upcoming years.

Anyhow, currently merchandised “3D” employs only a pure stereovision – only two views (left and right) are delivered in order to provide depth impression, typically with use of special glasses worn by the viewer.

This work is related to a new generation of 3D video systems which would go beyond the currently applied stereoscopic solutions and their limitations.

The considered features of the next generation of 3D video systems include providing **better impressions of depth**, **better reproduction of the 3D scene** structure and higher level of **interaction** with the user.

The exemplary applications of the next generation of 3D video technology are free viewpoint navigation and glasses-free 3D.

In a **free viewpoint navigation system** (Fig. 1) the viewer can virtually move through the scene and interactively choose a point of observation (view). The selected view, as seen by a virtual camera, is synthetically generated and provided to the user with a classical monoscopic or stereoscopic display. Television systems with such feature are often referred to as Free viewpoint TeleVision (FTV).

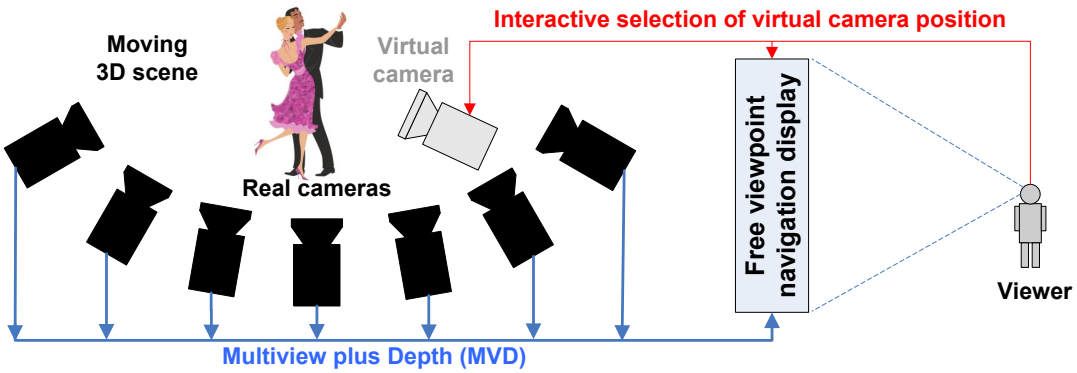


Fig. 1. Free-viewpoint navigation. Depicted “virtual camera” is a camera that does not actually exist in the scene but its content is synthesized from the content of existing ones.

In a **glasses-free 3D video system** (Fig. 2) the depth sensation are provided without requiring the user to wear a special pair of glasses. The effect of three-dimensional impressions is attained by usage of autostereoscopy, where multiple views are displayed simultaneously. The user can slightly change the point of view, resulting in seeing different pair of views shown on the autostereoscopic display. With the currently used technology, change of the position is limited to horizontal parallax only.

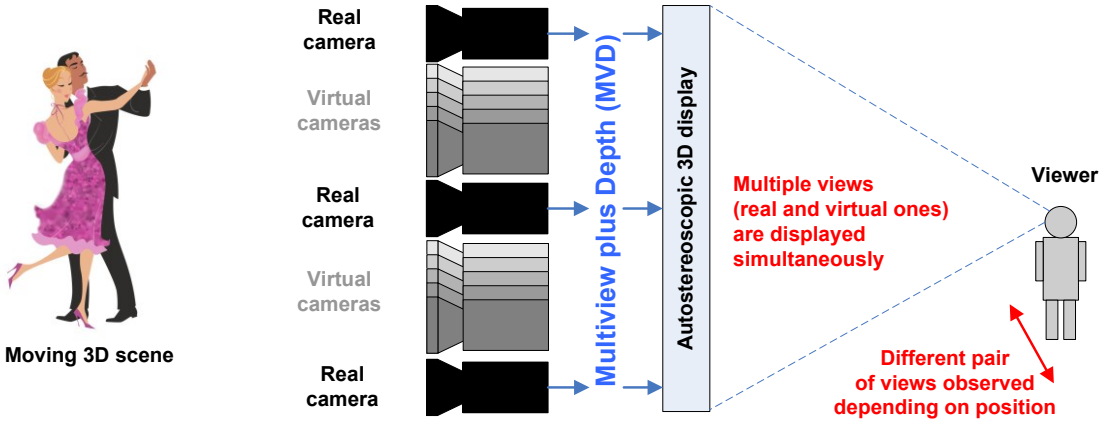


Fig. 2. Glasses-free 3D. Depicted “virtual cameras” are cameras that do not actually exist in the scene, but their content is synthesized basing on the content from existing ones.

The recent works of research laboratories all around the world show that one of the most important aspects of the upcoming 3D video technology is a **method for representation of the 3D scene**, suitable for **efficient coding** and transmission of the 3D video. The current generation of video coding technology broadly available [112][113][130][137], that is applicable for 3D, employs scene representation by means of **multiview video**. In such, the content of the 3D scene is represented by a number of views, observing the scene from different angles and positions. Typically, those views are coded and transmitted in simulcast or with use of simple inter-view predictive schemes that unfortunately do not provide satisfying efficiency of compression. Recent works [2][7][76][148][149] on compression efficiency for delivery of multiview video (e.g. composed of 22 views for an exemplary modern autostereoscopic display) report about asymptotical 30% bitrate reduction related to simulcast videos. The resultant bitrate for all such views is far too high to be accepted by neither the broadcasters nor the market. This stimulates question to arise on **how to achieve a better compression performance**.

An alternative 3D scene representation type, which recently has gained a lot of interest among the researchers [1][2][11][12][130][131][132][133][134], is **Multiview Video plus Depth (MVD)**. In the case of MVD, video streams from multiple camera positions are **transported** along with corresponding information about the depth of the scene in a form of **depth maps**, which carry information about three-dimensional structure of the scene.

Considered in the dissertation, **depth maps** are matrices of values reflecting distances between the camera and points in the scene. Typically, depth maps are presented as gray-scale video (Fig. 3), where the closer objects are marked in high intensity (light) and the far objects are marked in low intensity (dark). **The depth estimation and coding are the fundamental problems in this work.**

a) b)

Fig. 3. The original view (a) and corresponding depth map (b) of a single frame of “Poznan Carpark” [85] 3D video test sequence. In the depth map, the closer objects are marked in high intensity (light) and the far objects are marked in low intensity (dark).

Usage of video and depth maps in MVD representation is an idea considered around the world, because it provides an ability to generate a synthetic view as seen by a virtual camera. Such virtual camera can be placed in an arbitrary position (e.g. impossible in real world) or in position of another real camera. The latter case can be used to predict content of a given view, basing on content in other views [124][125][138][139]. This feature is a key technique [140] for new generation of 3D video framework, where **MVD allows significant reduction of number of views that are directly transported** e.g. instead of 22 videos (required by an exemplary modern autostereoscopic display) only 3 videos with corresponding depth maps are transported. The content of the remaining (not transported) views can be then reconstructed, at the decoder or in the display, basing on compact MVD representation.

This dissertation is related to transmission (or transport) of video bitstreams. Although, in a common sense, these words refer to delivery done through a telecommunication channel (e.g. cable-link, WiFi, terrestrial radio link etc.), the results of this work also apply to storage

as well (a file on CD, DVD, Blu-ray, hard-drive etc.) and therefore, in such a context these words will be used.

The scientific problems related to Multiview Video plus Depth (MVD) instantiate a general scope of the dissertation. Those are as follows:

1. In MVD, additionally to the video**, depth maps have to be transmitted in an efficient way.** Coding of depth maps differs from coding of natural scenes and for now it has been found that usage of classical video coding is inefficient in that case. This comes mainly from the two facts. First, depth maps are matrices composed of scalar values. Although depth maps often are represented as videos, such videos are gray-scale and less textured than natural ones. Secondly, depth maps are more vulnerable to degradation of the edges than natural images or videos, where importance of very sharp edges for quality of subjective sensations is only moderate. Therefore, much work on development of depth-specific coding tools is still required, which is **one of the subjects of the dissertation** (see Chapters 4 and 5).

2. **High quality[1] depth maps are needed** for production of the content and for creation of test sequences. There are many ways to attain depth maps but all have some problems. E.g. for natural scenes, depth maps can be acquired with use of a special depth-sensing cameras. Unfortunately usability of such depth-sensing cameras is handicapped to indoor scenes mostly, due to limited range (often only about 5m) and due to the physical phenomena used (e.g. illumination of infra-red light).

   A more general solution is to algorithmically estimate depth maps basing on images from multiple views, e.g. from stereoscopic pair. Although many solutions are known, **still, algorithmic estimation of the depth is a demanding task,** both with respect to the quality of the estimated depth and computational complexity of the algorithms, **which constitutes another subject of the dissertation** (see Chapters 6 and 7).

3. **Temporal consistency of the depth** is a subject which relates mainly to depth estimation but also negatively impacts performance of the depth coding. Temporal inconsistency of the depth manifests typically as annoying flickering in the video which is synthesized from the input video and the corresponding depth maps. Improvement of temporal consistency of the estimated depth is yet **another goal of the dissertation** (more details on this in Subchapter 2.6).

[1] The meaning of 'quality' of a depth map is considered further in Subection 1.5.3.

The solutions for the above-mentioned issues will be studied in the dissertation in a context of multiview 3D video systems. In this dissertation, by a 3D video system is understood by a structure presented in Fig. 4.

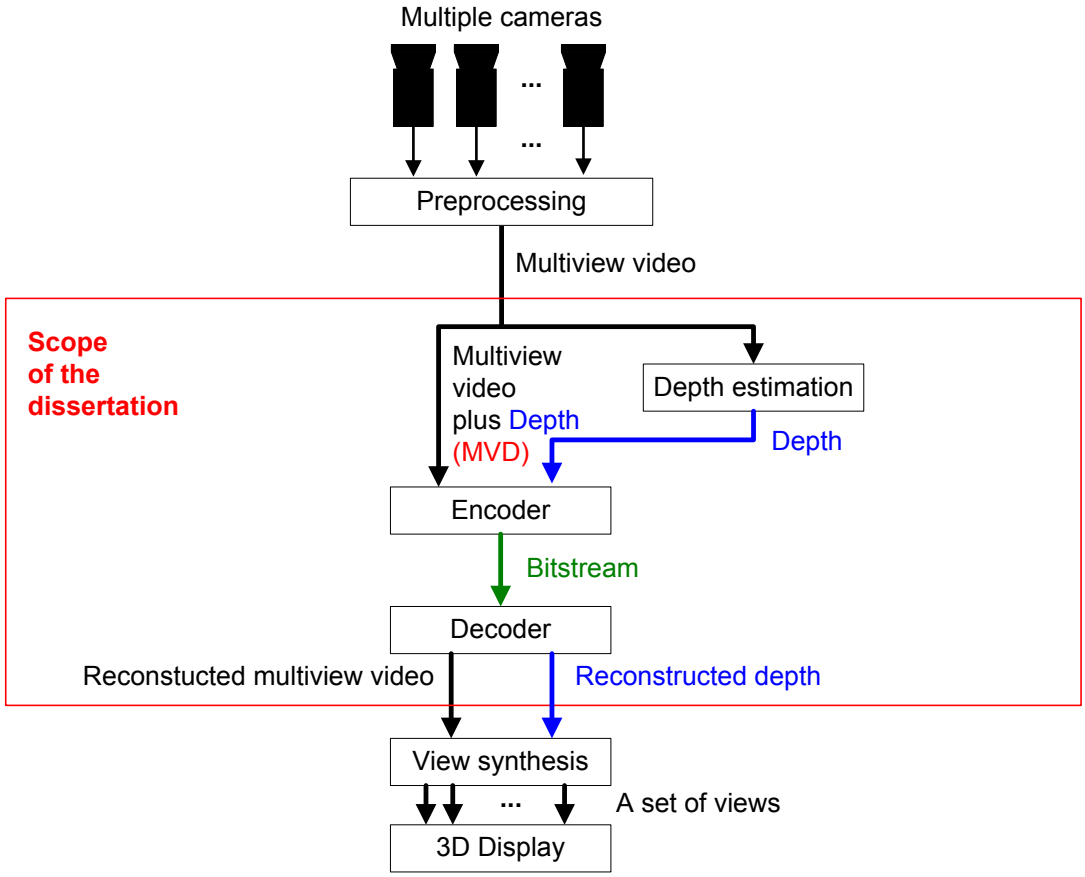


Fig. 4. The scope of the work (marked in red, dotted frame) as a part of the whole 3D video system.

The first stage of processing (Fig. 4) is acquisition of videos from multiple cameras. In the dissertation, no assumptions about the number of cameras is made, but it seems that in practical cases, number of cameras may vary around 3 to 10 [146]. The content of the acquired videos is preprocessed which incorporates: image distortion removal [150][151] rectification [152][154] and color correction [154]. Then, basing on the preprocessed videos, depth maps are estimated with use of depth map estimation algorithm. The next step is lossy compression of the videos (together with the depth data) and coding into a binary stream (bitstream), which is transported to the decoder. The decoder reconstructs the videos along with the depths and then feeds them to the view synthesis algorithm, which generates a set of views that are displayed on a 3D display.

The scope of dissertation within such a 3D video system covers **depth estimation and coding**, which is marked in Fig. 4, inside red-dotted frame.

1.2. The goals and the theses of the dissertation

The goal of this work is to cope with the problems related to development of upcoming generation of 3D video systems. In particular Multi View plus Depth (MVD) scene representation is considered. New proposals for **depth map estimation** with stress on depth quality, disparity precision and accuracy and also temporal consistency will be presented. For **depth map coding** a new proposal for depth representation and compression will be presented.

The theses of the dissertation are as follows:

T1) Depth estimation can be improved by usage of modeling of the cost function based on maximization of a posteriori probability.

T2) Precision and accuracy of estimated depth maps can be improved in post-processing with iterative insertion of intermediate values, controlled using view synthesis quality.

T3) Temporal consistency of estimated depths can be improved using noise removal from input multiview video.

T4) Non-linear representation of depth can be employed in order to improve compression efficiency of depth maps in 3D video systems.

The results for theses T1-T3 are shown in Chapter 3. The results for thesis T4 are shown in Chapter 5.

1.3. The overview of the dissertation

The dissertation is organized as follows. In **Chapter 1,** an introduction to the subject of multiview and 3D video systems is provided. The methodology that has been used during the works on the dissertation is shown. A focus is given to the need of performing experimental verification of the proposals. For that purpose, presented is a set of multiview video sequences, broadly used around the world as test material. A special highlight is given to the author's participation in production of some of those broadly used video test sequences.

Chapter 2 presents current achievements in area of depth estimation with focus on the subjects that are important for the dissertation and further considerations. In particular, global methods are introduced with particular attention given to optimization functions ($DataCost$ and $TransitionCost$) and their probabilistic inclinations, which are further subject of the dissertation.

Chapter 3 describes research performed by the author in area of depth estimation. First, a theoretical model based on Maximum A posteriori Probability is considered. This model is then verified empirically with use of the test sequences and the conclusions are drawn. Basing on the conclusions a novel approach to depth estimation is proposed.

Finally, two more novel algorithms for depth estimation are proposed. The first one, Mid-level Hypothesis algorithm, is aimed at improvement of precision and accuracy of the estimated disparity maps. The second one is aimed improvement of depth temporal consistency with use of noise reduction techniques.

Each of these three achievements is concluded with evaluation of their performance and the experimental results.

In **Chapter 4**, depth coding techniques are discussed with focus on the state of the art directly related to the subjects considered in the dissertation.

Chapter 5 presents research that has been conducted by the author in area of depth representation and coding. A novel idea of non-linear depth representation is presented. First a proof-of-concept idea with use of a simple non-linear function is presented. Then, an original theoretical derivation for non-linear representation of depth is provided. The proposed non-linear depth representation is highlighted as a tool for improvement of compression performance. Experimental verification and study of compatibility with existing coding technology is presented. Finally, adoption of proposed non-linear depth representation to international video coding technology standards developed by ISO/IEC MPEG group (and recommendations of ITU) is highlighted.

Chapter 6 presents achievements of the author that are related to the development of technology for 3D video coding with use of the depth. A 3D video codec is presented that has been developed by Chair of Multimedia Telecommunication and Microelectronics, Poznan University of Technology, as a proposal for "Call for Proposals for 3D Video Coding Technology" [129] issued by ISO/IEC MPEG. The evaluation of the proposal is shown, along with author's share in the work.

In **Chapter 7**, summary and conclusions of the dissertation are presented. The chapter lists the original results of the dissertation.

1.4. The methodology of work

The goal of the dissertation is to study whether it is possible to **improve efficiency in coding of depth** data and whether is it possible to improve **the quality[2] of algorithmically estimated depth maps.**

In both of these problems, theoretical evaluation of the proposals is nearly impossible, because, in order to provide a fair evaluation, the proposed tools (for depth estimation of depth coding) should be evaluated along with several other advanced tools, known from state-of-the-art solutions and proposals (for depth estimation of depth coding respectively).

Therefore, the only reliable way to evaluate advantages and disadvantages of the proposals is by performing series of experiments with multiview video test sequences. Only such allows empirical measurement of coding efficiency and evaluation of the quality of the estimated depth. In order to do that, the author has implemented and integrated the proposed techniques for depth estimation and coding into the following software packages.

For the reasons clearly presented in Chapter 2 the algorithms implemented in ISO/IEC MPEG **Depth Estimation Reference Software (DERS**) and in View Synthesis Reference Software (VSRS) have been used as reference for experimentation in area of depth estimation. Therefore, author's proposals have been implemented and integrated in MPEG reference software (DERS version 5.1 [126] and VSRS version 3.0 [124][125] respectively) and the results have been compared against the original performance of unmodified versions of DERS and VSRS.

In the second part of the dissertation, related to depth coding, the results are presented on the basis of MVC+D [114][115][116] and AVC-3D [117][118][119] video coding technologies (not yet described in any standards when the works have been conducted) and

[2] The meaning of 'quality' of a depth map is considered further in Subsection 1.5.3.

also on the basis of HEVC-based coding technology, co-developed by the author, submitted by **Poznan University of Technology to the "Call for Proposals for 3D Video Coding Technology**" [129] issued by MPEG. Therefore, the tools proposed by the author have been implemented in ISO/IEC MPEG test model software – **3D-ATM** [120] (for MVC+D and AVC-3D) and **HTM** [123] (for HEVC-based).

In both depth estimation and depth coding, the mentioned software packages (in the original versions and with integrated author's proposals) have been used to perform thoughtful experimentation. The results allowed the author to perform examination of the performance of the proposals both by objective manners (like with usage of **PSNR** values or **Bjøntegaard** measures) and subjective manners (subjective test and **Mean Opinion Score** ratings).

Basing on that, the conclusion have been drawn which provided directions for further works.

1.5. Multiview video test sequences

As mentioned in the previous Section, a reliable way to assess performance of algorithms in the two fields related to the dissertation, which are depth estimation and depth coding, is performing series of experiments with multiview video sequences.

It is a general problem, as reliable evaluation of performance is needed in research works which relate to algorithms and tools currently known, developed, or e.g. submitted as proposals for adoption in international coding standards. There are two sides of this problem:

- What test data, in form of test sequences or images, should be used to allow common ground and reference for comparisons?
- How to express and assess the quality of algorithmically generated depth maps?

In the area of evaluation of depth estimation algorithms, an interesting scientific undertaking is related with webpage of Middlebury University [142]. The site is a repository for computer vision datasets and evaluations of related algorithms. Also, the site present results of multiple state-of-the-art depth estimation algorithms evaluated under the same conditions. Unfortunately, the methodology proposed by the authors can be found inadequate for experimentation with multiview and 3D television:

First of all, the site evaluates quality of the depth, basing on still images, which disallows observation of temporal effects and artifacts, which are very important in case of moving pictures, considered in multiview and 3D video system.

Secondly, the methodology of evaluation of depth estimation algorithms used in the webpage is based on comparison with ground-truth depth maps. Conformance with the real-depth, although important in case of many research fields (like computer vision, robotics etc.) in not the primary goal in the case of 3D video systems, where the depth maps are used mainly for the sake of virtual view synthesis.

Moreover, the datasets in Middlebury webpage do not contain objects with specular reflections, glossy surfaces (e.g. mirrors), partially transparent surfaces (windows) etc. Such effects occur in real world natural scenes and lack of such examples in Middlebury data set belittles its usefulness.

Therefore, currently the most adequate known methodology of evaluation depth estimation algorithms has been developed during the works of ISO/IEC MPEG group. The author of the dissertation is an active contributor to this works. In particular, he participated in creation of multiview video test sequences adopted to multiview video sequences set [129], currently, broadly used for test purposes in experiments on development of 3D-related technologies [137][236][237][238]. This mentioned evaluation method developed in works of MPEG, used as main objective assessment method thorough the dissertation, will be described further in Subsection 1.5.3. Before that, first, in Subsection 1.5.1, the production of multiview video sequences at Poznan University of Technology, Chair of Multimedia Telecommunications and Microelectronics will be provided, in which the author had strong participation. Later, selected multiview video test sequences set will be shown in Subsection 1.5.2.

1.5.1. Production of the test material at Poznan University of Technology

For research purposes and for production of multiview video test material, Chair of Multimedia Electronics and Telecommunications, has been built an experimental framework for works on future 3D television.

Fig. 5. A set of 9 Canon XH-G1 cameras used in multiview system, developed at Poznan University of Technology, Chair of Multimedia Telecommunications and Microelectronics.

The system consists of 9 cinematic Canon XH-G1 cameras (Fig. 5) placed on a mobile (wheeled) metal rig (Fig. 6). The rig has been manufactured exclusively to provide special mounting pads that allow precise alignment of the cameras. The output video signal is HDTV (1920x1080) and is provided via SDI interface. All streams are temporally synchronized with use of a GenLock and captured by a PC cluster. The whole post-processing is done offline.

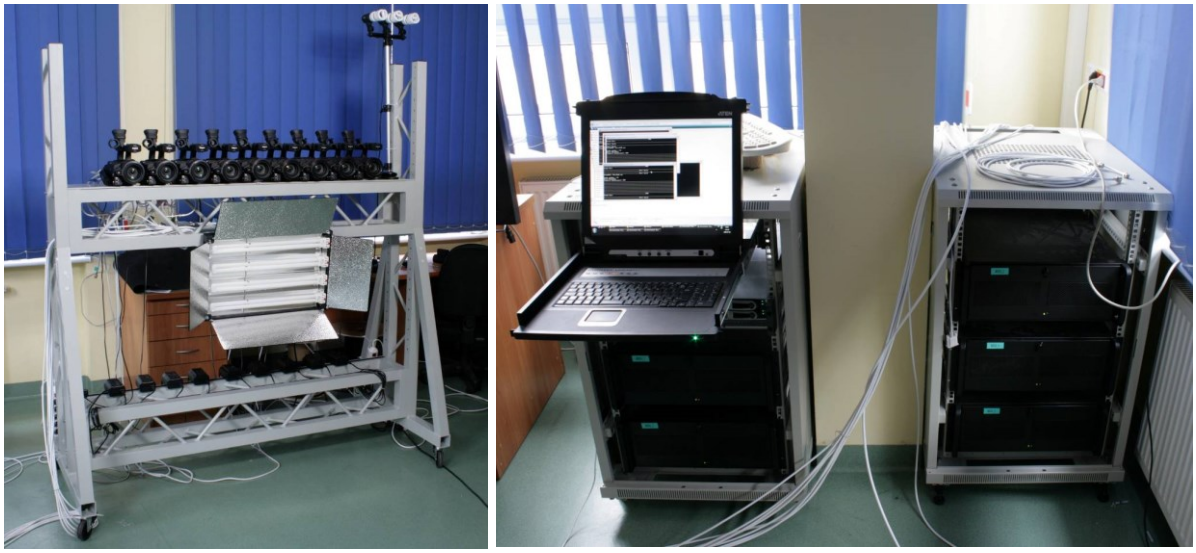


Fig. 6. Multi-camera rig (left) and recording system (right), both developed at Poznan University of Technology, Chair of Multimedia Telecommunications and Microelectronics, Poland.

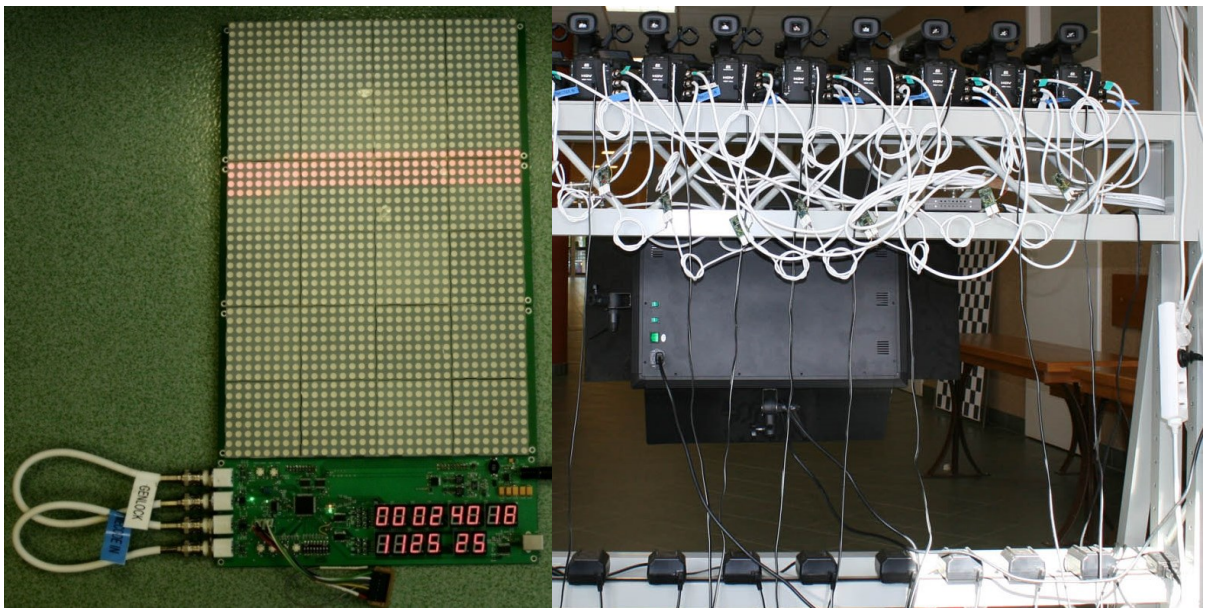


Fig. 7. Electronic board for multi-camera synchronization testing (left) and synchronization circuitry connected to the cameras (right), both developed at Poznan University of Technology, Chair of Multimedia Telecommunications and Microelectronics, Poland.

In order to estimate accuracy of synchronization a special calibration board (Fig. 7) has been used. The calibration board, designed at the Chair of Multimedia Telecommunication and Microelectronics, contains a control system, which is synchronized with the same signal that is propagated to cameras by GenLock and TimeCode inputs. The board presents the current time marker, number of frame lines and frame frequency on seven-segment displays. Simultaneously, the diodes corresponding to a single image line are lightened. Each diode emits light only in a single interval.

The board is placed in front of the cameras. The view from every camera should show the same diodes switched on (number of lighting diodes depends on exposition time). If cameras are not synchronized the view is different. The board also allows for observing the camera synchronization process. Usually camera adjusts its inner clock to external synchronization signal in about 1 second.

Fig. 8. Preparation for production of “Poznan Hall” and “Poznan Carpark” and “Poznan Street” sequences [85] at Poznan University of Technology, Chair of Multimedia Telecommunications and Microelectronics, Poland.

After development of the system hardware it was possible to start production of multiview view sequences. The goal of such works was to develop a set of professional test sequences suitable for research works in the area of future 3D television. However a set of such sequences has already been maintained by ISO/IEC MPEG group, but is has turned out to be too simple (both in terms of the content and the depth of the scene) to meet the demands of research concerning, inter alia, algorithms for determining the depth maps.

For example, the sequences [240] are characterized by uncomplicated textures and little motion. In addition, those sequences [240] present indoor scenes with low dynamic range of the depth and uncomplicated 3D structure (such as “Alt Moabit”) or external scenes with high dynamic range of depth, but laminar (planar) structure of the depth. The “Newspaper” sequence [239] in turn, although it contains more motion and has a more complicated depth structure, its usefulness is limited due to overexposure. Sequences “Champagne Tower” and “Pantomime” [242] present interesting motion and are also challenging in terms of transparent objects, however, those are indoor sequences, filmed against a black background which occupies most of the stage. The “Lovebird” (1 and 2) [241] are outdoor sequences but they lack complex movement therein and do not present complicated structure of the depth in the scene. Furthermore, all of the sequences were filmed using a fixed, not-moving, set of CCTV cameras.

Lack of availability of satisfactory test sequences composed of good quality, caused MPEG- FTV group to announce a Call [135] to provide new, more advanced three-dimensional video content. In particular, desirable were multiview video sequences that would meet the following demands:

- diversification of the filmed content,
- high dynamic range of the depth and complexity of its structure,
- good lighting (overexposure/underexposure),
- movement of the camera,
- complex motion in the scene,
- reflective and transparent surfaces,
- presentation of both indoor and outdoor natural scenes.

Therefore, the goal of production of multiview video sequences at Chair of Multimedia Telecommunication and Microelectronics was to meet the above-mentioned requirements.

Three test sequences have been produced – “Poznan Street”, “Poznan Carpark” and “Poznan Hall”:

The first one, “Poznan Street”, presents an outdoor scene moving people and driving cars (Fig. 9), produced near the office of Poznan University of Technology, Faculty of Electronics and Telecommunications, located at Polanka Street, Poznań, Poland.

The scene of the second one, “Poznan Carpark”, is located in the backyard of the mentioned office. The sequence also presents moving cars and walking persons.

The third sequence, “Poznan Hall”, has been produced in the hall of the mentioned office building. It presents an indoor scene with moving camera and complicated motion, including walking persons, rotating umbrella etc.

All of the produced test sequences have been acquired along with all data required for calibration of the system. Basing on that, rectification and color calibration of the sequences have been performed [81].

The sequences have been submitted [85] to MPEG as a response to the Call [135]. During the forthcoming considerations and works, depth maps for those sequences have been algorithmically estimated [82]. This was crucial because the generated depth maps inclined to serve as ground truth data in the future. This task has consumed hundreds of hours of experiments with finding the optimal settings and creating manual data for semi-automatic depth estimation. The author of the dissertation had one of the biggest shares in this work, which has been done also in cooperation with other research centers [136].

Two of the sequences produced by Chair of Multimedia Telecommunication and Microelectronic have been adopted to the test sequence set used by MPEG, and also broadly around the world, for research on technologies related to 3D. Specifically “Poznan Street” and a part of “Poznan Hall” sequence (named “Poznan Hall 2”) have been adopted.

The author of the dissertation had strong influence on production of these sequences, both on the content and on the technical and scientific sides related to them.

The adoption of these sequences as MPEG test material is a strong indication that the requirements of the Call [135] have been met.

1.5.2. Test sequences used in the dissertation

Along with the sequences produced by Chair of Multimedia Telecommunication and Microelectronics, other sequences maintained by ISO/IEC MPEG group as 3D test material have been selected for test purposes in this dissertation. All of the sequences can be used for research purposes [85][81][82][129][137][236][239][240][241][242][243][244].

“Poznan Street” sequence [85]

“Poznan Hall” sequence [85]

“Poznan Carpark” sequence [85]

“Undo Dancer” sequence [244]

“Balloons” sequence [242]

“Kendo” sequence [242]

“Lovebird 1” sequence [241]

“Newspaper” sequence [239]

Fig. 9. Exemplary frames from multiview video test sequences.

Table 1. Multiview video sequences used for test purposes in the dissertation and their characteristics.

<table>
<tr><th>Sequence name</th><th>Width x Height</th><th>Frame rate [frames/s]</th><th>Camera</th><th>Total frames</th><th>Views numbers</th><th>Depth data available for views</th><th>Coded views in the CfP [129]</th></tr>
<tr><td>Poznan Carpark</td><td rowspan="3">1920 x 1088</td><td rowspan="3">25</td><td rowspan="2">Canon XH-G1, 3-CCD camera</td><td rowspan="2">250</td><td rowspan="2">0…8</td><td rowspan="2">3,4,5</td><td>-</td></tr>
<tr><td>Poznan Street</td><td>3,4,5</td></tr>
<tr><td>Poznan Hall</td><td>200</td><td>0…8</td><td>5,6,7</td><td>5,6,7</td></tr>
<tr><td>Lovebird1</td><td rowspan="4">1024 x 768</td><td rowspan="4">30</td><td>Point Grey Flea camera (CCD), Moritex ML-0813 lenses</td><td>240</td><td>0..8</td><td>3,5,7</td><td>3,5,7</td></tr>
<tr><td>Newspaper</td><td>Point Grey Research Flea camera (CCD) with 1/3-inch Sony lenses</td><td>300</td><td>0…8</td><td>2,4,6</td><td>2,4,6</td></tr>
<tr><td>Balloons</td><td rowspan="2">XGA CMOS, 8-bit RGB-Bayer camera</td><td rowspan="2">300</td><td rowspan="2">0…6</td><td rowspan="2">1,3,5</td><td rowspan="2">1,3,5</td></tr>
<tr><td>Kendo</td></tr>
<tr><td>GT Fly</td><td rowspan="2">1920 x 1088</td><td rowspan="2">25</td><td rowspan="2">Computer-generated sequences</td><td rowspan="2">250</td><td rowspan="2">1,2,3,5,9</td><td rowspan="2">1,2,3,5,9</td><td rowspan="2">1,5,9</td></tr>
<tr><td>Undo Dancer</td></tr>
</table>

There are in total 8 sequences in the test set (Table 1), presenting various scenes, both natural and computer-generated (Fig. 9). The test set is provided with ground truth depth data. For natural test sequences, the depth has been algorithmically estimated from the video. In some cases, additional manual help was needed, and therefore it is said that those depth maps have been estimated semi-automatically. For synthetic sequences (GT Fly, Undo Dancer sequences) the ground truth depth maps have been computer-generated along with the video.

In this dissertation, the views of the cameras are consequently numbered from 0 (see: Table 1). Therefore, in the case of Lovebird1 sequence, the original camera indices (which were starting from 1) have been renumbered to range 0..8.

1.5.3. Assessment of the quality of depth maps

As mentioned above, currently the most adequate known methodology of evaluation depth estimation algorithms has been developed by ISO/IEC MPEG group. **In the dissertation, author has decided to employ it as a main objective assessment method.**

MPEG methodology for evaluation of quality of depth maps has been constituted as a part of 3D framework [137]. It employs view synthesis for evaluation of quality of depth maps, which can be used to evaluate depth estimation algorithm itself.

During the evaluation, three views are explicitly considered – *A, B* and *V* (Fig. 10). First for view *A* and view *B* depth maps are estimated. Typically, this is performed with implicit use of some side views. Depth estimation may employ many views (e.g. views *A-1, A* and *A+1* for depth estimation of view *A*). The estimated depths of view *A* and view *B*, along with their original images, are used to synthesize a virtual view in position of middle view *V*. The original image of view *V* is used for reference and comparison, which provides indirect evaluation the depth map estimation algorithm used. **Therefore, the quality of the depth is assessed indirectly by evaluation of quality of synthesized view.**

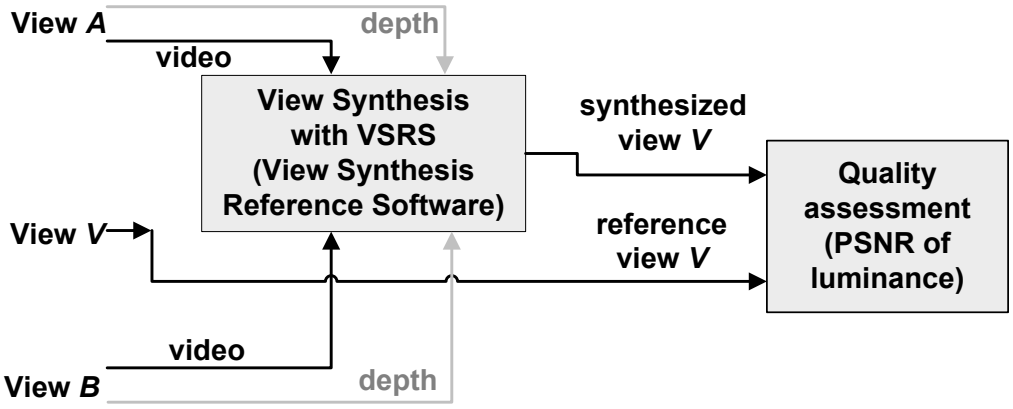


Fig. 10. Depth map assessment procedure developed by ISO/IEC MPEG and used in the dissertation.

The synthesis of a virtual view, employed in the mentioned methodology, can be performed by means of Depth Image Based Rendering (DIBR) [141]. In DIBR, pixels from given input views are shifted, with respect to their depth, to different spatial positions in the target view. Depending on the structure of the 3D scene, some pixels may become occluded by others. On the other hand, some pixels in the target view cannot be rendered because they are occluded in the input views. Such pixels are inpainted [124].

In methodology developed by ISO/IEC MPEG group, for the sake of synthesis of virtual views, usage of View Synthesis Reference Software (VSRS) [124][125] is recommended. **For the purpose of view synthesis, also in this dissertation VSRS is used.**

Table 2. Specification of views selected for evaluation of depth estimation (Fig. 10) for multiview test sequences used in the dissertation.

Sequence name	Resolution	Ground truth disparity maps available for views	Views used for depth estimation (View A and B)	Synthesized view (view V) used for quality evaluation
Poznan Carpark	1920x1088	3,4,5	3,5	4
Poznan Street				
Poznan Hall		5,6,7	5,7	6
Lovebird1	1024x768	3,5,7	3,5	4
Newspaper		2,4,6	4,6	5
Balloons		1,3,5	3,5	4
Kendo				

As can be noticed in Table 1, presented on page 24, views of computer-generated sequences (Undo Dancer, GT Fly) are placed in irregular spatial positions. Therefore, no fair depth estimation can be performed for reasonable comparison with other sequences. Moreover, computer-generated ground depth data is available which makes depth estimation impractical for such cases. Thus, those computer-generated sequences have not been used in experiments with depth estimation and are used mainly for reference purposes.

The described above **virtual view synthesis-based depth map quality evaluation methodology** is used thought the dissertation, both in parts related to depth estimation and depth coding.

Chapter 2. State of the art in depth map estimation

This chapter provides an introduction and overview of techniques of stereoscopic depth map estimation. Sections 2.1-2.2 provide an introduction to the basics of depth estimation and Sections 2.3-2.6 focus on the methods that are directly related to the author's proposals in this dissertation.

2.1. Depth estimation fundamentals

Algorithmic estimation of the depth is a long-lasting scientific problem. The first works on depth estimation go back to 1950's and, although many years of works, the current state of the art is still far away from satisfying level in many applications, especially in case of new generation of 3D video systems. This regards both to the quality of resultant depth maps and complexity of the algorithms.

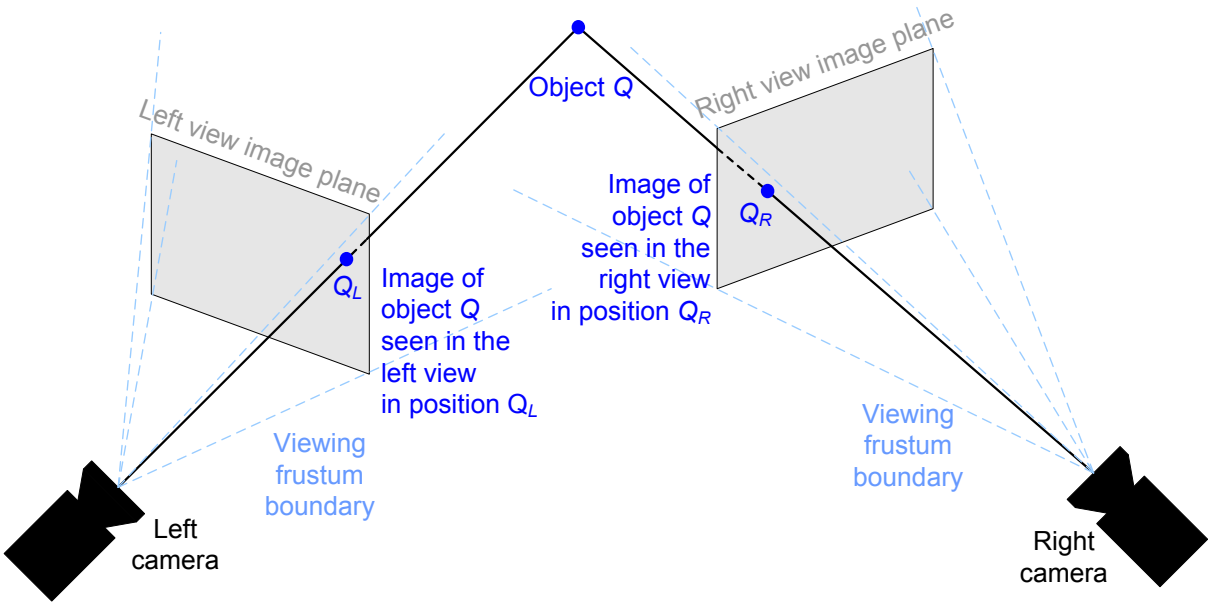


Fig. 11. Object Q projected onto image planes. Different positioning of the cameras leads to different projected positions in left and right view. Only objects within viewing frustum of given camera are visible.

The basic principle in algorithmic depth estimation is usage of two views of the same 3D scene. In each of the views, a given object is seen from different angle and position (Fig. 11) and therefore its observed position is different. Most of depth estimation algorithms determine depth by finding correspondence between pixels across the views. For example, for a given projected position Q_L (Fig. 11) in the left view, the algorithm searches for projected position Q_R in the right view, so that both pixels, at positions Q_L and Q_R, correspond to the same point

of object Q. Such potential correspondences are geometrically inclined to lay along so called epipolar lines, which can be derived from a pin-hole model of the camera [155][156].

Positions and orientations of epipolar lines and indicated by the locations of the cameras, their orientations and other parameters, like focal length, angle of view etc. Typically, all of those parameters are gathered in a form of intrinsic and extrinsic camera parameter matrices [157][158]. In a general case, epipolar lines may lay along arbitrary angles.

An important case in depth estimation (Fig. 12b) is a setup where the cameras of multiview video system are arranged linearly, so that the axes of the viewpoints are all parallel. **Such setup, considered in the dissertation, is called linear arrangement of the cameras.** Linear arrangement can be attained both by precise physical positioning of bodies of the cameras or by post-processing of images captured by other arbitrary setup of the cameras, e.g. angular (Fig. 12a), with use of rectification techniques [152][154] along with distortion removal [150][151].

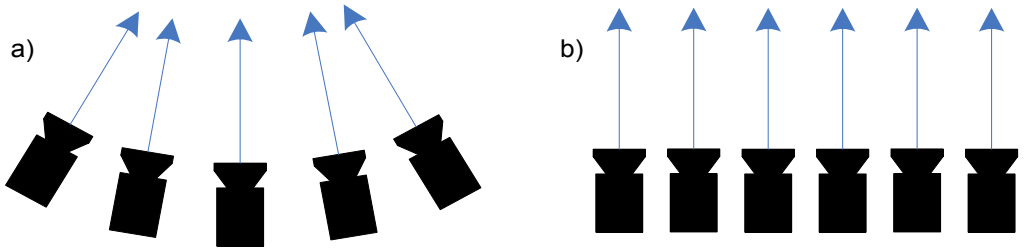


Fig. 12. Various arrangements of cameras: a) angular camera arrangement, b) linear arrangement of the cameras with parallel axes of the cameras (considered in the dissertation)

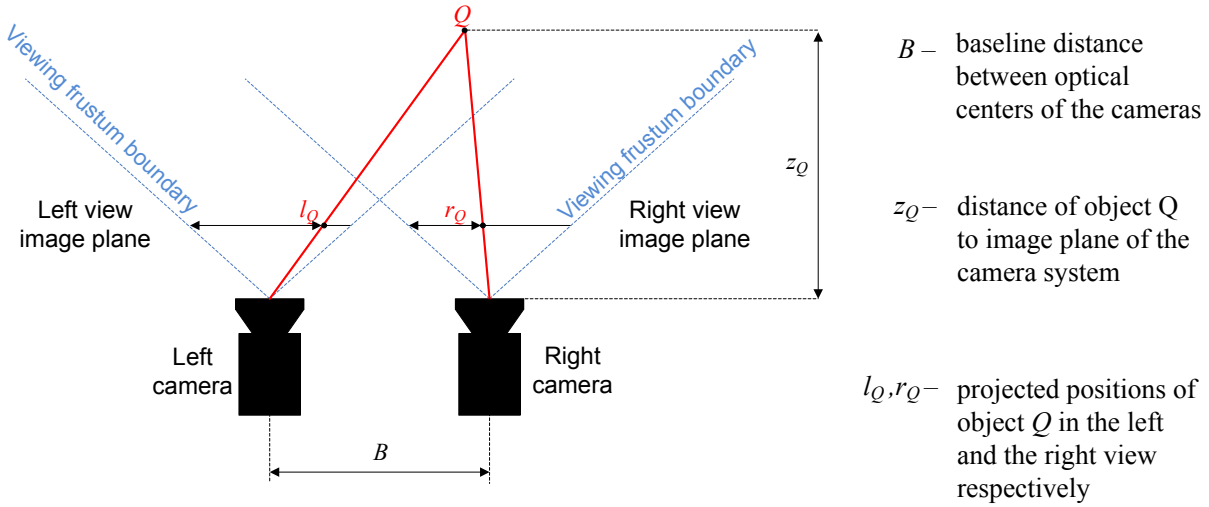


Fig. 13. Exemplary objects Q projected onto image planes of two cameras that are aligned horizontally with parallel axes of the cameras.

In considered case of linear arrangement of the cameras, the image planes of all views coincide (Fig. 13) and the epipolar lines are all aligned with the horizontal axes of the images. Therefore, the differences in observed positions of objects become disparities along horizontal rows. Due to the projective nature of such video system with linear camera arrangement, the further the object is from the camera, the closer is its projection to the center of the image plane (Fig. 13).

Basing on projected positions of objects observed in different views, their 3D position can be mathematically determined. The question of the depth can thus be determined by finding disparity value for each point. For a stereoscopic pair of cameras, horizontally aligned, distant by the length of baseline B, the depth z_Q of point Q of can be calculated as follows:

$$z_Q = f \cdot B \cdot \frac{1}{d_Q} \quad , \tag{1}$$

where: z_Q - distance of object Q from the image plane of the camera system (depth), B - baseline distance between optical centers of lenses of the pair of cameras, d_Q - disparity of point Q, which simply is the difference between observed projected positions for given stereo pair. For the cases presented in Fig. 13:

$$d_Q = r_Q - l_Q \ , \tag{2}$$

It is important to note that a search range is typically required to be defined as an input to the disparity search process. Once the disparity for a pixel is determined, often [159], depth z (or disparity d) is stored in form of a normalized disparity δ:

$$\delta = \delta_{max} \cdot \left(\frac{d - d_{min}}{d_{max} - d_{min}} \right), \tag{3}$$

which is often presented in the form of:

$$\delta = \delta_{max} \cdot \left(\frac{1}{z} - \frac{1}{z_{far}} \right) / \left(\frac{1}{z_{near}} - \frac{1}{z_{far}} \right), \tag{4}$$

where z_{near} and z_{far} are distances to the closest and the farthest object in a scene (corresponding to d_{max} and d_{min} disparity values) and δ_{max} is the maximal value for given representation of normalized disparity – e.g. 255 for 8-bit representation.

As depth is unambiguously related to disparity by mathematical equations (1), (3) and (4), colloquially, the terms depth (z), disparity (d) and normalized disparity (δ) are often used interchangeably. I.e. the term depth estimation is used where in fact, disparity estimation is considered [124][134][139][142][147][159][218][220][223][224].

Also, although this dissertation relates to "depth" estimation and "depth" coding it is worth to notice that, in fact, most of the presented considerations relate to disparity d and normalized disparity δ.

With use of equations (1), (3) and (4), the problem of depth map estimation can be expressed as problem of disparity estimation of disparities. Yet, even is such stripped form, it is still is a complex and challenging task.

In general, two classes of depth estimation methods can be distinguished – sparse depth estimation (which are also referred to as indirect methods) and dense depth estimation (which are also referred to as direct methods).

The scope of this dissertation lays within the latter case of dense depth estimation.

Out of scope of this dissertation, are sparse methods, which search for visual features, such as corners or edges, and then match corresponding features between frames/views (e.g. [160][161][162][163][164]). It can be noted though that some of indirect methods, although yield only with sparse information about the depth, are targeted at providing ultimately a complete 3D reconstruction of the scene [165][166].

In **dense depth estimation,** considered in the dissertation, the depth is often expressed in form of a depth map which is a matrix composed of all depth values in given view. Yet even more often, the depth is expressed in a form of normalized disparity map, presented as gray-scale image (Fig. 3b, Fig. 14c).

The dissertation focuses on the most efficient and most commonly used methods known from literature [142][143], based on block-matching with local or global optimization of the generated disparity maps. **The goal of the dissertation is improvement of those methods.** Some of them are discussed below, up to the level which is required for the comprehension of the ideas presented in the dissertation, in the following Subchapter 2.2.

2.2. Local estimation methods

Dense depth estimation methods, known from literature, find depth by means of disparity estimation, which can further be converted to depth (1). In order to find disparities between corresponding points, those methods employ calculation of similarities between fragments of the processed images. Typically used similarity metrics are Sum of Absolute Differences (SAD) or Sum of Squared Differences (SSD), between pixels or blocks of pixels, or normalized cross-correlation [167]. Some works propose more advanced solutions, like use of binary matching cost [168], non-parametric local transforms like "rank” or “census” [173] or even approaches that incorporate mixtures of transforms [17].

a)

b)

c)

Fig. 14. A single frame of Poznan Carpark multiview test sequence (a) and the corresponding disparity maps (b) and (c). In the case of (b) ground truth depth map is presented and in the case of (c) depth map estimated with use of Winner-takes-all technique (with 1×1 pixel block size, SAD similarity metric).

In the most basic approach, disparity is estimated for all pixels in the image independently. As a result, for each pixel, the most similar pixel in the second image is found, yielding a resultant disparity. Such approach is called “winner-takes-all” (WTA) as other possible correspondences, apart from the best ones, are never selected.

Unfortunately, winner-takes-all approach does not provide satisfactory results [142][143] - as it can be seen in Fig. 14c, the depth map produced with WTA method is very noisy and often does not reflect the real depth of objects in the scene. This results from the fact that often the sought disparity cannot be determined locally on per-pixel basis as there are many similar pixels in second images resulting in equally good disparity candidates (Fig. 15). Typically, this occurs in case of correspondence search performed in homogenous regions (single colored homogenous regions - Fig. 15 black arrow) or glossy regions (e.g. reflections on the glass, windows or windscreen - Fig. 15 white arrow) of the images. In such cases, the disparity has to be determined basing on not only local features of the images (like only pixel to pixel similarity) but also using neighborhood of given pixel. This is based on an assumption that some of neighboring pixels contain texture that can be used for reliable correspondence search (Fig. 15 gray arrow). In the simplest solution, blocks of pixels are used in order to aggregate information from surrounding pixels. In such case, SAD, SSD similarity metrics (or more advanced ones), over all pixels that reside inside the given block, are calculated.

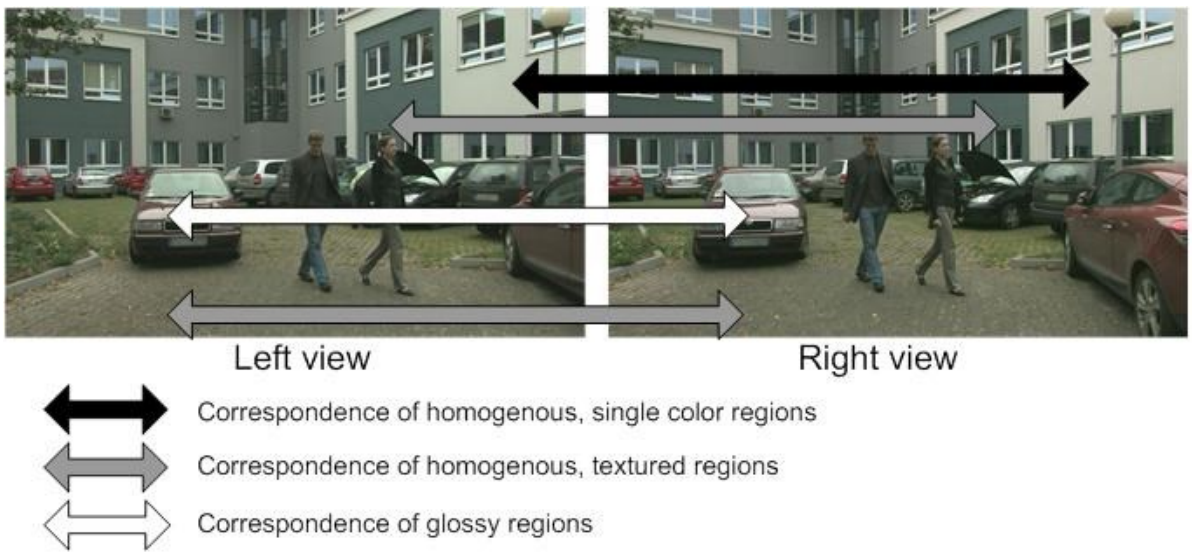


Fig. 15. Correspondence of features which is the main idea behind algorithmic depth estimation. Examples of regions which are problematic are shown.

Another interesting extension of simple stereo matching with SAD, SSD similarity metrics, is usage of certainty/reliability maps. Those are matrices composed of values which indicate whether corresponding disparity value is certain (has been estimated with high likeness) or uncertain (e.g. in cases of untextured regions - Fig. 15). A similar approach is proposed in paper [171]. A stereo matching algorithm for disparity estimation is presented as a method for retrieving a 3D reconstruction of the observed scene. Along with the implemented estimation algorithm itself, a method of its verification is proposed, by means of generating for each disparity map, a respective certainty map which holds information about likeliness that the respective disparity value was chosen well. In turn, in work [172] a reliability map is used to allow fast depth estimation performed on GPU (Graphics Processing Unit) on graphics card in PC computer. In the mentioned work “rank” transform is used [173].

In more sophisticated approaches to local depth estimation, so called fuzzy segmentation/soft segments [15][86][84][174][175] are used. In such, authors extend block matching scheme with a 2-dimensional weighting function which is calculated basing on the content of the images. This function defined the weights with which similarity of particular pixels is aggregated into the resulting similarity metric. A similar approach is to employ guided filters [176] that are variation over content-adaptive filters. In work [177], authors not only show very good quality of depth estimation but also claim that their algorithm is capable of performing in real-time. Work [178] also presents algorithm which is by the authors, described as near-real-time, based on guided filters, showing good performance on Middlebury website [142], among other state-of-the-art depth estimation techniques. In [179], even more advanced solution is proposed which combines usage of guided filters with information about gradients in analyzed images.

Independently from the usage of exact method, aggregation of information from many pixels yields much more coherent results than in case of single pixel-to-pixel matching (Fig. 14b). The main disadvantage of such methods, which aggregate the information from the neighborhood, is limitation of range of the neighboring pixels that can be considered, due to at least two reasons. First, usage of wider neighborhood significantly increases computational complexity of such methods and thus they are ineffective respective to other solutions [142][143]. Secondly, the size of the neighborhood must not exceed the size of the objects that are matched. Therefore, there is a kind of uncertainty principle, because usage of small blocks allows for localization of small objects but with limited accuracy, and usage of greater size of blocks provides good accuracy in case of large objects but yields with bad results in case of small objects. Typically, size of matched regions ranges from 16 to 64 pixels. Also,

methods with variable, adaptively selected block size are known [169][170][180][181]. In work [206] authors propose depth estimation method which aggregates information from neighboring filters, developed basing on probabilistic model with diffusion. Also enhancement of the optimization algorithm, with application of Gauss-Seidel method, is proposed, so that the convergence of optimization algorithm can be attained faster.

Usage of motion information for enhancement of depth estimation is also commonly considered subject. Authors of [182] have proposed usage of depth refinement technique which estimated ego-motion of the camera in order to attain depth maps with sub-pixel accuracy. Frame-by-frame prediction of 3D scene is performed by tracking of feature point coordinates and thus the proposed method allows depth estimation in a video without the need for disparity computations in each single frame.

In paper [8] usage of motion field estimation is proposed. Calculated motion vectors are then used to extend pixel similarity metric, which is originally based on SAD. The authors adopt optical flow Classic+NL technique [183], based on classical formulation of optical flow by Horn and Schunck.

Often, usage of more than two views in depth estimation is proposed [184]. Such approach allows reduction of problems related with occlusions, texture-less regions or shadows [185][186][187][188][189]. Typically in such cases [126][184], instead of similarity metric calculated between a single pair of views, similarity metrics are calculated between all available views and the processed view, and the minimal (the best) one is used for optimization.

In some works [185][186][206], occlusions are explicitly modeled by marking pixels that are believed to be hidden by other pixels, basing on the current stage of computations.

Although local estimation methods are not very often used as stand-alone in state-of-the-art depth estimation techniques, their concepts are exploited by more advanced **global optimization methods**, described below.

2.3. Global optimization methods

Another, far more interesting approach that aspires to find a global optimum instead of local one (like in case of local methods described in Section 2.2) is to redefine the disparity/depth estimation as a problem which can be solved with use of generic optimization methods. In order to do so, an energy function $Fitness$ over a depth map is formulated:

$$Fitness = \sum_{P} FitCost_P \quad , \tag{5}$$

where $FitCost_p$ depicts sub-component of $Fitness$ function for particular pixel p in the considered disparity map. Such function is often related to as "energy", "goal function" or "performance index" in other optimization applications.

Because such $Fitness$ function is formulated on per-pixel basis, it can be used in variety of generic optimization algorithms. Among many known (like genetic optimization) only a few of them have found application in field of depth map estimation, due to the fact that the amount of considered disparity values is relatively large (e.g. hundreds).

The most commonly used optimization algorithms are graphs cuts (GC) and belief propagation (BP) [197]. However, the description of those algorithms is out of scope of the dissertation, some brief introduction has been provided in Subsection 2.3.3 and 2.3.4 respectively.

In the case of both GC and BP algorithm, the function $FitCost_P$ is typically modeled as a sum of two sub-functions: $DataCost$ and $TransitionCost$ for each pixel (6):

$$FitCost_P = DataCost_p(d_q) \quad + \sum_{q \in (\text{neighborhood of } p)} TransitionCost_{p \to q}(d_p, d_q), \tag{6}$$

where:

p – pixel (point) for which $FitCost$ is evaluated,

d_p – assumed disparity of pixel p,

q – some pixel (point) in neighborhood of pixel p,

d_q – assumed disparity of pixel q.

$DataCost_p(d_p)$ - models the direct correspondence between pixels and express how given pixel p is similar to those pointed by its disparity d_p in other images. This is further described in Subsection 2.3.1.

$TransitionCost_{p,q}(d_p, d_q)$ - penalizes disparity maps that are not smooth. If a given pixel p has vastly different disparity d_p than its neighbors (pixels depicted by q) it gets high $TransitionCost$ penalty. This is further described in Subsection 2.3.2.

Of course, more advanced approaches than presented in equation (6) are known [190][191][192][195][196][197], where higher order $FitCost$ function is defined, but their application is not very common [142][143].

The usage of $DataCost$ and $TransitionCost$ is a common idea of all global optimization methods like belief propagation or graph cuts. Depending on the approach those are defined as probabilities [202][203][204][205] or in terms of energy [193][197][198]. Some works [215] use mathematic concept of partition function, related to Boltzmann probability distribution, in order to exchange energy formulation into probability, and vice versa. Unfortunately, there is lack of empirical verification of whether such operations are justified.

The problem of definition and formulation of $DataCost$ and $TransitionCost$ functions is one of the main subjects of the dissertation for which results are shown in Chapter 3, Sections 3.1 to 3.6.

Below, in Subsections 2.3.1 and 2.3.2, an introduction to typical formulations of $DataCost$ and $TransitionCost$ functions will be provided.

2.3.1. *Data Cost* function

The $DataCost$ function models the direct correspondence between pixels and express how given pixel p is similar to those pointed by its disparity d_p in other images. The higher the difference between those pixels is, the higher is the value of $DataCost_p(d_p)$.

The most commonly $DataCost$ is defined in terms of energy related to similarity metrics between fragments of images, calculated in pixels or blocks. Typically, Sum of Absolute Differences (SAD) [180] or Sum of Squared Differences (SSD) [181][204] metrics are used. Some state-of-the-art works which relate to $DataCost$ function propose usage of "rank" or "census" [173] for calculation of better similarity metric. Work [17] proposes a more advanced approach, where mixture of various similarity metrics is incorporated in order to attain better quality in depth estimation, but theoretical foundations are missing.

In paper [204], which in the most related to the dissertation in area of depth estimation, authors provide a similar derivation of $FitCost$ function based on MAP assumptions. Unfortunately, the authors have omitted the consequences of this derivation related to

$DataCost$ and have limited their work to consideration of Gaussian model (corresponding to Sum of Squared differences energy formulation). The authors do not provide any verification of whether such assumptions are correct.

Similarly, work [205] employs posteriori probability for modeling of $FitCost$ function. Authors consider a more advanced model for $DataCost$ which incorporates Generalized Gaussian model with arbitrary power exponent. Therefore, in for value of 2 Gaussian model is considered and for value of 1, Laplace model is considered. Also here, any verification of whether such assumptions are correct is not provided, apart from theoretical considerations.

In work [216] authors have proposed usage of truncated-linear $DataCost$ function which actually responds to Absolute Difference similarity metric, with is limited so that is does not exceed some given maximal level. Apart from the concept being very scientifically interesting and giving promising results, the authors have not supported their proposal with empirical data verifying their assumptions. Moreover, no analysis of noise nor cross-correlation between matched images have is performed.

In work [217] authors thoughtfully analyze probabilistic model of correspondence in 3D space. Instead of Maximum A posteriori rule, a different approach for evaluating entropy and mutual information, called EMMA, is proposed. Authors claim, that one of advantages of EMMA is that it does not require a prior model for the functional form of the distribution of the data, and that the entropy can be maximized (or minimized) efficiently using stochastic approximation. Unfortunately, the method is presented in context of 3D modeling and not depth map estimation itself which disallows comparison with other state-of-the-art methods in the field of the dissertation.

2.3.2. *Transition Cost* function

The $TransitionCost$ is a term of $FitCost$ function which penalizes disparity maps that are not smooth. Its role is regularization of the resultant depth/disparity map. The higher are the differences between disparity d_p of pixel p and disparity values d_q of all neighboring pixels d_q, the higher is the value of $TransitionCost_{p\to q}(d_p, d_q)$.

Typically, $TransitionCost_{P\to Q}(d_P, d_q)$ is defined independently from pixel positions p and q and thus it can be simplified to $TransitionCost(d_p, d_q)$. Also, very often, $TransitionCost$ is not defined as function of d_p and d_q independently, but as a function of $|d_p - d_q|$ only: $TransitionCost(|d_p - d_q|)$.

Among the most commonly known are three models for $TransitionCost$ function – Potts model, linear model and truncated-linear model:

a) Potts model [198]

$$TransitionCost(|d_p - d_q|) = \begin{cases} 0 & if\ |d_p - d_q| = 0 \\ \alpha & otherwise \end{cases} . \quad (7)$$

b) Linear model [126][190]

$$TransitionCost(|d_p - d_q|) = \gamma \cdot |d_p - d_q| . \quad (8)$$

c) Truncated-linear model [216]

$$TransitionCost(|d_p - d_q|) = \min(\gamma \cdot |d_p - d_q|, \alpha) . \quad (9)$$

Used notation:

p – pixel for which $FitCost$ function is evaluated,

d_p – assumed disparity of pixel p,

q – some pixel in neighborhood of pixel p,

d_q – assumed disparity of pixel q,

α, γ – constant parameters.

In general, TransitionCost functions incorporate some sort of constant parameters, like γ or α coefficients. The main purpose of such constant parameters is to provide weighting to the relation with DataCost function, to which it is added to formulate FitCost function (6). The most commonly used parameter γ of linear and truncated-linear models, is widely called **"Smoothing Coefficient"** as its value sets how much depth maps that are not smooth are penalized by FitCost function. Usage of small values of Smoothing Coefficient results in sharp depth maps which are similar to those attained with local depth estimation methods. Usage of large values of Smoothing Coefficient results in generation of very smooth, even blurred depth maps. The selection of Smoothing Coefficient is typically done manually (the depth estimation is thus supervised) which is an **important problem in practical usage** of depth estimation methods based on belief propagation or graph cuts in applications, where unsupervised operation is expected.

All of the mentioned models (Potts, linear and truncated-linear) are typically used because they are simple and provide some additional advantage in the case of belief-propagation algorithm, because they allow reduction of computational complexity of execute of particular

steps from $O(D^2)$ polynomial to $O(D)$ linear time, where D is the number of disparity considered values. As typically D ranges from 40 to 100, this provides vast reduction of real computational complexity.

In work [216**],** authors have proposed usage of truncated-linear-shaped $TransitionCost$ function for depth estimation and have compared it against other state-of-the-art techniques. Although the results are promising, the foundations of the proposal are not given.

In papers [204][205] authors consider derivation of $TransitionCost$ function based on Maximum a Posteriori rule, similar to the approach in this dissertation. Basing on this, Markov Random Field model for stereoscopic depth estimation is formulated by means of belief-propagation algorithm. Unfortunately, the work proposes only an approximation of $TransitionCost$ function.

The lack of works, which provide theoretical analysis of application of **Maximum A posteriori Probability (MAP)** optimization rule to formulation of $\boldsymbol{DataCost}$ **and** $\boldsymbol{TransitionCost}$ for depth estimation, along with empirical experimentation which would support formulation of such theoretical models, is **one of the motivations of the dissertation**.

In Subsections 2.3.3 and 2.3.4 below, graph cuts and belief propagation algorithms are presented. Those are used in the dissertation solely as tools for optimization of depth maps with regards to $FitCost$ function (and thus $DataCost$ and $TransitionCost$ functions). Therefore, as the dissertation does not relate directly to those algorithms and thus in-depth knowledge about them is not needed, the presented introduction will be very short. A more comprehensive and detailed description can be found in the references.

2.3.3. Graph Cuts

In this Subsection a brief introduction to graph cuts (GC) algorithm is provided. The dissertation is not related to the GC algorithm itself. Rather than that, the graph cuts algorithm is used as a reference technique in the experiments related to depth estimation. Therefore, as in-depth understanding of graph cuts is not needed for reading of the dissertation, only a short description and survey of state-of-the-art techniques is shown. Further details and detailed description of GC algorithm can be found in literature mentioned below.

In general, graph cuts is an algorithm which solves energy minimization problems by reducing them to instances of the maximum flow problem [190][195][198] in a graph. In such cases, max-flow/min-cut theorem [193][194][195][196] can be used in order to efficiently

find a minimal cut of the graph which corresponds to finding an optimal solution of a problem represented in the considered graph.

One of the applications of graphs cuts algorithm is solving the problem of labeling of an image. **In depth estimation it relates to assigning disparity values (labels) to particular pixels (points) of the input image**. It has been proven [190][193][194] that in binary cases (where only two different labels are defined) can be solved optimally by the GC algorithm.
If there are more than two labels (which is the case in depth estimation) the produced solutions usually lay near the global optimum [191][192], targeted by $FitCost$ energy function.

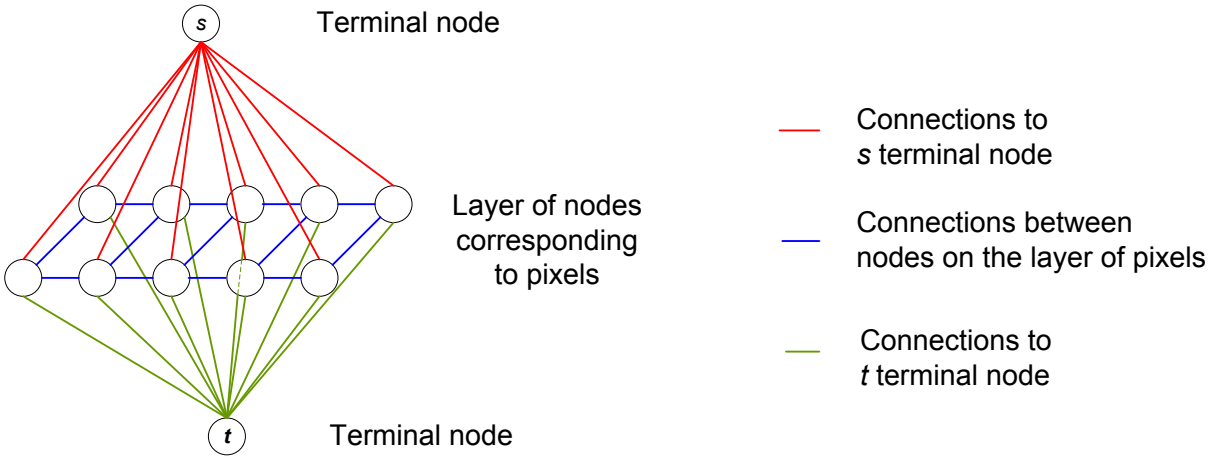


Fig. 16. An exemplary graph used in graph cuts algorithm for depth estimation.

In depth estimation based on graph-cuts algorithm [190][192][195][198][199][200] a graph is defined (example in Fig. 16), whose nodes model pixels and whose edges model $FitCost$ components associated with them. Thus, the structure of the graph reflects the definition of $DataCost$ and $TransitionCost$ function. There are many graph structures known from the literature [190][192][195][198] with different complexity and performance.

Although the discussion of details of specific graph-cuts methods is above the scope of the dissertation, it is worth to notice that the most commonly used variant of graph-cuts solver is so called alpha-expansion [195]. In such, multi-labeling problem is solved iteratively, where, in each iteration only a binary labeling problem is solved. In particular, two alternative disparity maps are considered, the current one and a hypothetical one (filled with a single disparity value) proposed at the current iteration. A graph is created, whose edges represent $FitCost$ function compounds related to both of the alternative disparity maps. Then, an optimal cut between the current and the alternative disparity map is sought in this graph (basing on max-flow min-cut theorem) which yields a better fitted (energy minimal) disparity

map, composed of fragments of the current and the alternative disparity map. This is repeated with different alternative disparity maps (typically for all possible disparity values) until the current disparity map is optimized to the point of convergence.

2.3.4. Belief Propagation

In this Subsection a brief introduction to belief propagation (BP) algorithm is provided. Although the dissertation is not related to the BP algorithm itself, a short description and survey of techniques is shown for the reference of the state-of-the-art. Further details can be found in the literature, although in-depth knowledge about belief propagation algorithm is not needed for understanding of the dissertation.

The belief propagation (BP) algorithm [201][202][203][204][205] can be seen as an extension and generalization of the well-known Viterbi algorithm (dynamic programming). The Viterbi algorithm operates on graph describing a lattice of observations with one-directional connections which constitute a 1-dimensional field of nodes (Fig. 17a).

Specifically, belief propagation applied in depth estimation, extends this scheme to a 2-dimensional field of nodes, where the nodes typically correspond to structural elements of the image, like pixels (Fig. 17b). The most often, bidirectional connections between the nodes in BP are considered and such variant of the algorithm is called loopy-belief-propagation [207].

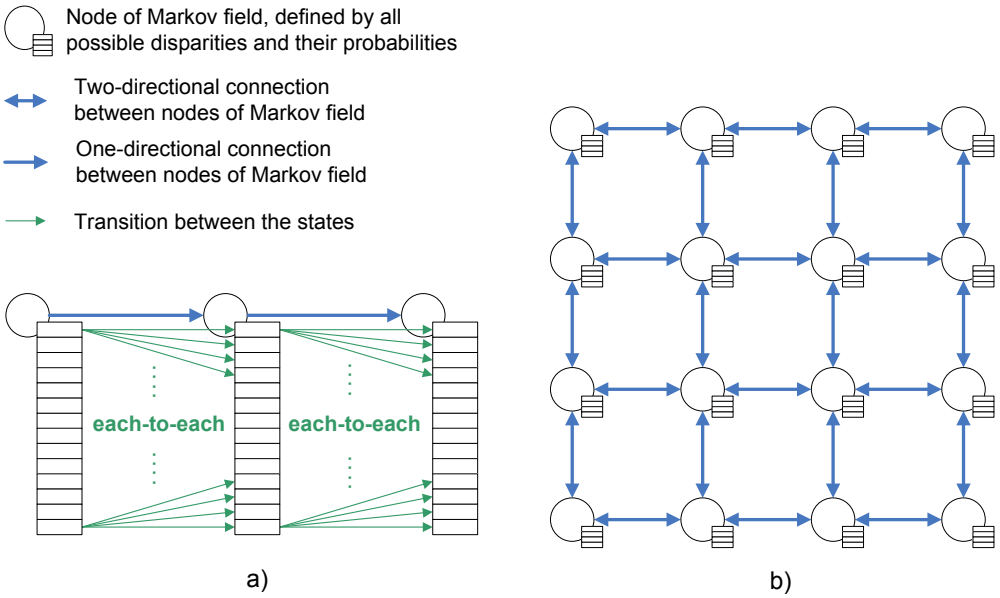


Fig. 17. Illustration of Viterbi algorithm (a) and belief propagation (BP) algorithm (b). In the depicted BP example, considered is Markov Random Field with 4-way (left, right, top and bottom) neighborhood of nodes and bidirectional connections between the nodes.

Those nodes communicate with their neighbors by sending messages. Each such message is composed of beliefs of the source node about probabilities of disparities that are considered (and are possible) in the target node.

Therefore, BP algorithm iteratively estimates probabilities associated with considered disparity values and basing on that to choose the most probable disparity for each node.

Although it is not a strict requirement, it is often assumed that in BP algorithm the sought disparity map is modeled by a 2-dimensional Markov Random Field (MRF) [208]. In MRF defined for the problem of depth estimation, each node of the field is defined by all possible disparities and corresponding probabilities. By analogy to the nomenclature known in Hidden Markov Models (HMMs) the possible disparities correspond to hidden states, the $TransitionCost$ corresponds to transition probability and the $DataCost$ corresponds to the output probability. In such MRF-based belief propagation formulation $FitCost$ optimization function (and thus also $DataCost$ and $TransitionCost$) are expressed as log-probabilities (in logarithmic scale). Belief propagation can optimize marginal probability [209][210] or maximum a posteriori probability [211] of optimal selection of disparities, both which are defined by $FitCost$ function. The latter case is analogous to graph cuts algorithm [197].

2.5. Accuracy and precision of disparity values

An important aspect of dense depth estimation algorithms, considered in the dissertation, is the degree of accuracy of estimated disparity values, which is limited by the employed disparity representation precision.

Apart from the techniques that are based on optical flow [18][218], where practically continuous range of disparities is outputted, both local and global methods (mentioned in Section 2.2 and 2.3 respectively) assume a discrete set of disparities. This can be defined by three parameters: d_{min} and d_{max} which correspond to minimum and maximum possible disparities between the views (disparity range) and d_{step} which is a minimal step between each of consecutive disparity values, expressed as a multiple of the spatial sampling period in images (quantization step of disparity representation).

For example (Fig. 18a - Kendo sequence), if we consider a set of disparity values in range from $d_{min} = 1$ and $d_{max} = 25$ with $d_{step} = 1$ (all expressed as multiples of the spatial sampling period in images, thus disparity representation is "full-pixel" precise), generated are 25 discrete disparity values which are mapped with equation (3) to normalized disparities.

Histogram of such normalized disparity map is sparse (Fig. 18a) as only 25 discrete disparity values exist.

In example depicted in Fig. 18b (Poznan Carpark sequence), $d_{min} = 1$ to $d_{max} = 80$ with $d_{step} = 0.5$ (thus representation of disparity is "half-pixel" precise), the histogram is more dense as total of $\frac{80}{0.5} = 160$ discrete disparities values are considered. In computer-generated examples of Undo Dancer and GT Fly and sequences (Fig. 18cd) d_{step} has been set to a such value, so after mapping to normalized disparity exactly all values are possible in the histogram (Fig. 18cd).

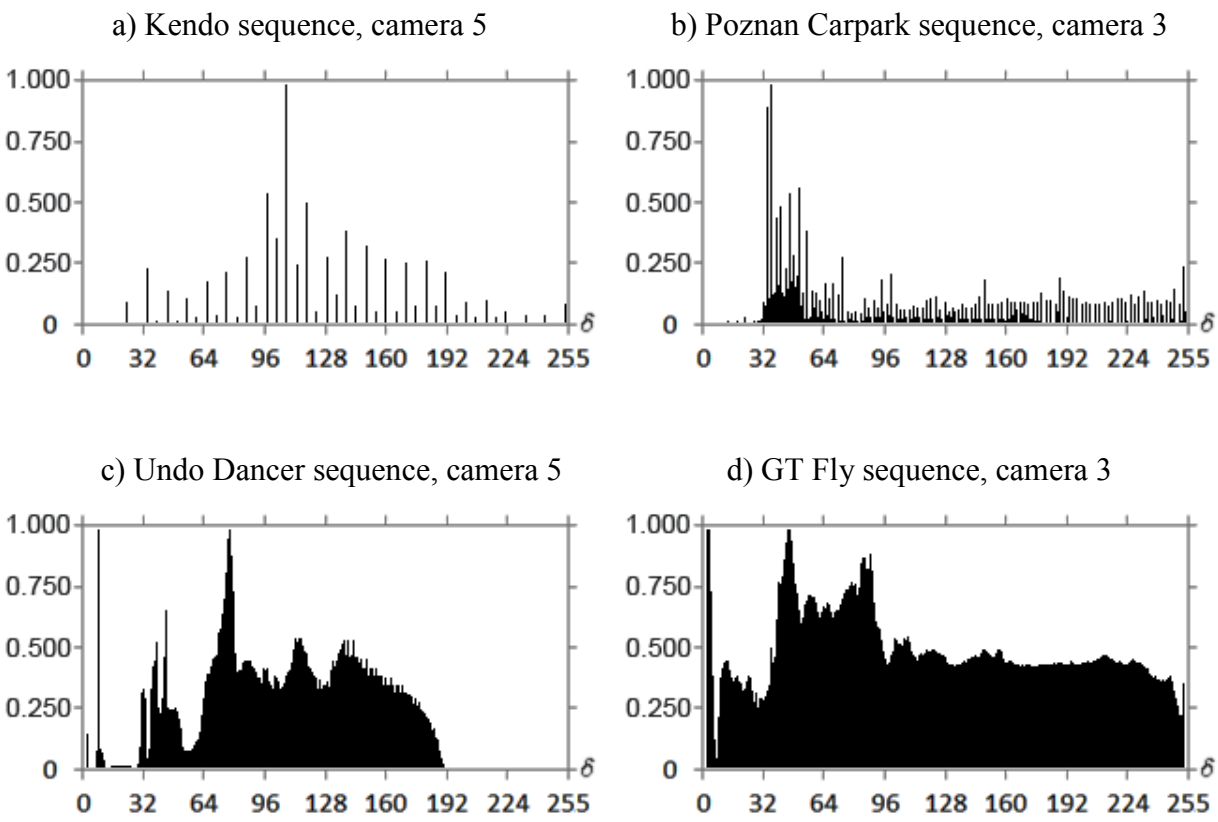


Fig. 18. Histograms of normalized disparity δ values in ground truth disparity maps. The graphs have been vertically normalized to range [0;1].

Usage of a small d_{step} (e.g. "half-pixel" precision or "quarter-pixel" precision) leads to higher precision of the estimated depth map (and likely, higher accuracy) but increase the computational complexity of the depth estimation algorithm as the number of total considered disparity labels is increased.

Usage of a large d_{step} (e.g. "every-two-pixel" precision) leads to lower precision of the estimated disparity map (and likely, lower accuracy) but allow computational complexity savings.

The dependency on computational complexity, especially in the case of belief-propagation or graph-cuts algorithms (Section 2.3), typically enforces usage of only full-pixel precision [142][143] corresponding to $d_{step} = 1$.

There are a few papers that consider sub-pixel precise depth maps.

In [182] depth map is refined for sub-pixel accuracy with use of ego-motion of the camera. Motion parameters of the cameras system are estimated by tracking of feature points. The paper brings an additional benefit, because the depth in the video can be estimated without the need for disparity computations in each single frame. Unfortunately, as the method stereoscopic video sequence with motion of the camera, it is not suitable for generic cases of processing.

Other authors [219] have proposed an algorithm for depth map improvement by anisotropic diffusion. This method provides smooth, high-precision disparity maps, but unfortunately it does not preserve depth discontinuities over the edges of the objects.

Authors of [18] propose an additional precision refinement step, analogic to the idea proposed in the dissertation. Usage of optical flow algorithm on the top of belief-propagation is proposed, so that the finally outputted disparity range is continuous.

The idea of evaluation-by-hypothesis, similar to proposed in dissertation, is employed in paper [220]. A variational segmentation model which intends to decompose an image into distinct regions, using piecewise smooth functions, is employed to compute a smooth depth map, basing on multiple depth hypotheses obtained from different matching algorithms. The certainty of the depth is not considered though, which is a drawback of that proposal.

Regardless of the works mentioned above, currently, there is lack of fast post-processing techniques that could improve the precision of a disparity map and well preserve spatial edges. This observation is one of motivations of this dissertation. **The author proposes a post-processing algorithm that increases precision of generated disparity maps**, preserves spatial edges and is not computationally expensive. The original results for such are presented in Section 3.7.

2.6. Temporal consistency of the depth

The straight-forward method for depth estimation in video sequences is to estimate depth map for each frame independently. Such an approach is simple and also allows for parallel generation of depth map in in consecutive frames [126]. Unfortunately, estimation of depth independently for each of consecutive frames in the video yields with depth maps which are not temporally consistent due to noise. This manifests as random fluctuation of depth values, even of objects that are still. Such fluctuations are adverse, because they lead to occurrence of artificial movement in 3D representation. Desired depth map temporal consistency means that changes of the depth of objects in time are correlated with actual motion of the objects and do not vary from frame to frame in a random way. Therefore, one of the biggest of challenges in this research area is how to provide depth maps that are consistent in time.

Typically, depth data for video is estimated independently for each frame of the sequence. Majority of state-of-the-art techniques that tackle temporal consistency, in various ways expand depth estimation algorithms into time domain. For example, in [221] authors propose to extend standard 4-neighborhood belief propagation depth map estimation scheme [222] to 6-neighborhood scheme by addition of temporal neighbors: from previous and from next frame. These neighbors are obtained by motion estimation. Therefore, depth value is optimized with respect to depth value in subsequent frames. In turn, authors of [223] propose segment-based approach. In order to provide temporally consistent depth value, apart from traditionally used spatial matching of segments, also temporal segment matching is performed. Such approach increase complexity of the whole depth estimation process, which already is computationally expensive.

In work [224] a method for estimating temporally and spatially consistent dense depth maps in multiple camera setups is presented. Authors propose that for this purpose, initially, depth estimation is performed for each camera with the piece-wise planarity assumption and Markov Random Field (MRF) based relaxation at each time instant independently. Then, moving pixels are identified and MRF formulation is updated by the additional information from the depth maps of the consequent frames through motion compensation. For the solution of the MRF formulation for both spatial and temporal consistency, Belief Propagation approach is utilized. The results presented by the authors indicate that the proposed method provide reliable dense depth map estimates both in spatial and temporal domains. Unfortunately, the method comprises substantial modification of belief propagation algorithm, which lowers its usability.

Although the works related to matter of temporal consistency, including the mentioned above, describe a huge variety of techniques, still, generating maps that are consistent in time is still a scientific problem that requires further investigation. This is one of the author’s goals of the dissertation. In particular, **the goal is to further study and develop proposal given the author** in [3] and [20], where problem of **temporal inconsistency** is tackled by elimination of its cause which is existence of noise in video sequences. **Therefore, noise removal techniques are employed.**

Noise reduction is a well-known and widely recognized technical field. Wide variety of examples of noise reduction techniques can be found e.g. in [225][226][227] [228][229][230][231][232][233][234][235]. Classical noise reduction techniques aim to provide a denoised image directly to the audience. In the case of depth estimation, more artifacts are allowed, because denoised version of the image is only to be used for depth estimation. Thus, wider range of techniques can be considered. Though, the problem of the noise reduction itself is not a subject of the dissertation.

The author’s results and developments related to estimation of temporally consistent depth by noise reduction are presented in Chapter 3.8.

Chapter 3. Proposed methods for depth map estimation

In this chapter, novel tools for depth estimation are proposed. First, a theoretical formulation for depth map estimation based on **Maximum A posteriori Probability** (MAP) optimization rule is presented. It is shown what assumptions are required in order to attain classically used Absolute Differences [180] or Squared Differences [181][204] pixel similarity metrics in formulation of $DataCost$ function. Those assumptions are then **verified** and on the basis of attained results, a more general solution and a **formulation of $\boldsymbol{DataCost}$ is proposed**. The parameters of the proposed probabilistic model are measured **empirically** with use of the test sequence set.

Then, similarly, a formulation of $\boldsymbol{TransitionCost}$ **function is proposed** on the basis of a probabilistic model. Also for this model, parameters are measured **empirically** with use of the test sequence set.

Next, assessment of the proposed depth estimation technique with use of proposed $DataCost$ and $TransitionCost$ function is performed. The attained gains are highlighted.

In Section 3.7 a novel tool for refinement of depth map with Mid-Level Hypothesis algorithm is presented which increase the **precision** and **accuracy** of the generated disparity map. The gains in terms of PSNR and computational complexity are shown.

Finally, a tool for improvement of **temporal consistency** of estimated depth maps with use of noise reduction is presented along with its evaluation.

The following Section 3.1 starts with derivation of $DataCost$ function, based on Maximum A posteriori Probability (MAP) optimization rule which will used for formulation of proposed depth estimation algorithm,

3.1. Proposed *Data Cost* derivation based on MAP

This Section starts with derivation of $DataCost$ function, based on Maximum A posteriori Probability (MAP) optimization rule. Attainments of this Section will further be used for formulation of proposed depth estimation algorithm.

As mentioned in the introduction, one of the most crucial aspects in depth estimation is usage of pixel correspondences in the views. Basing on similarity metrics between pixels, the best matching pixel pairs are chosen and used to derive disparity/depth.

In most of the works related to block/image matching (and depth estimation in particular) no theoretical foundation is provided to the problem of optimal selection of the best match

[15][17][126][180][181][204]. Surprisingly, often simple Sums of Absolute or Squared Differences (SAD or SSD in blocks) similarity metrics are considered [180][181][204] without in-depth studies or consideration.

Such empirical approach, without theoretical formulation, is easy, but has disadvantages:

- It does not provide scientific foundation for the considerations,
- As there is no mathematical model, the probability that the chosen match is the best is unknown.
- Thus it is difficult to incorporate empirical proposals as a part of broader framework, like optimization algorithms, where apart from pixel similarity metric (referred to as $DataCost$), also other terms are used ($TransitionCost$).

Therefore, in this dissertation **a theoretical formulation, based on Maximum A posteriori Probability (MAP) is derived.**

Let us consider disparity estimation in a case of two cameras which are perfectly horizontally aligned with parallel optical axes. The views are rectified [152][154] and the distortions [152][154] are assumed to be removed. Therefore epipolar lines are aligned with horizontal rows in the images.

Images from the left view $L_{x,y}$ and from the right view $R_{x,y}$ have the same widths W and the same heights H.

For given row of pixels with coordinate y in both views, observed are pixel luminance values in the left view and in the right view:

$L_{1,y}, L_{2,y}, \dots, L_{W,y}$ – luminance values in the left view,

$R_{1,y}, R_{2,y}, \dots, R_{W,y}$ – luminance values in the right view (both indexed from 1 to W).

All of those are random variables, considered to have been observed and thus these variables constitute our a posteriori observation set.

We search for disparity value $d_{x,y}$ for each pixel at coordinates x, y (in the right view) which would maximize probability of $p(d_{x,y})$ under the condition of a posteriori observations of luminance values in both views. This probability will be demarked as $p_{x,y,d}$:

$$p_{x,y,d} \equiv p\left(d_{x,y} \middle| \left(L_{1,y}, L_{2,y}, \dots, L_{W,y}, R_{1,y}, R_{2,y}, \dots, R_{W,y}\right)\right), \tag{10}$$

where $\left(L_{1,y}, L_{2,y}, \dots, L_{W,y}, R_{1,y}, R_{2,y}, \dots, R_{W,y}\right)$ is overall conditional expression of observation of luminance values.

Therefore MAP rule for selecting optimal disparity value $d_{x,y}^{*}$ can be formulated as follows:

$$d_{x,y}^{*} = \underset{\arg d}{\max}\left(p_{x,y,d}\right). \tag{11}$$

In order to allow the depth estimation algorithm to use the MAP rule (11), the term $p_{x,y,d}$ has to be modeled basing solely on values that are known after the observation (a posteriori), e.g. luminance values in the left view $L_{1,y}, L_{2,y}, \dots, L_{W,y}$ and in the right view $R_{1,y}, R_{2,y}, \dots, R_{W,y}$.

We will transform equation (10) with use of Bayes rule:

$$p(A,B) = p(A) \cdot p(B|A) = p(B) \cdot p(A|B), \tag{12}$$

expressed in the form below:

$$p(B|A) = \frac{p(B) \cdot p(A|B)}{p(A)}. \tag{13}$$

Thus we get:

$$p_{x,y,d} = \frac{p\left(\left(L_{1,y}, L_{2,y}, \dots, L_{W,y}, R_{1,y}, R_{2,y}, \dots, R_{W,y}\right)|d_{x,y}\right) \cdot p(d_{x,y})}{p(L_{1,y}, L_{2,y}, \dots, L_{W,y}, R_{1,y}, R_{2,y}, \dots, R_{W,y})}, \tag{14}$$

which, by rearrangement of $(\dots)|d_{x,y}$ term for each luminance separately, can be written as:

$$p_{x,y,d} = \frac{p(L_{1,y}|d_{x,y}, L_{2,y}|d_{x,y}, \dots, L_{W,y}|d_{x,y}, R_{1,y}|d_{x,y}, R_{2,y}|d_{x,y} \dots, R_{W,y}|d_{x,y}) \cdot p(d_{x,y})}{p(L_{1,y}, L_{2,y}, \dots, L_{W,y}, R_{1,y}, R_{2,y}, \dots, R_{W,y})}. \tag{15}$$

Assumed is presence of noise which has independent realizations in each of the views. Therefore, each of pixel luminance values in the left view $L_{l,y}$ (at coordinates l, y) is independent from each of pixel luminance values in the right view $R_{r,y}$ (at coordinates r, y).

Moreover, when considering the denominator of (15), it can be assumed that also pixel luminance values in the left view $L_{1,y}, L_{2,y}, \dots, L_{W,y}$ are independent from each other, as do pixel luminance values in the right view $R_{1,y}, R_{2,y}, \dots, R_{W,y}$. Specifically, this also holds true for the sought pair of pixels matched by disparity $d_{x,y}$, as denominator of equation (15) does not consider any specific matching or correspondence of pixels, as those probabilities are not conditional with respect to $d_{x,y}$. Therefore, we can simplify the denominator of (15) as:

$$p\left(L_{1,y}, L_{2,y}, \dots, L_{W,y}, R_{1,y}, R_{2,y}, \dots, R_{W,y}\right) = \prod_{l=1..W} p(L_{l,y}) \cdot \prod_{r=1..W} p\left(R_{r,y}\right). \tag{16}$$

A similar simplification could be done in the case of the nominator of (15), but here, on the contrary, probabilities of $L_{l,y}|d_{x,y}$ and $R_{r,y}|d_{x,y}$ are conditional, because are considered under the condition of occurrence of $d_{x,y}$.

Such condition of $d_{x,y}$ means that in the given pixel with with coordinates x, y, for which we calculate $p_{x,y,d}$, a disparity value $d_{x,y}$ is assumed, so that two pixels, in the left and in the right view, correspond to each other. Such pair of pixels is not independent, and therefore probabilities of their luminance values $\mathrm{p}(\mathrm{L}_{\mathrm{l,y}},)$ and $\mathrm{p}(\mathrm{R}_{\mathrm{r,y}},)$ cannot be simplified as in (16). Such exception occurs, when coordinate l in the left view correspond to the same pixel in the right view with coordinate r, which is true when l and r are linked by disparity $d_{x,y}$:

$$\begin{aligned} &r = x \quad (x \text{ expresses the coordinate in the right view for which } d_{x,y} \text{ is considered}),\\ &l = x + d_{x,y}. \end{aligned} \tag{17}$$

For other pairs of pixels (not corresponding to each other), random variables describing their luminance values are independent, like in the case of (16). Therefore, we can express $p_{x,y,d}$ from (15) as:

$$p_{x,y,d} = \frac{\prod_{l=1..W, l\neq x+d_{x,y}} p(L_{l,y}|d_{x,y}) \cdot \prod_{r=1..W, r\neq x} p(R_{r,y}|d_{x,y})}{\prod_{l=1..W} p(L_{l,y}) \cdot \prod_{r=1..W} p(R_{r,y})} \cdot p\left(\left(L_{x+d_{x,y},y}, R_{x,y}\right)|d_{x,y}\right) \cdot p(d_{x,y}). \tag{18}$$

Also, with the exception for the mentioned case (17), the probability distributions related to $p(L_{l,y}|d_{x,y})$ and $p(R_{r,y}|d_{x,y})$ are independent from $d_{x,y}$ (because those random variables represent pixels that are not connected by disparity $d_{x,y}$) thus:

$$p_{x,y,d} = \frac{\prod_{l=1..W, l\neq x+d_{x,y}} p(L_{l,y}) \cdot \prod_{r=1..W, r\neq x} p(R_{r,y})}{\prod_{l=1..W} p(L_{l,y}) \cdot \prod_{r=1..W} p(R_{r,y})} \cdot p\left(\left(L_{x+d_{x,y},y}, R_{x,y}\right)|d_{x,y}\right) \cdot p(d_{x,y}). \tag{19}$$

It can be noticed, that all $\prod(\dots)$ terms in the nominator can be simplified with $\prod(\dots)$ terms in the denominator of (19). This applies to all l and r, with except for the case (17):

$$p_{x,y,d} = \frac{1}{p(L_{x+d_{x,y},y}) \cdot p(R_{x,y})} \cdot p\left(\left(L_{x+d_{x,y},y}, R_{x,y}\right)|d_{x,y}\right) \cdot p(d_{x,y}) \quad . \tag{20}$$

It can be further seen, that term $p(L_{x+d_{x,y},y})$ is probability distribution of luminance values in the left view, which is independent from the corresponding disparity value $d_{x,y}$ and therefore can be expressed as $p(L_{x,y})$. We finally get:

$$p_{x,y,d} = \frac{1}{p(L_{x,y}) \cdot p(R_{x,y})} \cdot p\left(\left(L_{x+d_{x,y},y}, R_{x,y}\right) | d_{x,y}\right) \cdot p(d_{x,y}) \quad . \tag{21}$$

The derivation of **formula (21) is one of the key achievements of the dissertation.** It describes probability $p_{x,y,d}$ that given pixel with coordinates x, y has disparity $d_{x,y}$ under the condition of a posteriori observations of luminance values in both views.

Therefore, selection of $d_{x,y}$ which maximizes $p_{x,y,d}$ fulfills Maximum A posteriori Probability (MAP) rule (11). In Section 3.4 it will be used in order to propose a novel depth estimation method. In the meanwhile, in Section 3.2, it will be shown how the mentioned equation (21) can be simplified in order to attain classical Squared Differences (and thus Sum of Squared Differences for blocks – SSD) and Absolute Differences (and thus Sum of Absolute Differences for blocks – SAD) pixel similarity metrics that are commonly used in depth estimation algorithms.

3.2. Simplification of *Data Cost* to classical SSD and SAD similarity metrics

Let's now analyze the equation (21), derived in the previous Section, expressing probability $p_{x,y,d}$ that given pixel with coordinates x, y has disparity $d_{x,y}$, basing on Maximum A posteriori Probability (MAP) rule. In this Section, a simplification of (21) is shown, which can be used to attain classical pixel similarity metrics: Squared Differences (and thus Sum of Squared Differences for blocks – SSD) and Absolute Differences (and thus Sum of Absolute Differences for blocks – SAD). The presented simplification is interesting as it **shows the set of conditions (resulting from assumptions)** which, if are met in a practical case, indicate that usage of SAD or SSD is optimal from Maximum A posteriori Probability optimization point of view. Therefore, it will be shown in what cases, usage of SAD or SSD is optimal. It can be noted though, that the presented reasoning does not limit the application of SAD or SSD pixel similarity metrics to the presented cases only.

Terms $p(L_{x,y})$ and $p(R_{x,y})$ are probability distributions of luminance values in the left and right view respectively. They can simply be measured as histograms of the left and the right view. The interpretation of these terms is that correspondence between pixels with luminance values that occur more often is more probable. The mentioned terms are omitted by the state-

of-the-art pixel similarity metrics proposals. Such corresponds to a situation, where histograms of the compared images are flat.

Similarly, $p(d_{x,y})$, probability distribution of disparity values $d_{x,y}$ can be estimated as histogram. It can be imagined, that this brings some quality to the distinction between depth planes (e.g. foreground vs. background). This also is omitted by the state-of-the-art pixel similarity metrics proposals, which corresponds to a situation, where all disparities are equally probable.

The term $p\left(\left(L_{x+d_{x,y},y}, R_{x,y}\right)|d_{x,y}\right)$ is a probability that luminance value $\mathrm{L_{x+d_{x,y},y}}$ of pixel in the left view and luminance value $R_{x,y}$ of pixel in the right view will occur, on the condition that those pixels are corresponding to each other and the occurred disparity is $d_{x,y}$.

Again, according to Bayes rule in form (13), the term $p\left(L_{x+d_{x,y},y}, R_{x,y}\middle|d_{x,y}\right)$ can be expressed alternatively as either:

$$p\left(L_{x+d_{x,y},y}, R_{x,y}\middle|d_{x,y}\right) = p(L_{x+d_{x,y},y}) \cdot p\left(R_{x,y}\middle|L_{x+d_{x,y},y}, d_{x,y}\right) \quad \text{or as} \quad (22)$$

$$p\left(L_{x+d_{x,y},y}, R_{x,y}\middle|d_{x,y}\right) = p(R_{x,y}) \cdot p\left(L_{x+d_{x,y},y}\middle|R_{x,y}, d_{x,y}\right) \quad . \quad (23)$$

Those forms are equivalent and lead to similar formulation, so the work will focus on the latter (23) only. Term $p(R_{x,y})$ simplifies with the term in the denominator of (21) shown on page 51 :

$$p_{x,y,d} = \frac{p(d_{x,y})}{p(L_{x,y})} \cdot p\left(L_{x+d_{x,y},y}\middle|R_{x,y}, d_{x,y}\right) \quad . \quad (24)$$

In order to understand the interpretation of usage of SAD or SSD similarity metric as a model for $p\left(L_{x+d_{x,y},y}\middle|R_{x,y}, d_{x,y}\right)$, we have to do the following assumptions:

- The presence of **additive noise**, the same in both of the views (in particular, with equal standard deviation σ).
- **Lambertian model of reflectance** in the scene, which means that the observed light intensity of given point in the scene is independent from the angle of viewing, and thus is equal amongst the views.
- **Compatible color profiles** of the cameras, so that given light intensity is represented as the same luminance value Y among the views (in the consideration, for given pair of corresponding pixels $L_{l,y}$ in the left view and $R_{r,y}$ in the right view).

Gaussian distribution of the noise

Let us first consider the presence of **Gaussian noise**. For such, the conditions mentioned above can be mathematically expressed as:

$$L_{l,y} \sim Gaussian_{(Y,\sigma)} \quad , \tag{25}$$

$$R_{r,y} \sim Gaussian_{(Y,\sigma)} \quad . \tag{26}$$

where $Gaussian_{(Y,\sigma)}$ is normal probability distribution, with mean value Y and standard deviation σ.

The term $p\left(L_{x+d_{x,y},y}\middle|R_{x,y},d_{x,y}\right)$ is considered and thus random variable $R_{x,y}$ is assumed to be a posteriori observation with given, concrete value (also as $d_{x,y}$ is considered conditionally too), thus $Y = R_{x,y}$. Therefore, the pixels are assumed to correspond to each other and thus both random variables have the same expected value $Y_{x,y}$. Moreover, the difference in luminance between $L_{x+d_{x,y},y}$ and $R_{x,y}$ results only from the probability distribution $Gaussian_{(R_{x,y},\sigma)}\left(L_{x+d_{x,y},y}\right)$ of the noise, where both $R_{x,y}$ and $L_{x+d_{x,y},y}$ are our a posteriori observations:

$$p\left(L_{x+d_{x,y},y}\middle|R_{x,y},d_{x,y}\right) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{\left(L_{x+d_{x,y},y}-R_{x,y}\right)^2}{2\sigma^2}\right) \quad , \tag{27}$$

therefore we get:

$$p_{x,y,d} = \frac{p(d_{x,y})}{p(L_{x,y})} \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{\left(L_{x+d_{x,y},y}-R_{x,y}\right)^2}{2\sigma^2}\right) \quad . \tag{28}$$

We are looking for Maximum A posteriori Probability and thus we search for the best matching disparity d which has the highest (maximal) probability $p_{x,y,d}$. It is equivalent to finding d with maximal $log\left(p_{x,y,d}\right)$. After natural logarithm on both sides of the equation is taken:

$$log\left(p_{x,y,d}\right) = log\left(p(d_{x,y})\right) - log\left(p\left(L_{x,y}\right)\right) - log\left(\sigma\sqrt{2\pi}\right) - \frac{\left(L_{x+d_{x,y},y}-R_{x,y}\right)^2}{2\sigma^2} \quad . \tag{29}$$

It can be noticed that if all terms except the last one (on the right) are omitted, the equation (29) simplifies to SSD formula for pixel similarity metric:

$$log\left(p_{x,y,d}\right) = -\frac{1}{2\sigma^2}\left(L_{x+d,y} - R_{x,y}\right)^2. \tag{30}$$

The terms omitted in such way, $p(d_{x,y})$ and $p(L_{x,y})$ and $log(\sigma\sqrt{2\pi})$ correspond to: probability distribution of disparity values, probability distribution of luminance values in the left view and constant offset, respectively. Such omission could be justified if all of those terms were constants which would be true if both of the mentioned probability distributions were uniform.

We can thus conclude, **that usage of SSD (Sum of Squared Differences) metric is** optimal (from Maximum A posteriori Probability point of view) for the case of presence of **additive Gaussian** noise, **independent** between the views, **uniformity** of distributions of **disparities** and **luminance** values and **Lambertian model of reflectance**.

Laplace distribution of the noise

Now, let us consider the presence of **Laplace distribution of the noise.** If such is assumed, similarly as in the case of Gaussian, we can define the following:

$$L_{l,y} \sim Laplace_{(Y,b)} \quad , \tag{31}$$

$$R_{\mathrm{r,y}} \sim Laplace_{(Y,b)} \quad , \tag{32}$$

where $Laplace_{(C,b)}$ is Laplace probability distribution with mean value Y and the attenuation parameter b.

Analogously to the case of Gaussian distribution above, we can come to conclusion that if the probability distribution is in form of Laplace function:

$$p_{x,y,d} = \frac{p(d_{x,y})}{p(L_{x,y})} \cdot \frac{1}{2 \cdot b} \cdot exp\left(-\frac{\left|L_{x+d_{x,y},y} - R_{x,y}\right|}{b}\right) \quad , \tag{33}$$

and with use the same trick (as in the case of Gaussian noise) with taking logarithm of both sides of (33):

$$log(p_{x,y,d}) = log(p(d_{x,y})) - log\left(p(L_{x,y})\right) - log(2 \cdot b) - \frac{\left|L_{x+d_{x,y},y} - R_{x,y}\right|}{b} \quad . \tag{34}$$

Here, we can see that if all terms except the last one (on the right) are omitted, the equation (34) simplifies to SAD formula for pixel similarity metric:

$$log(p_{x,y,d}) = -\frac{1}{b}\left|L_{x+d_{x,y},y} - R_{x,y}\right| \quad . \tag{35}$$

Again, the omitted terms, $p(d_{x,y})$ and $p(L_{x,y})$ and $log(2 \cdot b)$ correspond to: probability distribution of disparity values, probability distribution of luminance values in the left view

and constant offset, respectively. Such omission could be justified if all of those terms were constants which would be true if both of the mentioned probability distributions were uniform.

We can thus conclude, **that usage of SAD (Sum of Absolute Differences) metric is** optimal (from Maximum A posteriori Probability point of view) for the case of presence of **additive Laplace** noise, **independent** between the views, **uniformity** of distributions of possible **disparities** and **luminance** values and **Lambertian model of reflectance**.

The abovementioned theoretical derivations are novel, mainly because they show a set of conditions, which if are met in a practical case, indicate that usage of SAD or SSD is optimal from Maximum A posteriori Probability optimization point of view. Of course, the presented reasoning does not limit the application of SAD or SSD pixel similarity metrics to the presented cases only and thus usage of SAD or SSD can be found to be optimal in other cases and under optimization on different basis than Maximum A posteriori Probability.

The question arises, whether those conditions (resulting from the assumptions made during the derivation of SSD-based and SAD-based (equations (30) and (35)) pixel similarity metrics for $DataCost$) **are met in practical cases**. This will be verified in Section 3.3 below.

3.3. Verification of the assumptions

In Section 3.1, equation (21) has been derived (see page 51), expressing probability $p_{x,y,d}$ that given pixel with coordinates x, y has disparity $d_{x,y}$, basing on Maximum A posteriori Probability rule. In Section 3.2, under some assumptions, this equation (21) has been simplified to forms related to SSD and SAD (equations (30) and (35) on page 53 and 54) similarity metrics, respectively. In this Section a verification of those assumptions presented in Section 3.2 will be presented, on basis of multiview video test sequence set (Fig. 9 on page 23, Table 1 on page 24).

Let's remind what assumptions have been made in Section 3.2:

- Independence of the noise in the frames (in time domain and in inter-view domain).
- Gaussian (or Laplace) distribution of the noise, the same among all of the views.
- Uniform distributions of luminance values in the views.
- Uniform distributions of disparity values in the views.
- Lambertian model of reflectance in the scene.
- Color profile compatibility among the views.

The assumptions will be verified in the mentioned order. First independence of noise in the frames will be checked, then the shape of the noise distributions (Gaussian or Laplace).

In order to perform the analysis of the noise existing in a practical cases, the noise from the multiview video test sequences will been extracted with use of some proposed method. **The description of this noise extraction method in presented in the following subsection.**

Then, basing on the extracted noise, its characteristics will be evaluated and presented. Finally, a model for the noise will been developed, which is then used in the further part of the dissertation.

3.3.1. Noise extraction technique used for the analysis

The goal of Section 3.3 is to verify whether the conditions assumed in Section 3.2 are met in a practical case of multiview video sequences set, developed by ISO/IEC MPEG and used in the dissertation (Fig. 9 on page 23, Table 1 on page 24). This requires that the noise which is present in the video sequences has to be extracted for analysis.

The noise existing in a video sequence can be simply attained as a difference between the original and a denoised version of the given sequence. There are many methods for noise reduction in video sequences e.g. [228][229][230][231][232][233][234][235], all of which can be used for noise extraction. The most advanced ones include spatial filtering, temporal filtering, Kalman filtering, motion estimation and background extraction techniques. In this chapter though, the purpose of noise reduction is in fact solely noise extraction for further analysis. Therefore, there is no need for use of advanced techniques, and thus **the simplest and the most straight-forward one is used**.

For the noise analysis in this chapter, a very straight-forward technique is exploited, in which fragments of sequences representing **only still scenes** (without any movement) are used. It is assumed that each frame of each sequence represents the same image altered only with different noise. Therefore, the image without noise can be easily retrieved, as an average of the frames.

Denoised image $Denoised(x, y)$ can thus be interpreted as expected value estimator $E[Frame_i(x, y)]$ of many different realizations of random process $Frame_i(x, y)$:

$$Denoised(x, y) = \frac{1}{N} \sum_{i=0}^{N-1} Frame_i(x, y) \tag{36}$$

where $Frame_i(x, y)$ would express luminance value at coordinates x, y in frame with index i.

The sought noise value $Noise_i(x,y)$, can be thus simply calculated as:

$$Noise_i(x,y) = Frame_i(x,y) - Denoised(x,y) \quad . \tag{37}$$

In a case, when frames $Frame_i(x,y)$ are independent from each other, estimator (36) is unweighted. Otherwise, estimator (36) is weighted but it can be noted that the error of the estimation decreases with N. Therefore, even for moderate values of N (like the length of the considered sequences), the error is negligible.

It can be noticed that the noise values $Noise_i(x,y)$ that result from such operation (37) are real numbers. Therefore, they cannot be represented as integers in a classical 8-bit data format without loss. To avoid that, the resulting noise values have been used in their original form (real numbers, stored with single floating-point precision). However, at some part of the presented considerations below, the noise values are analyzed with use of histograms. In such experiments, the probability distributions of noise values $Noise_i(x,y)$ are measured by counting in bins, with bin size equal to $1/16$ of quantization step of the luminance values (e.g. in $Frame_i(x,y)$, where luminance values are stored in 8-bit data format).

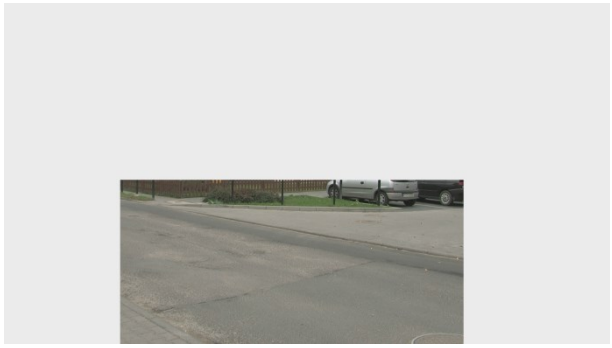
Therefore, given value of the noise $Noise_i(x,y)$ is counted in histogram in bin identified by index $NoiseBinIdx_i(x,y)$ identified as follows:

$$NoiseBinIdx_i(x,y) = \lfloor Noise_i(x,y) \cdot 16 \rfloor \,, \tag{38}$$

where $\lfloor \cdot \rfloor$ depicts the floor rounding operator.

The presented noise extraction method has been applied on the sequences selected for the test set in the dissertation (Fig. 9 on page 23, Table 1 on page 24). Those are mostly moving sequences, which would disallow usage of the presented algorithm. Therefore, only regions and frame ranges that are still for a considerable amount of time (Table 3) have been considered. It is worth to notice, that Poznan Street, Poznan Hall and Poznan Carpark multiview video test sequences have been used in their original, full-length (uncut) version (versions included in ISO/IEC MPEG multiview video test are cut to frame range where there was some movement).

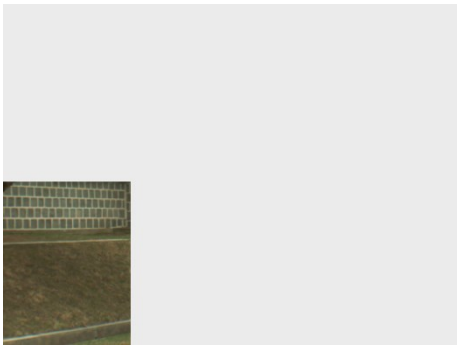
In all of the used sequences, range of frames and spatial regions, in which the scene is still, have been extracted manually. The selection of used regions is summarized in Fig. 19 (spatial positions of selected regions) and in Table 3 (frame ranges and the area of the marked regions).



”Poznan Street” sequence

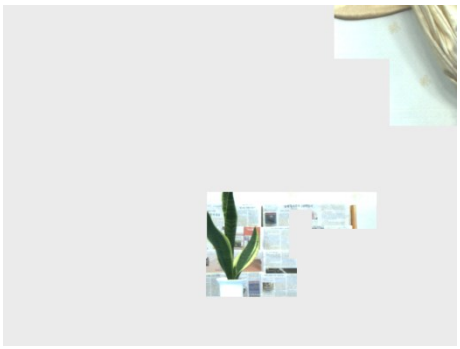
”Poznan Carpark” sequence

”Poznan Hall” sequence



”Lovebird1” sequence

”Balloons” sequence



”Newspaper” sequence

Fig. 19. Regions in the multiview video sequence set that has been manually marked as still for the sake of estimation of the noise. Unused regions have been marked in gray.

In column E in Table 3, number of pixels that have been marked as still (Fig. 18) for each of the sequences is shown. In general, sequences with lower resolution (Column B) have lower number of pixels per view and per frame than sequences with higher resolution. On the other hand, Column F in Table 3 summarizes the total number of pixels in regions marked as still (Fig. 19) per view, but in the whole analyzed range of frames (Table 3 column D).

Those numbers are provided here in Table 3, because they mentioned during **verification of statistical hypotheses**, presented further in Subsections 3.3.2 to 3.3.5.

Computer-generated sequences – GT Fly and Undo Dancer – have not been used for the analysis of noise, clearly because there is no noise in those sequences.

Table 3. Multiview video sequences used for analysis of the noise.

A.	B.	C.	D.	E.	F.
Sequence name	**Resolution**	**Camera**	**Frames used**	**Area of regions marked as still (number of pixels / view / frame)**	**Total number of pixels marked as still per view**
Poznan Street	1920 x 1088	Canon XH-G1, 3-CCD camera	32	563 047	18 017 504
Poznan Carpark			64	587 030	37 569 920
Poznan Hall			32	1 384 287	44 297 184
Lovebird1	1024 x 768	Point Grey Flea camera (CCD), Moritex ML-0813 lenses	64	105 984	6 782 976
Newspaper		Point Grey Research Flea camera with 1/3-inch Sony lenses	32	71680	2 293 760
Balloons		XGA CMOS, 8-bit RGB-Bayer camera	48	27 648	1 327 104
Kendo			Moving seq. – no still regions		
GT Fly	1920 x 1088	Computer-generated sequences – no noise			
Undo Dancer					

For Kendo sequence it was impossible to estimate the noise with use of the technique described in 3.3.1, as the whole scene is moving. Instead of that, a sequence recorded with the same camera system (Table 3) has been used – the Balloons sequence.

After the noise has been extracted with the method described above, the verification of the assumptions made in Section 3.2 could be started. Starting with verification of the assumption about independence of the noise in subsequent frames in the multiview video test sequences, the results are presented below.

3.3.2. Independence of the noise in the subsequent frames

This Subsection presents verification of the assumption of independence of noise in time domain, particularly in subsequent frames in analyzed video sequences. First, the noise has been extracted from the original test sequences (Table 1 on page 24) with technique mentioned in Subsection 3.2.1.

In order to verify the hypothesis of noise independence, the following metrics have been used:

- Pearson-correlation (linear) calculated for each pair of subsequent frames,
- 2-dimensional histograms of noise values for each pair of subsequent frames,
- chi^2 independence test of noise values for each pair of subsequent frames.

As a first check, linear correlation coefficient has been evaluated for each pair of subsequent frames in each of considered (Table 3) test sequences:

$$PCC_{i,i-1}(x,y) = \frac{\sum_{y=1}^{H}\sum_{x=1}^{W}(Noise_i(x,y) - \overline{Noise_i}) \cdot (Noise_{i-1}(x,y) - \overline{Noise_{i-1}})}{\sqrt{\sum_{y=1}^{H}\sum_{x=1}^{W}(Noise_i(x,y) - \overline{Noise_i})^2 \cdot \sum_{y=1}^{H}\sum_{x=1}^{W}(Noise_{i-1}(x,y) - \overline{Noise_i})^2}} \quad . \qquad (39)$$

The results for all pairs of frames in form of graphs are attached in Appendix, in Fig. 81 to Fig. 86. It can be noticed, that for some pairs of frames, the linear correlation coefficient is higher and reaches even level of about 0.1295 (Balloons) – e.g. in Fig. 20. Apart from the value itself, it can be noticed that all of the graphs are quite random, fluctuating on both positive and negative values, which indicate that there is no linear correlation between noise in subsequent frames.

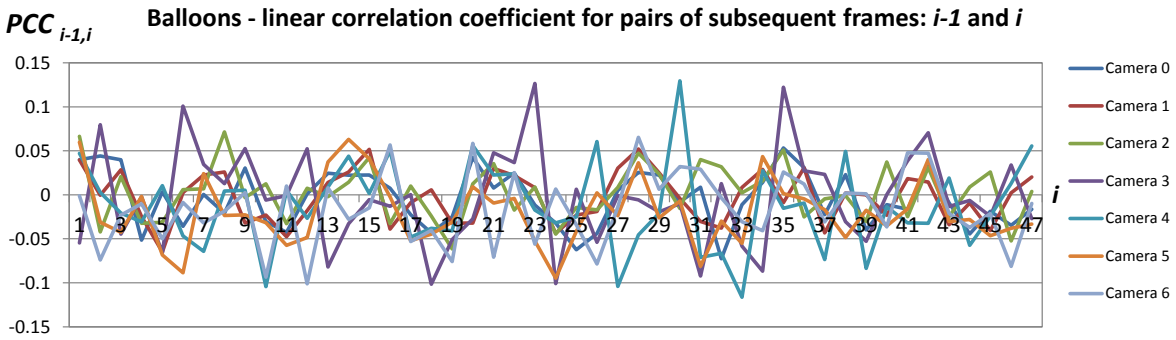


Fig. 20. Linear correlation coefficient measured for two subsequent frames, index $i-1$ and index i, for Balloons sequence.

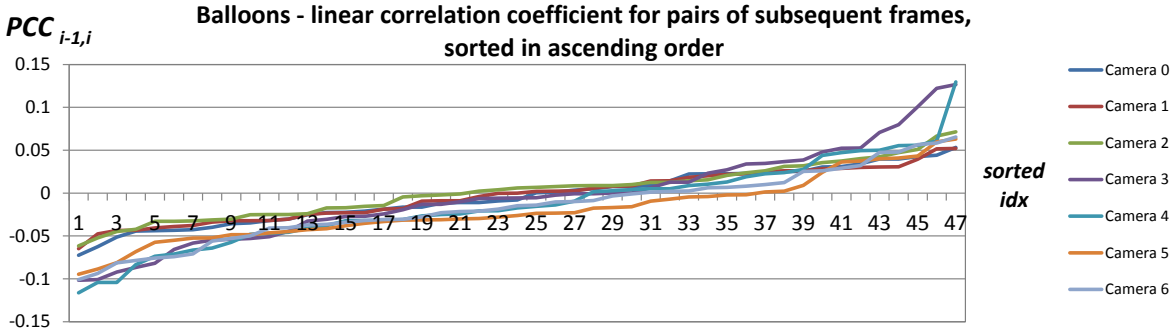


Fig. 21. Results from Fig. 20 sorted in ascending order ($sorted\ idx$). Linear correlation coefficient measured for two subsequent frames for Balloons sequence.

As the results in Fig. 20 are difficult to interpret, the values of correlation coefficient have been sorted in ascending order (from the lowest PCC value to the hightest) and again, in that

new form, presented in Fig. 21. It can be seen, that high levels of PCC values occur rarely and that in the most of frames, PCC ranges from -0.05 to 0.05 which is negligible.

Moreover, even higher values (like 0.1295 for Balloons sequence or about 0.0569 for some Poznan sequences) do not provide evidence that the random variables are linearly correlated. Also levels of the measured correlation coefficient for other sequences (Table 4) indicate that the noise in subsequent frames is not linearly correlated. In average the linear correlation coefficient is about ±(0.01-0.03), which is negligible.

Table 4. Linear correlation coefficient calculated over pairs of frames in test sequence set.

Sequence name	$\max_i \lvert PCC_{i,i-1} \rvert$	$\text{average}_i \lvert PCC_{i,i-1} \rvert$	$\min_i PCC_{i,i-1}$	$\max_i PCC_{i,i-1}$
Poznan Street	0.0569	0.0142	-0.0569	0.0344
Poznan Carpark	0.0561	0.0158	-0.0561	0.0533
Poznan Hall	0.0211	0.0084	-0.0211	0.0138
Lovebird1	0.0396	0.0117	-0.0396	0.0339
Newspaper	0.0360	0.0113	-0.0360	0.0293
Balloons	0.1295	0.0325	-0.1163	0.1295

As mentioned above, the performed experiments indicate that the noise in subsequent frames **is not linearly correlated.** Of course linear correlation is not the only existing form of dependence between random linear variables.

A simple and robust method of verifying whether there is any dependence between two random variables, e.g. α and β is the usage of formula:

$$p(\alpha, \beta) = p(\alpha) \cdot p(\beta) \quad , \tag{40}$$

which is true only if such two variables α and β are independent.

In order to perform such verification, the two-dimensional histograms of $Noise_i(x, y)$ vs. $Noise_{i-1}(x, y)$ have been measured. An exemplary histogram has been presented in Fig. 22. One axis of the histograms relates to the values of noise in the frame i, and the other axis relates to the values of noise in frame $i - 1$. The analogous histograms for other sequences, averaged over all frames have been gathered in Fig. 23 (in the same presentation form as in Fig. 22).

If there would be any dependence between those two random variables (modeling noise in frame i and $i - 1$), there would be an asymmetry in the graph, related to the fact that (40) is not meet.

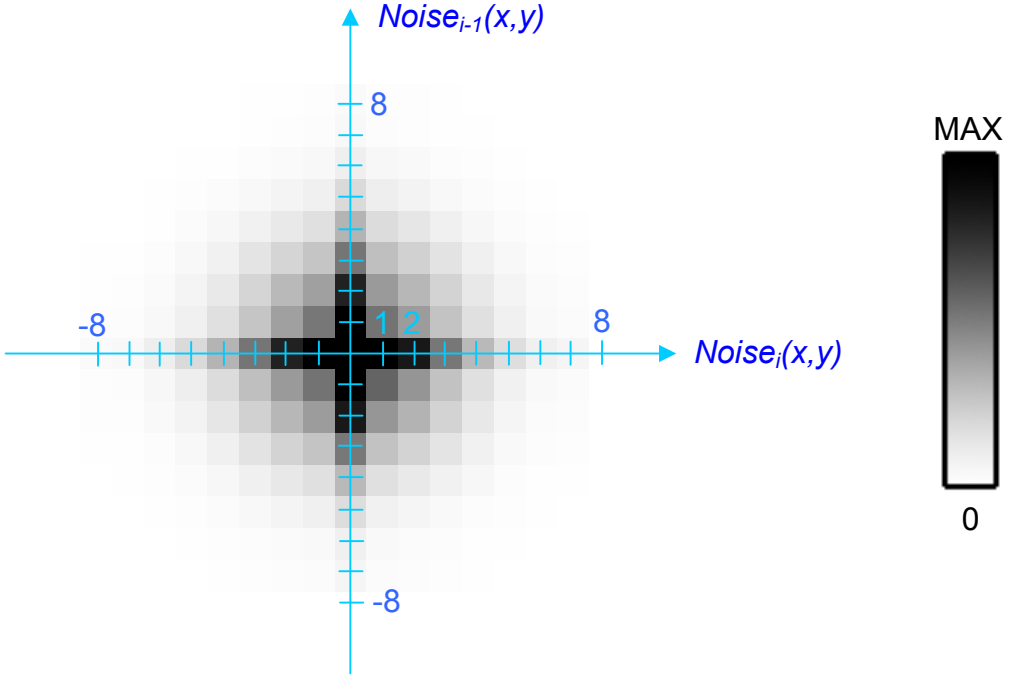


Fig. 22. Exemplary plot of two-dimensional histogram of $Noise_i(x, y)$ vs. $Noise_{i-1}(x, y)$ for Poznan Street test sequence, camera 0, frame 0. One axis of each histogram relates to the values of noise in the frame i, and the other axis relates to the values of noise in frame $i - 1$. The same visualization method has been used in Fig. 23.

\

Camera index / Sequence	0	1	2	3	4	5	6	7	8
Poznan Street (cameras 0..8)									
Poznan Carpark (cameras 0..8)									
Poznan Hall (cameras 0..8)									
Lovebird1 (cameras 0..8)									
Newspaper (cameras 0..8)									
Balloons (cameras 0..6)								-	

Fig. 23. Plots of two-dimensional histogram of $Noise_i(x, y)$ vs. $Noise_{i-1}(x, y)$ for various test sequences and cameras, averaged over frames. The plots are presented in the same way as in Fig. 22 but **for the sake of brevity, visualization of axes has been omitted**.

As can be noticed (Fig. 23), the graphs are separable (40) in both dimensions which indicates that the random variables are independent and thus that the noise in subsequent frames is independent.

Apart from such visual verification, the independence has been tested mathematically. Given normalized 2-dimensional histogram (depicted by $H[\cdot]$ operator, with bins described in (38) on page 57) of noise values in subsequent pair of frames i and $i-1$:

$$H[Noise_i(x,y), Noise_{i-1}(x,y)\,](\alpha,\beta) = h_{i,i-1}(\alpha,\beta), \tag{41}$$

histogram of noise values in frame i:

$$H[Noise_i(x,y)]\,(\alpha,\beta) = h_i(\alpha) \tag{42}$$

and histogram of noise values in frame $i-1$:

$$H[Noise_{i-1}(x,y)]\,(\beta) = h_{i-1}(\beta), \tag{43}$$

we assume that those normalized histograms correspond to probability distributions of noise in the corresponding cases. If the noise distribution are independent between the frames, then according to (40), the expected distribution of $h_{i,i-1}{}'(\alpha,\beta)$ will be:

$$h_{i,i-1}{}'(\alpha,\beta) = h_i(\alpha)\cdot h_{i-1}(\beta) \quad . \tag{44}$$

The energy of difference, between the expected distribution $h_{i,i-1}{}'(\alpha,\beta)$ and the observed one $h_{i,i-1}(\alpha,\beta)$ has been used in order to perform ***chi*^{2} independence test.**

$$\chi_{ind.}{}^2 = \sum_{\alpha\in\Phi}\sum_{\beta\in\Phi}\frac{\left|h_{i,i-1}{}'(\alpha,\beta) - h_{i,i-1}(\alpha,\beta)\right|^2}{h_{i,i-1}{}'(\alpha,\beta)} \quad . \tag{45}$$

Φ is a set of possible noise values. Range [-8…8] has been selected in order to cover the whole usable range on noise values Fig. 22 and at the same time, to avoid small number of sampels in histogram bins, which is desired in case of chi^2 test. Therefore, $-8 \le \alpha,\beta \le 8$, which results in total $\varphi = 17$ of values.

The working null hypothesis is that the observed distributions are dependent.

The working alternative hypothesis is that the observed distributions are independent.

Now, number of degrees of freedom will be calculated, which is equal to the number of cells $\varphi\cdot\varphi$, minus the reduction in degrees of freedom df_{red}. As the expected distribution has

been estimated (it is not known from a theoretical model), the number of degrees of freedom have to be reduced by $df_{red} = \varphi + \varphi - 1$ (the number of rows and cols in (45) is equal to φ). Finally, the number of degrees of freedom is:

$$df_{ind.} = \varphi \cdot \varphi - (\varphi + \varphi - 1) = 256. \tag{46}$$

The confidence level has been assumed to be 0.05 and thus the corresponding ${\chi_{ind.}}^2{}_{critical}$ value, calculated from left-tailed χ^2 distribution, is:

$${\chi_{ind.}}^2{}_{critical} = 294.32067. \tag{47}$$

For each of the sequences and each of views ${\chi_{ind.}}^2$ statistic has been calculated and compared to ${\chi_{ind.}}^2{}_{critical}$ value:

$${\chi_{ind.}}^2{}_{ratio} = \frac{{\chi_{ind.}}^2}{{\chi_{ind.}}^2{}_{critical}}. \tag{48}$$

As left-tailed χ^2 distribution is used, value of ${\chi_{ind.}}^2{}_{ratio}$ which is greater or equal than 1 (measured ${\chi_{ind.}}^2$ statistic is greater/equal than ${\chi_{ind.}}^2{}_{critical}$) means that the null hypothesis cannot be rejected and thus the observed distributions of the noise values may be dependent. Value of ${\chi_{ind.}}^2{}_{ratio}$ which is smaller than 1 (measured ${\chi_{ind.}}^2$ statistic is lesser than ${\chi_{ind.}}^2{}_{critical}$) means that the null hypothesis must be rejected and thus the observed distributions of the noise values are independent (at the given confidence level).

From the results presented in Table 5, it can be seen that ${\chi_{ind.}}^2{}_{ratio}$ is definitely below 1 (ranges from 0.0145 to 0.0387 which is negligible). **This leads to a conclusion that the null hypothesis has to be rejected. Finally, this provides evidence that the noise in subsequent frames is independent.**

Table 5. Results of chi^2 independence test, for pairs of succesive frames of the test sequences. The results have been averages over time and over cameras.

Sequence name	${\chi_{ind.}}^2{}_{ratio}$
Poznan Street	0.0145
Poznan Carpark	0.0249
Poznan Hall	0.0194
Lovebird1	0.0387
Newspaper	0.0269
Balloons	0.0307

In Fig. 23 it can be noticed that in the case of Lovebird 1 sequence, camera 2 of has slightly different graph than other cameras. It can be supposed that this particular view has been acquired with different camera settings. Similar phenomenon can be observed in results presented in the next Subsection.

Now, it has been proven that the noise in succesive frames of tested multiview video test sequences is independent. This gives positive verification of the first of the assumption made in Section 3.2. Next, verified will be the second condition which consists in assumption about Gaussian or Laplace distribution of the noise in the video sequences. The results are provided below.

3.3.3. Probability distributions of the noise

In Section3.2 it has been shown that, from Maximum A posteriori Probability point of view, apart from other conditions, usage of SSD metric is optimal when condition of presence of additive Gaussian noise is meet and that usage of SAD metric is optimal when condition of presence of additive Laplace noise is meet. Here, those conditions will be verified on practical example of multiview test sequences. From those sequences, noise has been extracted with use of the technique mentioned in Subsection 3.3.1. In the previous Subsection it has been proven that the realizations of noise in subsequent frames of the tested sequences are independent. Therefore the sought probability distributions of the noise can be estimated with use of histograms calculated over all frames of each sequence. If the noise was not independent between the frames, averaging over the frames would be statistically incorrect.

For the reasons stated in Subsection 3.3.1, the histogram analysis of the noise is performed with use of bins $NoiseBinIdx_i(x, y)$ defined in (38) on page 57, with bin size of $1/16$ of the normal quantization step of the luminance values (the smallest representable luminance value difference).

The results have been presented in Fig. 24 - Fig. 30 in form of average (over all cameras) for visualization. The source data can be found in Appendix in Fig. 87 - Fig. 102.

As can be noticed, none of measured distribution of the noise, extracted from the test sequence set, represents a Laplace distribution. This means that usage of formula (35) (presented on page 54), which connects Laplace noise distribution with SAD pixel similarity metric **cannot be justified** basing on Maximum A posteriori Probability (MAP) rule. **This is a very important result**, because most of the state-of-the-art depth estimation algorithms

[15][126][142][143][185][192][188][201][204] use some form of SAD pixel similarity metric.

In general, it can be said, that the measured noise distributions are visually very similar to Gaussian (normal) distribution. For the visual comparison, in Fig. 24 - Fig. 30 (and in Fig. 87 - Fig. 102 in the Appendix), apart from the measured data (marked in continuous blue line), on the same figures, also Gaussian (normal) distribution has been depicted (marked in dotted-red line). The visualized Normal distribution has the same parameters: μ and σ (Table 6). There are some exceptions for this mentioned "similarity" to Gaussian distribution though, described below.

In the case of **Poznan Street, Poznan Carpark and Poznan Hall** sequences, the measured distribution is slightly skewed in such a way, that the maximum of the distribution is at position of about 0.4. This may be a results of internal noise reduction algorithm implemented in the Canon XH-G1 camera or a results of internal non-linear processing of data from the camera sensor. Standard deviations are very similar among the views (Table 6), but there are little differences among the sequences. Those are 2.45 (Poznan Street,), 2.28 (Poznan Carpark) and 2.01 (Poznan Hall).

In the case of **Lovebird1** sequence, standard deviations are the lowest in the whole test set and are very similar across all of the cameras – at level of about 0.66. The only exception is camera 2, where the standard deviation is about 2.5 times higher – it has been measured to be about 1.65. This might be evidence that this particular view has been acquainted with different parameters – e.g. the exposure time has been shorter, which has been corrected with higher amplification gain, which also amplified the noise. Apart from that anomaly, the Gaussians are well-symmetric and centered at value of 0. This means that the distribution of the noise in such example is well-centered.

The probability distribution of the noise in **Newspaper** sequence is very similar to Gaussian distribution in all of the cameras. The standard deviations are very similar among the views at a level of about 1.23.

The distributions of **Balloons** sequence strictly follow Gaussian "bell" shape. Also here, standard deviations are very similar among the views, at level of about 1.01.

For **Kendo** sequence it was impossible to estimate the noise with use of the technique described in 3.3.1, as the whole scene is moving. Instead of that, a sequence recorded with the same camera system has been used – the Balloons sequence.

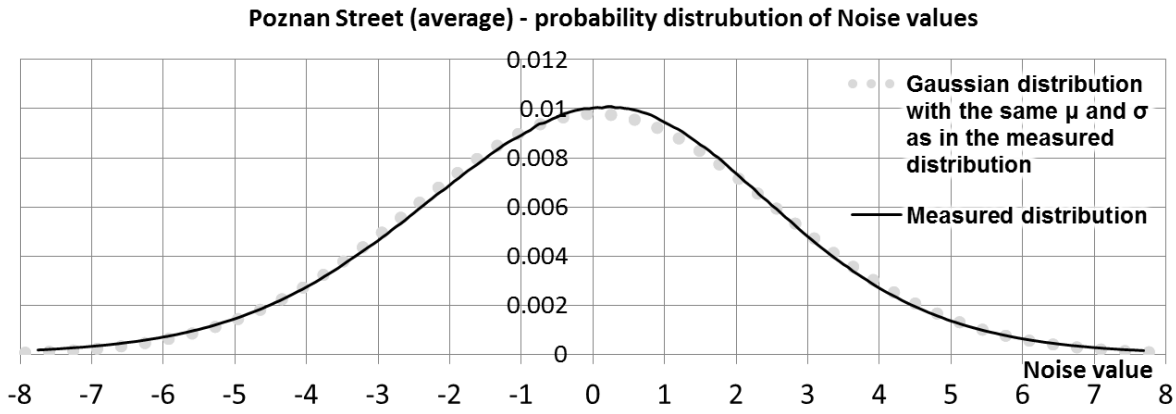


Fig. 24. Measured probability distribution of noise values in Poznan Street sequence (averaged over all views), estimated with histogram bin size of $1/16$.
See (38) on page 57 for more details.

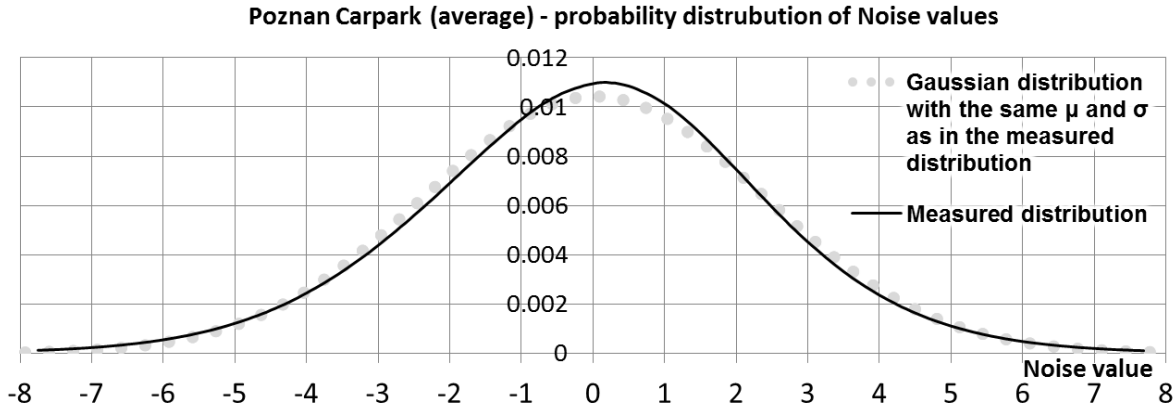


Fig. 25. Measured probability distribution of noise values in Poznan Carpark sequence (averaged over all views), estimated with histogram bin size of $1/16$.
See (38) on page 57 for more details.

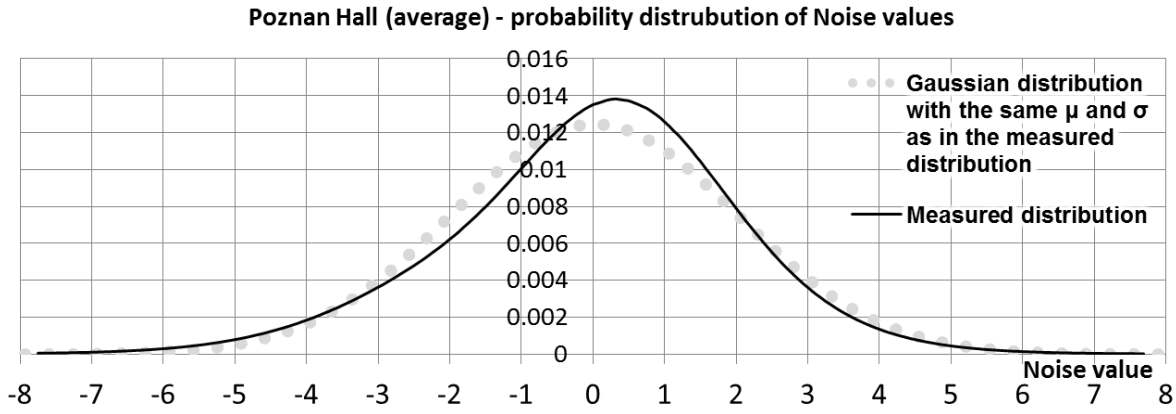


Fig. 26. Measured probability distribution of noise values in Poznan Hall sequence (averaged over all views), estimated with histogram bin size of $1/16$.
See (38) on page 57 for more details.

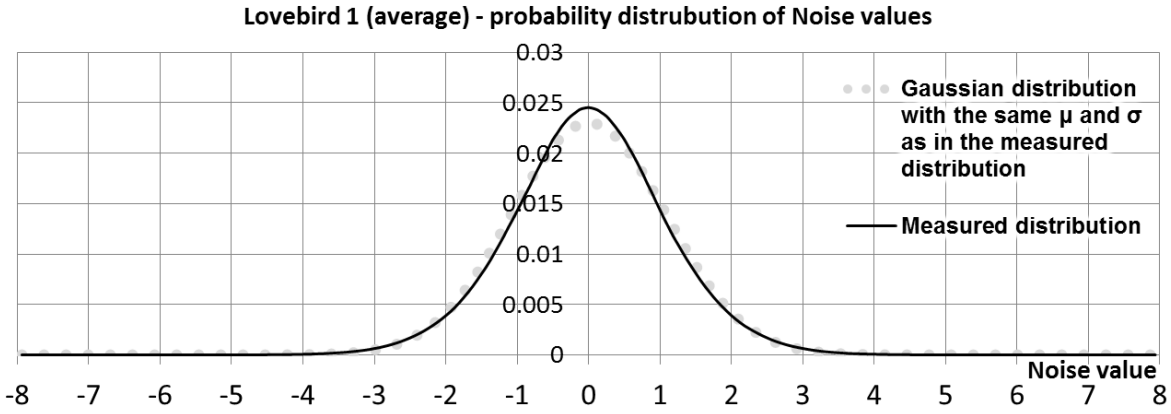


Fig. 27. Measured probability distribution of noise values in Lovebird 1 sequence (averaged over all views), estimated with histogram bin size of $1/16$.
See (38) on page 57 for more details.

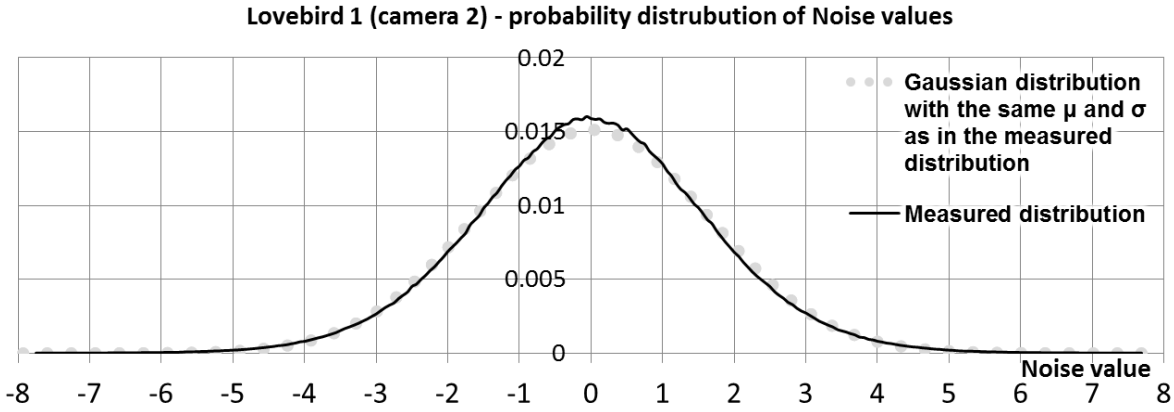


Fig. 28. Measured probability distribution of noise values in Lovebird 1 sequence (camera 2), estimated with histogram bin size of $1/16$(See (38) on page 57 for more details).
In the case of this camera, the standard deviation is about 2.5 times higher than in other cameras of Lovebird 1 sequence..

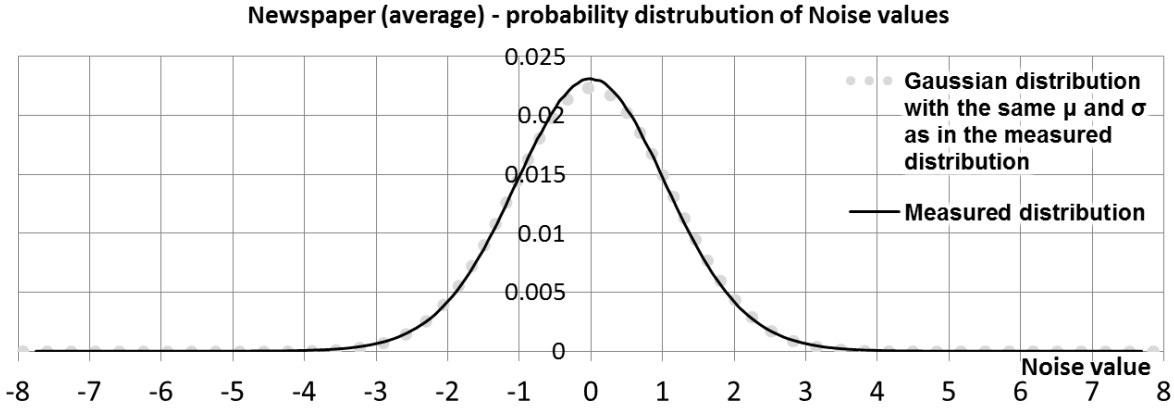


Fig. 29. Measured probability distribution of noise values in Newspaper sequence (averaged over all views), estimated with histogram bin size of $1/16$.
See (38) on page 57 for more details.

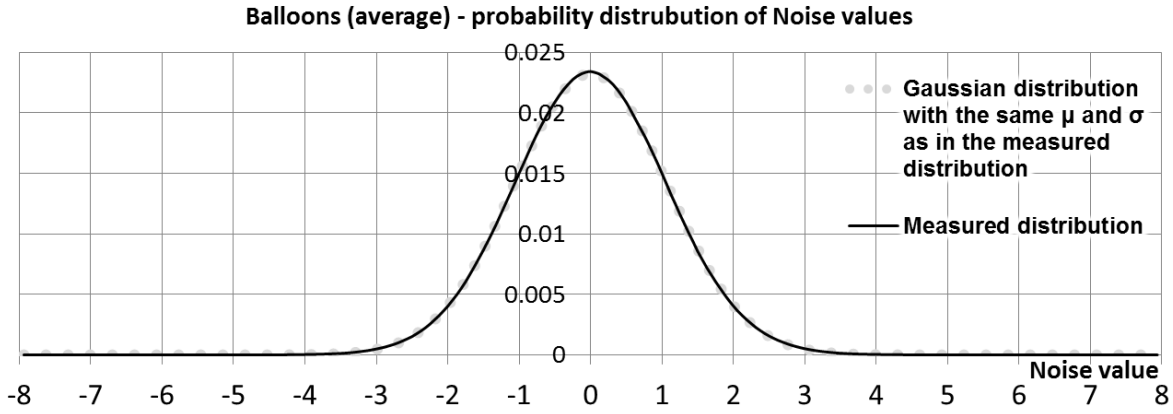


Fig. 30. Measured probability distribution of noise values in Balloons sequence (averaged over all views), estimated with histogram bin size of $1/16$.
See (38) on page 57 for more details.

Table 6. Summary of the Gaussian model - parameters of the noise distributions in the test sequences.

Sequence Name	Standard deviation	Maximum point of distribution, related to EX	Notes
Poznan Street	2.45	0.41	Measured distribution is skewed
Poznan Carpark	2.28	0.42	
Poznan Hall	2.01	0.51	
Lovebird1, w.o. cam.2	0.66	0.02	Camera 2 of Lovebird1 sequence has vastly different noise profile
Lovebird1, camera 2	1.65	0.01	
Newspaper	1.11	-0.02	
Kendo	1.01	0.01	Kendo is a moving sequence – values taken basing on Balloons sequence only
Balloons			
Undo Dancer	Computer-generated sequences – no noise		
GT Fly			

In this Subsection, probability distributions of the noise measured in the test sequences has been presented. Visually, it was noticed that undoubtedly those distributions are not Laplace distributions which violates one of the possible assumptions made in Section 3.2.

Although it was also noticed that the measured distributions are visually very similar to shape of Gaussian function, the statistical proof for that was not provided. This will be the goal of the next Subsection, where assumption about Gaussian shape of the measured probability distributions will be verified statistically with use of chi^2 goodness-of-fit statistical test.

3.3.4. Chi-square test for Gaussian probability distribution of the noise

In previous Subsection it has been shown that the distributions presented in Fig. 24 - Fig. 30 (and also in detail in Appendix in Fig. 87 - Fig. 102) undoubtedly are not Laplace distributions, but in general follow the shape of Gaussian function. Yet it has not been proven whether those distributions are indeed Gaussians or not. First at all, the shape of noise distribution slightly varies among the views. Also, in some of the sequences (Poznan Street, Poznan Carpark, Poznan Hall) the distribution is skewed, such that its maximum point is displaced in relation to the expected value (Table 6).

Therefore, to provide a proof, a statistical test has to be performed. As measured histograms of the considered distributions are available, statistical chi^2 goodness-of-fit statistical test ${\chi^2}_{gof}$ has been used.

In fact, the following reasoning will show that, in spite of the visual similarity, the measured distributions are not Gaussians:

The working null hypothesis is that the observed distribution is normal (Gaussian).

The working alternative hypothesis is that the observed distribution is not normal (Gaussian).

As stated in Subsection 3.3.1, the histogram analysis of the noise is performed with use of bins $\mathrm{NoiseBinIdx_i}(x,y)$ defined in (38) on page 57, with bin size of $1/16$ of the normal quantization step of the luminance value (which correspond to $1/16$ of the smallest representable luminance value difference)). For the practical reasons, each of the distributions (for all sequences and all views) has been observed with use 256 bins. As the dynamic range of noise values is [-255…255] (extreme values), the observed range of noise values in measured histograms is [-8;8], because $\frac{255}{16} \approx 8$.

In the analyzed case of chi^2 goodness-of-fit, the observed distribution thus will be histogram of noise in given view/sequence and the expected distribution is Gaussian. The standard deviation and the mean of expected distribution has been estimated (are not known from a theoretical model) the number of degrees of freedom is:

$$df_{gof} = 256 - 1 - 2 = 253 \quad . \tag{49}$$

The confidence level has been assumed to be 0.05 and thus the corresponding ${\chi_{gof}}^2{}_{critical}$ value, calculated from right-tailed χ^2 distribution, is:

$${\chi_{gof}}^2{}_{critical} = 291.10174 \quad . \tag{50}$$

For each of the sequences and each of views ${\chi_{gof}}^2$ statistic has been calculated and compared to ${\chi_{gof}}^2{}_{critical}$ value.

$$ {\chi_{gof}}^2{}_{ratio} = \frac{{\chi_{gof}}^2}{{\chi_{gof}}^2{}_{critical}} \quad . \tag{51} $$

As right-tailed χ^2 distribution is used, value of ${\chi_{gof}}^2{}_{ratio}$ which is smaller than 1 (measured ${\chi_{gof}}^2$ statistic is lesser than ${\chi_{gof}}^2{}_{critical}$) means that the null hypothesis cannot be rejected and thus the observed distribution may be Gaussian. Value of ${\chi_{gof}}^2{}_{ratio}$ which is greater or equal than 1 (measured ${\chi_{gof}}^2$ statistic is greater/equal than ${\chi_{gof}}^2{}_{critical}$) means that the null hypothesis must be rejected and thus the observed distribution is not Gaussian.

The results of equation (51) calculated for the test sequences are gathered in Table 7. It can be noticed that for the most of the cases, the ratio between ${\chi_{gof}}^2$ and ${\chi_{gof}}^2{}_{critical}$ is of magnitude of about 10^1 – 10^2 proving that the distributions are not Gaussians. The only exception is the Balloons sequence, where ${\chi_{gof}}^2{}_{ratio}$ fluctuates around 1 (the presented multiplied showing the level of magnitude of 10^0). Thus, depending on a particular camera of the Balloons, the hypothesis that the distributions are Gaussians must be rejected (marked in white in Table 7) in or may not be rejected (marked in gray in Table 7).

Therefore, in spite of the visual impression that the observed probability distributions are Gaussian-like, generally it can be concluded that **for most of the sequences, the null-hypothesis must be rejected and almost none of them is Gaussian** (at given confidence level).

Table 7. $\chi^2 ratio$ results for all views of the tested sequences. Values that are less than 1.0 (marked in gray) indicate that given cases pass the χ^2 test.

Sequence \ Camera index		0	1	2	3	4	5	6	7	8
Name	**Multiplier**	**$\chi^2 ratio$, scaled by the multiplier**								
Poznan Street (cameras 0..8)	**$10^1\times$**	7.93	7.65	6.71	6.82	7.00	4.90	5.54	5.51	5.11
Poznan Carpark (cameras 0..8)	**$10^2\times$**	3.89	3.56	3.03	3.18	3.03	3.33	3.31	2.02	1.89
Poznan Hall (cameras 0..8)	**$10^3\times$**	2.12	1.66	1.84	1.75	1.64	2.08	1.76	1.55	1.28
Lovebird1 (cameras 0..8)	**$10^2\times$**	0.50	1.49	0.46	1.84	1.95	1.56	1.08	0.86	1.33
Newspaper (cameras 0..8)	**$10^1\times$**	1.30	1.38	1.03	2.07	1.92	1.24	2.03	1.84	2.65
Balloons (cameras 0..6)	**$10^0\times$**	1.03	1.42	1.16	0.88	0.94	1.90	0.69	-	

Both the graphs presented in Fig. 24 - Fig. 30 (and in Fig. 87 - Fig. 102 in the Appendix) and Table 7, refer to the question, whether the probability distributions of the noise in tested multiview video sequences are Gaussians. A comparison of the visual impressions that can done, basing on the mentioned figures (that the distributions are similar to Gaussians), and the results of statistical analysis (that almost none of the distributions are Gaussian) show that there is discrepancy between those two methods.

This discrepancy (between the visual impressions and results of ${\chi^2}_{gof} test$) can be explained on the basis of number of observed samples. As number of samples increase, the χ^2 *test* becomes more discriminating. With a large number of observed samples (millions in the experiment - Table 3 column G, on page 59), the measured distribution should be almost exactly Gaussian in order to pass through the chi^2 test, while the measured distributions still have slight variations (Table 6 on page 69, Fig. 24 - Fig. 30 and also in Appendix in Fig. 87 - Fig. 102).

As almost none of the test sequences have passed the performed chi^2 test and the null-hypothesis (that their noise distributions are Gaussian) have been rejected, a conclusion about formula (30), presented on page 53 in Section 3.2, which connects Gaussian noise distribution with SSD pixel similarity metric can be drawn. Basing on Maximum A posteriori Probability (MAP) rule, the usage of this formula **cannot be justified**. **This is an important result,** because many state-of-the-art depth estimation algorithms [191][192][195][194] use some form of SSD pixel similarity metric.

The results described above conclude the part of the dissertation aimed at verification of the assumptions about the noise in the multiview video sequences. In the following Subsections, other assumptions will be tested. First, verification of the assumption about the uniformity of distributions of luminance values will be provided.

3.3.5. Uniformity of probability distributions of luminance value

In this Subsection a short verification of uniformity of distribution of luminance values is provided. This was one of the assumptions made during the derivations in Section 3.2.

The desired distributions have been measured as simple histograms of luminance, averaged over all frames of the tested sequences in each view separately – a total number of 256 bins have been used.

The examples are presented in Fig. 31 while the detailed results can be found in Appendix (Fig. 113 to Fig. 120). The graphs have been normalized to range [0;1] for the sake of visualization.

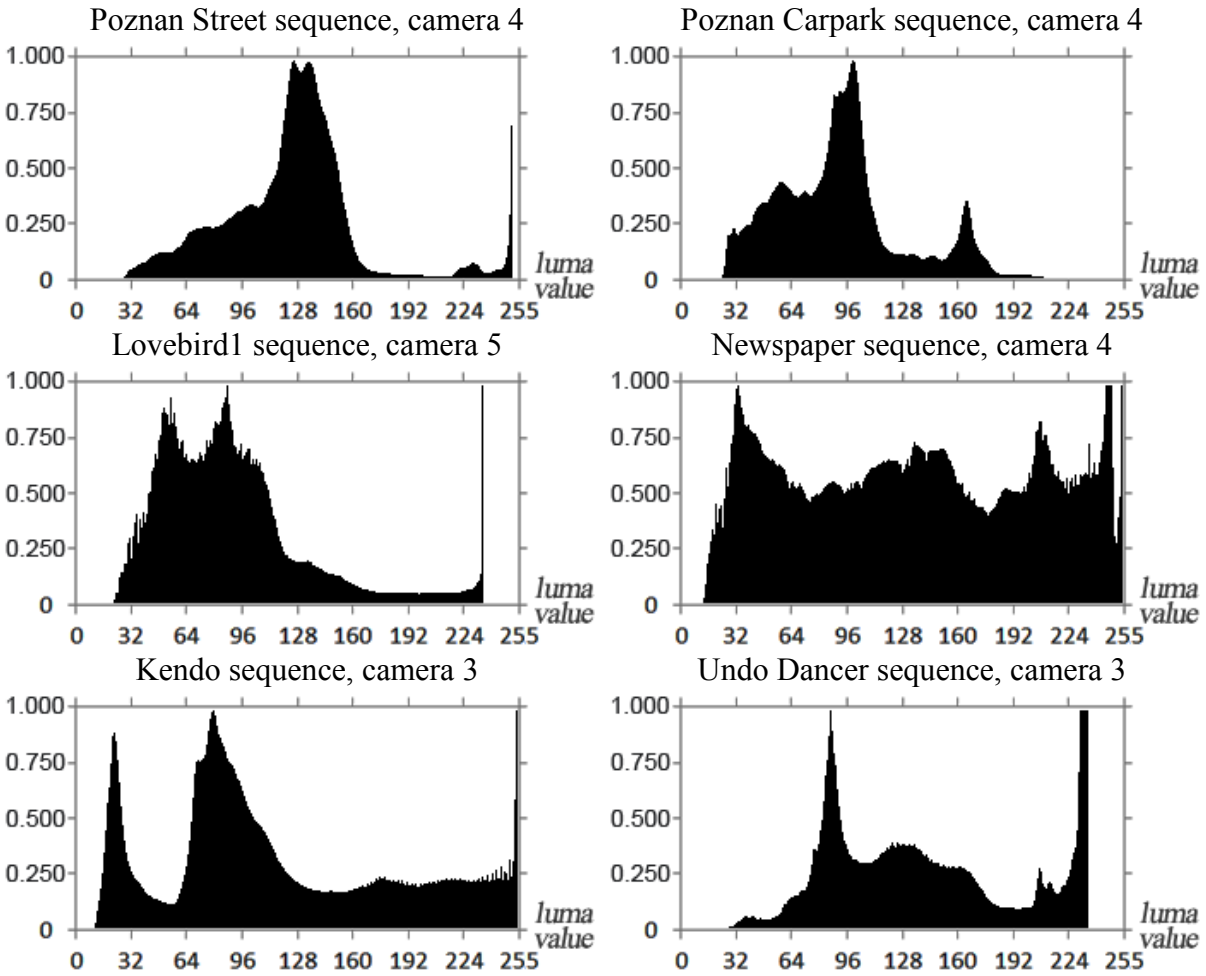


Fig. 31. Histograms of luminance values of pixels in exemplary views of the test sequences. The graphs have been normalized to range [0;1].

Even a short visual verification shows that the **luminance distributions are not uniform** and that more detailed analysis would be redundant. In fact, the figures are brought into the dissertation, only because **they will be used further** in the work.

It can be concluded that yet **another of assumptions is violated**, from the ones made in Section 3.2 (uniformity of luminance distributions) while performing simplification of the formula (21) (page 51) to forms connected with SSD (30) and SAD (35) (pages 53 and 54).

This confirms that the usage of the mentioned formulas **cannot be justified** basing on Maximum A posteriori Probability (MAP) rule.

In the next Subsection, similarly to this Subsection, verification of another assumption, about uniformity of distributions of disparity values, is performed.

3.3.6. Uniformity of probability distributions of disparity value

In this Subsection a verification of uniformity of distribution of disparity value is provided. This was one of the assumptions made during the derivations in Section 3.2.

The estimation of parameters related to depth values for depth estimation is a chicken-egg problem. Therefore apart from performing the depth estimation, the sought distributions have been measured with use of ground truth depth maps provided with the multiview test sequence set (see: Section 1.5). The depth value distributions have been calculated as simple histograms. Those have been averaged over all frames of the tested sequences in each view separately – a total number of 256 bins have been used, as histograms of depth (normalized disparity) have been used.

As it can be seen, the histograms of ground truth depth maps for natural sequences are sparse - only some of the disparity values are present in the depth maps. It results from a fact that those ground truth depth maps have been estimated with use of a predefined limited set of disparities – e.g. only pixel-precise disparities were considered. As those limited set of disparities is scaled into normalized-disparities (depth) as mentioned in Chapter 2, with use of equations (3) or (4), the resultant depth distribution is sparse.

In case of the computer-generated sequences (Undo Dancer and GT Fly), the histograms are dense, because all of the disparities in normalized range 0..255 are present.

Similarly to the previous Subsection, even a short visual verification shows that the **depth distributions are not uniform** and that more detailed analysis would be also redundant. Again, the figures are brought into the dissertation, only because **they will be used further** in the dissertation.

This Subsection can be concluded that once again **one of the conditions made in Section 3.2 is not meet.** Again, it confirms that the usage of the formulas mentioned on page 53 and page 54 (related to SSD-related formula (30) and SAD-related formula (35), respectively) **cannot be justified** basing on Maximum A posteriori Probability (MAP) rule.

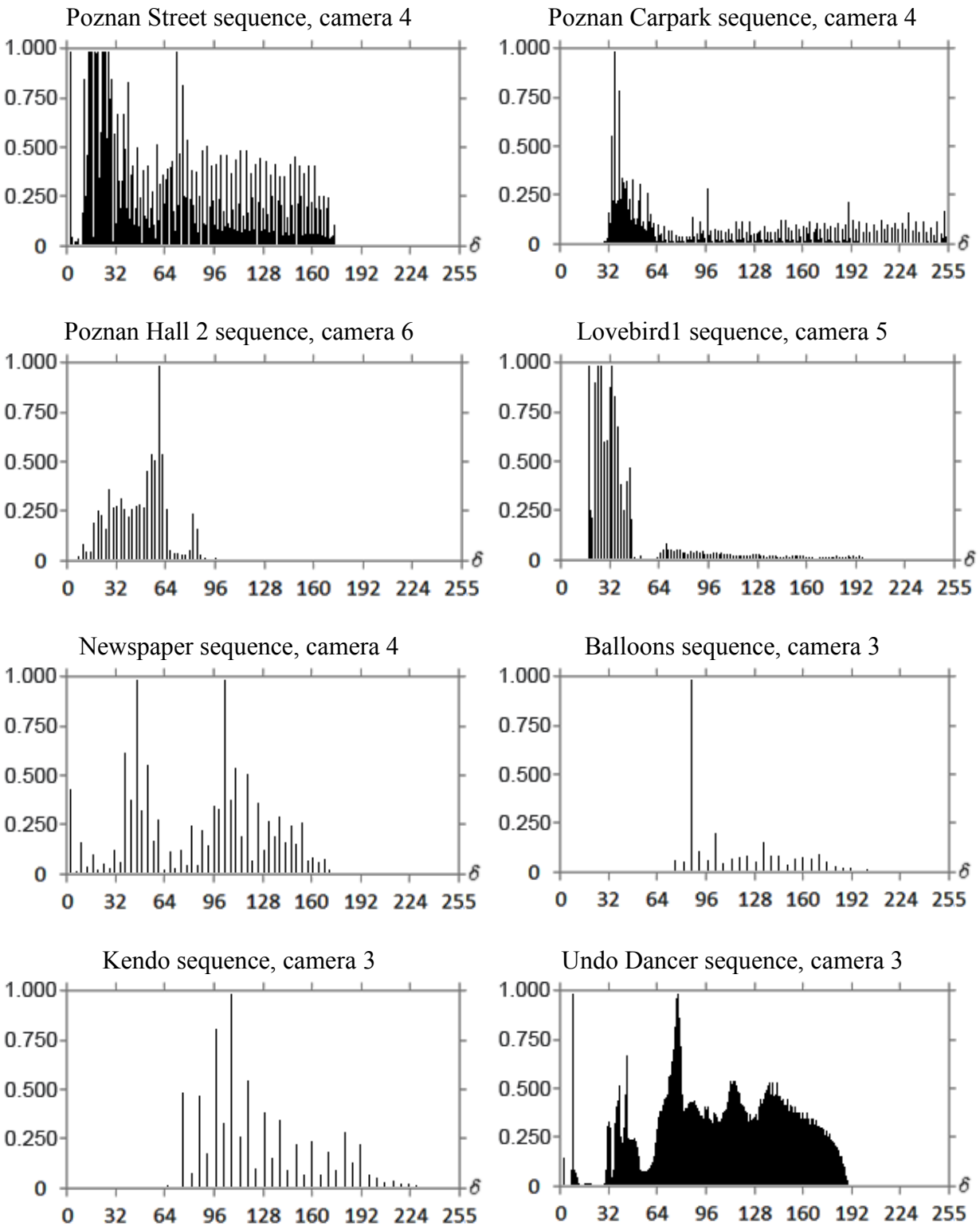


Fig. 32. Histograms of normalized disparity values of pixels in depth maps. The graphs have been normalized to range [0;1].

3.3.7. Lambertian model of reflectance and color profile compatibility among the cameras

In this subsection, the last two of assumptions from Chapter 3.1 that have been made, will be verified. Those are assumption about Lambertian model of reflectance in the scene and assumption about the compatibility of the color profiles among the views. Verification of those two assumptions is provided together, because it will be performed on the same basis.

The most important consequence of **Lambertian model of reflectance** in the scene, for the considerations in the dissertation, is that the observed light intensity of a given point in the scene is independent from the angle and the position from which it is observed by some camera. Thus, the observed light intensity is equal amongst the views.

On the other hand, **color profile compatibility** means that given light intensity is represented as the same luminance value Y among the views. To met that, all parameters and elements used in the cameras (lenses, exposure time, sensor, post-processing) during the acquisition must be calibrated.

Both of those assumptions, if are met, sum to a condition where given point of scene is observed with the same light intensity (Lambertian model of reflectance) which then is represented as the same luminance (color profile compatibility). Therefore, both of those issues can be verified by a test whether the luminance of given pixel is the same in all of the views (which differentiate by position, settings and sensor characteristics) and whether the small differences can be explained purely by the existence of noise (which as has been shown in Subsection 3.2.2 - is independent among the views).

In order to test that an experiment has been performed (Fig. 33), in which values of luminance of corresponding pixels in two views have been compared in a 2-dimensional histogram in which one of the axes is related to pixel value in picture from one camera (called view *X*) and the other axis is related to pixel value in picture from second camera (called view *Y*).

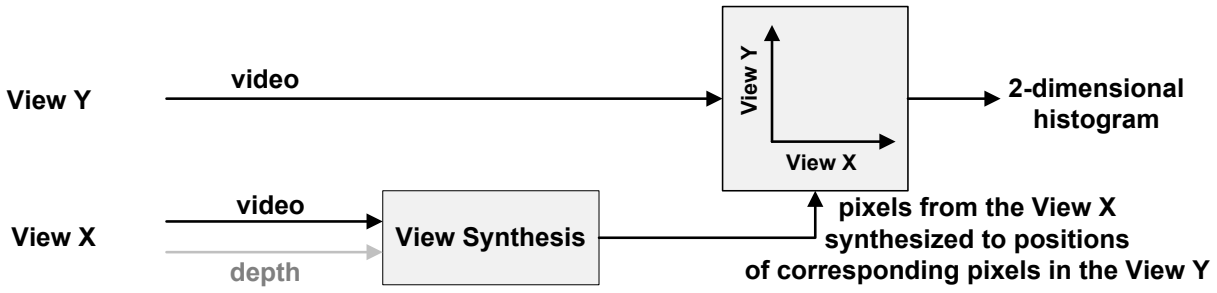
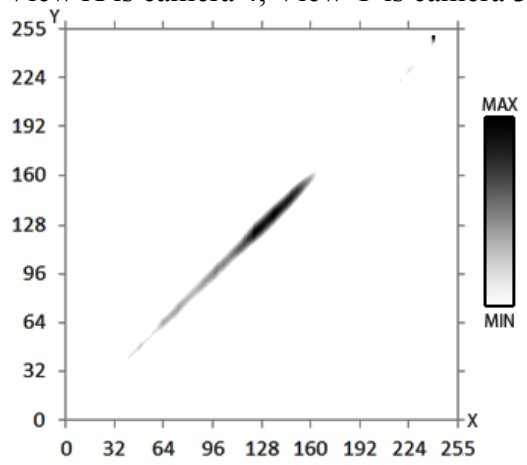
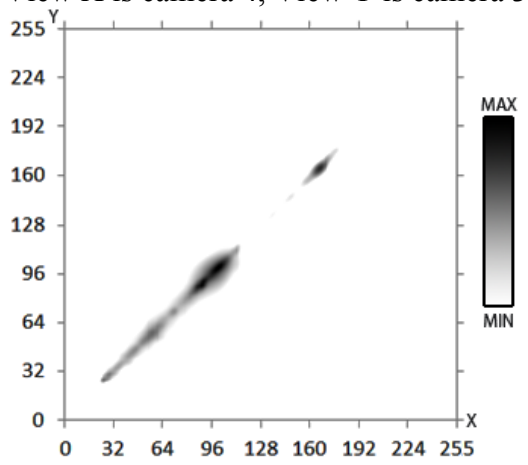


Fig. 33. Scheme of the experiment for verification of Lambertian model of reflectance and color profile compatibility assumptions.

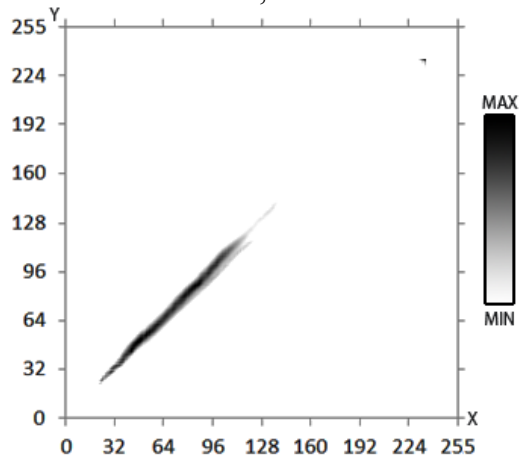
Poznan Street sequence.
View X is camera 4, View Y is camera 3



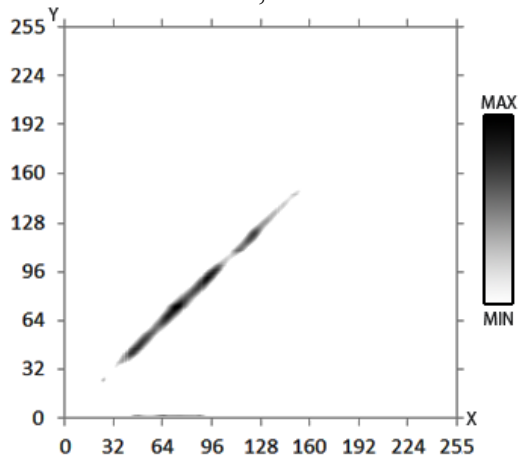
Poznan Carpark sequence.
View X is camera 4, View View Y is camera 3



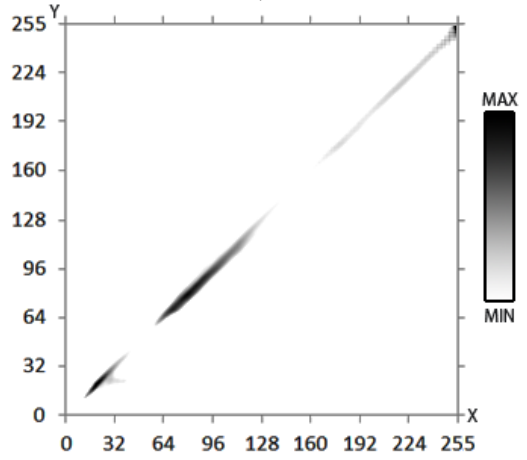
Lovebird1 sequence.
View X is camera 5, View Y is camera 4



Newspaper sequence.
View X is camera 4, View Y is camera 3



Kendo sequence.
View X is camera 3, View Y is camera 2



Computer-generated Undo Dancer sequence.
View X is camera 3, View Y is camera 2

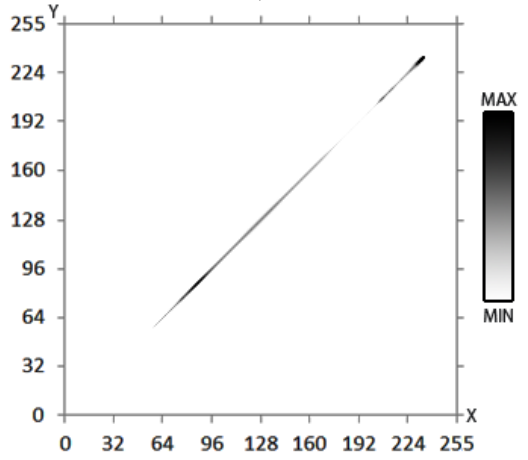
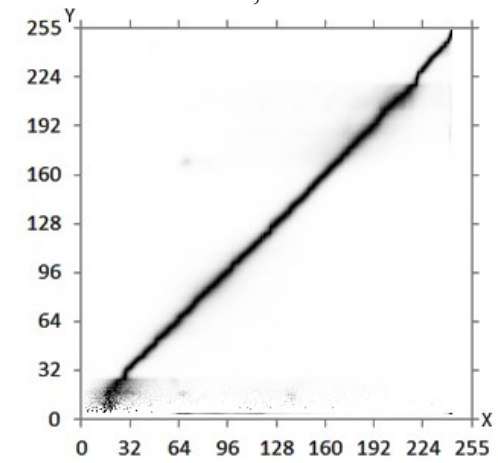
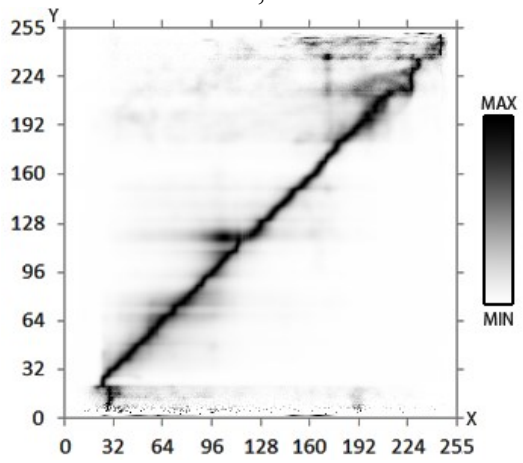


Fig. 34. Graphs of 2-dimensional histograms of luminance values (in logarithmic gray-level scale) of corresponding pixels in the views X and Y for some of the tested sequences.

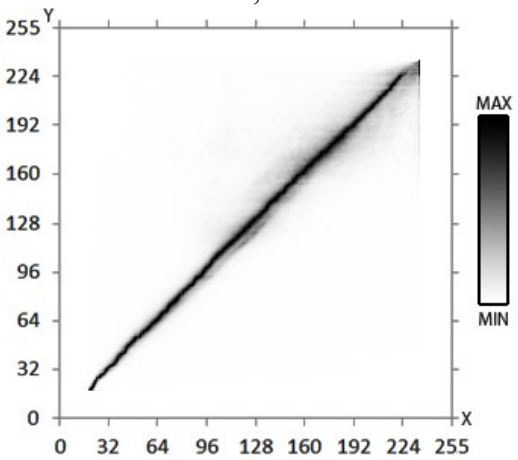
Poznan Street sequence.
View X is camera 4, View Y is camera 3



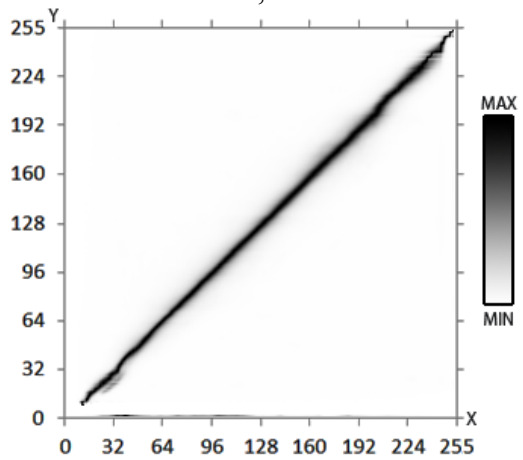
Poznan Carpark sequence.
View X is camera 4, View Y is camera 3



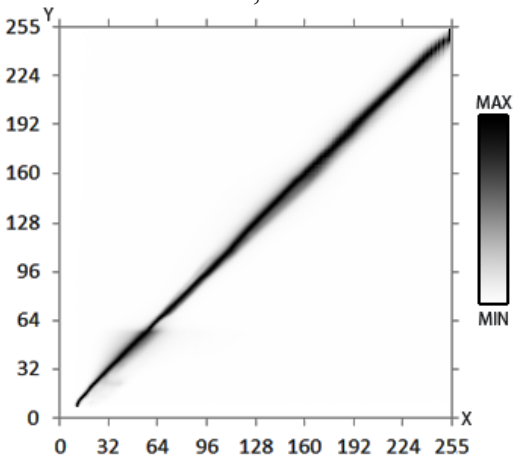
Lovebird1 sequence.
View X is camera 5, View Y is camera 4



Newspaper sequence.
View X is camera 4, View Y is camera 3



Kendo sequence.
View X is camera 3, View Y is camera 2



Synthetic Undo Dancer sequence.
View X is camera 3, View Y is camera 2

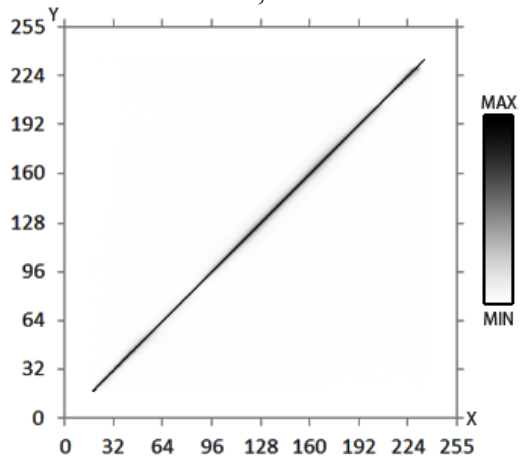


Fig. 35. Graphs of 2-dimensional histograms of luminance values (in logarithmic gray-level scale) of corresponding pixels in the views X and Y for some of the tested sequences.

In the experiment, the correspondence between the pixels in views has been based on ground truth depth maps available for the tested sequence (see: Subsection 1.5.2, Table 1 on page 24). Pixels that have no correspondence in the second view (e.g. are occluded) have been omitted. The test sequences and views required by the settings described in Table 2 on page 26 have been used.

Fig. 34 presents exemplary results of measured 2-dimensional histograms of luminance values of corresponding pixels in views *X* and *Y*. The rest of attained graphs can be found in Appendix – Fig. 103 - Fig. 111. It can be noted that in case of natural sequences, luminance values do not lay strictly on the diagonal of the graph. The curve is slightly distorted which suggests that the color profiles in the considered cameras are not strictly compatible and thus that the color calibration and correction have not been done perfectly. Moreover, the width of the curve is changing over the luminance values, which suggests that the relation between the luminance values in the considered views is not strictly resulting from Lambertian model of reflectance in the scene or that the amplitude of the noise varies with the luminance level.

It can be seen that some of the regions of plots Fig. 34 and Fig. 103 - Fig. 111 that do not occur often are invisible due to normalization. Therefore for presentation purposes, the graphs have been normalized with respect to the probability of given luminance to occur – by luminance histograms of view X. Those histograms used as normalization factor have already been presented in Fig. 24 - Fig. 30 (on page 67) and also in Appendix in Fig. 87 - Fig. 102.

The 2-dimensional histograms of luminance values of corresponding pixels in view X and view Y, normalized in the way mentioned before, are presented in Fig. 35 (examples) and in detail in Appendix - Fig. 121 - Fig. 129.

It is worth to notice, that in case of computer-generated sequences (e.g. Fig. 35 – Undo Dancer or GT Fly) the graphs of presented normalized 2-dimensional histograms are straight lines with unitary width. This confirms that:

- there is no noise for computer-generated sequences,
- the color profiles are compatible among the views of computer-generated sequences,
- the Lambertian model of reflectance has been used for rendering of those for computer-generated sequences.

These abovementioned statements are not true for the natural sequences. The curves for natural sequences are not straight lines (they are curved), they have width (there is noise in the sequences) and the overall shape of the histogram is randomly distorted (which may indicate that the Lambertian model of reflectance is at least not strictly valid).

This is an **important result,** because it shows that two more of the assumptions made in Chapter 3.1 are not true. Therefore it can be concluded, that for the natural sequences, the usage $DataCost$ model, simplified to the form of SSD (30) or SAD (35) pixel similarity metrics, cannot be justified basing on Maximum A posteriori Probability (MAP) rule.

As practically most of the state-of-art depth estimation techniques use some form of SSD or SAD pixel similarity metric [126][142][143][191][192][195], question arise whether their performance could be improved if the knowledge about the distributions measured in this Chapter was known. Answer to this question will be provided in the following Section 3.4.

3.4. The proposed probability model for *Data Cost* function

In previous Section 3.3 a verification has been provided for the assumption made in Section 3.2, during the simplification of formula (21) (see page 51) to classical Squared Differences (and thus SSD - Sum of Squared Differences, formula (30), page 53, for blocks) and Absolute Differences (and thus SAD - Sum of Absolute Differences, formula (35), page 54, for blocks) pixel similarity metrics. It has been shown that at least some of the required conditions are not meet. Therefore it was concluded that the usage of such simplifications cannot be justified basing on Maximum A posteriori Probability (MAP) rule.

In this Section, a **novel depth estimation method will be proposed**.

The main idea of the proposal is that instead of performing the mentioned over-simplification, the derived formula (21) will be used directly as formulation for $DataCost$ function.

To remind, the formula (21) describes Maximum A posteriori Probability that for given pixel with coordinates x, y disparity i has occurred:

$$p_{x,y,d} = \frac{p\left(\left(L_{x+d_{x,y},y}, R_{x,y}\right) | v_{x,y,d}\right) \cdot p(d_{x,y})}{p(L_{x,y}) \ \cdot \ p(R_{x,y})} \quad . \tag{52}$$

In order to use this formula directly, all of the terms of probability in eq. (52) have to be modeled. Fortunately, all of the required terms have already been measured during the verification of the mentioned assumptions in Section 3.3. In particular:

Probability distribution of luminance values in the left view $p(L_{x,y})$ and in the right view $p(R_{x,y})$ have been seamlessly calculated as histograms of the input pictures, as those terms do not depend on pixel correspondence related to disparity $d_{x,y}$. The results for that have already been shown in Fig. 31 (on page 73) and in Appendix: Fig. 113 to Fig. 120.

On the other hand, probability distribution of disparity $p(d_{x,y})$, and probability of corresponding luminance values in the left and the right view $p\left(\left(L_{x+d_{x,y},y}, R_{x,y}\right)|d_{x,y}\right)$ depend on disparity $d_{x,y}$. Having a ground truth disparity map for given scene, both of those terms can be directly modeled:

- $p(d_{x,y})$, which is probability distribution of disparity $d_{x,y}$, has been estimated as a histogram of the given ground truth disparity map (see: Fig. 32 and in Appendix: Fig. 130 - Fig. 134).
- $p\left(\left(L_{x+d_{x,y},y}, R_{x,y}\right)|d_{x,y}\right)$ is a 2-dimensional probability distribution that has been estimated as a 2-dimensional histogram of luminance values $L_{x+d_{x,y},y}$ and $R_{x,y}$ of pixel pairs, which are known to correspond to each other, basing on given disparity value $d_{x,y}$ from the ground truth disparity map. These also have already been shown in Fig. 34 and in Fig. 103 - Fig. 111 in the Appendix.

Finally, having all of the terms measured, we can express $DataCost$ for pixel p (with coordinates x, y) to be equal to the expression presented in equation (52) in logarithmic scale. Usage of logarithmic scale is a common trick used in formulation of energy and probability functions for optimization algorithms [191][197] (see Subsections 2.3.3 and 2.3.4).

We retrieve $DataCost$ to be as follows:

$$DataCost_{x,y}(d_{x,y}) = -10 \cdot \log(p_{x,y,d}) \tag{53}$$

which can be simplified as:

$$\begin{aligned} DataCost_{x,y}(d_{x,y}) = {} & 10 \cdot \log\left(p\left(\left(L_{x+d_{x,y},y}, R_{x,y}\right)\middle|d_{x,y}\right)\right) + \\ & -10 \cdot \log\left(p(d_{x,y})\right) + 10 \cdot \log\left(p(L_{x,y})\right) + 10 \cdot \log\left(p(R_{x,y})\right). \end{aligned} \tag{54}$$

The final formulation of DataCost defined in equation (54) is expressed is logarithmic scale, because state-of-the-art depth estimation algorithm use it for calculations [142][143]. Therefore, such formulation has allowed for direct application of the proposal in graph cuts algorithm implements in MPEG Depth Estimation Reference Software [126].

The results of the proposed depth estimation method, attained with use of the estimated model are presented further. Before that, a formulation for $TransitionCost$ function will be proposed in Section 3.5, so that the results will be reported jointly in in Section 3.6.

3.5. The proposed probability model for *Transition Cost* function

In Chapter 2.2 it has been mentioned, that in the state-of-the-art depth estimation techniques, $TransitionCost$ function between disparities d_p and d_q of neighboring pixels p and q is denoted as $TransitionCost_{p\to q}(d_p, d_q)$. In most of the state-of-the-art depth estimation techniques, $TransitionCost$ typically simplified as a function of a single argument: $|d_p - d_q|$. Examples are (see page 35): Potts model [198] in eq. (7), linear model [126][190] in eq. (8) or truncated-linear model [216] in eq. (9).
Such usage of those models is arbitrarily, due to at least two reasons:

1. The relation between probability of disparities between neighboring nodes is typically not measured empirically and therefore, assumption about the correctness of given $TransitionCost$ model can be verified only by performing the depth estimation.
2. All of the mentioned $TransitionCost$ models incorporate constant parameters (e.g. α and γ in equations (7), (8) and (9) on page 35). Those constant are typically chosen experimentally which is done with limited precision (for example, only four different values of α are tested).

In this dissertation a probabilistic model for $TransitionCost$ is proposed. Similarly as in Chapter 3.1, a theoretical formulation will be shown, which then will be verified with use of empirical estimation basing on the ground truth data.

The proposal employs assumption that $TransitionCost_{p\to q}(d_p, d_q)$ can be modeled basing on probability that given two neighboring pixels p and q have disparities d_p and d_q respectively. This will be denoted as 2-dimensional probability distribution $p_{2D}(d_p, d_q)$, for the sake of brevity and distinction between pixel p and 1-dimensional probability distribution $p_{1D}(\cdot)$ that will be introduced later.

It is assumed that considered probability distribution $p_{2D}(d_p, d_q)$ is independent from position of pixels p and q in the image and the only constraint is that pixels p and q are directly neighboring.

Therefore, as in Section 3.4, we can express $TransitionCost$ in logarithmic scale, so that it could be used directly inside of state-of-the-art depth estimation algorithms [126]:

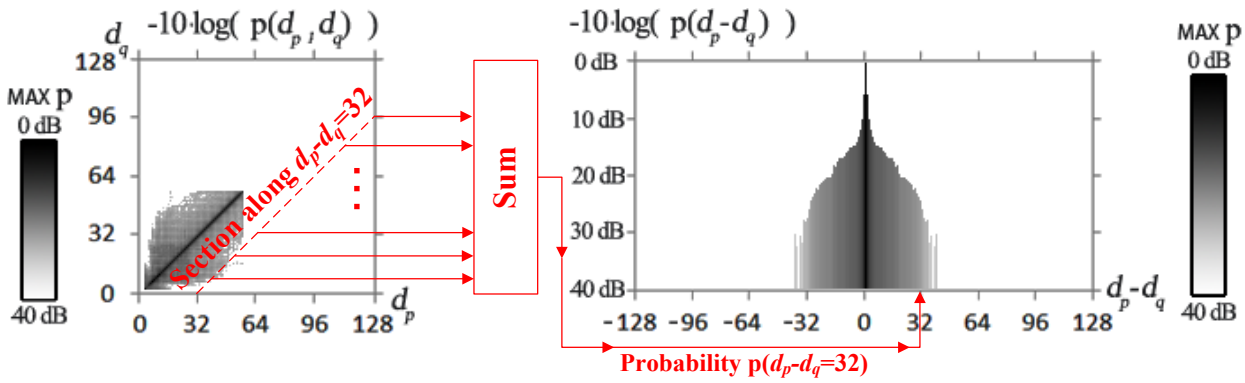
$$TransitionCost_{p\to q}(d_p, d_q) = -10 \cdot \log\left(p_{2D}(d_p, d_q)\right) \ . \tag{55}$$

The main idea of the proposal, similarly as in Section 3.4, is that instead of making assumptions about the shape of $TransitionCost$ function, it will be measured empirically, basing on the ground-truth data available for the test sequences.

The formulation of $TransitionCost$ defined in equation (55) depends on probability distribution $p_{2D}(d_p, d_q)$. For real data it can be measured as 2-dimensional histogram of disparity value pairs d_p and d_q of neighboring pixels p and q. In the dissertation, this has been performed over all frames of all used test sequences and all views for which ground truth depth data is available (sequences listed in Table 1 on page 24).

Some of the results (exemplary histogram per sequence) are presented in Fig. 37 (left column) and Fig. 38 (left column). The rest of the gathered data is provided in Attachment - Fig. 135 to Fig. 143 (left columns). The results are presented in a form described in Fig. 36.

As the distribution of normalized disparity δ is sparse (E.g. in Fig. 32), for the sake of visualization, the plots have been done in domain of disparity d. Therefore, the attained 2-dimensional distribution is dense and monotonic. It can be noticed that the maximum of the curves lay approximately along the diagonal but also there are strong bands on both sides. Such strong band in the histogram means that for given pixel p with disparity d_p, in any neighboring pixel q, a value of disparity d_q is probable to occur if it lays within the probability band of disparity d_p.



Distribution of probability $p_{2D}(d_p, d_q)$ that neighboring pixels: p and q, in the ground truth disparity map have disparity values d_p and d_q , as plot of 2-dimensional histogram.

Distribution of probability $p_{1D}(d_p - d_q)$ that neighboring pixels: p and q, in the ground truth disparity map have difference of disparities $d_p - d_q$, as plot of 1-dimensional histogram, calculated with use of (56). Exemplary calculation for $p_{1D}(d_p - d_q = 32)$ has been shown in red.

Fig. 36. Explanation of plots Fig. 37 - Fig. 38 and Fig. 135 - Fig. 143, showing probability distributions of disparity values d_p and d_q of neighboring pixels p and q. The both histograms are presented in logarithmic scale and in the same shading, where black reflects the maximum probability value, and white reflects 40dB of attenuation of the probability.

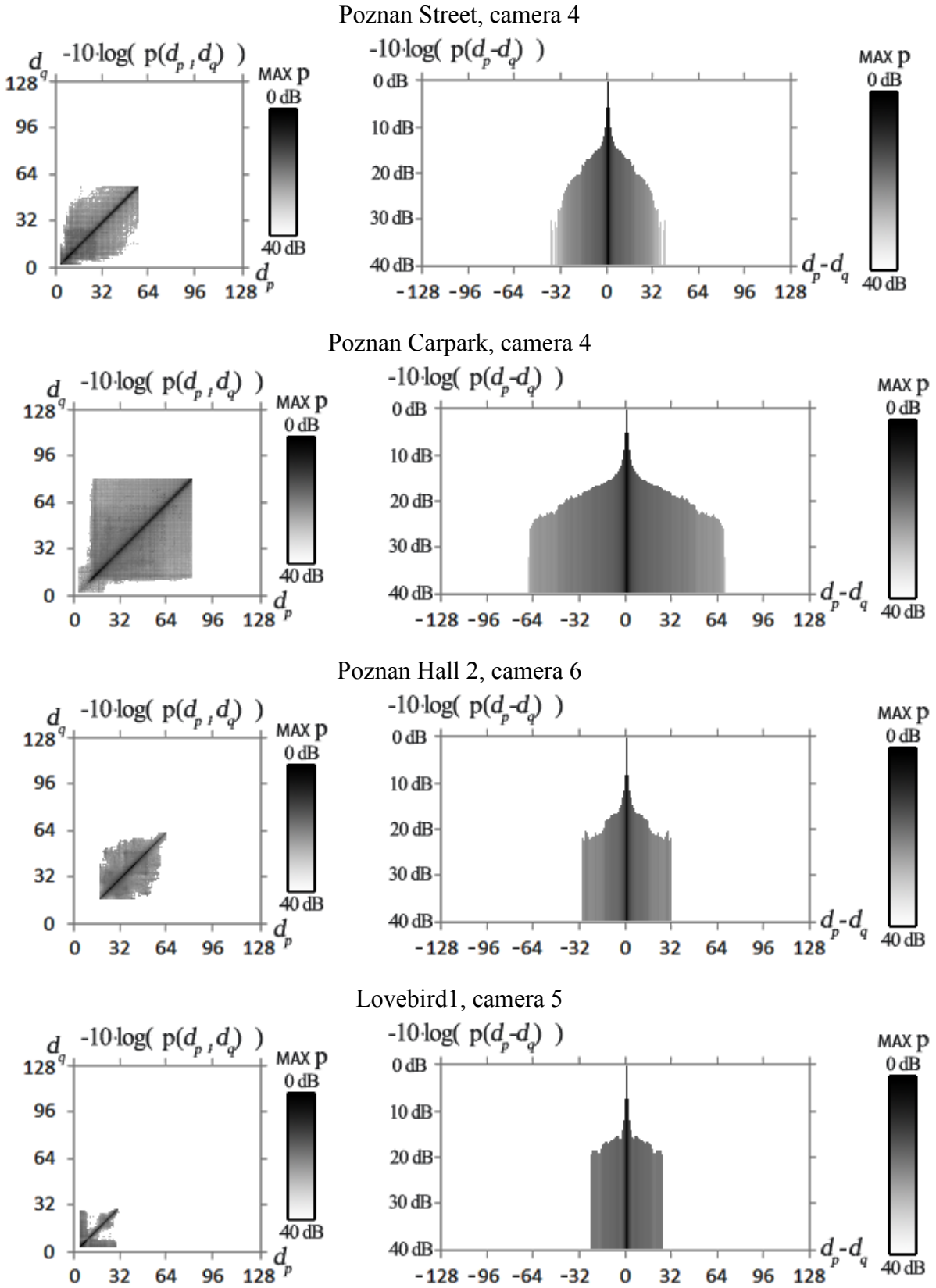


Fig. 37. Histogram of neighboring disparity values: d_p and d_q in ground truth disparity maps for some of the test sequences. The histograms have been visualized as 2D plots (left) and histograms in domain of $d_p - d_q$ disparity difference (right). All plots are presented in logarithmic scale and in the same shading. See Fig. 36 for explanation.

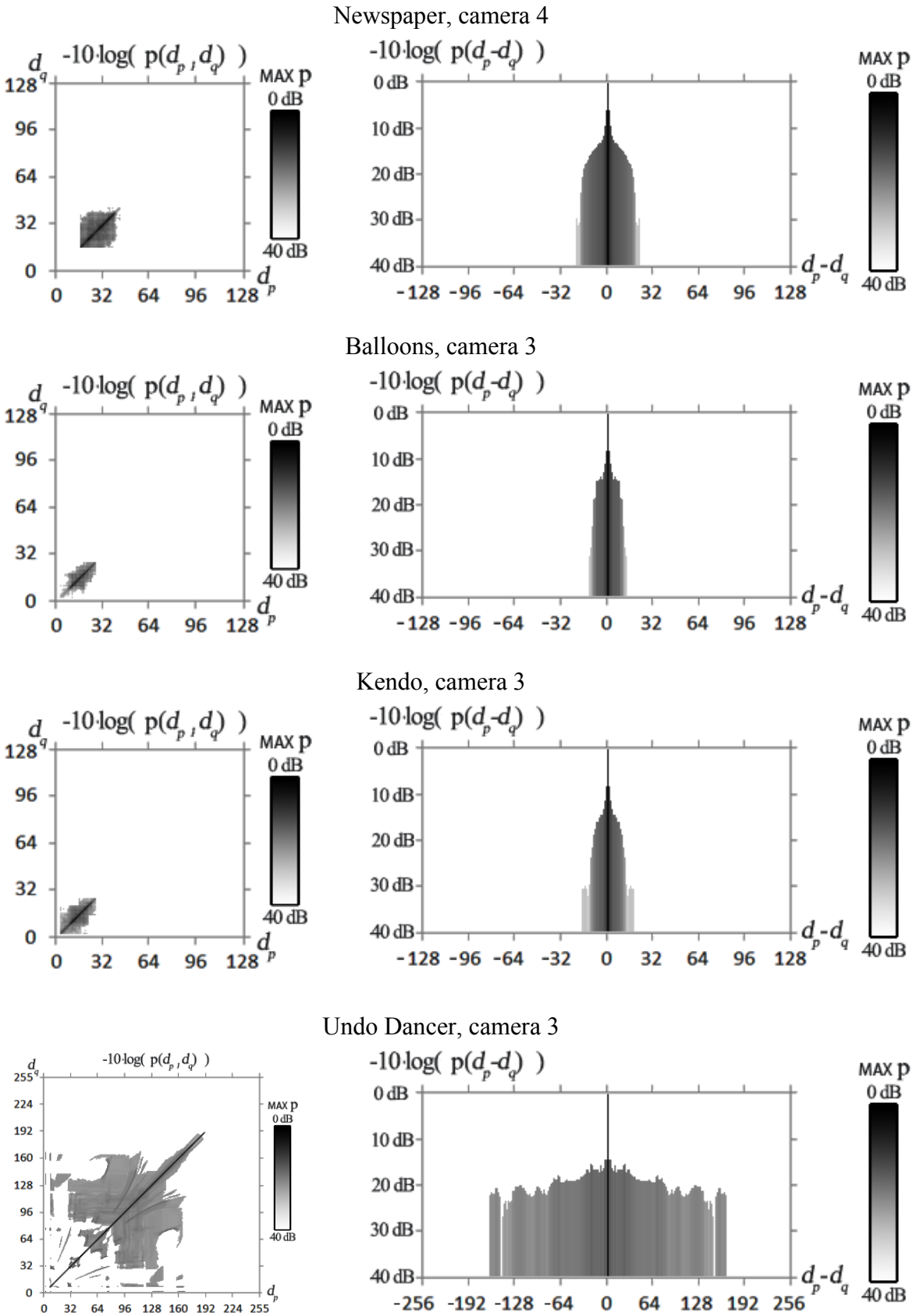


Fig. 38. Histogram of neighboring disparity values d_p and d_q in ground truth disparity maps for some of the test sequences. The histograms have been visualized as 2D plots (left) and histograms in domain of $d_p - d_q$ disparity difference (right). All plots are presented in logarithmic scale. See Fig. 36 for explanation.

Because often, $TransitionCost$ is expressed as a function of a single argument $|d_p - d_q|$, instead of two independent arguments d_p, d_q - e.g. see page 35: eq. (7) [198] or eq. (8) and eq. (9) [126][190][216], it is interesting to also see whether such formulation is justified. In order to do that, apart from figures presenting $p_{2D}(d_p, d_q)$ as 2-dimensional plots (e.g. in Fig. 37 on the left), also 1-dimensional plots of $p_{1D}(d_p - d_q)$ probability of given disparity difference $d_p - d_q$ have been visualized such that (also, see Fig. 37):

$$p_{1D}(d_p - d_q) = \sum_{i,j\ such\ that:\ i-j=d_p-d_q} p_{2D}(i,j) \quad . \tag{56}$$

The results are shown on the right sides of the mentioned figures (Fig. 37 - Fig. 38 and Fig. 135 - Fig. 143 in the Appendix). Having a look on these presented 1-dimensional distributions of $d_p - d_q$ (expressed in logarithmic scale) one can notice that the plots are firstly falling approximately linearly and then they reach plateau until the limits of the histogram plot. Such plots resemble the shapes (examples presented in Fig. 39) of linear model (8)-page 38 and truncated-linear model (9) for $TransitionCost$.

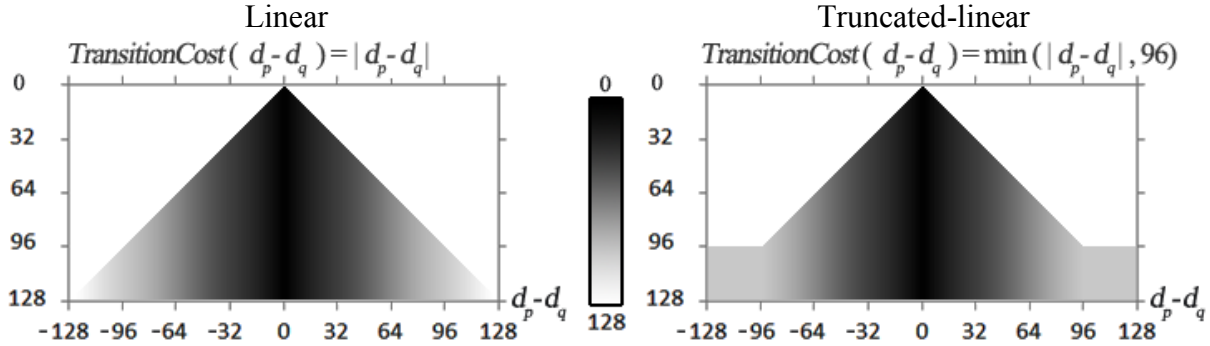


Fig. 39. Exemplary graphs of classical $TransitionCost$ functions: linear (left) and truncated-linear (right). Figure supplied for comparison with graphs presented in the right columns of Fig. 37 and Fig. 38 (detailed results: Fig. 135 - Fig. 143 in the Appendix).

Therefore we can conclude that those classical models (linear and truncated-linear) may be adequate for the case, when the $TransitionCost$ express probability in a logarithmic scale (in which $TransitionCost$ has been depictd in figures). What is important – in case of each sequence, $TransitionCost$ has different scale. Without the knowledge coming from empirical analysis of the $TransitionCost$, performed likewise as in the dissertation, this scale would have to be calibrated manually of experimentally (e.g. with use of Smoothing Coefficient in DERS). **This is an important advantage of the proposal presented in the dissertation.**

3.6. Experimental results for the depth estimation with the proposed *FitCost* model

In the previous Sections 3.4 and 3.5 probabilistic models has been proposed, for $DataCost$ abd $TransitionCost$ respectively. The functional advantages of the proposals has been presented, which include lack of need for manual calibration of parameters.

In this Section an experimental assessment for those models will be provided. Those two proposals together a complete model for $FitCost$ function, which, as mentioned in equation (6) on page 35, is a sum of $DataCost$ and $TransitionCost$ functions. Such $FitCost$ function model will be used in the experimental assessment described below.

The proposed $DataCost$ and $TransitionCost$ models has been implemented into MPEG Depth Estimation Reference Software (DERS), version 5.1 [126]. The tests have been performed, following the evaluation methodology described in Subsection 1.5.3. The used test sequences and view settings have been described in Table 2 on page 26. The model parameters for $DataCost$ and $TransitionCost$, which has been estimated in Section 3.4 and in Section 3.5, have been used.

The original (unmodified) DERS algorithm is a supervised algorithm in a sense, that special control parameter – Smoothing Coefficient – has to be given. Therefore, a wide range of Smoothing Coefficient has been tested. For the sake of brevity, the best and the worst performing settings for each sequence has been identified.

The overall results are presented in Table 8. The more detailed plots can be found in Appendix – Fig. 144 and Fig. 145. It can be seen that the results of DERS with the proposed probabilistic model are very similar to the best case of the original (unmodified) DERS in most of the cases and are very little better in some cases.

In average over the tested sequences, the proposed method provides about **0.08dB** gain over the best identified case generated by the original (unmodified) DERS (with manually crafted Smoothing Coefficient per sequence) and **about 2.79dB** gain over the worst case generated by DERS.

The most important thing to notice is that the proposed depth estimation technique does not require any manual settings (usage of such depth estimation is thus unsupervised). The employed fit-cost function model, based on Maximum A Posteriori rule is inhered from the knowledge coming from analysis of the $TransitionCost$. Therefore, the proposed depth map estimation method has been tested only once in one configuration.

\

Table 8. Gains attained with joint usage of the proposed $DataCost$ and $TransitionCost$ models, related to the best and the worst results attained by the original (unmodified) DERS, depending on Smoothing Coefficient parameter setting.

Sequence Name	PSNR [dB] – virtual view versus the original view. Virtual view was synthesized with use of disparity maps with "full-pixel" precision, estimated with use of:		
	Original (unmodified) DERS - the worst setting of Smoothing Coefficient	Original (unmodified) DERS the best setting of Smoothing Coefficient	Proposed probabilistic model implemented in DERS
Poznan Street	27.56	31.98	32.02
Poznan Carpark	29.05	30.71	30.95
Poznan Hall 2	32.17	32.85	32.81
Lovebird1	27.09	29.80	29.83
Newspaper	27.86	31.91	31.95
Balloons	29.95	32.94	32.98
Kendo	33.02	35.46	35.69
Average	**29.53**	**32.24**	**32.32**
Average gain of the proposed method related to given case	**+2.79**	**+0.08**	**-**

In Sections 3.1 to 3.6 a complete probabilistic model for $FitCost$ ($DataCost$ and $TransitionCost$) has been proposed. The part of the dissertation first started with general theoretical derivation of $DataCost$ based on Maximum a Posteriori Probability rule (Section 3.1). Then, the derived formula (21) from page 51 has been thoughtfully analyzed with respect to simplification (Section 3.2) to classical forms related with to SSD or SAD (equations (30) and (35) on pages 53 and 54) along with verification of the conditions that have to be met for such simplification. It has been shown that at least some of the conditions are not meet in a practical case of multiview test sequences (Section 3.3) and basing on that another formulation for $DataCost$ model has been proposed (Section 3.4). A method for estimation of parameters of this model has been shown on an example of the test sequences. Next, a probabilistic model for $TransitionCost$ has been proposed (in Section 3.5) also with a method for estimation of parameters of this model. This Section concludes these considerations with results presented above.

In the following Sections 3.7 and 3.8 other subjects of the dissertation, related to depth estimation, will be studied. First, the subject of disparity **precision and accuracy refinement** and then the subject of **temporal consistency** will be shown. Each of Sections will self-contain the achieved results.

3.7. Depth refinement by Mid-Level Hypothesis

This Section shows the authors achievements in area of depth estimation related to **precision and accuracy refinement of depth maps**. First, the problem is stated, then, an original proposal of algorithm is presented. Finally, the results conclude this Section.

As mentioned in the Chapter 2, modern state-of-the-art disparity estimation techniques comprise optimization using iterative algorithms like Belief Propagation or Graph Cuts[190][191][201]. These algorithms are robust, but their complexity increases vastly with requested size of the outputted disparity map. In particular, the complexity of disparity map estimation increases with:

- The resolution of matched images - e.g. estimation of Full-HD depth frame takes about 4-times the same time as of XGA frame.
- The disparity search range d_{min} to d_{max} (the computational complexity increases approximately with linear proportion to the width of the selected range).
- The expected precision d_{step} of disparity values - e.g. estimation of disparity map with "half-pixel" precision (d_{step}=0.5) takes approximately twice the time of needed for estimation of "full-pixel" precision (d_{step}=1.0) disparity map.

Fig. 40. Quality (a) and computation complexity (b) of disparity estimation vs. precision of the disparity map, as a function of d_{step} (d_{step} is expressed as a multiple of the spatial sampling period in images).

Unfortunately, complexity of depth estimation increases faster than the growth of benefits from attained higher precision. As research done by the author (Fig. 40) has revealed, increasing the number of disparity levels vastly increases the complexity of disparity estimation, but the fidelity of 3D scene model **tends to saturate**. Results shown in Fig. 40, have been generated with use of ISO/IEC MPEG test sequences [85][240] for which disparity maps have been estimated with algorithm implemented in Depth Estimation

Reference Software (DERS) [126]. Various disparity precision settings of d_{step} have been used - from quarter-pixel precision ($d_{step} = 0.25$) to four-pixel precision ($d_{step} = 4$). With use of the methodology mentioned in Section 1.5, the quality of estimated disparity maps has been measured indirectly (Fig. 10 on page 25) by assessment of quality of a synthesized view, by PSNR related to the original view (PSNR values are depicted in Fig. 40).

The results of the abovementioned experiments are rationale of the statement, that **it is not efficient to estimate high-precision depth maps in a regular single-step process**, because the costs such higher-precision estimation do not provide proportional, desired gains.

In spite of the mentioned saturation effect of fidelity of 3D scene model, accuracy and precision of disparity estimation is an important issue for 3D video representations. Such applications require that disparity is estimated with accuracy to fractions of the spatial sampling periods in images, which in turn means that the precision of the estimated disparity should be sub-pixel. Due to computational complexity, sub-pixel disparity estimation could be difficult in the context of future real-time applications. Therefore, in typical scenario, disparities are estimated with only full-pixel precision [142][143] corresponding to $d_{step} = 1$.

Unfortunately, full-pixel precision is insufficient for most applications related to 3D video. Such lack of precision (and thus also lack of accuracy) is especially noticeable in the case of continuous flat surfaces that are nearly (not exactly) perpendicular to the optical axis of the camera (Fig. 41). In the corresponding disparity map, there exists a false contour as a result of only full-pixel precision of disparity. Such a false contour may be observed as a unit-step edge (Fig. 41a) that results in severe artifacts by reconstruction of a 3D scene. These artifacts could be substantially reduced by refining the disparity map to sub-pixel precision. In the case of half-pixel refinement, a unit-step false edge would be replaced by two half-step edges (Fig. 41b). This would yield significantly reduced artifacts in the reconstructed 3D scene.

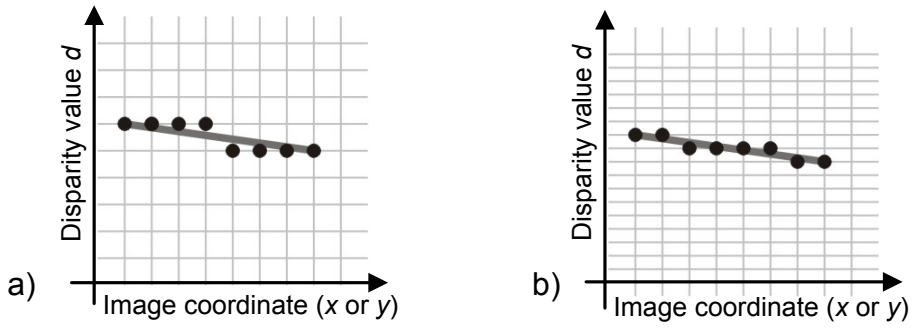


Fig. 41. False contours in disparity maps for a surface nearly perpendicular to the camera axis: a) full-pixel precision ($d_{step} = 1$) , b) half-pixel precision ($d_{step} = 0.5$).

In order to avoid the above mentioned problems, the author proposed usage of two-step disparity estimation (Fig. 42) with original refinement technique called Mid-Level Hypothesis (MLH) technique. In the first step, disparity is roughly estimated - usually with precision to the sampling period (full-pixel precision). In the second step the disparity map is refined in order to attain sub-pixel precision and accuracy.

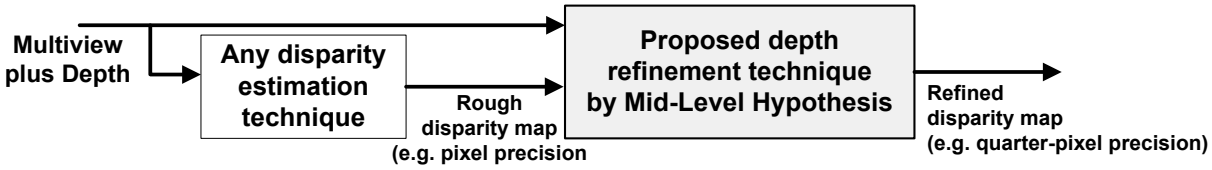


Fig. 42. The idea of Mid-Level Hypothesis depth refinement technique.

This MLH technique identifies the false edges in a rough disparity map. Then, at individual pixels, the technique introduces mid-level (the intermediate level) values of disparity in order to reduce the false contours of disparity maps. In that way the quantization step of the disparity values is halved and thus the precision is doubled. Of course, this technique may be used iteratively. After n iterations, the disparity quantization step is reduced by factor of 2^n and thus the precision of disparity representation in also enhanced by factor of 2^n.

3.7.1. Idea of depth refinement by Mid-Level Hypothesis algorithm

At the input, there is a disparity map with limited precision. Such may come from any disparity estimation technique, but in the experiments DERS algorithm has been used. Moreover, at the input of the algorithm, there is a set of input views. The *basic view* is the view that corresponds to the disparity map being processed. The other views will be called *side views*. Those side views are used to refine disparity map of the base view.

From the basic view and the rough disparity map, a synthetic side views are obtained. The more accurate is the disparity map, the more similar are the synthetic side views to the actual side views. Therefore, increasing similarity of the two versions of each of the side views may be used as an indicator of the increasing accuracy of the depth map.

At the beginning of the process, the edges that correspond to the disparity quantization step are identified. For full-pixel disparity map, these are *unit-step edges*. For the sake of simplicity this name will be used. The disparity map is processed only locally along those potentially false contours (Fig. 43). The potential improvement may be done by introduction of the mid-level values into the disparity map.

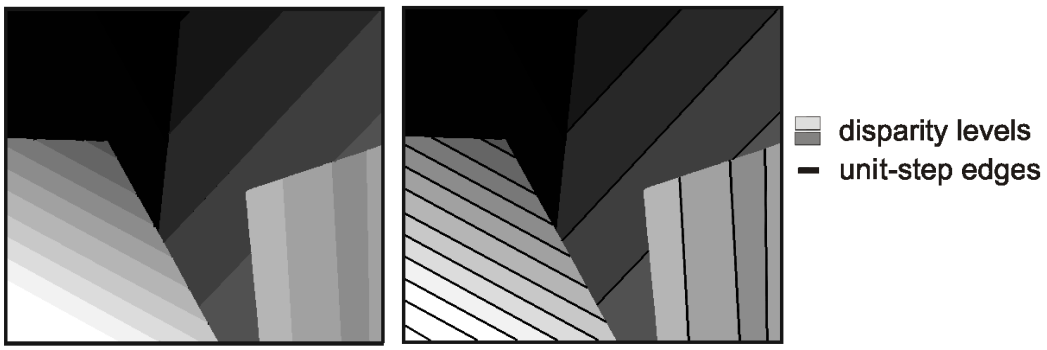


Fig. 43. Disparity map (left) and the same disparity map with marked unit-step edges (right) for exemplary "Venus" image [142]. The image has been selected for illustration, because it is composed from simple objects and the depth edges can be easily noticed.

False edges occur when an inaccurate rough estimation is forced to quantize disparity levels. The unit-step edges may be false contours or may also represent actual depth differences. Therefore there is an uncertainty to be resolved. In our refinement algorithm it is done by verification of the mid-level hypothesis. At first, the algorithm assumes that each pixels neighboring to unit-step edges (Fig. 44) in the disparity map, should have the intermediate disparity level. Then this hypothesis is verified for each pixel.

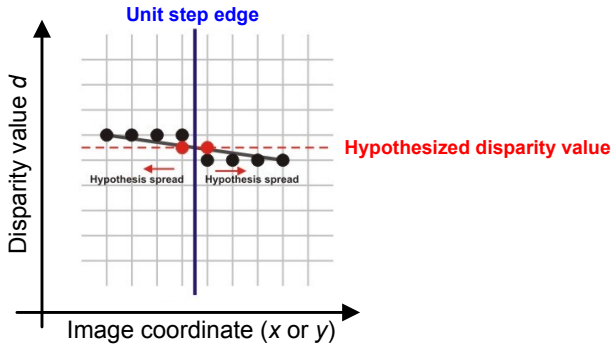


Fig. 44. Spreading of mid-level hypothesis starting from a unit-step edge.

Therefore, along those potentially false edges, a question arises that has to be answered at each individual pixel: *Should the algorithm change the disparity value to a mid-level value or not?* Hypothesis that this question can be answered positively is verified by comparison of the two alternatives. For those two alternatives, local synthesis of side views is performed twice (Fig. 45) – once with the unmodified disparity value and once with assumed change to the hypothesized mid-level value. The synthesized contents of the side-views are compared to the real contents of the side views. The hypothesis which provides higher level of similarity is assumed to be true.

Thus, the hypothesis of the disparity mid-level value is verified positively if the newly synthesized contents in the side views are more similar to the real side views as compared to the contents in the side views synthesized using the input rough disparity map.

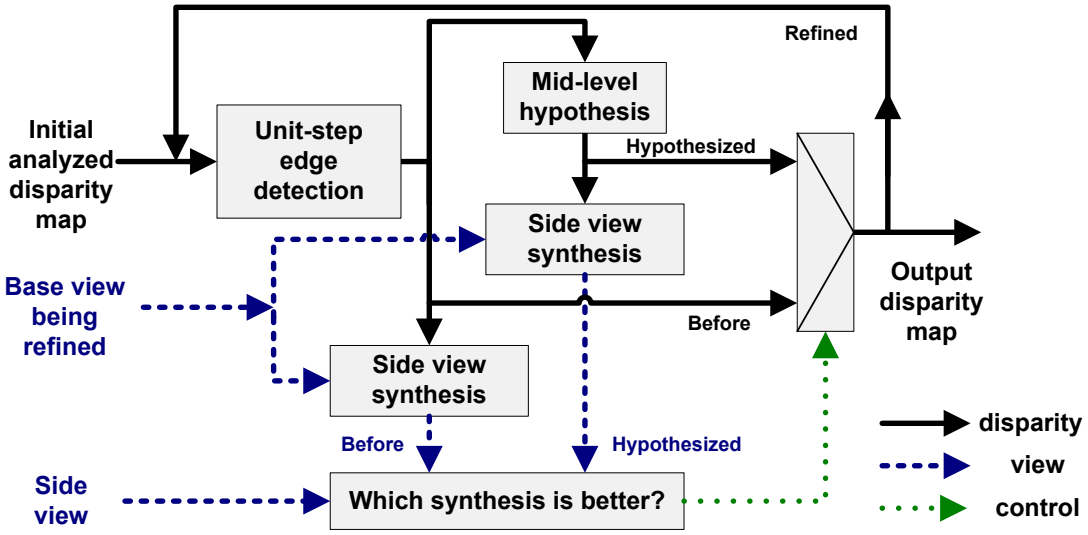


Fig. 45. Scheme of the proposed disparity refinement algorithm. For the sake of brevity, the case of only a single side view is presented.

So, in principle, "mid-level hypothesis" is assumed to be true at each edge pixel, i.e. set the pixel value to intermediate value. Then, the hypothesis is verified by checking if the synthetic size view is more similar to the original side view.

Then, iteratively, mid-level hypothesis is spread from each unit-step edge. Spreading stops when no point passes the verification test (Fig. 45).

Therefore, the precision of the disparity map is improved by insertion of intermediate disparity levels in-between of existing levels. The proposed technique never degrades the processed disparity map, because the verification mechanism does not allow that to happen. Thus, also the accuracy of the processed disparity map is enhanced.

Below, implementation of the MLH algorithm is shown, with particular steps that are performed, highlighted.

3.7.2. Implementation of the algorithm

Steps of the depth refinement by Mid-Level Hypothesis algorithm are presented below. The algorithm continues until loop-exit condition is reached, which has been formulated in Step 5. If the loop-exit condition is not meet, the algorithm goes back to Step 1.

Step 1. Detection of unit-step edges

The proposed technique detects unit-step edges in the disparity map by simple comparison of disparity values in the neighboring pixels. Pixels, whose disparity labels differ by 1 (and thus at the current precision their disparity values differ by d_{step}) from neighboring pixels are classified as belonging to a unit-step edge. Those pixels are marked for further processing (Fig. 46). They potentially belong to a false contour in the disparity map.

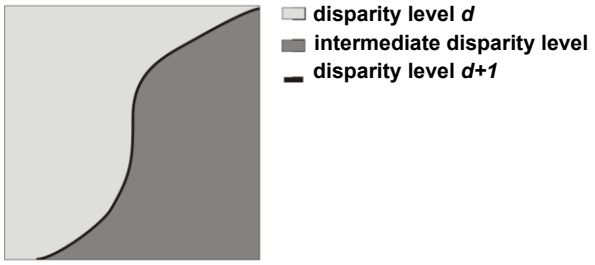


Fig. 46. Detection of unit-step edges.

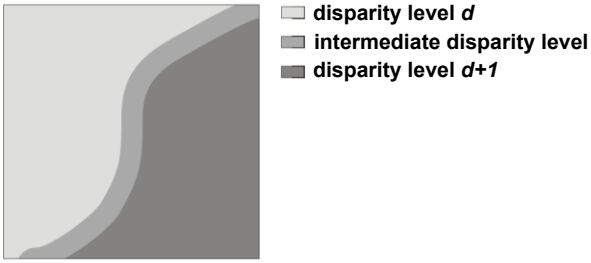


Fig. 47. Intermediate level hypothesis.

Step 2. Introduction of intermediate disparity levels

It is supposed that the marked pixels (Fig. 47) should have intermediate values of disparity. So, pixels on both sides of a unit-step edge are processed.

Step 3. Verification of intermediate level hypothesis

Unit-step edges may occur in two distinct cases: they may represent actual edges in the scene but they also may result from rough disparity quantization. That decision ambiguity is resolved by verification of hypothesis of intermediate level (Fig. 48).

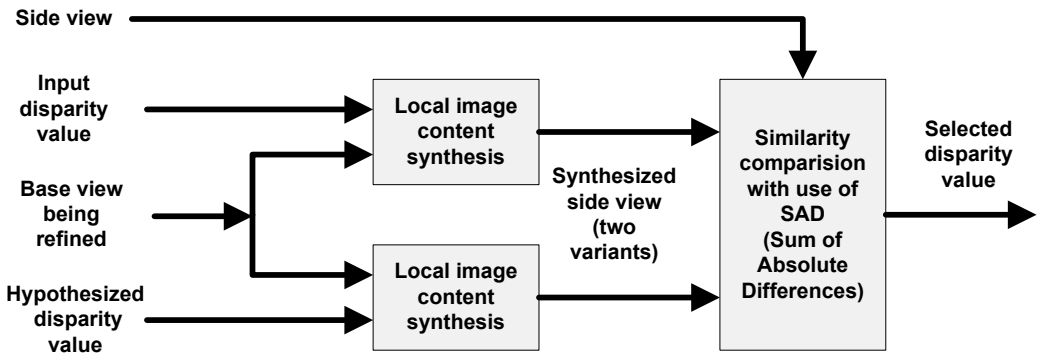


Fig. 48. Scheme of the verification step. For the sake of brevity, the case of only single side view is presented.

Assumed disparity value is verified by comparison of the quality of the two synthesis variants of the side view: one obtained from the input disparity values and the second one obtained with the assumed intermediate disparity value. The disparity value that provides better synthesis of the side view (measured by Sum of Absolute Differences - SAD) is selected as a resultant disparity value.

Step 4. Spreading of the hypothesis

The pixels, that have passed the verification, retain their intermediate disparity values. Then, the mid-level hypothesis is assumed for the neighboring pixels. Thus, the hypothesis is spread to all neighboring pixels within 8-connectivity neighborhoods. These pixels are also marked for further processing (Fig. 49). The mid-level hypothesis is tested for all those pixels.

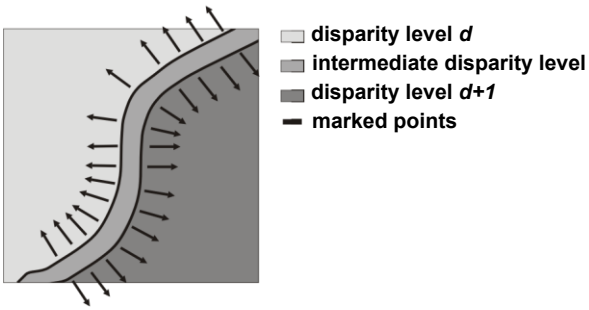


Fig. 49. Spreading direction of intermediate level hypothesis.

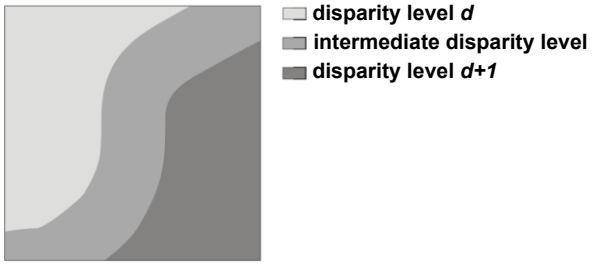


Fig. 50. A disparity map refined with MLH algorithm.

Step 5. Loop-exit condition

If there are still marked pixels, algorithm loops to step 2. The algorithm stops when there is no pixel marked for processing. The result of the algorithm is an improved disparity map with new intermediate disparity levels (Fig. 50). Note that usually only a portion of all pixels is processed, i.e. the mid-level hypothesis is verified in the selected pixels only. This observation is closely related to the low complexity of the technique.

3.7.3. Experimental results for depth refinement

For the evaluation of the proposed depth refinement by Mid-Level-Hypothesis algorithm, first, the reference data has been generated. Pixel-precise and quarter-pixel precise disparity maps have been estimated for the test sequences and view settings described in Section 1.5

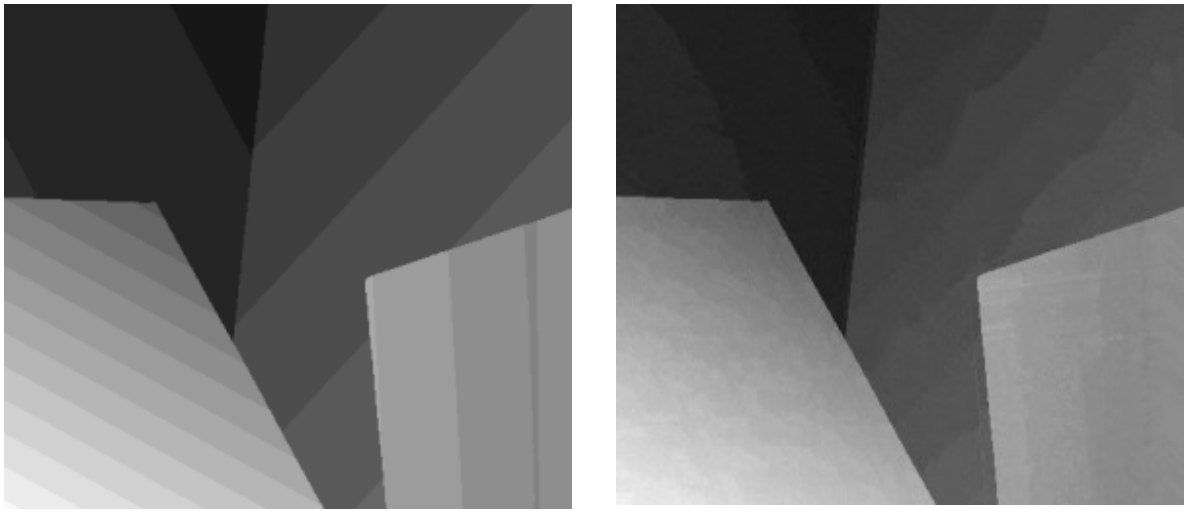
(Table 2 on page 26) with use of original (unmodified) DERS algorithm. A wide range of Smoothing Coefficient has been tested and the best and the worst performing settings for each sequence have been identified.

After that, the pixel-precise disparity maps have been refined with use of MLH algorithm in order to generate quarter-pixel precise disparity maps. Therefore – two versions of quarter-pixel precise disparity maps were available – generated directly by original (unmodified) DERS and generated by MLH algorithm on the base of input pixel-precise disparity maps.

All of the resultant disparity maps have been used to synthesize virtual views which have been then compared to the original views (according to the methodology described in Subsection 1.5.3).

The results are gathered in Table 9. In average over the tested sequences, the proposed method provides even better results than pure quarter-pixel precision for **about 0.28dB**. This extra gain comes from the fact that MLH algorithm not only provides enhanced precision (from full-pixel to quarter pixel) but also refines the disparity values with usage of information from the neighboring side views. Comparing to pixel-precise disparity maps, the gains are even higher and are **about 2.01dB in average** over the test set.

For illustration purposes, the MLH algorithm has also been launched on a "Venus" image set [142], where the improvement can be seen by unarmed eye (Fig. 51).



a) before refinement: full-pixel precise disparity map

b) after refinement: quarter-pixel precise disparity map

Fig. 51. Results of proposed Mid-Level Hypothesis precision refinement algorithm used on exemplary "Venus" image [142], composed from simple objects, thanks to what the disparity edges can be easily noticed.

Table 9. Averaged virtual view synthesis quality of the proposed MLH refinement technique compared to original (unmodified) DERS technique, for the test sequences, evaluated according to methodology described in Subsection 1.5.3.

Sequence name	**PSNR [dB] (vs. the original view) of the virtual view synthesized with use of depth maps estimated with:**		
	DERS (quarter-pixel precision)	**DERS (full-pixel precision)**	**DERS (full-pixel) + Proposed MLH (result is quarter-pixel precise)**
Poznan Street	35.53	31.98	35.49
Poznan Carpark	31.22	30.71	31.84
Poznan Hall 2	35.21	32.85	35.39
Lovebird1	31.13	29.75	31.26
Newspaper	33.07	31.91	33.25
Balloons	34.28	32.94	34.87
Kendo	37.27	35.46	37.57
Average	**33.96**	**32.23**	**34.24**
Avg. ΔPSNR gain of the proposal related to given reference	**+ 0.28**	**+ 2.01**	

Table 10**.** Average frame computation time of the proposed MLH refinement technique compared to original (unmodified) DERS technique, for the test sequences, evaluated according to methodology described in Section 2.5.

Sequence name	**Average frame computation time [s]**		
	DERS (quarter-pixel precision)	**DERS (full-pixel precision)**	**Proposed MLH + DERS (full-pixel) (result is quarter-pixel precise)**
Poznan Street	18235	4538	4538 + 424 = 4962
Poznan Carpark	17983	4325	4325 + 384 = 4709
Poznan Hall 2	18129	4363	4363 + 413 = 4776
Lovebird1	932	237	237 + 33 = 270
Newspaper	955	246	246 + 72 = 318
Balloons	912	239	239 + 84 = 323
Kendo	976	254	254 + 87 = 341
Average (rounded)	**8303**	**2029**	**2029 +** 214 = **2243**
Avg. speed-up of the proposal related to given reference	**3.7 ×**	**0.9 ×**	

What is worth to notice is the fact that the mentioned **average gain of 2.01dB** is attained at much lower computational cost that the direct quartet-pixel precise disparity estimation. Computational complexity of the proposed MLH refinement technique was compared against

the direct usage of DERS technique with use of PC computer (3,6GHz processor). The results prove that in the tested conditions, the direct full-pixel estimation and quarter-pixel refinement is **about factor of 3.4× to 3.8× (about 3.7× in average)** faster than direct quarter-pixel estimation using the original (unmodified) DERS technique. Of course, MLH algorithm used with prior full-pixel precise depth estimation, works at speed of about 0.9× (is about 10% slower) than pure usage of sole full-pixel precise depth estimation.

Therefore it can be said, that advantages of MLH algorithm can be seen from two alternative points of view. Firstly, it can be seen as precision and accuracy refinement technique that brings average gain of **about 2dB** related to input pixel-precise depth maps and 0.28dB related to input quarter-pixel precise depth maps. Secondly, it can be seen as speed optimization technique which allows for about factor of **3.7×** computational cost reduction, related to application of direct quarter-pixel precise depth map estimation.

In the Section above, a novel algorithm for precision refinement of disparity maps has been proposed. In the next section, another subject of the work will be analyzed, related to the temporal consistency of the estimated depth.

3.8. Temporal consistency improvement of the depth by noise reduction

This Section shows the developments of the author in area of depth estimation related to enhancement of the **temporal consistency** of stereoscopic depth maps. The problem has been already stated in the introduction (Section 2.6). Here, an original proposal for estimation of temporally consistent depth maps is presented.

The main idea of the proposed approach is that the estimated depth can be more temporally consistent if a noise reduction technique is applied to the input video a priori to the depth estimation. The presented approach extends the previous authors works [3][20].

Although, as it has been mentioned, there are many methods for noise reduction in video sequences, all of which could be used, during the author works, two methods have been developed. First one is **Still Background Noise Reduction (SBNR)** and the second one is **Motion-Compensated Noise Reduction with Refinement (MCNRR).**

Both of those methods are presented below in Subsections 3.8.1 and 3.8.2. The considerations are summarized in Subsection 3.8.3 by experimental results.

3.8.1. Still Background Noise Reduction (SBNR) technique

This technique us based on early works of the author [3][20] in which noise is reduced by filtration of regions that are still, and presumably belong to the background. The filtration is performed in time and independently for each view of test sequences.

The first version of the algorithm [20] was working in rectangular blocks. Each block has been classified as moving or still, with respect to differences between consecutive frames and processed reference frame. Moving blocks were left unchanged during processing and would be ignored in the course of noise analysis. Blocks classified as still were linearly filtered with respect to previous frames and would be used for the further analysis of noise. Such a nature of the processing was resulting in blocky-effect similar to the one known from compression.

Therefore, a second version of the algorithm [3] working with single pixels has been proposed and has been used in this dissertation. It originally consists of three main steps:

- Motion detection, where pixels are classified as moving or steady.
- Noise filtering, where steady pixels are filtered in time.
- Artifact removal, where errors of motion detection stage are repaired.

Because the purpose for application of this noise reduction technique in works related to this dissertation is not to achieve a subjectively pleasant denoised video sequence, but to extract and analyze the noise, the third step from the original proposal has been omitted. It would be useless, as only pixels classified as steady are used in noise analysis.

Stage 1. Motion detection

The role of motion detection (Fig. 52) is to classify pixels from input frame as moving or as steady. Result of this classification is combined into a binary map, called motion map.

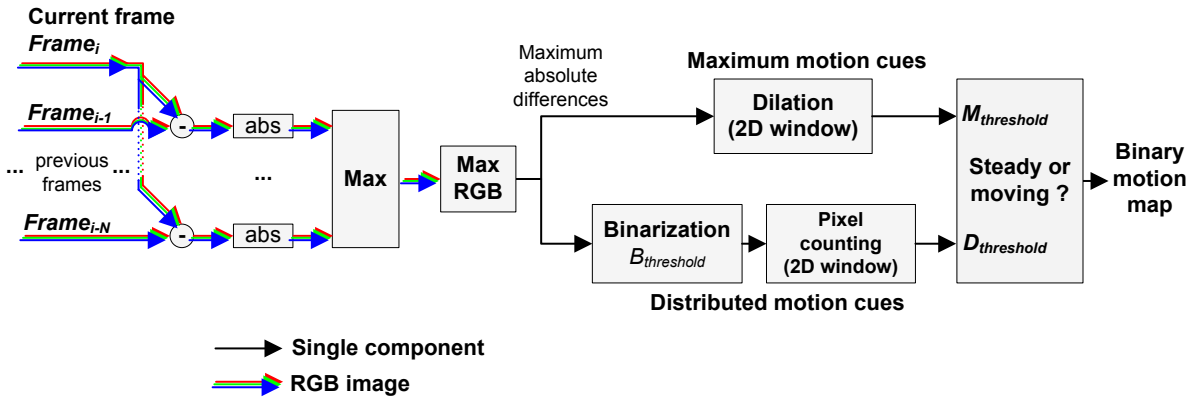


Fig. 52. Block scheme of motion detector in Still Background Noise Reduction (SBNR) algorithm.

Each pixel of input frame is compared with corresponding pixels of N previous frames by means of absolute differences. These absolute differences are then maximized between frames and over RGB color components. Resulting maximum absolute differences are feed to two parallel paths (Fig. 52) which provide (for each pixel) different cues about motion that occur in neighboring pixels – one provides cues about maximal motion, and the other one provides cues about distributed motion:

- Maximal motion cue map (top path in Fig. 52) is obtained with use of dilation filter. Dilation is performed with square mask.
- Distributed motion cue map (bottom path in Fig. 52) is obtained by counting of pixels that exceed certain level (binarization/thresholding) in window surrounding each pixel.

Fig. 53. Motion map (bottom) obtained for exemplary frame (top) of Poznan Carpark sequence (white pixels – moving, black pixels – steady).

Output of the motion detection - binary motion map (Fig. 53) - is produced by combining of motion cue maps from top and bottom path. Pixel is marked as moving (demarked in white

in Fig. 53) if any of motion cues indicates movement (exceeds certain level). Otherwise, pixel is marked as steady (demarked in black in Fig. 53).

Stage 2. Noise filtering

Pixels classified as moving are left unchanged (are not modified by the algorithm) and are ignored in the course of noise analysis.

Pixels classified as steady are assumed to be stationary in time and thus all changes in the signal are considered as noise. As shown in Section 3.2.2, the noise is independent in the subsequent frames and thus filtration can be used in order to attain statistically unweighted estimation of the expected real value. From the signal processing point of view, only the DC component should pass thought the filtering process. Therefore, the cut-off frequencies were set as low as possible for given filter structure, practically at level of about 2 Hz.

The tested filters were:

- FIR (Finite Impulse Resp.) filters (Parks-McClellan, equiripple) of order: 5, 10, 50, 100,
- IIR (Infinite Impulse Response) Butterworth filters of order: 1, 3,
- IIR (Infinite Impulse Response) elliptic filters of order: 1, 3.

All of the tested filters yielded with very similar results and thus the simplest, the most computationally efficient has been used - first order Butterworth IIR filter.

Arbitrary parameters used in the noise reduction process, like window sizes and threshold levels depend on image resolution and camera system. These were optimized for experiments empirically. The values that have been userd are gathered in Table 11.

Table 11. Thresholds used in SBNR algorithm for the test sequences.

Sequence name	FPS	Window size	IIR filter cut-off frequency	$B_{threshold}$	$D_{threshold}$	$M_{threshold}$
Poznan Street	25	9×9	2.5 Hz	8	20	15
Poznan Carpark	25	9×9	2.5 Hz	8	20	15
Poznan Hall	25	9×9	2.5 Hz	8	20	15
Lovebird	30	9×9	3.0 Hz	8	20	15
Newspaper	30	9×9	3.0 Hz	8	20	15
Balloons	30	9×9	3.0 Hz	8	20	15
Kendo	30	9×9	3.0 Hz	8	20	15
GT Fly	25	Computer-generated sequences – no noise				
Undo Dancer	25					

Stage 3. Artifact removal

Noise removal scheme in SBNR technique, consisting in motion detection and noise filtering steps is simple and computationally efficient. Unfortunately, it may be a cause of artifacts resulting from hard-decisive classification of pixels as steady or moving.

Fig. 55 shows three trajectories of exemplary pixel: original value (a), filtered value (b) and value after artifact removal (c). At the beginning (segment I), the pixel is classified as steady. It varies due to noise, which is filtered (filtered (b) is the same as (c)). Then (II), pixel value starts to change significantly and is classified as moving. As a result of that, the filtering phase is omitted: (a), (b) and (c) are the same. Up to this moment, there are no artifacts.

In segment III, pixel is classified as steady, because its value changes very slowly. Filtered pixel trajectory changes even slower, resulting in discrepancy between trajectories (Fig. 55), which is lesser than threshold of motion detector. After a while, the discrepancy rise beyond threshold and pixel is instantaneously classified as moving in segment IV. Filtering switches off, and thus trajectories are updated to original, which causes another steady segment V. Rapid switching causes visual artifact in the output image.

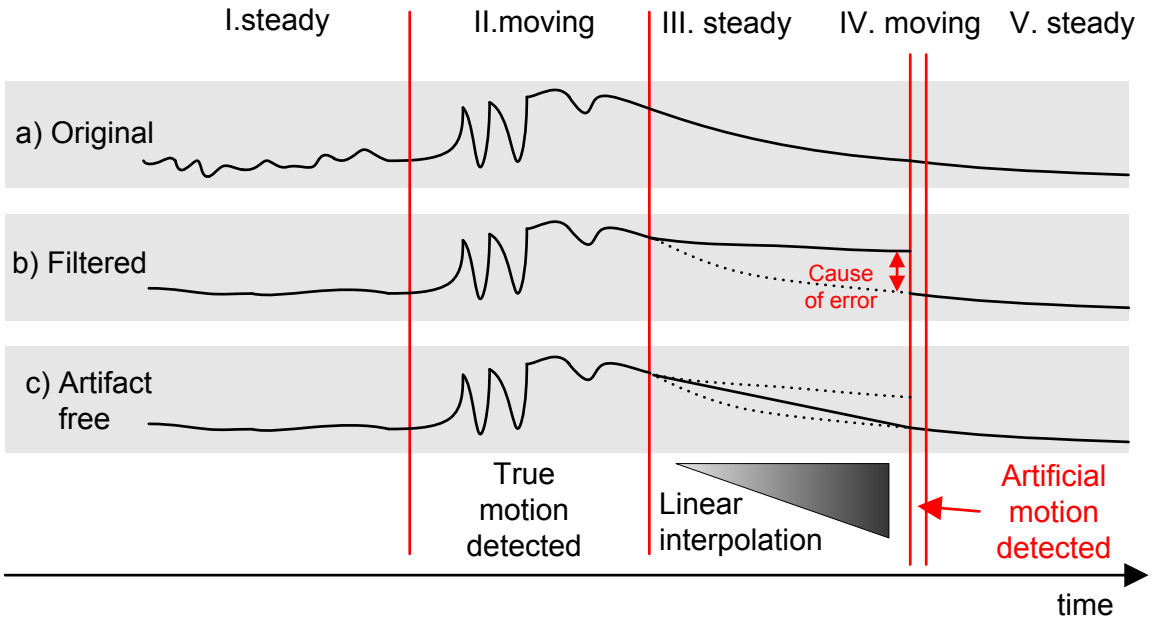


Fig. 54. Artifact removal in Still Background Noise Reduction (SBNR) technique on example of trajectories of exemplary pixel values.

Therefore, an additional step of artifact removal is proposed, where errors of motion detection stage are repaired. First, rapid changes of pixel classification (steady or moving) are

predicted. If such rapid change is predicted, pixel value is linearly interpolated (Fig. 55) between original (a) and filtered (b) trajectories before the change occurs.

The results for SBNR are provided further in Subsection 3.8.3. In the meanwhile, another noise reduction technique developed by the author will presented below.

3.8.2. Motion-Compensated Noise Reduction with Refinement (MCNRR) technique

The technique described above (Still Background Noise Reduction) performs well over not-moving background regions. Unfortunately, it omits moving regions which disallows noise analysis over the whole scene. In order to overcome this drawback, a motion compensated noise reduction technique has been used.

This technique consists in two main stages. The first one is **Motion-Compensated Noise Reduction** and the second one is **Refinement**. Both constitute MCNRR noise reduction method, described below.

Stage 1. Motion-Compensated Noise Reduction

For the sake of work savings, author has decided to use an already developed motion compensation package called "mv-tools" [245], which is a plug-in for VirtualDub/AviSynth video scripting framework [246]. As this package is designed for single-view processing, each view of a multiview sequence is processed separately – no inter-view correspondences are used.

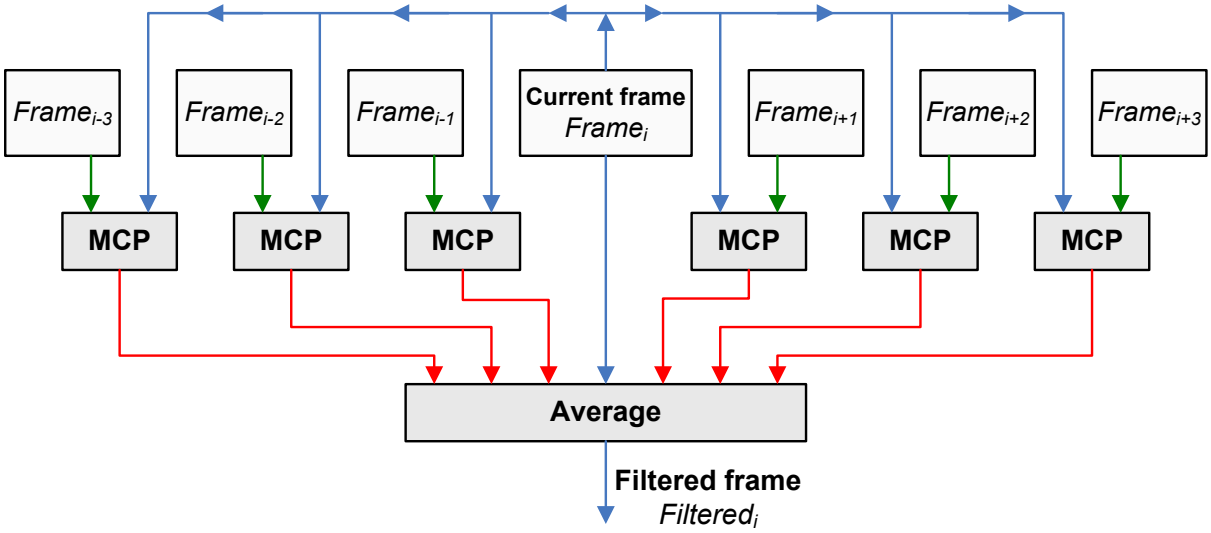


Fig. 55. Noise reduction scheme used in Motion-Compensated Noise Reduction (MCNRR) algorithm. The MCP block depicts motion compensated prediction presented in Fig. 56 in detail.

The proposed algorithm is as follows. Each frame of each view is processed independently. Block-based motion estimation (Fig. 55) is performed in order to find motion vectors pointing to frames neighboring in time (3 previous and 3 following ones). Then, low-pass filtering is performed on matched blocks under the form of simple average which is estimation for expected value (see Subsection 3.1.1).

It can be noticed, that not all blocks from motion estimation are used in the average. The compensated block is firstly compared with the original contents of the current frame (Fig. 56). Only if the best candidate found by motion estimation is similar enough (basing on Sum of Squared Differences criterion) it is feed to the average block. Otherwise it is omitted. Therefore, average may be performed on various numbers of blocks, from 1 (only the current frame) to 7 (the current frame, 3 previous and 3 next frames).

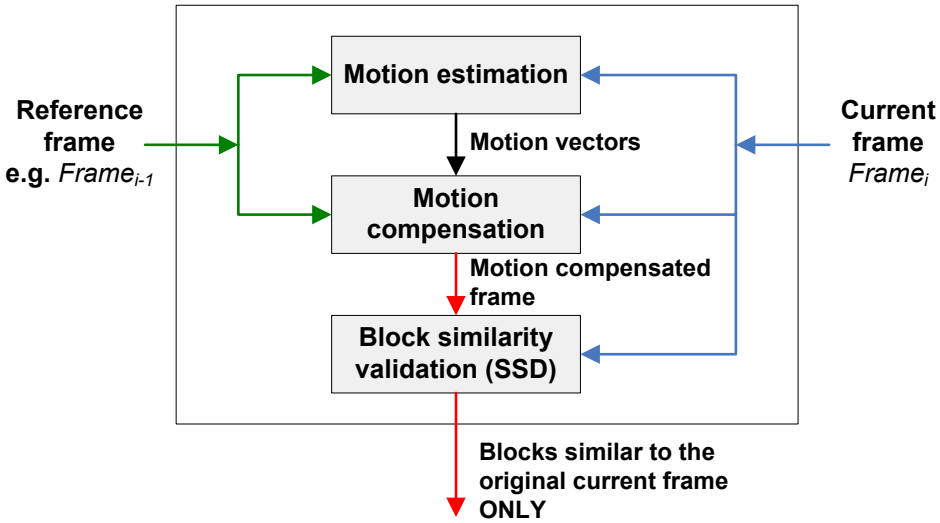


Fig. 56. Motion compensated prediction used in MCNRR algorithm.

Stage 2. Refinement

The result of noise reduction depicted in Fig. 55 are filtered frames $Filtered_i(x, y)$. Although subjectively the results are satisfying, the analysis of suppressed noise shows that this process is vulnerable to errors and produces artifacts in form of edges which are not perfectly matched at the motion estimation stage. Moreover, blocks with those edges are not discarded in the similarity validation stage (Fig. 57). As a result of that, the edges of fast moving objects are slightly blurred.

Therefore, the author of this dissertation proposes a refinement stage in which those artifacts are reduced (Fig. 58).

a)

b)

Fig. 57. Exemplary artifacts generated in extracted noise, generated by mv-tools motion-compensated noise reduction technique: a) the original frame $Frame_i(x, y)$ of Poznan Street sequence and b) difference between the denoised frame $Filtered_i(x, y)$ and the original $Frame_i(x, y)$, denoted $SumRGB_{\mathrm{i}}(\mathrm{x}, \mathrm{y})$ in Fig. 58, showing artifacts on the edges of fast-moving objects. Gray level represents zero noise value (no difference).

First, the filtered frames $Filtered_i(x, y)$ are compared with the original (not processed) frames $Frame_i(x, y)$ with respect to Absolute difference measure, performed on each RGB channel independently, giving $AbsDif_{\mathrm{i}}(\mathrm{x}, \mathrm{y})$. Then, sum of those differences $SumRGB_{\mathrm{i}}(\mathrm{x}, \mathrm{y})$ is calculated and feed to a noise gate, where values lesser than with threshold $G_{threshold}$ are zeroed. The result is processed with a 2-dimensional dilation filter, which leads to the spatial extension of regions which are non-zero in the processed images. Then, each value is

normalized, relatively to standard deviation σ_{DifRGB} calculated in parallel, basing on $SumRGB_{\mathrm{i}}(\mathrm{x},\mathrm{y})$ signal. After that, the normalized values are feed to another noise gate, where values lesser than with threshold $N_{threshold}$ are zeroed. Then, directly neighboring pixels, that are non-zero, are gathered into segments. Segments, which have relatively small area, lesser than $S_{threshold}$ pixels, are deleted (zeroed).

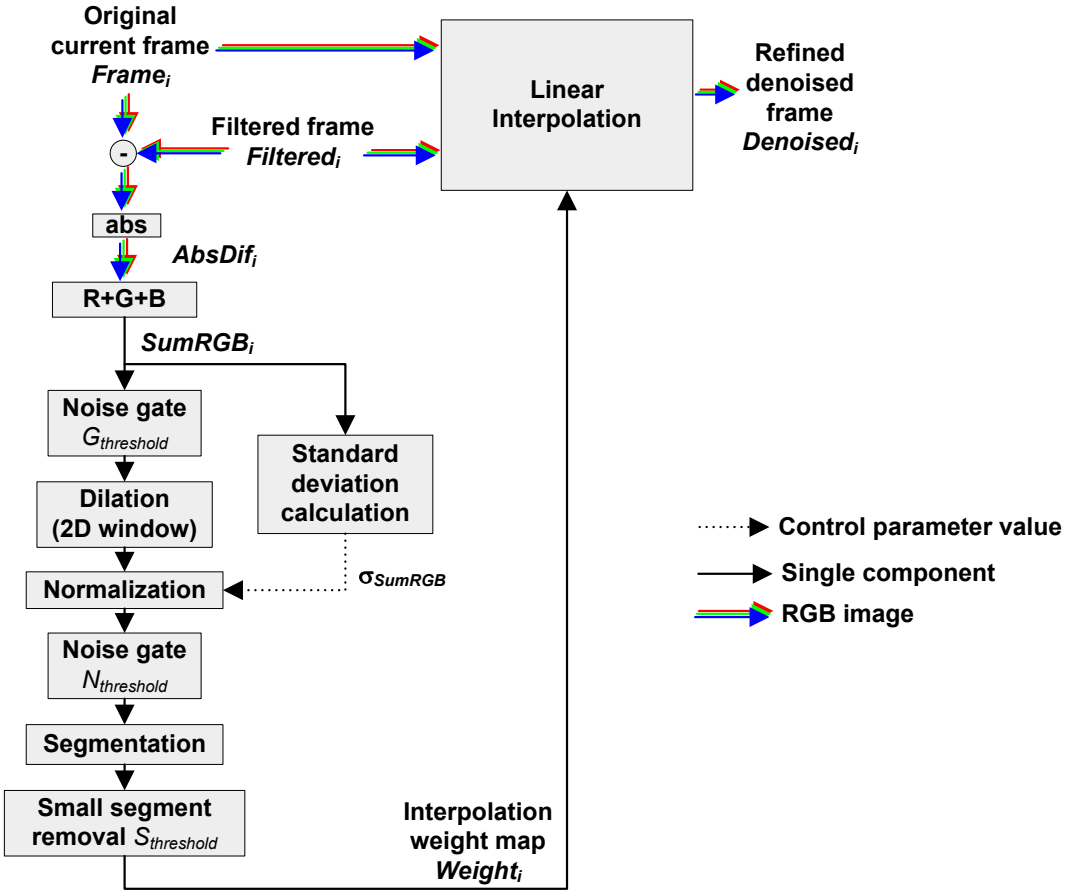


Fig. 58. Scheme of the refinement stage in MCNRR algorithm.

In the experiments, the thresholds values were set to $G_{threshold} = 1$, $N_{threshold} = 1$, $S_{threshold} = 1$, uniformly for all sequences and views.

The idea behind calculation of $Weight_{\mathrm{i}}(\mathrm{x},\mathrm{y})$ signal is to detect the regions that suffer from artifacts introduced by application of motion-compensated noise reduction algorithm (Fig. 57).

In regions where the artifacts occur, high values (up to 1.0) of $Weight_{\mathrm{i}}(\mathrm{x},\mathrm{y})$ are generated. On the other hand, in regions, where there are no artifacts, low values (around 0) of $Weight_{\mathrm{i}}(\mathrm{x},\mathrm{y})$ are generated.

The finally attained signal $Weight_{\mathrm{i}}(\mathrm{x},\mathrm{y})$ used for linear interpolation between the filtered frames $Filtered_i(x,y)$ and the original (not processed) frames $Frame_i(x,y)$. Thanks to that, the resultant, refined denoised frame $Denoised_{\mathrm{i}}(\mathrm{x},\mathrm{y})$ is practically free from artifacts.

For now, two algorithms for noise reduction, developed by the authors for works on estimation of temporally consistent depth maps, have been presented. In the Subsection below, experimental results of depth estimation for those two methods are presented.

3.8.3. Experimental results for temporal consistency improvement

In order to experimentally asses the proposed approach, the two noise reduction algorithms, developed by the author, have been used (Fig. 59) on the multiview video test sequences set (Table 1 on page 24). This includes usage of Still Background Noise Reduction (SBNR) and Motion-Compensated Noise Reduction with Refinement (MCNRR).

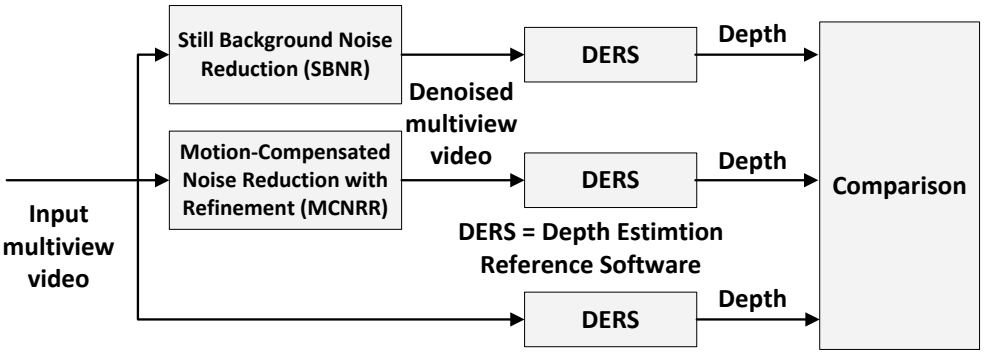


Fig. 59. Scheme of the experiments for assessment of the techniques related to improvement of temporal consistency by noise reduction.

In Fig. 60, exemplary visual results attained with and without use of the proposed noise reduction techniques. As can be noticed on Fig. 60a,b, moving objects (people) are left unchanged while background (wall and cars) is significantly denoised. It is worth to notice that denoised images are not blurred, because only temporal filtering is employed. Although quality of depth maps (Fig. 60c,d) has not changed, temporal consistency expressed as difference between frames (Fig. 60e) is vastly improved. As shown, background remains static (black means no changes) and thus is consistent is time. Of course, there is no improvement over moving objects, because they are not filtered.

After the visual examination of the results of noise removal, application of SBNR technique for depth estimation has been tested [3]. Basing on the denoised view, depth maps have been generated, which have been evaluated with respect to their quality (Fig. 59). In particular, it has been done indirectly (Fig. 10 on page 25) through assessment of quality of virtual views, synthesized with use of depth maps, generated basing of the denoised videos.

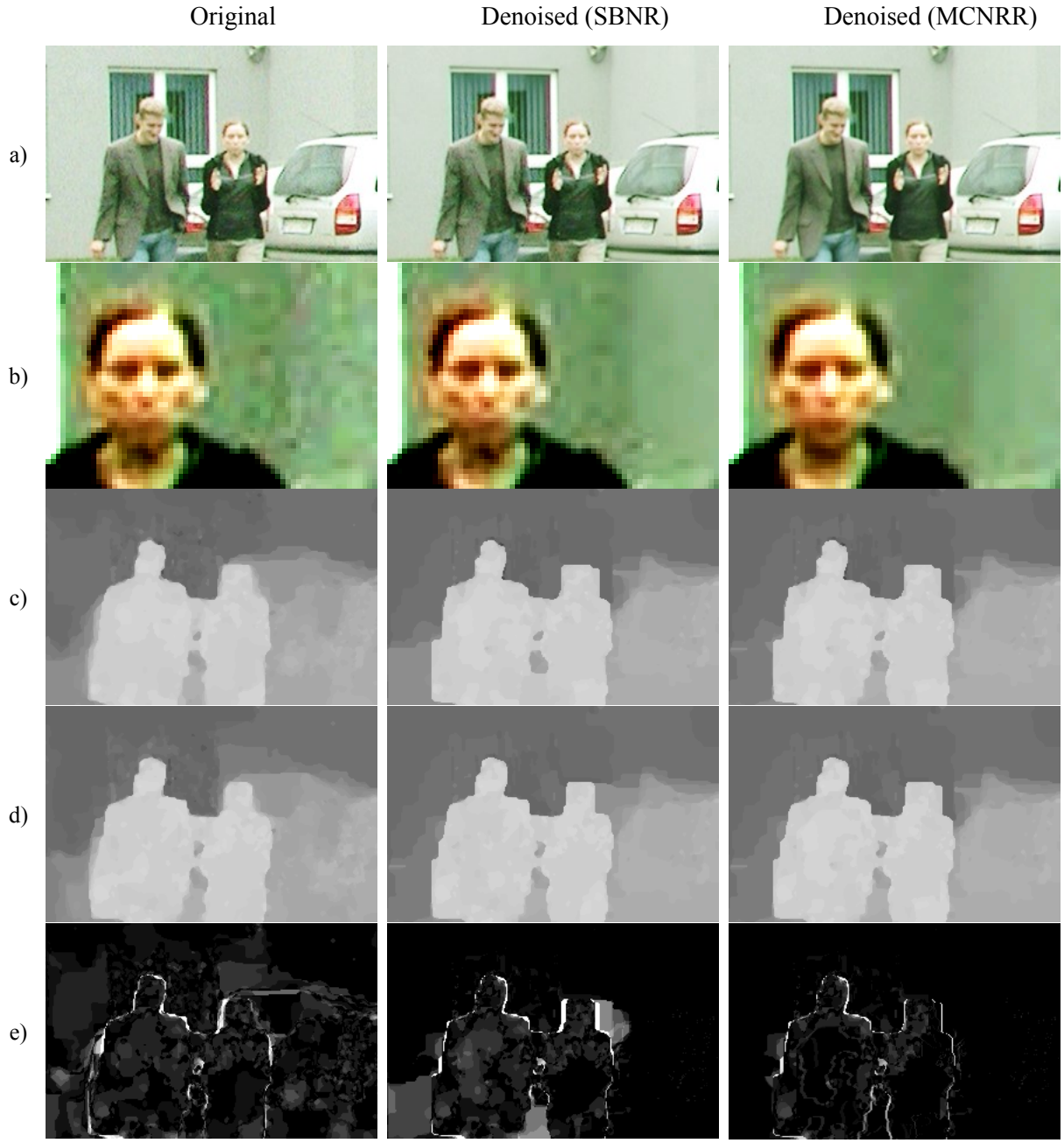


Fig. 60. Exemplary results of proposed technique: original (left), denoised with use of SBNR technique (center) and denoised with MCNRR technique (right).
The images has been intensified for better reproduction of the differences:
a,b) original image, c,d) depth maps for two consecutive frames,
e) difference between depth maps.

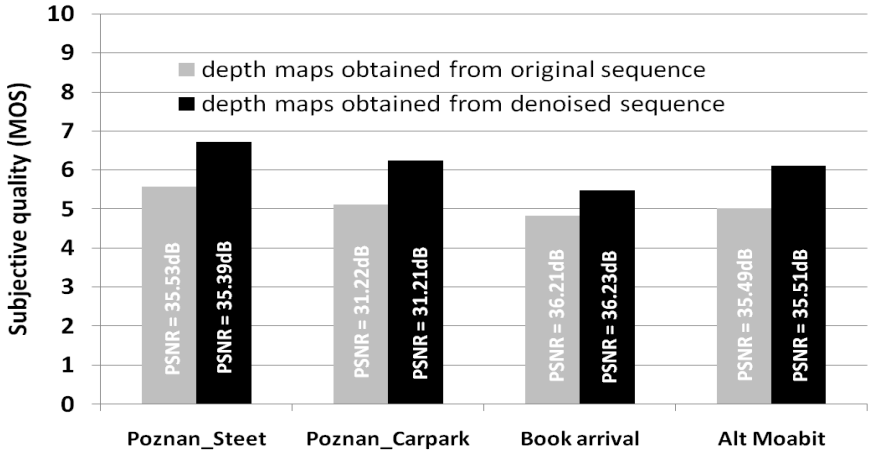


Fig. 61. Subjective evaluation results for SBNR technique, quoted from [3]. Please note that the used sequence set is different from the set used in the dissertation. Also in work [3] quarter-pixel precise depth estimation has been used which is the source of discrepancy between PSNR values in the Table 12, whereas PSNR gains/losses are quite similar in both cases.

Table 12. Averaged virtual view synthesis quality of the proposed depth estimation with noise reduction in the input video, compared to original (unmodified) DERS technique, for the test sequences, evaluated according to methodology described in Subsection 1.5.3.

Sequence Name	**PSNR [dB] (vs. the original view) of the virtual view synthesized with use of depth maps estimated with pixel-precision basing on:**		
	Views denoised with:		**Reference (original views)**
	Proposed SBNR technique	**Proposed MCNRR technique**	
Poznan Street	31.93	31.92	31.98
Poznan Carpark	30.74	30.79	30.71
Poznan Hall 2	32.78	32.83	32.85
Lovebird1	29.79	29.78	29.80
Newspaper	31.90	31.91	31.91
Balloons	32.91	32.93	32.94
Kendo	35.41	35.39	35.46
Average	**32.21**	**32.22**	32.24
Average ΔPSNR gain related to the reference	**- 0.03**	**- 0.02**	-

Both objective and subjective evaluation have been performed with use of four test sequences [85][240] – Poznan Street, Poznan Carpark, Book Arrival and Alt Moabit. This limited set of sequences has been chosen, because only a small number of subjects available for subjective testing were available (15 persons). The results are shown: objective PSNR

(Fig. 61 as values on vertical bars) and subjective Mean Opinion Score (MOS) (Fig. 61), both in comparison to the original views. In the study, MOS is expressed by a 10-point continuous scale. Rating of the quality was in range from 1 ("very bad with annoying impairments/artifacts") to 10 ("excellent, artifacts are imperceptible").

It can be seen that application of SBNR technique for noise removal from the tested videos, used then for depth estimation, provides gain of **about 0.7 to 1.2 MOS point**. It can also be noticed that PSNR levels have not changed. The latter is not surprising, because PSNR measure is not designed to assess quality of temporal consistency.

After such initial assessment of SBNR technique (described in more detail in [3]) on limited set of sequences, both proposed noise reduction techniques (SBNR and MCNRR) have been assessed in a similar way (Fig. 10 on page 25), but with usage of all of the test sequences selected in the dissertation. View synthesis settings described in Table 2 on page 26, have been used. The results are presented in Table 12.

It can be noted, that there is discrepancy between PSNR values in the Table 12 and the ones presented previously in Fig. 61 as values on vertical bars. This discrepancy comes from a fact that in [3] quarter-pixel precise depth estimation has been used and in the newly presented case (Table 12) full-pixel precision has been used. Apart from that, PSNR gains/losses are quite similar in both cases and fluctuate around zero - in average, there are practically no gains or losses of PSNR. Again, this is not surprising, because PSNR measure is not designed to assess quality of temporal consistency.

Table 13. Averaged linear correlation coefficient between depth values in subsequent frames.

Sequence name	Linear correlation coefficient averaged over each sequence					
		Depth maps estimated from:				
	A. Ground truth depth maps	***B.*** Original views	***C.*** Views denoised with SBNR technique	$\frac{C}{B}100\%$	***D.*** Views denoised with MCNRR technique	$\frac{D}{B}100\%$
Poznan Street	0.9496	0.9552	0.9558	100.06%	0.9562	100.10%
Poznan Carpark	0.9607	0.9393	0.9486	100.99%	0.9547	101.64%
Poznan Hall 2	0.9264	0.9225	0.9257	100.35%	0.9319	101.02%
Lovebird1	0.9911	0.9608	0.9751	101.49%	0.9799	101.99%
Newspaper	0.9972	0.9941	0.9964	100.23%	0.9967	100.26%
Balloons	0.9960	0.9622	0.9789	101.74%	0.9796	101.81%
Kendo	0.9651	0.9544	0.9651	101.12%	0.9560	100.17%
Average	**0.9694**	**0.9555**	**0.9637**	**100.85%**	**0.9650**	**100.99%**

The results presented in Table 13 show that application of the proposed noise removal techniques for depth estimation provide gains in a form of increase of correlation between subsequent depth frames in given view. Column A presents linear correlation coefficient, averaged over all frames and views, calculated between subsequent frames of ground truth depth maps of given sequence. Columns B, C and D present similar results but calculated for depth estimated with use of the original views, views denoised with SBNR technique and views denoised with MCNRR technique, respectively. It can be seen that although the gains in linear correlation coefficient increase are small (up to 1,81%, about 0.06% - 1.99% in average, related to column B) it must be taken into perspective that the improved regions are mostly edges of the objects that cover only a small portion of the whole scene (see Fig. 60) and sometimes, differences even between the ground truth (column A) are very small – e.g. Newspaper sequence which is already highly correlated (the most the scene is not moving background).

Table 14. Bjøntegaard gains in bitrate (negative numbers are bitrate savings) and PSNR (positive numbers denote quality increase) - results of MVC compression of depth maps estimated with use of DERS basing on denoised test sequences, related to compression of depth maps estimated with use of DERS basing on the original test sequences (anchor).

Sequence name	Proposed SBNR technique		Proposed MCNRR technique	
	dBR, %	dPSNR,dB	dBR, %	dPSNR,dB
Poznan Street	-31.47	1.34	-35.14	1.53
Poznan Carpark	-46.57	2.01	-45.19	1.85
Poznan Hall 2	-26.44	1.54	-29.01	1.70
Lovebird1	-34.12	1.10	-34.91	1.17
Newspaper	-33.64	1.34	-33.42	1.33
Balloons	-23.96	0.93	-21.99	0.86
Kendo	-0.02	0.00	-5.26	0.24
Average	**-28.03**	**1.18**	**-29.28**	**1.24**

Finnaly, another test has been performed. The estimated depth maps, resulting from conderations above, have been coded with use of MVC [112][113] video codec. The compression performance has been measured and depicted in form of Bjøntegaard metric. The results are shown in Table 14 below. It can be seen, that application of the proposed noise reduction techniques on the input video, have seriously influenced the estimated depth maps, because their compression performance has vastly changes.

The coding performance of such (compared to the original depth maps estimated with modified DERS basing on the original multiview video) is about **28.03% higher** in the case of SBNR (which relates to increase of PSNR of about 1.18dB) or about **29.28% higher** in the case of MCNRR (which relates to increase of PSNR of about 1.24dB).

In general it can be said, that the average compression performance gain over the tested set is **about 30% of bitrate reduction,** while providing the same quality of synthesized views (the bitrate reduction has been measured with Bjøntegaard metric over PSNR of synthesized views). **This provides as strong indication that the temporal consistency of the estimated depth has been vastly improved**, as one one the main compression tools in coding technology implemented in MVC is temporal prediction. The more the subsequent frames are correlated, the higher compression performance can be attained.

3.9. Summary of the achievements in the area of depth estimation

In this Chapter, author's achievement and results related to the area of depth estimation have been presented, related to theses T1-T3 of the dissertation. The main covered topics are proposal of probabilistic model based on Maximum A posteriori Probability optimization in depth estimation, proposal of accuracy refinement technique based on Mid-Level Hypothesis and improvement of temporal consistency in the estimated depth maps.

In Sections 3.1 to 3.6 a complete probabilistic model for $FitCost$ function (composed of $DataCost$ and $TransitionCost$) has been proposed. This part of the dissertation has been started with general theoretical derivation of $\boldsymbol{DataCost}$ based on **Maximum a Posteriori Probability** rule (Section 3.1). Then, the derived formula (21) from page 51, has been thoughtfully analyzed with respect to simplification (Section 3.2) to classical forms related with to SSD or SAD (equations (30) and (35) on pages 53 and 54) along with verification of the conditions that have to be meet for such simplification. It has been shown that at least some of the conditions are not meet in a practical case of multiview test sequences (Section 3.3) and basing on that a novel formulation for $DataCost$ model has been proposed (Section 3.4). A method for estimation of parameters of this model has been shown on an example of the test sequences. Next, a probabilistic model for $\boldsymbol{TransitionCost}$ has been proposed (in Section 3.5) also with a method for estimation of parameters of this model. In the end of the considerations, the results have been shown.

The attained results show average **gain of about 0.08dB to 2.8dB** in average over tested sequence set**,** calculated with respect to PSNR of virtual views, synthesized with use of depth

maps generated with the proposed method, over the reference. As reference, the original unmodified Depth Estimation Reference Software (DERS) has been used with manual calibration of Smoothing Coefficient per sequence. For the case of selection of the worst checked (yet rational) Smoothing Coefficient value for the original DERS, the average gain is about 2.79dB. For the case of selection of the best found Smoothing Coefficient, the average gain is only about 0.08dB, but it can be noted that the proposed technique attained that without manual of such coefficient.

This constitutes **one of the biggest advantages** of the proposed depth estimation method – it does not require arbitrary manual calibration of coefficients. All required model parameters can be algorithmically estimated like it was shown on example of the tested sequences in Section 3.5.

In Section 3.7 author's achievements in area of depth estimation related to **accuracy refinement of depth maps** have been shown. A novel depth refinement by Mid-level Hypothesis technique has been shown. The proposed method provides an alternative solution for generating sub-pixel precise depth maps, without necessity to increase complexity of the core depth estimation algorithm - in the proposal, the depth in refined in a post-processing step.

The presented results show that the quarter-pixel-precise depth maps, generated with use of the proposed technique, **provide gains of about 0.3dB of PSNR in average** (over the tested sequences)**,** related to the quarter-pixel-precise depth maps generated with the original unmodified Depth Estimation Reference Software (DERS). Comparing to full-pixel-precise depth maps generated with the original unmodified DERS, **the gains are even higher and are about 2dB of PSNR** in average. All of the PSNR gains have been measured as quality of virtual view synthesized generated with use of the given depth map over the tested sequence set. As for complexity, the proposed technique provides **reduction of about 3.7× of computation time**, related to direct quarter-pixel-precise depth estimation using the original (unmodified) DERS technique.

In Section 3.8 developments of the author in area of depth estimation related to enhancement of the **temporal consistency** of stereoscopic depth maps have been shown. A novel approach is proposed in which temporal consistency of the estimated depth is increased by application of noise reduction technique in the input video, a priori to the depth estimation itself. Two noise reduction techniques has been developed by the author in order to provide proof of the presented concept - Still Background Noise Reduction (SBNR) and Motion-Compensated Noise Reduction with Refinement (MCNRR). Although the developed

noise reduction techniques are fairly simple, they have provided evidence that the proposed approach brings substantial gains.

In results it has been shown that **visual quality of the input video has been improved**. Further, results of experiment with depth estimation have been shown. It has been noticed that although the proposal does not provide gains in case of objective PSNR metric, the subjective evaluation, of the views synthesized basing on the depth maps generated with use of the proposed noise reduction technique, shown that application of the proposal **provides gain of about 0.7 to 1.2 MOS points** (Mean Opinion Score). Finally, temporal consistency of generated depth maps has been verified with use of Pearson Linear correlation coefficient and with use of video coding (MVC). Both of the tests has shown that the application of the proposed noise reduction methods increase temporal consistency of the estimated depth. For instance, **the gains in compression of depth maps are about 30%** (Bjøntegaard metric, in average over the tested sequence set, retaining the same synthesized views quality) related to the compression of depth maps estimated with the original, unmodified Depth Estimation Reference Software.

The abovementioned summary end the part of the dissertation related to depth estimation. Further, Chapter 4 and 5 will focus on area of depth coding.

Chapter 4. State-of-the-art in depth map coding

The 3D video systems are evolving from simple stereoscopic systems to sophisticated second-generation systems that provide more realistic perception of the 3D space [247]. Prospective applications of the second-generation 3D video systems include autostereoscopic displays, variable-baseline-distance systems as well as the free-viewpoint television [250][251]. The second-generation 3D video systems need efficient representations of 3D scenes. Practical description of a 3D scene is multiview video plus depth (MVD) [252], i.e. multi-viewpoint video together with the corresponding depth maps estimated during the process of content production.

In this chapter an overview of coding tools that involve depth will be provided. Then, techniques that are directly related to the dissertation, which is representation of the depth for coding, will be highlighted.

4.1. Coding tools that involve depth

The new generation of 3D video is a subject of research in many laboratories around the World and is a very fast developing field. From 2006, Motion Picture Experts Group (MPEG) of ISO/ITU founded a new multiview/FTV/3DV activity [248][249][144][250][251] targeted at developing a framework and technology that will be a part of a future 3D standard. Thanks to MPEG, many research centers around the world cooperate in order to develop an agreed technology. These works include multiview coding, depth estimation depth coding etc. It can be said, that works of MPEG (and further JCT-3V) participants reflect the current state-of-the-art in the area of multi view video and 3D-TV.

The first step of MPEG towards 3D video was Multiview Video Coding (MVC) technology completed in 2009. MVC, described as annex H of ISO/IEC 14496-10:2012 and ITU Rec. H.264 video coding standard, is an extension of AVC (Advanced Video Coding) technology, extending it with ability to transmit multiple videos in a more efficient way that exploits inter-view similarities. This is attained by usage of well-known motion-compensation prediction mechanism, adapted as inter-view disparity compensation. Such approach is a balanced compromise between higher codec complexity and compression efficiency. The gains of MVC over AVC simulcast are about 30% for the dependent views [2][7][76][148][149]. The base (non-dependent) view of MVC is coded in the same way as in AVC and thus this view single view of MVC is backward compatible with AVC.

The MVC standard itself does not enable coding of supplementary depth data. The works of MPEG group on development of AVC/MVC extensions, allowing coding of depth maps, have been proceeding in parallel to the development of this dissertation. In particular, "MVC Extension for Inclusion of Depth Maps" (MVC+D) coding technology has recently been included in Annex I of ISO/IEC 14496-10:2012 and ITU Rec. H.264 video coding standard. Another, more advanced AVC/MVC extension, called "AVC compatible video-plus-depth extension" is currently (end of 2013) going through balloting stage of standardization in ISO/IEC and ITU committees. The description, as annex J of ISO/IEC 14496-10:2012 and ITU Rec. H.264, is expected to be finalized in 2014.

Another thread of scientific works in MPEG group is development of a new generation of 2D (monoscopic) video coding technology, named High Efficiency Video Coding (HEVC). HEVC provides substantial gains over AVC, ranging from 40% to 60% [253]. It is worth to notice that such gains are higher than those of MVC over AVC. In context of multiview/3D video coding, it means that it is more efficient to employ HEVC simulcast (for each view independently) rather than to use MVC [76].

On the other hand, usage of inter-view coding tools (like those in MVC) in HEVC can bring even higher gains. Therefore, multiview extension of HEVC (in a way similar as MVC is an extension of AVC) is long-time target of scientific works around the world and in the MPEG group.

During the above-mentioned works, the compression of depth was identified as an important research task. This task is somewhat different from the task of video compression where the goal is to compress visual data in such a way that the decoded video is possibly similar to the input video. On the contrary, depth is not watched by a viewer but it is used to synthesize virtual views needed for an autostereoscopic display or in a free-viewpoint television system. So, mostly the decoded depth quality is expressed by the quality of the synthesized views.

The developments in area of depth compression include coding methods which have various levels on compatibility with legacy technologies defined by standards.

In work [254] authors analyze coding of depth with use of wavelet transform with use of techniques developed for JPEG2000. Although the technique could be extended with mechanism applicable for efficient compression of depth video sequences instead of still pictures, at the presented stage such feature was missing. This lack of research has been filled in work [255], where authors propose complete coding system, along with analysis of impact of wavelet compression on image-based rendering synthesis of virtual views. The results are

promising, yet still, authors did not address the problem of compatibility with existing coding standards.

In work [256] it is suggested that wavelet coding with motion compensated prediction can provide good efficiency when used for coding of mixture of homogenous regions and sharp edges, like in the case of depth maps. Similar concepts are developed in works [257][258], where usage of adaptive wavelets with tree-based partitioning is proposed. Although the authors mention promising compression performance, the proposed approach is not a compelling solution, because it uses coding technology which is incompatible with existing video coding standards, both regarding to the syntax and technology.

In work [259], depth coding in AVC is considered, which incorporates skip-mode selection decision based on distortion analysis. The authors show that thanks to their approach not only the coding performance is improved but also temporal consistency of the reconstructed depth is enhanced.

4.2. State-of-the-art directly related to the proposals in the dissertation

The techniques that are more aligned with state-of-the-art in video compression include platelets [260][261][262] wedgelets [265][266]. In work [260] platelets have been proposed for medical applications. Authors of paper [261] have proposed platelets as an efficient depth coding tool and have developed this approach over years [262][263]. In such, depth coding is integrated with existing coding pipeline based on macroblocks and coding units as a special coding mode. The general idea is that inside of given coding region (e.g. macroblock, coding units) the depth is modeled as a flat plane called platelet. Independently, usage of platelets is considered in [133] in context of MVC codec.

Wedgelets are extension of the idea of platelets, where given block is represented with more planes, separated by edges. The foundations of such idea have been given in [264], where partitioning schemes have been theoretically considered. The application of platelets to depth coding can be found in [265] and further developed form in [266]. In those works, depth coding is also integrated with coding pipeline as a coding mode. In particular, the depth can be modeled as a single plane (platelet) or two planes separated by a discrete edge whose location is signaled in the bitstream. In work [267] a contour-based coding of depth is proposed. In such, edges in the depth maps are identified and then their positions are signaled in the bitstream. The proposal relates to intra-coding only, but provides interesting gains.

In the standardization expert groups like MPEG, VCEG and JCT-3V, there is strong expectation to use the existing coding tools as much as possible for coding of depth.

The basic compression tools are usually capable of processing 8-bit samples, and they use uniform quantization, i.e. the quantization with the constant quantization step that might be changed for some data structures like slices and macroblocks or coding units. Such uniform quantization is characteristic for both basic modern video standards: the AVC [111] and the new one – the HEVC [121]. On the other hand, it can be intuitively understood that exact depth is very important for foreground objects while small depth degradations in far background are mostly well tolerated by the human visual system. Therefore, non-uniform quantization would probably be appropriate for depth coding. Therefore, in order to preserve conformance with the standards like AVC and HEVC, we propose to process the depth values using a non-linear function. Such processing together with uniform quantization is equivalent to the requested non-uniform quantization.

The idea of depth processing using non-linear transformation of the depth-sample values is not a new one. It was already considered in [268] but with no particular relation to compression. The authors consider the influence of the depth representation on the attained visual quality only.

In [269], a non-linear transformation of sample values was used to obtain finer depth quantization in the background, i.e. a non-linear transformation was used in the opposite way to that proposed in this dissertation. Although, the authors show that the overall objective quality of a virtual view synthesized with use of their proposal is increased, it is missed that it is the gains are coming from the background areas (for which the depth is represented more precisely) and no analysis on impact of the foreground objects and thus on the visual quality is done.

In paper [270] author also propose depth coding tool abased on non-uniform representation but unlikely in the dissertation, the depth prediction signal is transformed instead of the depth values.

Chapter 5. Proposed non-linear depth representation for coding

In this Chapter, research conducted by the author in area of depth coding will be presented. In particular the concept of depth representation will be studied. First a proof-of-concept idea will be presented, using a simple non-linear function. Then an original theoretical derivation for non-linear representation of depth will be provided. The result will be shown in form of both subjective and objective assessment. Finally, adoption of proposed non-linear depth representation to video coding technology standards developed by ISO/IEC MPEG group is highlighted.

5.1. The idea of non-linear depth representation

The straightforward approach to depth map transmission is to use uniformly quantized disparity values, normalized to range 0…255, called depth (3). That would also be a quite good method if the application of the transmitted data is unknown. However, in the case of next generation 3D video systems, considered in this dissertation, the transmitted depth map is used to synthesize virtual views. Therefore, mostly, the decoded depth quality is expressed by the quality of the synthesized views.

Mentioned, straight-forward linear representation of depth with uniform quantization of disparity, unfortunately does not match the properties of the **human visual system that is more tolerant to disparity errors in the background of a synthesized scene** (Fig. 62) than to errors in the foreground. Therefore, the author has developed a coding scheme which resembles non-uniform quantization, so that distant objects are quantized more roughly than the closer ones.

In the standardization expert groups like MPEG, VCEG and JCT-3V, there is strong expectation to use the existing coding tools as much as possible for coding of depth. The basic compression tools, is characteristic for both basic modern video standards: the AVC [111] and the new one – the HEVC [121], are usually capable of processing 8-bit samples, and they use uniform quantization, which is not optimal for coding of depth for the purpose of virtual view synthesis. Therefore, in order to preserve conformance with the standards like AVC and HEVC, it is proposed to process the depth values using a non-linear function. Such processing together with uniform quantization is equivalent to the requested non-uniform quantization (Fig. 63).

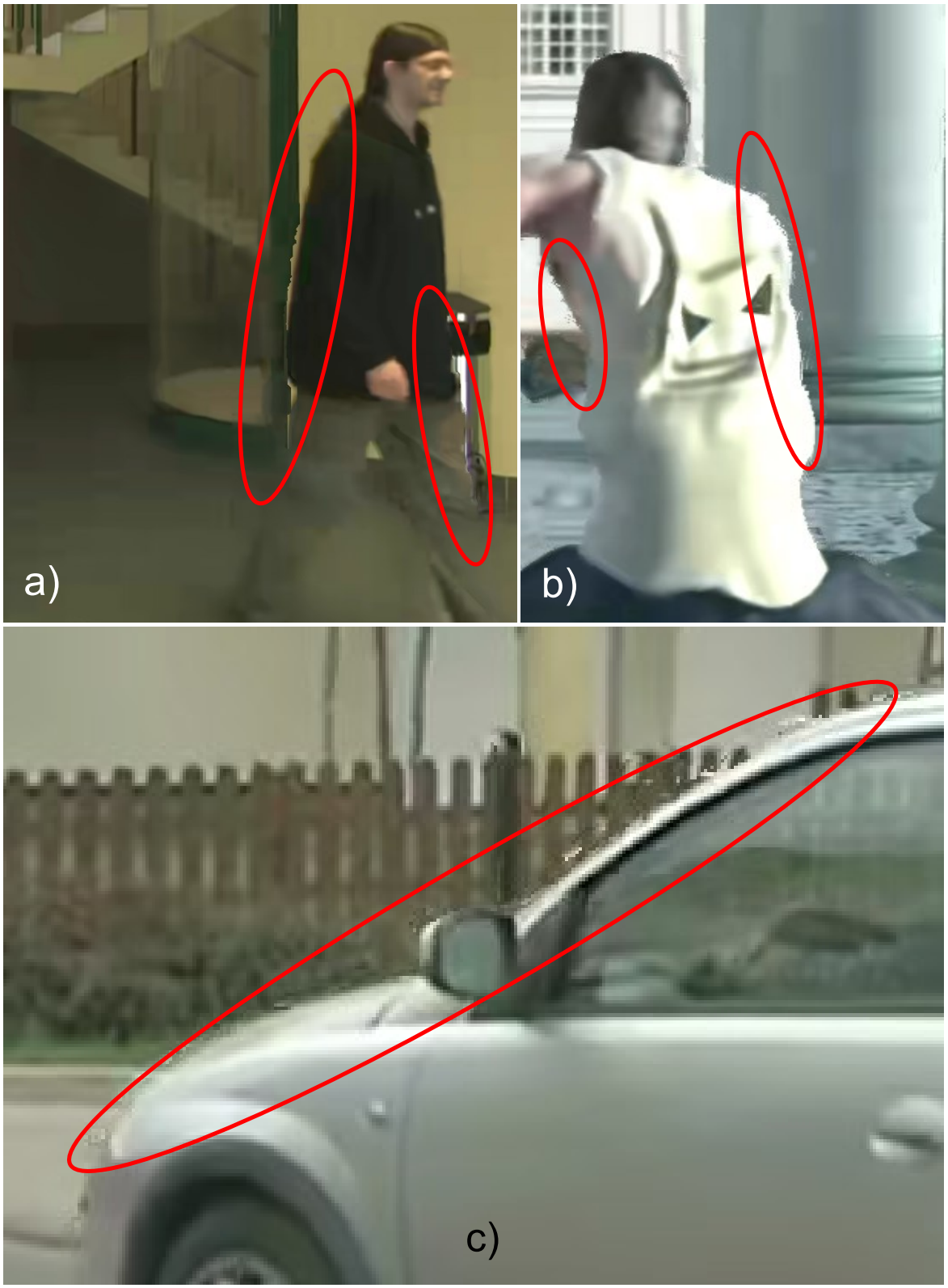


Fig. 62. Exemplary artifacts resulting from linear quantization in coding of depth maps, especially visible for objects in the foreground, marked in red.
a) “Poznan Hall 2” sequence, b) “Undo Dancer” sequence, c) “Poznan Street” sequence.

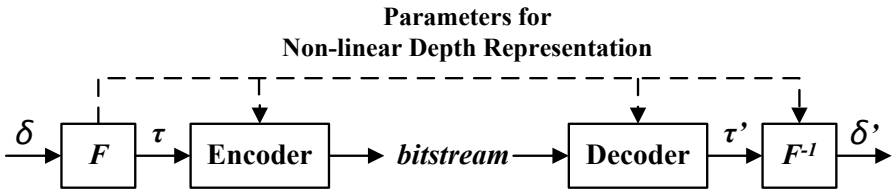


Fig. 63. Non-uniform quantization realized by a transform F performed on depth values δ on the input of the coded and inverse transform F^{-1} on the output of the codec.

Assume that distance to a point on a real object is z. Practical limitations yield that for all objects in a scene the depth values are within a final interval, i.e. $z_{near} < z < z_{far}$ where z_{near} and z_{far} are the distances to closest and the farthest object in a scene. The depth data are usually stored as disparity d which is proportional to the inverse of z [159]:

$$\delta = \delta_{max} \cdot \left(\frac{1}{z} - \frac{1}{z_{far}}\right) / \left(\frac{1}{z_{near}} - \frac{1}{z_{far}}\right) \quad , \tag{57}$$

where $\delta_{\max}$ is 255 for typically used 8-bit depth samples.

For uniform quantization, this representation has the following advantage: a higher depth resolution of nearby objects is obtained. In order increase this effect, an additional non-linear transformation is proposed to be performed on the depth-sample values:

$$\tau = F[\delta] \quad , \tag{58}$$

where τ is the transformed depth and $F[\cdot]$ is a non-linear function, e.g. as shown in Fig. 64 for the most common case of 8-bit representations.

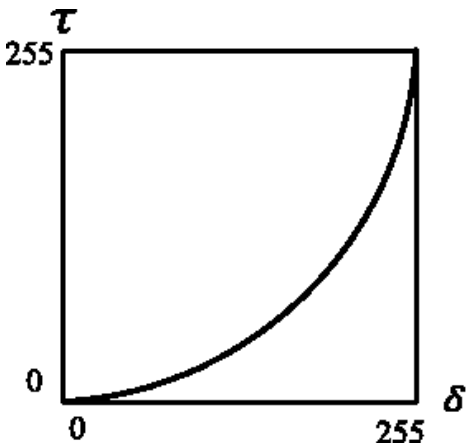


Fig. 64. Non-linear depth transformation performed before coding.

The transformation (58) is performed on depth samples before coding. Now, the depth coding itself is performed on the internal values τ instead of the external values δ. This non-linear transformation influences prediction errors and their linear transforms (mostly DCT-like) that are used in the course of the intra-frame and inter-frame coding. The transform samples are quantized and this process is influenced by the proposed non-linear depth transformation.

After transmission, the process is reversed with inverse non-linear function $F^{-1}[\cdot]$ so that, basing on reconstructed transformed depth τ', reconstructed linearly quantized normalized disparity δ' is retrieved:

$$\delta' = F^{-1}[\tau'] \tag{59}$$

In the next subsection a simple proposal for non-linear transformation (58) and inverse non-linear transformation (59) will be presented. This simple proposal will provide a proof of concept of non-linear depth representation idea.

5.2. Proof of concept proposal for non-linear transformation

The first proposal for non-linear depth transformation proposed by the author is based on the idea of gamma-correction well known from luminance coding in classical video systems. It is proposed to use non-linear disparity representation in the codec, so that each sample value is defined by the following power-law expression:

$$\tau = \left(\frac{\delta}{\delta_{max}}\right)^{\gamma} \cdot \tau_{max} \tag{60}$$

where δ_{max} and τ_{max} are the maximal values of δ and τ, respectively (e.g. 255 for 8-bit precision).

In such way, closer objects are represented more accurately than the distant ones and thus quantization is non-uniform. The defined non-linear transformation has been successfully employed in Poznan University of Technology response to "Call for Proposals for 3D Video Coding Technology" [129] issued by MPEG. It was shown experimentally that already simple choice $\gamma = 1.3$ (for small QP values) and $\gamma = 1.6$ (for large QP values) gives good results.

Obviously, a reciprocal operation is performed after decoding the disparity map:

$$\delta' = \left(\frac{\tau'}{\tau_{max}}\right)^{1/\gamma} \cdot \delta_{max} \tag{61}$$

The initial proposal for non-linear depth transform (60) has been implemented by author as a part of codec developed by Poznan University of Technology, Chair of Multimedia Telecommunications and Microelectronics, which has been submitted to "Call for Proposals on 3D Video Coding Technology" issued by MPEG group in 2011 [129]. This proposal has been rated very high among other proposals and was found to be one of the best performing proposals in HEVC category. The excellent results attained by the proposed codec, provoked a deeper analysis of the share of gains provided by particular tools [271][272], cited in the graphs below (Fig. 65 - Fig. 68).

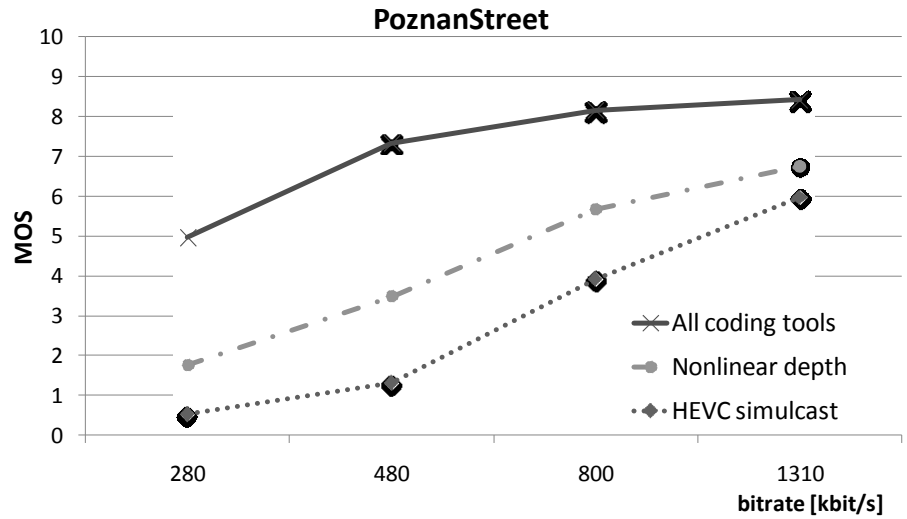


Fig. 65. Subjective test results for Poznan Street sequence.

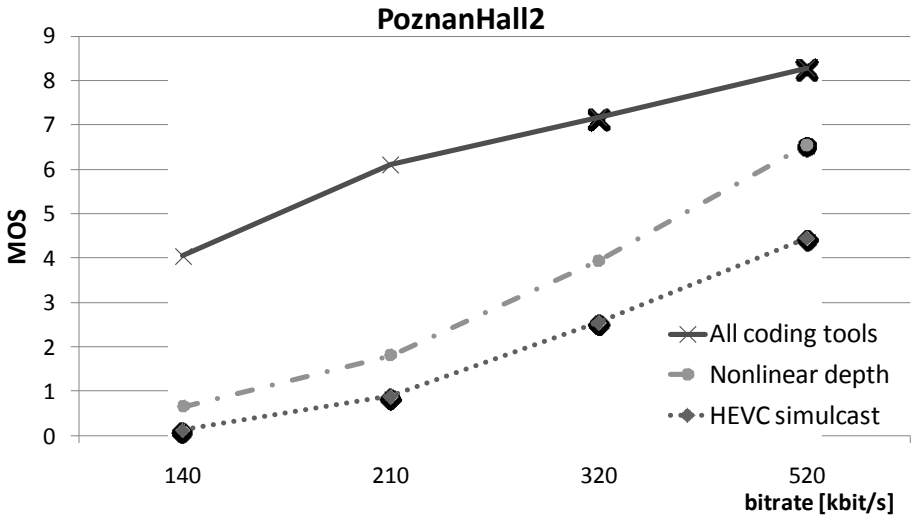


Fig. 66. Subjective test results for Poznan Hall 2 sequence.

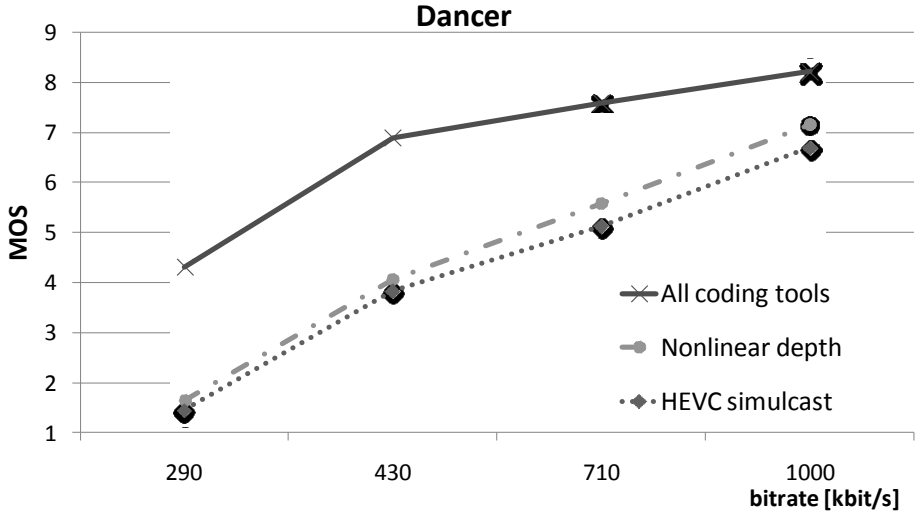


Fig. 67. Subjective test results for Undo Dancer sequence.

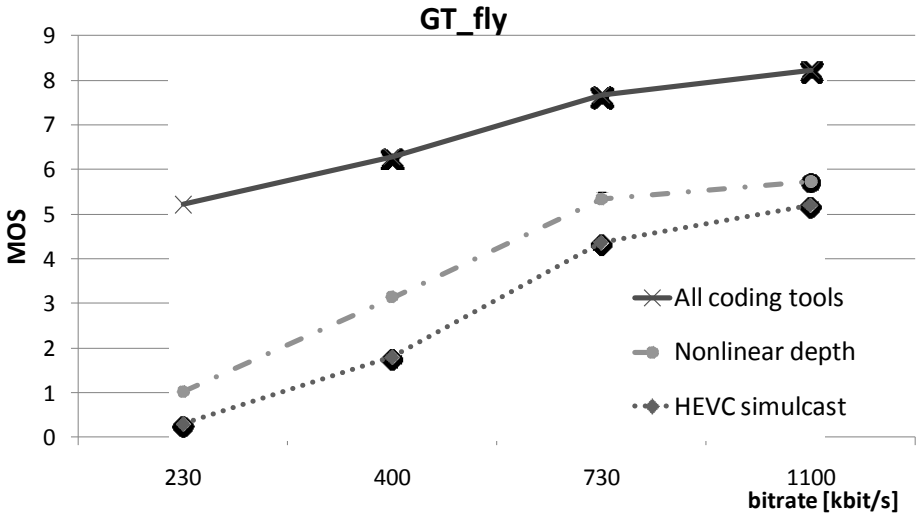


Fig. 68. Subjective test results for GT Fly sequence.

Results of this analysis (Fig. 65 - Fig. 68) show that sole application of non-linear depth transformation described by equation (60) leads to improvement of the subjectively assessed quality of the reconstructed video by about 1 MOS (Mean Objective Score point). It can be noticed that this also correspond to about one third of the overall subjective gains provided by the whole codec (with all coding tools) over HEVC simulcast coding of multiview test data.

Promising results of resolution of CfP [129] issued by ISO/IEC MPEG and results of further evaluation presented in the dissertation, provided a proof-of-the-concept for proposal of non-linear depth transformation. This has motivated author to do a more structured approach to the subject. The devised formulation is presented in the next subsection.

5.3. A theoretical approach to selection of non-linear transformation

In this subsection a theoretical derivation for non-linear depth transformation will be provided. The requirement is that distant objects are quantized more roughly than the closer ones. Therefore, the assumption is that the quantization step $s(\tilde{\delta})$ decreases with increasing of disparity δ. Also, the expectation is that the quantization step $s(\tilde{\delta})$ changes uniformly across the whole disparity range.

For the sake of simplicity of derivation, instead of considering values of δ and τ (within range 0 to δ_{max} and 0 to τ_{max} respectively) and transform function F, let's consider transform function $\tilde{F}$ and values of $\tilde{\delta}$ and $\tilde{\tau}$ normalized to range [0..1] so that:

$$\tilde{\delta} = \frac{\delta}{\delta_{max}} \quad ; \quad \tilde{\tau} = \frac{\tau}{\tau_{max}}. \tag{62}$$

For small values of $\tilde{\delta}$ (far objects) the quantization step s is assumed to be large, while for large values of $\tilde{\delta}$ (close objects) the quantization is expected to be fine:

$$s(\tilde{\delta}) = A \cdot e^{-\alpha \cdot \tilde{\delta}}, \tag{63}$$

where α is a constant parameter (e.g. $\alpha = 0.5 \div 5.0$ is a typical choice used in the preliminary experiments). As the sum of all the quantization steps should cover the whole unit interval:

$$1 = \int_0^1 s(k)\, dk = A \cdot \int_0^1 e^{-\alpha \cdot k} dk = -\frac{A}{\alpha} \cdot (e^{-\alpha} - 1) \quad , \tag{64}$$

where k is the integration variable used instead of $\tilde{\delta}$ or $\tilde{\tau}$, and the parameter A:

$$A = -\frac{\alpha}{e^{-\alpha} - 1} \quad . \tag{65}$$

Thus, the inverse non-linear depth transformation:

$$\tilde{\delta} = \tilde{F}^{-1}(\tilde{\tau}) = \int_0^{\tilde{\tau}} q(k)\, dk = A \cdot \int_0^{\tilde{\tau}} e^{-\alpha \cdot k} dk = -\frac{A}{\alpha} \cdot (e^{-\alpha \cdot \tilde{\tau}} - 1) \quad . \tag{66}$$

and after some mathematical operations:

$$\tilde{\delta} = -\frac{A}{\alpha} \cdot (e^{-\alpha \cdot \tilde{\tau}} - 1) \quad , \tag{67}$$

$$-\frac{\alpha}{A}\cdot\tilde{\delta}+1=e^{-\alpha\cdot\tilde{\tau}} \quad , \qquad (68)$$

$$\ln\left(-\frac{\alpha}{A}\cdot\tilde{\delta}+1\right)=-\alpha\cdot\tilde{\tau} \quad , \qquad (69)$$

$$-\frac{1}{\alpha}\cdot\ln\left(-\frac{\alpha}{A}\cdot\tilde{\delta}+1\right)=\tilde{\tau} \quad . \qquad (70)$$

and by substitution of A and some further simplification:

$$\tilde{\tau}=-\frac{1}{\alpha}\cdot\ln\left(1-\frac{\alpha}{-\frac{\alpha}{e^{-\alpha}-1}}\cdot\tilde{\delta}\right) \quad . \qquad (71)$$

$$\tilde{\tau}=\tilde{F}\left(\tilde{\delta}\right)=-\frac{1}{\alpha}\cdot\ln\left(1-\tilde{\delta}\cdot\left(1-e^{-\alpha}\right)\right) \quad . \qquad (72)$$

Finally, if the original notation with (not scaled) variables δ, τ and F is used, we can get the desired forward transformation:

$$\tau=F(\delta)=-\frac{\tau_{max}}{\alpha}\cdot\ln\left(1-\frac{\delta}{\delta_{max}}\cdot\left(1-e^{-\alpha}\right)\right) \quad . \qquad (73)$$

The results attained with use of transformation (73) lead not only to comparable subjective gains as in the case of power-law-based expression (60) but also provide objective gains (measured by PSNR). This fact, has been brought to the attention of MPEG group [49] for consideration as a tool for a new generation coding technology standards.

5.4. Approximation of non-linear depth transformation

One of the requirements, considered by experts during evaluation of proposed tools, like non-linear depth representation tool by the author, is that standards should be defined in a flexible way. Therefore it would be not feasible to define a single arbitrary transform function, e.g. defined by (60) or (73). In order to fulfill this requirement, the author has proposed that the shape of the plot of function $F[\cdot]$ is directly given to the encoder, transmitted in the bitstream and then decoded by the decoder.

The author's proposal of non-linear depth representation has been accepted and it was decided (e.g. [114][117]) that the non-linear transformation function will be linearly approximated in the intervals. It was accepted that only a set of the equidistant deviations

from the diagonal will be signaled in the bitstream (see Fig. 69) [49]. These deviations are defined solely by the deviation vector: $w = [w_0, w_1, \ldots, w_N]$.

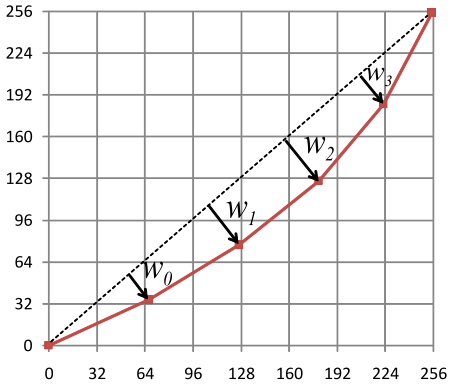


Fig. 69. Transformation definition by equidistant deviations w_i and the linear approximation in the intervals. In this example deviation vector w is 4 element vector.

The advantages of such approximation of non-linear depth transformation are as follows:

- Shape of the transform may be optimized individually by the encoder.
- Further developments may bring new ideas about the definitions of the transformation.
- The proposed approximation can be easily calculated using fixed-point arithmetic.
- The transformation can be switched off if in particular case it does not bring any gain.

In particular, it has been noticed that if the distribution of normalized disparity δ samples is concentrated around small part values of range 0 ... δ_{max} usually it is better to switch off the non-linear depth transformation. Such abnormal depth distributions are identified basing on expected value $E[\delta_{x,y}]$ of normalized disparity map $\delta_{x,y}$:

$$E[\delta_{x,y}] = \frac{1}{W \cdot H} \sum_{x \in 1..W} \sum_{y \in 1...H} \delta_{x,y} \quad , \tag{74}$$

where W and H correspond to the width and the height of the image, respectively.

Therefore it is beneficial to disable non-linear depth transformation, when $E[\delta_{x,y}]$ is below a predefined $E_{threshold}$ value (that was set to 100 in the experiments with 8-bit samples). This relatively simple condition can be used for automatic switching the tool on and off for individual sequence:

$$Enable\ NDR = \begin{cases} false & if\ E[\delta_{x,y}] < E_{threshold} \\ true & if\ E[\delta_{x,y}] \geq E_{threshold} \end{cases} \quad . \tag{75}$$

The mentioned advantage of flexibility of non-linear depth representation proposal, already have led to development done author and other research centers [273].

In the next subsection, experimental evaluation of the Non-linear Depth Representation coding tool, submitted by the author for consideration of ISO/IEC MPEG group, will be presented.

5.5. Experimental results for depth map coding

The non-linear depth representation tool (under the name of NDR) has been submitted to ISO/IEC MPEG group and has been evaluated by international experts in a series of Exploration Experiments e.g. [237][238][60][63]. These Exploration Experiments compare several tools related to given subject (e.g. depth coding) on the common ground environment, constituted in Common Test Conditions (CTC) [236] document. Each submitted result must be confirmed by at least one independent research center in order to be considered. Basing on such evaluation, only the best and the most interesting tools are adopted to the standard under development.

The experiments have been done for depth coding extensions of the AVC standard [114][115][116][117][118][119], which have been under development at that time – MVC+D and AVC-3D.

In all experiments, the transformation defined in equation (73) has been used with parameter $\alpha = 1.8$. The transformation was implemented by approximation with 41 nodes, i.e. deviations have been defined for 39 nodes (for two boundary nodes, for normalized disparity $\delta = 0$ and $\delta = 255$, the deviation is always 0). The deviation vector w (see Fig. 69) which has been used is as follows:

$$w = [2,4,7,8,10,12,14,16,17,19,20,21,22,23,24,25,26,26,27,27,27, \\ ,27,27,27,26,26,25,24,23,22,20,19,17,15,13,11,9,6,3]. \quad (76)$$

The final results have been submitted to MPEG in [57] and have been independently cross-checked by Samsung Corporation [274]. The Nonlinear Depth Representation (NDR) tool has been turned on in the case of three sequences from the set (GT Fly, Kendo and Balloons), while for others (Poznan Street, Poznan Hall 2, Undo Dancer, Newspaper) it has been turned off due to application of the rule described in equation (75) with value of the $E_{threshold}$ = 100.

The results are summarized in Table 15 and Table 16. The average gains in bitrate (dBR) as well as in PSNR (dPSNR) were calculated using Bjøntegaard measures [127]. 3 views and 3 depth maps are coded and results for respective bitstream components are presented.

The gains were calculated for the following cases (A, B, C and D), which are reflected in headings of Table 15 and Table 16:

A. Video-only bitrate (depth bitstream not included) and average PSNR for 3 views.
B. Depth-only bitrate and PSNR for depth (average for depth maps for 3 views).
C. Total bitrate for 3 views and 3 depth maps and average PSNR for 3 decoded views.
D. Total bitrate for 3 views and 3 depth maps and average PSNR for 3 decoded views and 6 views synthesized in spatial position in between the coded views (Fig. 70).

Table 15. Bjøntegaard gains in bitrate and PSNR due to application of non-linear depth representation in HP profile. 3 views and 3 depth maps are coded and results for respective bitstream components are presented.

Sequence	A. Multiview video coding		B. Depth coding		C. 3 views with depth maps		D. 3 views with depth maps and 6 synthesized views	
	dBR [%]	dPSNR [dB]	dBR [%]	dPSNR [dB]	dBR [%]	dPSNR [dB]	dBR [%]	dPSNR [dB]
GT Fly	0.00	0.00	-21.93	1.37	-1.25	0.05	-0.11	0.00
Balloons	-0.01	0.00	-25.87	1.34	-3.27	0.17	-2.59	0.13
Kendo	0.00	0.00	-25.76	1.64	-5.28	0.28	-4.13	0.20
Average (3 seqs.)	**0.00**	**0.00**	**-24.52**	**1.45**	**-3.27**	**0.17**	**-2.28**	**0.11**
Poznan Street	Zero gains - Nonlinear Depth Representation is disabled due to equation (75)							
Poznan Hall 2								
Undo Dancer								
Newspaper								
Average (all 7 seqs.)	**0.00**	**0.00**	**-10.51**	**0.62**	**-1.40**	**0.07**	**-0.98**	**0.05**

Table 16. Bjøntegaard gains in bitrate and PSNR due to application of non-linear depth representation in EHP profile. 3 views and 3 depth maps are coded and results for respective bitstream components are presented.

Sequence	A. Multiview video coding		B. Depth coding		C. 3 views with depth maps		D. 3 views with depth maps and 6 synthesized views	
	dBR [%]	dPSNR [dB]	dBR [%]	dPSNR [dB]	dBR [%]	dPSNR [dB]	dBR [%]	dPSNR [dB]
GT Fly	1.36	-0.05	-23.87	1.49	0.63	-0.02	0.74	-0.02
Balloons	0.44	-0.02	-21.79	1.13	-1.35	0.06	-1.18	0.05
Kendo	0.19	-0.01	-20.35	1.30	-3.41	0.15	-3.13	0.14
Average (3 seqs.)	**0.66**	**-0.03**	**-22.00**	**1.31**	**-1.38**	**0.06**	**-1.19**	**0.06**
Poznan Street	Zero gains - Nonlinear Depth Representation is disabled due to equation (75)							
Poznan Hall 2								
Undo Dancer								
Newspaper								
Average (all 7 seqs.)	**0.28**	**-0.01**	**-9.43**	**0.56**	**-0.59**	**0.03**	**-0.51**	**0.02**

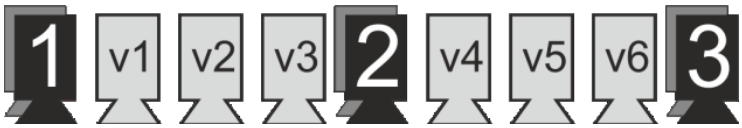


Fig. 70. The arrangement of the views: the v are marked in black while the views synthesized in the receiver are marked in gray.

The overall bitrate reduction with respect to the case D is up to 4.13% while there is no measurable increase of complexity. In average, the gains are **0.98%** (HP profile) and **0.51%** (EHP profile) when the has been done for all of the test sequences – also those that do not fulfill the requirements for the depth distribution as described in Chapter 5, eq. (64) (for these sequences the transformation was switched off). Only in the sequences in which the transform has been turned on are considered, the gains are 2.28% (HP) and 1.19% (EHP).

Please note that those mentioned gains are coming from coding tools for depth solely, while depth is about 10% of the whole bitstream. A more optimistic interpretation of the results, presented in column B of Table 15 and Table 16, shows that the bitrate of the depth itself has been reduced by **10.51% (HP)** and **9.45% (EHP)** when considering average over all test sequences and **24.52% (HP)** and **22.00% (EHP)** when the average is calculated only over the test sequences that use the proposed tool.

Also subjective tests [57] have been performed in order to compare visual quality of the synthesized views produced from the compressed depth maps both in the presence and in the absence ("anchor reference") of non-linear depth representation (for the same bitrate). For the tests, 32 subjects have assessed the quality of stereo clips (2 subjects needed to be rejected) using the single stimulus method.

The subjects have been presented a couple of tests in Double Stimulus Impairment Scale (DSIS) method [128]. First, the reference (stereo pair synthesized from uncompressed original) sequence was shown. Then, a tested case was shown - this could be randomly either one of:

- 3D-ATM [120] anchor, or
- 3D-ATM with proposed Non-linear Depth Representation.

The tested sequence has been always coded at constant bitrate: from the highest (R4) to the lowest (R1), reflecting Common Test Condition (CTC) [46][236] and general methodology developed by MPEG for exploration experiments (EE) [238].

The presented stereo pair was composed from two synthetic views, around the base view. After each test, subjects gave their scores reflecting quality. The sessions were performed during the MPEG meeting in Geneva.

The results for various bitrates (R1 – R4) are depicted in Fig. 71, together with 95% confidence intervals. The bitrates have been selected according to MPEG guidelines for individual test sequences [46][236]. The results show that non-linear depth transformation improves coding efficiency, although some of the confidence intervals overlap.

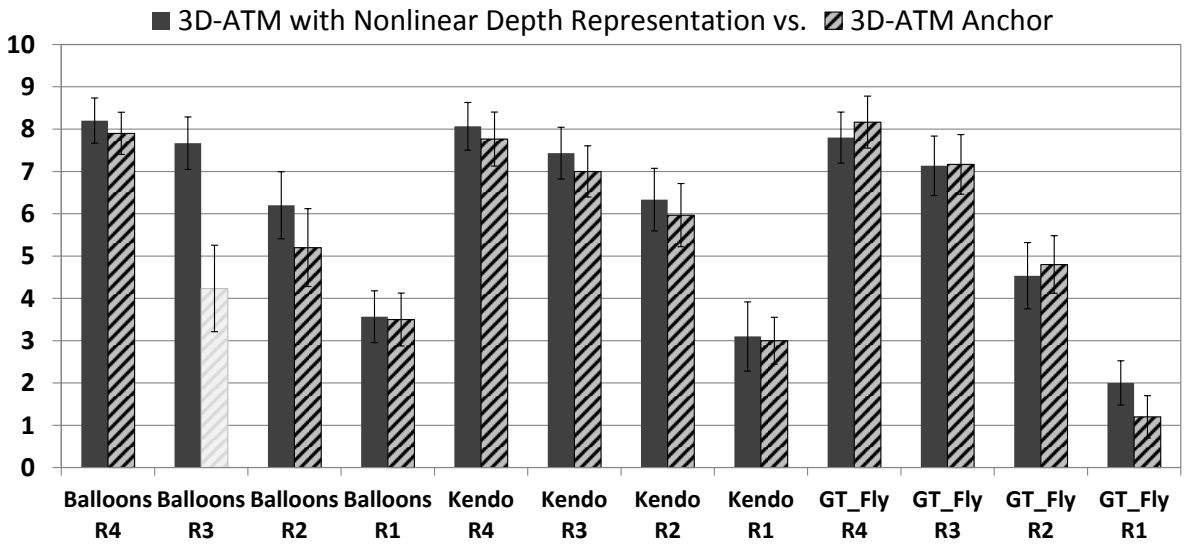


Fig. 71. Results of subjective evaluation of Nonlinear Depth Representation (NDR). One of the cases (Balloons sequence, R3 rate-point, Anchor) is shaded because the results is not reliable due to presentation problems that have occurred during the viewing.

One of the cases (Balloons sequence, R3 rate-point, 3D-ATM Anchor) is shaded because the results is not reliable due to presentation problems that have occurred during the viewing – one of the presented views from stereo-pair had artifacts not related to the experiment, which has influenced ratings given by the viewers inadequately.

5.6. Adoption of non-linear transformation in international coding standards

Each tool that is proposed to be a part of a standard under development is thoughtfully tested and questioned by the experts. The considered aspects include provided compression gains, computational complexity, configurability, implementability, etc.

One of the strongest expectations for coding of depth is to use the existing coding tools as much as possible. The proposed Non-linear Depth Representation (NDR) tool conforms that requirement in both of two major scenarios of depth coding that have been formulated until now:

1. Depth is compressed independently from multiview video in a sense that depth does not influence coding and decoding of multiview video. Such an approach is used in the depth coding extensions of the multiview coding techniques: MVC [113] and MV-HEVC (MHVC) [122]. For MVC, this approach is supported by ATM-HP test model software.
2. Depth values are used in the course of video coding and decoding, e.g. for view-synthesis prediction. Such an approach is used in the 3D video coding proposed as extensions of the AVC [117][118][119] and HEVC [123] standards. For AVC, this approach is supported by ATM-EHP test model software.

In the first approach (Fig. 72), the transformations do not influence any encoding or decoding process. Therefore the information about non-linear depth transformation may be transmitted in the SEI (supplementary enhancement information) messages. The depth coding extension [114][115][116] of MVC [112] has already incorporated depth representation information SEI message that optionally may be used to transmit the information about depth transformation.

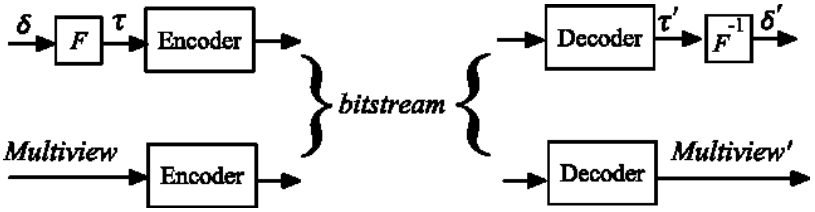


Fig. 72. Independent depth coding: δ and τ denote the original normalized disparity map and transformed, non-linearly represented values, δ' and τ' are decoded (reconstructed) values and $Multivew$ and $Multiview'$ are original and decoded multi-view video, respectively.

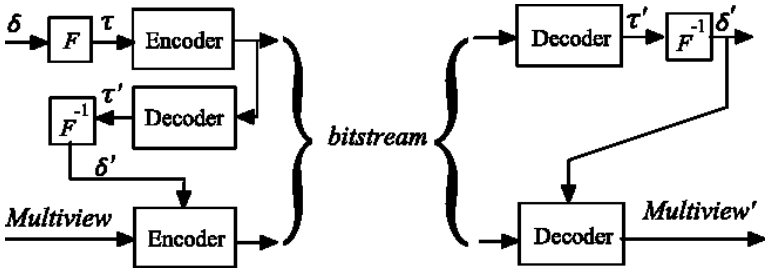


Fig. 73. Depth-dependent coding: : δ and τ denote the original normalized disparity map and transformed, non-linearly represented values, δ' and τ' are decoded (reconstructed) values and $Multivew$ and $Multiview'$ are original and decoded multi-view video, respectively.

In the second approach (Fig. 73), the encoding and decoding of multi-view video exploits the information about depth. A good example of such a depth-dependent operation is view-synthesis prediction. Such prediction needs the values of δ (normalized disparity) rather than τ (transformed representation). Therefore the values of the external representation δ must be used in the course of multiview video encoding and decoding (see Fig. 73).

Therefore, in both cases, the proposed Non-linear Depth Representation (NDR) tool is compliant with existing coding technology concepts and do not impose any overhead – both related do encoder nor decoder complexity.

The advantages of the proposed Non-linear Depth Representation (NDR) tool and positive verification, performed by independent research centers in MPEG group, resulted in **adoption of the proposed NDR tool to new 3D extensions of ISO/IEC 14496-10 and ITU Rec. H.264 international video coding standards**, describing new generation of 3D video coding technologies, known under names of MVC+D and AVC-3D. The first one has already been included as Annex I of AVC specification. The second one is expected to be finalized in 2014 and to be included as Annex J of AVC specification.

Table 18 and Table 17 show the syntax that has been adopted to those standards, describing to MVC+D [114][115][116] and 3D-AVC [117][118][119] video coding technologies, respectively. The parts related to adopted NDR proposal have been marked in gray.

In the case of MVC+D (Table 17), the coding is depth-independent Fig. 72 and therefore NDR tool has been adopted in form of a special Supplemental Enhancement Information (SEI) message. The value of 3 of field *depth_representation_type* in the bitstream indicates that Nonlinear Depth Representation tool is turned on. In such case, field *depth_nonlinear_representation_num_minus1* is signaled with the encoded size of deviation vector v. Next all elements of deviation vector v are transmitted. Thanks to that, the shape of non-linear transformation function can be reconstructed at the decoder (see Fig. 69).

In the case of AVC-3D (Table 18), the coding is depth-dependent (Fig. 73) and therefore NDR tools has been adopted in Depth parameter set unit. The syntax of the signaling is very similar as in the case of MVC+D with exception that the Depth Representation tool is turned on by value of *nonlinear_depth_representation_num* field greater than 0.

Table 17. Adopted syntax of Non-linear Depth Representation (marked in gray) in MVC+D coding technology in [114] Annex I "Multiview and Depth video coding" of ISO/IEC 14496-10:2012 and ITU Rec. H.264 video coding standards. I.13.1.3 - Depth representation information SEI message syntax.

Syntax element	C	Descriptor
depth_representation_info(payloadSize) {		
all_views_equal_flag	5	u(1)
if(all_views_equal_flag = = 0) {		
num_views_minus1	5	ue(v)
numViews = num_views_minus1 + 1		
} else		
numViews = 1		
z_near_flag	5	u(1)
z_far_flag	5	u(1)
if(z_near_flag || z_far_flag) {		
z_axis_equal_flag	5	u(1)
if(z_axis_equal_flag)		
common_z_axis_reference_view	5	ue(v)
}		
d_min_flag	5	u(1)
d_max_flag	5	u(1)
depth_representation_type	5	ue(v)
. . .		
if(depth_representation_type = = 3) {		
depth_nonlinear_representation_num_minus1	5	ue(v)
for(i = 1; i <= depth_nonlinear_representation_num_minus1 + 1; i++)		
depth_nonlinear_representation_model[i]	5	ue(v)
}		
}		

Table 18. Adopted syntax of Non-linear Depth Representation (marked in gray) in AVC-3D coding technology. The related specification is currently being finalized [118] as Annex I of ISO/IEC 14496-10:2012 and ITU Rec. H.264 video coding standards. J.7.3.2.13 - Depth parameter set RBSP syntax.

Syntax element	C	Descriptor
depth_parameter_set_rbsp( ) {		
depth_parameter_set_id	11	ue(v)
pred_direction	11	ue(v)
if(pred_direction = = 0 || pred_direction = = 1) {		
ref_dps_id0	11	ue(v)
predWeight0 = 64		
}		
if(pred_direction = = 0) {		
ref_dps_id1	11	ue(v)
pred_weight0	11	u(6)
predWeight0 = pred_weight0		
}		
num_depth_views_minus1	11	ue(v)
. . .		
depth_param_additional_extension_flag	11	u(1)
nonlinear_depth_representation_num	11	ue(v)
for(i = 1; i <= nonlinear_depth_representation_num; i++)		
nonlinear_depth_representation_model[i]	11	ue(v)
if(depth_param_additional_extension_flag = = 1)		
while(more_rbsp_data( ))		
depth_param_additional_extension_data_flag	11	u(1)
rbsp_trailing_bits( )		
}		

5.7. Summary of achievements in the area of depth coding

In this Chapter, a novel non-linear transformation has been proposed for representation and coding of the depth. First a proof-of-concept proposal has been presented, with use of a simple non-linear function, based on the idea of gamma-correction, well known from luminance coding in classical video systems. The verification of this preliminary proposal has been performed in with use of a codec developed by Poznan University of Technology, Chair of Multimedia Telecommunications and Microelectronics. This codec has been submitted to "Call for Proposals on 3D Video Coding Technology" (CfP) [129], issued by MPEG group in 2011, and has been rated very high among other proposals. Promising results of the codec in the resolution of CfP and also further results of subjective evaluation, created a motivation for more advanced proposal. The devised, theoretical approach yielded with a novel formulation of non-linear transformation for depth representation.

The final proposal has been evaluated experimentally with use of coding technology implemented in MPEG Reference Software for AVC-based 3D video coding technologies [120] – MVC+D and AVC-3D. It has been shown that application of the proposed non-linear depth representation provides substantial subjective gains of about **0.3 to 1 MOS point** (depending on the case in the tested sequence set) and considerable bitrate reduction – in average up to about **25%** bitrate reduction of the depth component of the bitstream.

Finally, **adoption of the proposed non-linear depth representation to new 3D extensions of ISO/IEC 14496-10 and ITU Rec. H.264 international video coding standards** has been highlighted. It is worth to notice, that the idea of proposed non-linear depth representation is a subject of pending patent by Poznań University of Technology in Poland [105] and in USA [98].

Chapter 6. A new 3D video coding technology

In 2011 ISO/IEC MPEG group has issued a "Call for Proposals on 3D Video Coding Technology" (CfP) [129]. The aim of the CfP, was to challenge the available state-of-the-art 3D compression technologies and to reveal the best one as a starting point for further works. Research centers and companies which has shown interest include: Microsoft, Samsung, Sony, LG, Qualcomm, Orange Labs, Nokia, Ericsson, Disney Research, Fraunhofer Institute for Telecommunications - Heinrich Hertz Institute, or universities like Aachen University (Germany), Nagoya University (Japan), National Institute of Information and Communications Technology (NICT, Japan), Gwangju Institute of Science and Technology (GIST, China), Zhejiang University of Science and Technology (China), Electronics and Telecommunications Research Institute (ETRI, Korea) and also Poznan University of Technology - Chair of Multimedia Telecommunications and Microelectronics (Politechnika Poznańska, Katedra Telekomunikacji Multimedialnej i Mikroelektroniki - KTMiM, Poland).

This Chapter describes **compression technology proposed by Poznań University of Technology** that has been presented in the response to the call of MPEG. The novelty of this proposal consists in new coding tools and in the selection of the tools resulting from extensive experiments. The technology has been described in MPEG document [66] and in conference contributions [11][12][9]. Some of the tools have been already described in the conference papers [14][10][282][283][284]. In paper [1] a detailed description has been provided.

The author of this dissertation was one of the members of the team that has developed the proposal from Poznan University of Technology. As in-depth description of the proposed codec has been shown in works mentioned above, **in this dissertation, only a brief description is provided**, with focus on the **original achievements of the author in that area**.

6.1. Comparison with other state-of-the-art codecs

In the "Call for Proposals on 3D Video Coding Technology" [129] two categories have been defined - AVC-compatible and HEVC-compatible. The most of the major MPEG group participants has provided their contributions - in total there were 12 contributions in AVC category and 11 in HEVC category. All of the proposals have been assessed both subjectively and objectively by independent research laboratories [275][276]. Results that has been shown in the end of 2011 revealed that there were three spectacular winners: Nokia Research Center [277] (winner in AVC category), Heinrich Hertz Institute (HHI) [278][279] in cooperation with Disney Research Labs [280] (winner in HEVC category) and Poznan University of Technology - Chair of Multimedia Telecommunications and Microelectronics ("Politechnika Poznańska, Katedra Telekomunikacji Multimedialnej i Mikroelektroniki" - KTMiM, Poland) – co-winner in HEVC category.

The proposal was rated **very high,** getting a **second place** just after technology provided by Fraunhofer Institute for Telecommunications - Heinrich Hertz Institute (HHI - in few configuration variants) - Fig. 74. Other participants of the competition (that ranked worse) decided not to reveal their exact identity, which remains hidden under "Pxx" code-names.

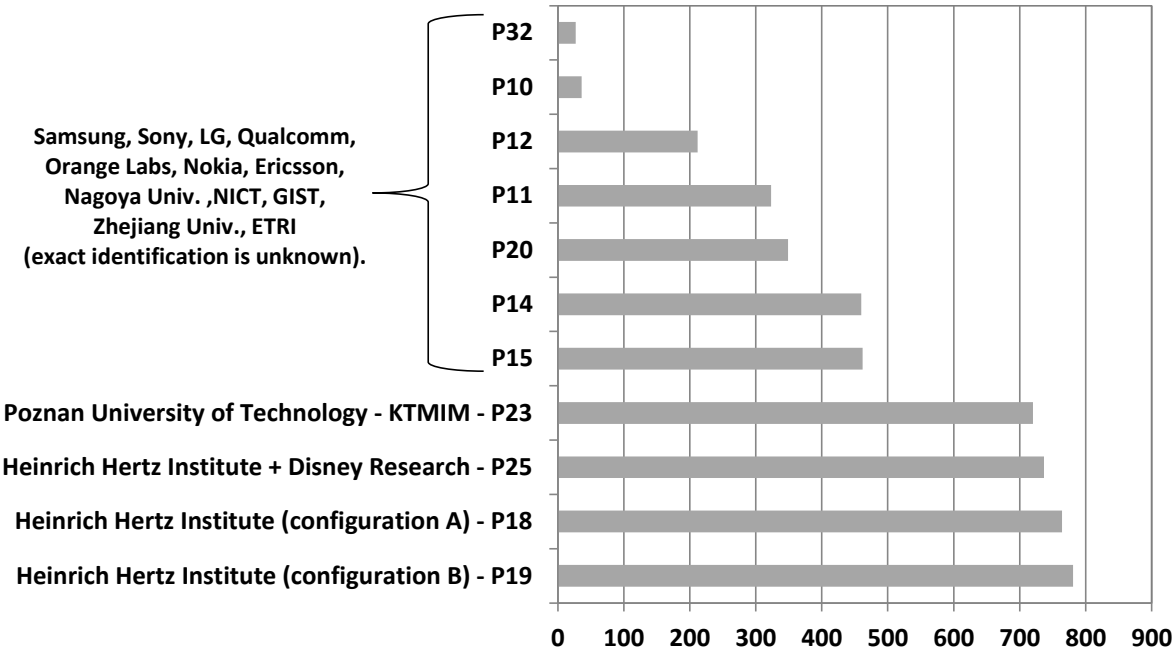


Fig. 74. Outranking of the CfP [129] proposals from various participants. Other participants of the competition that ranked worse did not revealed their identity which remains hidden under "Pxx" code-names.

In brief [276][281], all of the best performing proposals of the competition use very similar approaches. Differences in tools are summarized in Table 19. Those can be categorized into the following categories:

- **Base coding technology** - the technology that is used for backward compatible coding of a single base view. This is AVC or HEVC.
- **Disparity compensated prediction** - a tool which is used for prediction of dependent views with the reference to the base view, similar to well-known motion-compensated prediction. Typically, a toolset similar to MVC has been used.
- **Base view coding order** - order in which components (video and depth) of the base view are coded, which also constitutes the possible coding dependencies between them. Noticeably, in all of the proposals, video is coded before the depth.
- **Dependent view coding order** - order in which components (video and depth) of subsequent dependent views are coded, which also constitutes the possible coding dependencies between them (e.g. video-coded-first or depth-coded-first)
- **Depth image resolution** - spatial resolution in which depth component is coded. Full resolution means, that the depth is coded with the same spatial resolution as the base video. Reduced resolution means that the coded depth is decimated i.e. 2×.
- **Depth coding tools** - enhanced set of tools dedicated for coding of the depth.
- **View-synthesis prediction** - usage of view-synthesis for provision of additional prediction mechanisms.
- **Inter-view filtering** - filtering tool that processes video or depth data in inter-view domain. Those include tools that refine videos or depths with use of information exchange between the views.
- **Noise modeling** - tools that allow representation of a noise in the video as a separate layer.
- **Prediction between components of MVD** – existence of prediction mechanisms that operate between the component types - e.g. the depth is predicted from the video or the video is predicted from the depth.
- **Camera arrangement constraints** - constraints which are imposed on spatial 3D positions of the coded views. In many cases, only linear alignment of cameras along a straight-line (Fig. 12b) is allowed.

Table 19. Comparison of similar coding tools used in the winning technologies of the "Call for Proposals on 3D Video Coding Technology" [129].

Tool \ Proponent	**Nokia Research Center [277]**	**Poznan University of Technology - KTMiM [66]**	**Heinrich Hertz Institute (HHI) [278][279]**	**HHI + Disney Research Labs [280]**
Base coding technology	AVC	HEVC	HEVC	HEVC
Disparity compensated prediction	MVC	MVC-like	MVC-like	MVC-like
Base view coding order	Video, then Depth	Video, then Depth	Video, then Depth	No depth, Video only. Instead, so called "warps" are coded and the depth is derived at the decoder.
Dependent view coding order	Depth, then Video	Depth, then Video	Video, then Depth	
Depth image resolution	Reduced	Full	Full	
Depth coding tools	-	Non-linear Depth Representation	Wedgelets	
View-synthesis prediction	Prediction mode	Prediction mode, motion prediction and disocclusion coding	Disparity vector derivation	-
Inter-view filtering	Joint-view filter	Unified depth representation	-	-
Noise modeling	-	Spectral and spatial noise modeling	-	-
Prediction between components of MVD	-	Video-QP adjustment basing on Depth	Depth-map generation, Motion field inherence	-
Camera arrangement constraints	Linear arrangement only	Not constrained	Linear arrangement only	Linear arrangement only

In Table 19 it can be noticed that the promising results of proposal from Poznan University of Technology results from thoroughgoing selection of coding tools, comparable with those existing in other proposals. Also, it can be noticed that some of the tools are present solely in the proposal from Poznan – e.g. non-linear depth representation, noise modeling or not constrained camera arrangement. Some of those tools will be described in the following Sections of the dissertation.

6.2. The structure of the proposed 3D video codec

The proposed codec is compliant with the requirements that were defined by MPEG in the Call for Proposals (CfP) [129] [66] for HEVC-compatible category. These requirements resulted from studies of potential applications. In particular, one of the views – called the base view – is coded in compatibility with HEVC syntax, which allows extraction of a base view by a legacy decoder. The remaining views are called the side views. These side views and all depth data are coded with the use of new proposed tools [1].

Fig. 75 presents such scheme in example of coding of MVD data, composed of 3 video streams and 3 depth streams.

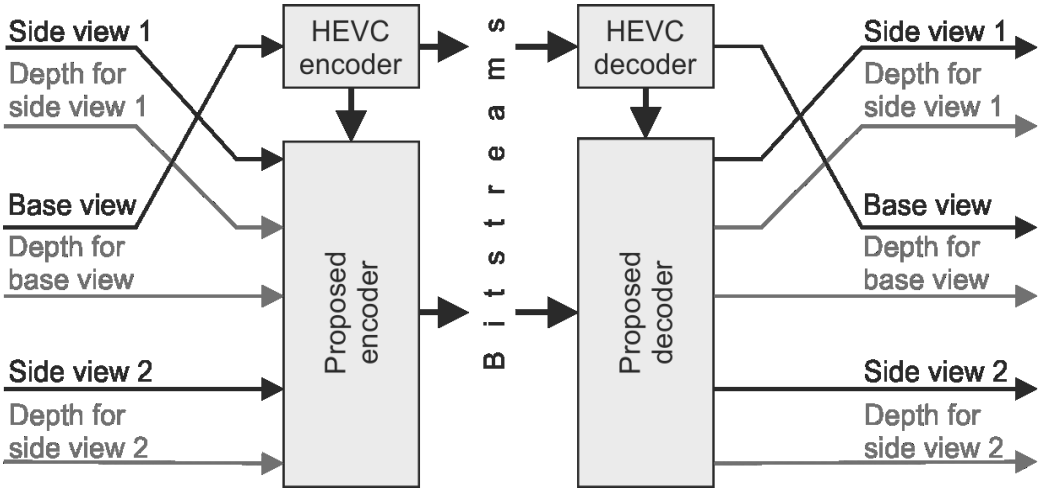


Fig. 75. Overall structure of the proposed 3D video codec, showing compatibility with HEVC syntax for the base view.

It may be pointed out that encoding and decoding of the side views and depth maps exploit information from the already coded views, which are used as references. Such hierarchical view coding structure, similar to Multiview Video Coding (MVC) [113], is used for both video and depth.

In fact, **only contents of the base view are coded as a whole**. In the side views, only a very small part of the image is transmitted at all. The remaining, not transmitted parts are reconstructed basing on the contents of the base view, with use of virtual view synthesis DIBR (Depth Image Based Rendering) technique. Such approach is called Disoccluded Region coding [1][12].

All of the proposed tools are integrated with the MVC structure and basic low-level HEVC compression tools like intra-frame prediction, inter-frame motion-compensated prediction, transform coding, in-loop filtering and others.

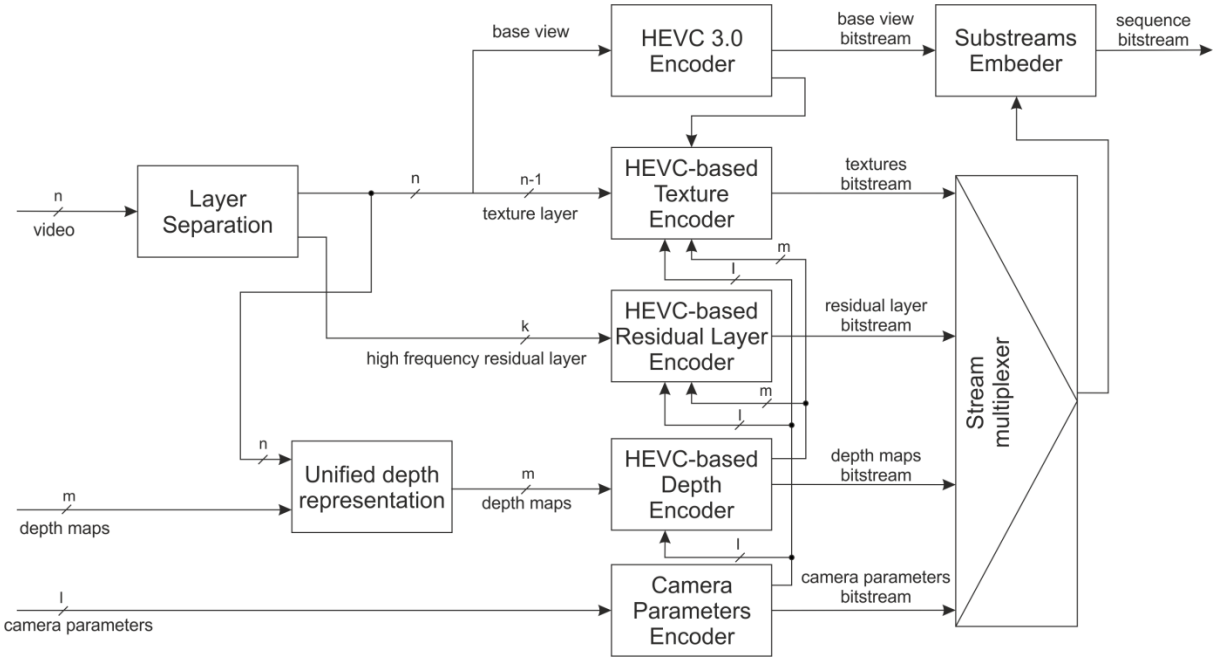


Fig. 76. The proposed 3D video encoder structure.

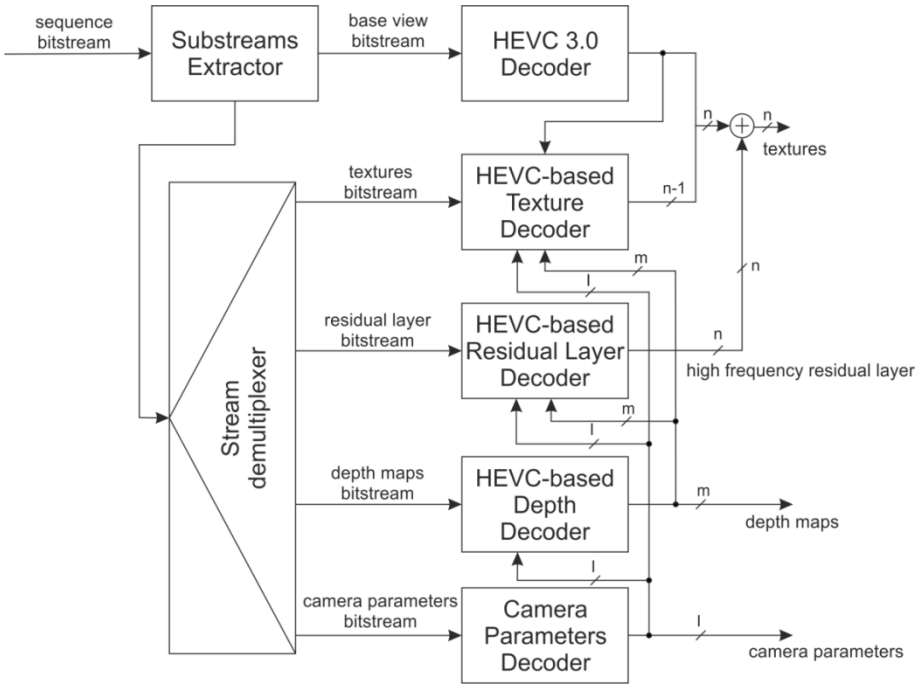


Fig. 77. The proposed 3D video decoder structure.

The detailed structure of the proposed coder and the decoder has been visualized in Fig. 76 and Fig. 77, respectively. These figures present a more general look on the codec, where arbitrary number of video streams, depth streams and other, can be transmitted.

The specific tools that are used in the 3D video codec are described below, with the achievements of the author of the dissertation highlighted.

Detailed explanation of the tools used in the codec can be found in [1][2][11][12].

6.3. Author’s contribution in the proposal of the new 3D video codec

The author of the dissertation has substantially contributed to the development of the 3D video codec in the following ways:

- The author contributed in formulation of the **overall structure of the codec**.
- The author had decisive voice **in the selection of the tools** that have been finally included in the finally proposd codec.
- The author co-developed the coding approach based on **Layer separation** by proposing the original idea and providing noise reduction algorithm (MCNRR, described in Subsection 3.8.2 of the dissertation).
- The author had participated in the proposal of **Unified Depth Representation** tool, among others by incorporating the Mid-Level Hypothesis algorithm, described in Subsection 3.7 of the dissertation.
- The author proposed and implemented **Non-linear Depth Representation** tool that already has been described in Chapter 5 of the dissertation.

6.3.1. Layer separation

The proposed 3D video coding technology use an approach, similar to Scalable Video Coding (SVC) or to wavelet coding, in which input video is spitted into layers (Layer Separation - Fig. 76) in the spatial frequency domain. Each layer presents different level of details and all layers represent the input video.

In the case of our proposal, the input video is split into two layers:

- low-frequency texture video **layer** (similar to base layer in SVC), which contains content that can be efficiently coded with classic predictive coding.
- high-frequency **residual layer**, which contains high frequency residual content that can be represented jointly for several views.

Both layers are transmitted to the decoder and after decoding are summed together in order to produce reconstructed video.

The separation of layers occurs at the very beginning of the processing as a result of noise reduction technique, already described in the dissertation (Subsection 3.8.2 of the dissertation) under the name of Motion-Compensated Noise Reduction with Refinement (MCNRR).

The process yields low-frequency texture layer which is fed to video texture encoder.

The high-frequency layer is modeled as a non-stationary random process. There are two components of the model that need to be encoded (Fig. 78): spatial energy distribution (SDE) and spectral envelope. The spatial energy distribution is estimated for each frame. For this purpose, a frame from the high-frequency layer is divided into rectangular non-overlapping blocks. In each of those blocks energy is measured. Energy values, associated with respective blocks, constitute a frame of spatial energy distribution, whose resolution is smaller than resolution of the input video.

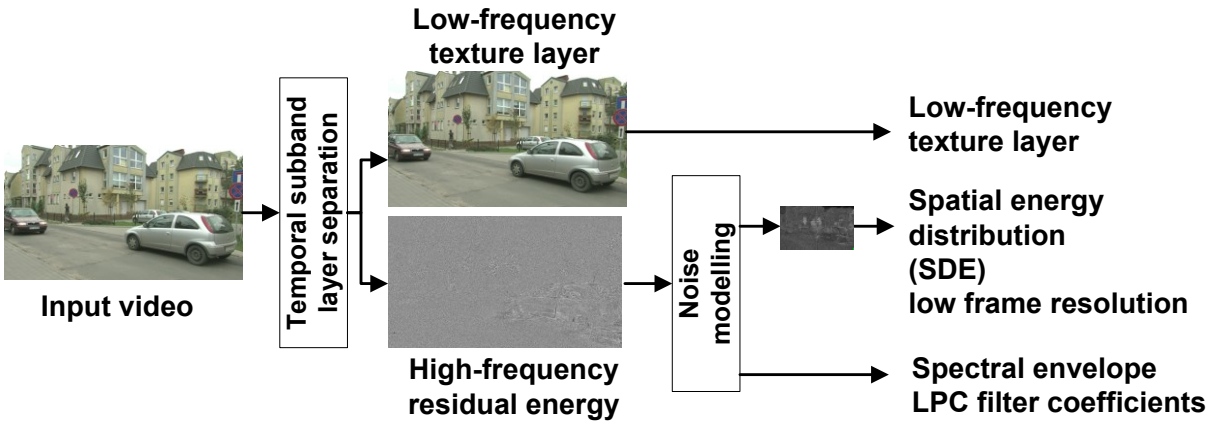


Fig. 78. High-Frequency Residual Layer Representation in the encoder.

This estimated spatial distribution of energy is used in order to normalize the high-frequency residual layer.

The second coded component of high-frequency residual layer is spectral envelope. It is estimated from energy-normalized high-frequency subband using a technique similar to LPC. The resulting set of separable IIR filter coefficients (in horizontal and vertical direction) is encoded using LAR coefficients (log-area-ratio [285]) with 8-bit representation. A set of such filter coefficients is estimated for each frame and transmitted to the decoder.

Parameters of the noise model are highly correlated among the views. The frames of the spatial distribution of energy of all views are mapped through view synthesis to a position of the base view, and then averaged. This operation results in only one joint spatial distribution

of energy (SDE). Similarly, the energy envelopes of all of the views are averaged, resulting in one joint spectral envelope.

In a decoder, pseudo-random white noise is generated and then modulated by the upsampled spatial energy distribution transmitted in the bitstream and then filtered with IIR filters that reflect the envelope of the original high-frequency layer spectrum. The resultant video, which resembles the original high-frequency subband, is added to the reconstructed low-frequency layer in order to restore the high-frequency components.

6.3.2. Unified Depth Representation

Another tools used in the encoder is Unified depth representation (Fig. 76). It is used to inter-change the depth information between the views.

As mentioned before, the idea behind the proposed technology is that only the base view (video and depth) is coded directly as a whole. In side views only the disoccluded regions are coded, while the remaining parts are reconstructed from the available views using DIBR technique. In such an approach, the amount of depth information in side-views is considerably reduced. Unfortunately, if the view synthesis algorithm uses inconsistent depth maps, it renders very annoying artifacts in the synthesized video. Because the amount of coded depth data is limited, it is necessary to adjust the input set of depth maps in such a way, that the single depth map related to the base view contains as much information as possible.

For that reason, the first step of the proposed 3D video compression algorithm is the depth map inter-view consistency refinement that produces Unified Depth Representation (named Consistent Depth Representation in [1]).

The refinement technique employs Mid-Level Hypothesis, described in the Subsection 3.7 and in [14]. This algorithm increases precision and accuracy of the artificially estimated depth maps and enhances alignment between the depth and the corresponding texture.

6.3.3. Non-linear Depth Representation

The 3D video codec from Poznan University of Technology has included the non-linear depth representation tool proposed by the author in a form that has been presented in Section 5.2 of the dissertation.

The results presented there (Fig. 65 - Fig. 68 on page 123) and also in [271][272] show that sole application of non-linear depth transformation described by equation leads to improvement of the subjectively assessed quality of the reconstructed video by about 1 MOS

point (Mean Objective Score). It can be noticed that this also correspond to about one third of the overall subjective gains provided by the whole codec (with all coding tools) over HEVC simulcast coding of multiview test data.

6.4. Experimental results for the new 3D video codec

The proposed 3D video coding technology has been submitted to "Call for Proposals on 3D Video Coding Technology" issued [129] by ISO/IEC MPEG. All of the proposals have been assessed both subjectively and objectively by independent research laboratories. The results of this evaluation have been shown in Fig. 74 at the beginning of Chapter 6.

In this subsection, experimental results that allow to estimate the overall compression performance of the technology, as well as to estimate the efficiency of the individual tools, are presented. This is done using subjective and objective tests of decoded video quality. The methodology resembles the one used by MPEG. A set of 8 MultiView plus Depth (MVD) test sequences has been used (Table 1 on page 24, Table 2 on page 26). Four of them were in 1920×1080 (Full-HD) resolution (Poznan Hall 2, Poznan Street, GT Fly and Undo Dancer). Other four test MVD sequences were in the 1024×768 (XGA) format. For the sake of brevity, results will be presented for the 1920×1080 sequences only.

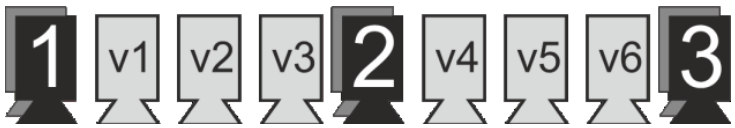


Fig. 79. The arrangement of the views: the views being coded are marked in black, while the views “v” being synthesized in the receiver are marked in gray.

For the experiments, 3 views from each sequence (video and depth) have been encoded at four different bitrates. Then the sequences have been decoded, and based on the decoded data six virtual views have been synthesized. These six virtual views (Fig. 79 – "v1", ... ,"v6") have been uniformly placed between the original views (Fig. 79– "1", ... ,"3") selected for coding (Table 2 on page 26). Similarly, six virtual views at the same spatial positions (Fig. 79 – "v1", ... ,"v6") from the uncompressed data have been synthesized in order to provide a reference for assessment. PSNR values (luminance) have been calculated (Table 20, Table 21, Fig. 80 left) and basing on them, average bitrate reductions versus HEVC simulcast were calculated using the Bjøntegaard formula [127]. It can be noted, that synthetic references were used instead of the video captured by real cameras because our aim was to assess the quality

degradation caused by the coding technology, not caused by the view synthesis algorithm itself.

For the view synthesis of the virtual views at positions "v1" to "v6" (Fig. 79), algorithm implemented in ISO/IEC MPEG Synthesis Reference Software [124] [125] has been used, with the default configuration.

In all cases, original (not pre-preprocessed) sequences have been used as references for quality measurement – both objective (PSNR) and subjective (MOS).

The subjective tests have been carried out [271][272] in accordance with the general rules of ITU Recommendation BT.500 [128]. A total number of 62 young persons were viewing each stereo pair (composed from virtual views "v3" and "v4", see Fig. 12) on a 46’’ Hyundai S465D polarization monitor. The Double Stimulus Method was selected for the subjective quality assessment that followed the rules used by the MPEG for evaluation of the proposals for the 3D video coding technology in 2011 [66].

In our experiments, the number of subjects involved was higher than in the official MPEG evaluation. The high number of subjects yielded that 95% confidence intervals were very small i.e. of order of ±(0.1÷0.25). Therefore, those intervals were not depicted on the plots (Fig. 80 right).

Fig. 80 in the left column, shows objective evaluation results (PSNR versus bitrate – BD-rates - in Table 20) and Fig. 80 in the right column, show subjective evaluation results (11-point MOS versus bitrate - BD-rates – Table 21), for virtual synthesized views for all four tested Full-HD sequences.

Please note that both subjective and objective quality assessments lead to somewhat similar conclusions. Application of Non-linear Depth Representation (Table 20 and Table 21 – column A) may result in more than 20% bitrate reduction.

View-synthesis inter-view prediction combined with MVC-toolset (Table 20 and Table 21 - column C) yields about 50–60% bitrate reduction. Disoccluded Region Coding implemented in a standard HEVC without the MVC toolset provides similar bitrate reductions (column B) of about 45%. The discrepancy between the results obtained by subjective and objective video quality assessment is the most noticeable in the case of Poznan Hall 2 sequence (S01) which probably results from low quality of associated depth maps.

The application of all devised tools except the Joint High-Frequency Layer Representation results in 50% or more of bitrate reduction (Table 20 and Table 21 - column D). When High-Frequency Layer coding is used, we can consider two cases: column E - when the high frequency layer is modeled but finally not reconstructed, which is more reliable for objective

evaluation (PSNR comparison with synthetic noise would be irrelevant) - the gains here are about 50% (objective quality measures) and over 60% (subjective quality assessment).

Table 20. Average bitrate reductions calculated as Bjøntegaard rates for luminance PSNR [dB] versus original (not preprocessed) sequences.

Test sequence	Average (over bitrates and sequences) Bjøntegaard rates [%] versus HEVC simulcast (negative numbers are bitrate savings)					
	A.	B.	C.	D.	E.	F.
	HEVC + Nonlinear Depth Representation	HEVC + Disoccluded Region Coding	MV HEVC + Disoccluded Region Coding	All but Joint High-Freq.Layer Repr. switched off	All but high-freq. layer not added	Proposed codec with all tools
Poznan Hall 2	-19.6	-20.3	-26.1	-14.7	-16.9	-23.7
Poznan Street	-27.2	-55.7	-56.8	-58.0	-62.8	-59.8
Undo Dancer	-29.1	-57.0	-58.0	-60.9	-61.1	-60.7
GT Fly	-23.2	-48.8	-49.4	-54.0	-55.4	-53.7
Average	**-24.8**	**-45.4**	**-47.6**	**-49.1**	**-49.1**	**-49.5**

Table 21. Average bitrate reductions calculated as Bjøntegaard rates for Mean Opinion Score (MOS) versus original (not preprocessed) sequences.

Test sequence	Average (over bitrates and sequences) Bjøntegaard rates [%] versus HEVC simulcast (negative numbers are bitrate savings)				
	A.	C.	D.	E.	F.
	HEVC + Non-linear Depth Representation	MV HEVC + Disoccluded Region Coding	All but Joint High-Freq.Layer Repr. switched off	All but high-freq. layer not added	Proposed codec with all tools
Poznan Hall 2	-24.5	-65.2	-67.2	-69.4	-70.1
Poznan Street	-35.7	-67.5	-72.2	-72.6	-74.8
Undo Dancer	-8.0	-52.3	-57.4	-61.4	-62.7
GT Fly	-29.6	-62.0	-69.0	-68.8	-67.2
Average	**-24.5**	**-61.7**	**-66.4**	**-68.1**	**-68.7**

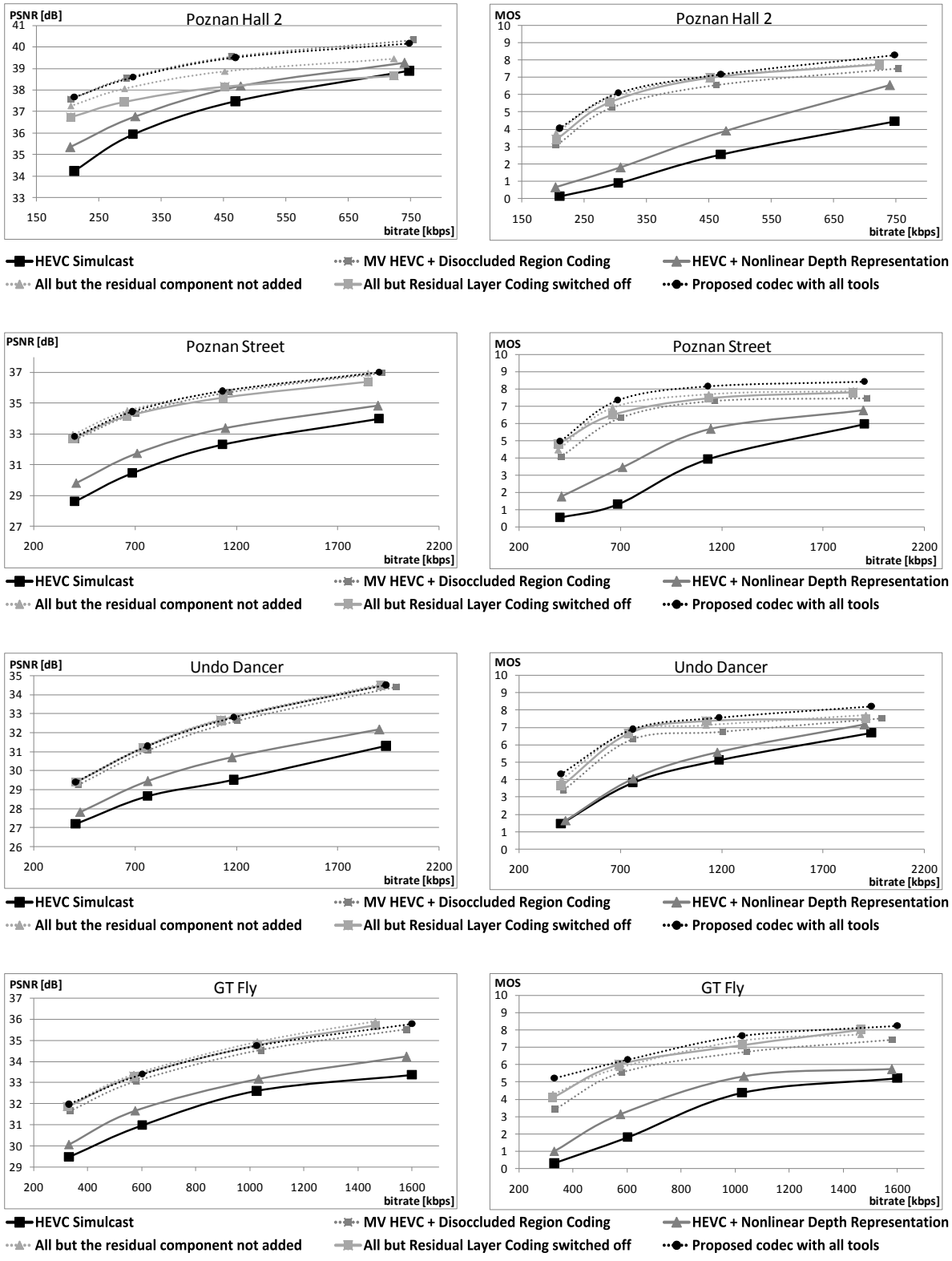


Fig. 80. Objective (left) and subjective (right) evaluation results for Full-HD test sequences.

In an another case - column F in Table 20 and Table 21- the synthetic noise is additionally summed with the output video (more important for subjective viewing), which improves the overall gains by about 1 percent. Such gain is not much, but also adequate to the bitrate cost of High-Frequency Layer (Fig. 80).

Therefore, the main gains of the whole proposal come from: Disoccluded Region Coding, Nonlinear Depth Representation and disparity-compensated prediction (MVC-like). It is also worth noting that the remaining tools together contribute a substantial gain of approximately 2÷4% of the overall bitrate as compared to simulcast HEVC (Table 20 and Table 21).

6.5. Summary of the achievements related to the new 3D video codec

The author of the dissertation has substantially contributed to the development of the presented 3D video code, by formulation of the overall structure of the codec. Also the author had decisive voice in the selection of the tools that have been finally included in the finally proposd codec. Author has proposed several coding tools, which include Layer separation, Unified Depth Representation and Non-linear Depth Representation. The author also have participated in implementation of the codec.

The proposed compression technology provides bitrate reduction of the order of 60% as compared to HEVC simulcast. This figure was obtained by systematic subjective tests. It proves high compression performance of the proposed technology that allows very efficient coding of the side views. The bitrate needed for even two side views with the corresponding depth maps is mostly below 50% of the bitrate for single-view video.

Chapter 7. Summary of the dissertation

In the Chapter 1.2 four theses T1-T4 have been as formulated. In the dissertation all of them have been considered and thoughtfully studied. The achievements and results for theses T1-T3 are shown in Chapter 3. The achievements and results for thesis T4 are shown in Chapter 5. **All of the theses of the dissertation have been proven** which is summarized below, in Section 7.1. Further, in Section 7.2 the **overview of the works** performed by the author during the development of the dissertation is given. Section 7.3 briefly reminds **the main achievements of the dissertation.**

7.1. Achievements related to theses

Ad. T1) Depth estimation can be improved by usage of modeling of the cost function based on maximization of a posteriori probability.

In the dissertation a complete probabilistic model for $FitCost$ function (composed of $DataCost$ and $TransitionCost$) has been proposed.

In Sections 3.1 to 3.6 a theoretical formulation based on Maximum A posteriori Probability (MAP) rule has been used to develop an improved depth estimation technique. This part of the dissertation has been started with general theoretical derivation of $\boldsymbol{DataCost}$ formulation based on MAP (Subsection 3.1). Then, the derived formula (21) (shown on page 51) has been thoughtfully analyzed with respect to simplification (Subsection 3.2) to classical forms related with to SSD or SAD (equations (30) and (35) presented on pages 53 and 54, respectively) along with verification of the conditions that have to be meet for such simplification. For that, analysis of the noise existing in 3D multiview video sequences has been performed which yielded with very vital results. It has been shown that at least some of the conditions are not meet in a practical case of multiview test sequences (Subsection 3.3) and basing on that a novel formulation for $DataCost$ model has been proposed (Subsection 3.4). A method for estimation of parameters of this model has been shown on an example of the test sequences. Next, a probabilistic model for $\boldsymbol{TransitionCost}$ has been proposed (in Subsection 3.5) also with a method for estimation of parameters of this model. In the end of the considerations, the results have been shown.

The attained results show average gain of **about 0.08dB to 2.8dB,** calculated with respect to PSNR of virtual views, synthesized with use of depth maps generated with the proposed

method, over the reference. As reference, original unmodified Depth Estimation Reference Software has been used with manual calibration of Smoothing Coefficient per sequence. For the case of selection of the worst checked (yet rational) Smoothing Coefficient value for the original DERS, the gain is about 2.8dB of PSNR, averaged over all of the tested sequences For the case of selection of the best found Smoothing Coefficient in original DERS software, the average gain is only about 0.08dB of PSNR, but it can be noted that the proposed technique attained that **without manual calibration of such coefficient**.

This constitutes **one of the biggest advantages of the proposed depth estimation method** – it does not require arbitrary manual calibration of parameters like Smoothing Coefficient. All required model parameters can be algorithmically estimated like it was shown on the example of the tested sequences in Subsection 3.5.

Ad. T2) Precision and accuracy of estimated depth maps can be improved in post-processing with iterative insertion of intermediate values, controlled using view synthesis quality.

In the dissertation a depth refinement algorithm has been presented which provides sub-pixel precision and accuracy to the estimated depth map.

In Section 3.7 a novel depth refinement by Mid-level Hypothesis (MLH) technique has been proposed. The devised method provides a solution for generating sub-pixel precise depth maps, without necessity to increase complexity of the core depth estimation algorithm - in the proposal, the depth in refined in a post-processing step. In MLH algorithms, hypothesized is intermediate depth value in the places where depth edges are identified. Such hypotheses are then verified basing on quality of virtual view synthesized with use of the considered depth map.

The presented results show that the quarter-pixel-precise depth maps, generated with use of the proposed technique, **provide gains of about 0.3dB of PSNR** related to the quarter-pixel-precise depth maps generated with the original unmodified Depth Estimation Reference Software (DERS). As for complexity, it has been shown that the proposed technique can provide **reduction of factor about 3.7× of computation time** in such case. Comparing to full-pixel-precise depth maps generated with the original unmodified DERS, **the gains are even higher and are about 2dB of PSNR**. All of the PSNR gains have been measured as quality of virtual view synthesized generated with use of the given depth map.

Ad. T3) Temporal consistency of estimated depths can be improved using noise removal from input multiview video.

In Subsection 3.8 developments of the author in area of depth estimation related to enhancement of the temporal consistency of stereoscopic depth maps have been shown. A novel approach is proposed in which temporal consistency of the estimated depth is increased by application of noise reduction technique in the input multiview video, a priori to the depth estimation itself. Two noise reduction techniques have been developed by the author in order to provide proof of the presented concept - Still Background Noise Reduction (SBNR) and Motion-Compensated Noise Reduction with Refinement (MCNRR). Although the developed noise reduction techniques are fairly simple, they have provided evidence that the proposed approach brings substantial gains.

In results it has been shown that visual quality of the input video has been improved. Further, results of experiment with depth estimation have been shown. It has been noticed that the proposal does not provide gains in terms of objective PSNR metric. On the other hand, the subjective evaluation, of the views synthesized basing on the depth maps generated with use of the proposed noise reduction technique, shown that application of the proposal provides **subjective gains of about 0.7 to 1.2 MOS** points (Mean Opinion Score). Finally, temporal consistency of generated depth maps has been verified with use of two methods. First, the correlation for pairs of subsequent depth frames has been analysed. It is shown that Pearson Linear correlation coefficient has been improved by about 1% in average. Secondly, temporal consistency of the depth has been measured in a case of usage in a 3D-video system, where the depth is also compressed and transmitted. In such case, it is shown that the compression performance has been substantially improved, and the application of the proposed methods give gains up to **about 30% of bitrate reduction (**Bjøntegaard metric, average over the tested sequence set, retaining the same synthesized views quality) related to the compression of depth maps estimated with the original, unmodified, stand-alone MPEG Depth Estimation Reference Software (DERS).

Ad. T4) Non-linear representation of depth can be employed in order to improve compression efficiency of depth maps in 3D video systems.

In Chapter 5, a novel non-linear transformation has been proposed for representation and coding of the depth. First a proof-of-concept proposal has been presented, with use of

a simple non-linear function, based on the idea of gamma-correction, well known from luminance coding in classical video systems. The verification of this preliminary proposal has been performed in with use of a codec developed by Poznan University of Technology, Chair of Multimedia Telecommunications and Microelectronics. This codec has been submitted to "Call for Proposals on 3D Video Coding Technology" (CfP), issued by MPEG group [129] in 2011, and has been rated very high among other proposals. Promising results the proposed coded in the resolution of CfP and also further attained results of evaluation, created a motivation for more advanced proposal. The devised, theoretical approach yielded with a novel formulation of non-linear transformation for depth representation.

The final proposal has been evaluated experimentally with use of coding technology implemented in MPEG Reference Software for AVC-based 3D video coding standards – MVC+D [114][115][116] and AVC-3D [117][118][119]. It has been shown that application of the proposed non-linear depth representation provides substantial subjective gains (0.3 to 1.0 MOS point, depending on the case) and considerable bitrate reduction – **up to about 25%** bitrate reduction of the depth component of the bitstream (in average over the tested sequences that employed the usage of the proposed tool).

Finally, **adoption of the proposed non-linear depth representation to new 3D extensions of ISO/IEC 14496-10 and ITU Rec. H.264 international video coding standards** has been highlighted. It is worth to notice, that the idea of proposed non-linear depth representation is a subject of **pending patents** by Poznań University of Technology in Poland [105] and in USA [98].

7.2. Overview of the performed works

In order to perform a fair and reliable assessment of the achievement of the dissertation, the author has done series of experiments with the test video sequences. For that sake, the author has implemented and embedded the proposed techniques of depth estimation and coding into existing software packages:

- ISO/IEC MPEG Depth Estimation Reference Software (DERS) for assessment of techniques related to depth estimation,
- ISO/IEC MPEG reference software for new AVC-based 3D coding standards,
- Poznan University of Technology 3D video codec implementing 3D-coding technology based on HEVC [121] compression.

The ISO/IEC MPEG Depth Estimation and View Synthesis Reference Software [124][126] contains about 15 000 lines of program code written in C++ programming language. Lines of program code, modified added by the author (including the proposed techniques) are about 1000.

The 3D-ATM reference software [120] contains about 230 000 lines of program code written in C++ programming language. Lines of program code, modified or added by the author (including all proposed techniques) are about 5 000.

The Poznan University of Technology 3D-video codec software [66] , based on ISO/IEC MPEG reference software for HEVC coding standard contains about 115 000 lines of C++ source code, from which about 60 000 has been added in development of the tools added by the team of Poznan University of Technology. Lines of code modified or added by the author (including the proposed NDR technique and other tools) are about 10 000.

Apart from that, for the sake of experimentation the author has creates a set of video and signal processing utilities, libraries and scripts that in total contain about 20 000 lines of C++ code and about 5 000 lines of code in Python language.

In total, the author has been working with about 385 000 lines of code from which about 41 000 lines of code has been originally created by the author.

The experiments, that the author has performed, consumed substantial amounts of computational costs. The experiments in area of noise analysis took equivalent of about 10 days of continuous computations of a single-threaded computer.

The experiments in area of depth estimation took equivalent of about of 5 000 days of continuous computations of a single-threaded computer.

The experiments in area of depth coding took equivalent of about of 1 600 days of continuous computations of a single-threaded computer.

In total, computations performed for experimentation in the dissertation consumed about **6 610 days** of continuous computations of a single-threaded computer.

Performing of such enormous amount of computation was possible only by employment of a multi-processor computational cluster which allowed for parallelization of works by a factor of about 72 (six work stations, 12 threads each). It is worth to notice, that a higher parallelization level was not possible due to memory limitations – each depth estimation pass consumes up to 2GB of memory and each coding pass consumes up to 1GB of memory.

Even with availability of computational cluster, the amount of preparations needed for launching such a wide range of experiments (e.g. partitioning of the work over machines and super-visioning of the computations) were a substantial part of the dissertation.

7.3. The primary original achievements of the dissertation

1. **Non-linear Depth Representation (NDR),** which can be used to improve compression efficiency of depth maps in 3D video systems. The proposal has been adopted to new 3D extensions of **ISO/IEC 14496 10 and ITU Rec. H.264 international video coding standards.** It is shown that application of NDR in those new coding technologies, known under names of MVC+D and AVC-3D, provides substantial subjective gains of about **0.3 to 1.0 MOS** point (depending on the case in the test set) and considerable bitrate reduction for depth component – up to about **24.52%** in the test sequences for which the tool has been employed.
2. A method for **generation of temporally consistent depth maps using noise removal from video.** It has been shown that with use of simple noise reduction techniques proposed in the dissertation (Still Background Noise Reduction – SBNR and Motion-Compensated Noise Reduction with Refinement – MCNRR) temporal consistency of estimated depth can be improved. The measured improvement in evaluation of the depth map quality by subjective virtual view synthesis-based methodology is **0.7 to 1.2 MOS** points. The measured improvement in compression performance of depth maps, expressing increase of temporal consistency is about **30% in average**.
3. **Depth refinement by Mid-Level Hypothesis (MLH)** algorithm, which provides sub-pixel precision and accuracy to the estimated depth map in post-processing. The advantages of MLH algorithm can be seen from two alternative perspectives. Firstly, MLH can be seen as precision and accuracy refinement technique that brings quality gains of **about 2dB of PSNR** related to input pixel-precise depth maps and 0.28dB related to input quarter-pixel precise depth maps. Secondly, MLH can be seen as speed optimization technique which allows for factor of about **3.7**× computational cost reduction, related to direct application of quarter-pixel precise depth map estimation.
4. **Participation in development of a new 3D video codec**, prepared at Chair of Multimedia Telecommunication and Microelectronics of Poznan University of Technology. The author had decisive voice in the selection of tools that have been incorporated in the codec. In particular, the employed tools include Mid-Level-Hypothesis and Non-linear Depth Representation, mentioned above. The proposed 3D codec has achieved a spectacular success in "Call for Proposals for 3D Video Coding Technology" issued by ISO/IEC MPEG group, showing gains of about 50%-60% of bitrate reduction, related to simulcast HEVC coding.

7.4. The secondary original achievements of the dissertation

1. A proposal of **depth estimation method** which employs empirical modeling of the cost function based **on maximization of a posteriori probability (MAP) rule**. The proposed methods allows for **unsupervised depth estimation** without need for usage of arbitrary settings or control parameters (like Smoothing Coefficient in Depth Estimation Reference Software) while giving comparable quality of generated depth maps to the case when supervised depth estimation is used, and such parameters are manually optimized. In the case, when sub-optimal settings of control parameters in supervised depth estimation with DERS is used, the proposed method provides gains **of about 2.8dB** measured in average PSNR quality of virtual views synthesized with generated depth maps in the tested sequence set.
2. The abovementioned algorithm has been presented along with a method for estimation of parameters of the model of cost function basing on empirical data. For this, **noise existing in multiview video** and **inter-view correlation** of corresponding samples **have been thoughtfully studied**.
3. The author of the dissertation had strong influence on production of **multiview test sequences**, developed at Chair of Multimedia Telecommunication and Microelectronics of Poznan, Poznan University of Technology. This participation relates to both the content of the sequences and the technical and scientific sides of their preparation. These sequences have been adopted to a multiview test sequence set maintained by ISO/IEC MPEG group, broadly used in experiments related to development of 3D-related technologies around the world.

7.5. Conclusions and future works

To conclude, the dissertation presents a thoughtful presentation of the works performed by the author during his recent scientific career. The presented achievements include:

- Improvement of depth estimation, expressed by depth precision, accuracy, temporal consistency and reduction of computational costs.
- Improvement of depth coding efficiency.
- Adoption of the proposed Non-linear Depth Representation tool to new 3D extensions of international video coding standards (ISO/IEC and ITU).

Moreover, the attained results have shown directions for future work. In particular, those include:

- Works on adaptation of Non-linear Depth Representation (NDR) technique to new 3D video coding technologies being currently developed, based on HEVC. Ultimately, adoption of NDR tool proposal to upcoming international standards (ISO/IEC and ITU) would be desired.
- Analysis of influence of noise removal from video on temporal consistency of generated depth maps. Experiments with more advanced noise reduction techniques, than simple examples shown in the dissertation, would substantially extend the scientific knowledge in that field.
- Further development of formulations of fit cost function of optimization algorithms used in depth estimation, based on Maximum A posteriori Probability rule. Basing on the empirical data, already gathered in the dissertation, even more efficient models can be proposed.

Some of the above mentioned ideas are already being investigated by the author. The results are expected to be attained in the upcoming months and years.

Bibliography

Author’s contributions

Papers in international journals and volumes

[1] M. Domański, O. Stankiewicz, K. Wegner, M. Kurc, J. Konieczny, J. Siast, J Stankowski, R. Ratajczak, T. Grajek, „High Efficiency 3D Video Coding Using New Tools Based on View Synthesis”, IEEE Transactions on Image Processing, Vol.22, No.9, pages 3517-3527, 2013.

[2] M. Domański, K. Klimaszewski, O. Stankiewicz, J. Stankowski, K. Wegner, „Efficient transmission of 3d video using MPEG-4 AVC/H.254 compression technology”, in S. Zeadally, E. Cerqueira, M. Curado, M. Leszczuk (eds.): Computer Communication Networks and Telecommunications, Lecture Notes in Computer Science, Springer-Verlag, Vol. 6157, pages 145-156, 2010.

[3] K. Wegner, O. Stankiewicz, „Generation of temporally consistent depth maps using noise removal from video”, in L. Bolc, R. Tadeusiewicz, and L.J. Chmielewski (eds.): Computer Vision and Graphics, Lecture Notes in Computer Science, Springer-Verlag, Vol. 6375, pages 292-299, 2010.

[4] M. Domański, K. Klimaszewski, J. Konieczny, M. Kurc, A. Łuczak, O. Stankiewicz, K. Wegner, “An experimental free-view television system”, Image Processing and Communications Challenges, R. Choraś, A. Zabłudowski (eds.), Academy Publishing House EXIT Warsaw, pages 169-176, 2009.

International competitions report

[5] O. Stankiewicz, J. Chmiel, M. Tłuczek, W. Świtała, “ReadIT: A portable text reading system for the blind people” - final report, 2nd place in World, IEEE Computing Society International Design Competition 2005, Washington D.C., July 2005.

Articles in materials of reputable worldwide conferences

[6] O. Stankiewicz, K. Wegner, M. Domański, „Nonlinear Depth Representation for 3D Video Coding”, IEEE International Conference on Image Processing (ICIP), Melbourne, Australia, Sept, 2013.

[7] M. Domański, T. Grajek, D. Karwowski, K. Klimaszewski, J. Konieczny, M. Kurc, A. Łuczak, R. Ratajczak, J. Siast, O. Stankiewicz, J. Stankowski, K. Wegner, „New coding technology for 3d video with depth maps as proposed for standardization within MPEG”, 19th International Conference on Systems, Signals and Image Processing (IWSSIP), Vienna, Austria, 11-13 April 2012.

[8] S. Cancino-Suarez, K. Klimaszewski, O. Stankiewicz, M. Domański, „Enhancement of stereoscopic depth estimation by the use of motion information", 44th Southeastern Symposium on System Theory (SSST), pages 94-98, Jacksonville, 2012.

[9] M. Domański, O. Stankiewicz, K. Wegner et al. "New Coding Technology for 3D Video within Depth Maps as Proposed for Standardization within MPEG", International Conference on Systems, Signals and Image Processing (IWSSIP 2012), pages 415-418, Vienna, Austria, April, 2012.

[10] M. Kurc, O. Stankiewicz, M. Domański "Depth map inter-view consistency refinement for multiview video", Picture Coding Symposium, Kraków, Poland, 2012.

[11] M. Domański, T. Grajek, D. Karwowski, K. Klimaszewski, J. Konieczny, M. Kurc, A. Łuczak, R. Ratajczak, J. Siast, O. Stankiewicz, J. Stankowski, K. Wegner, „Coding of multiple video+depth using HEVC technology and reduced representations of side views and depth maps", 29th Picture Coding Symposium PSC2012, Kraków, Poland, 2012.

[12] M. Domański, J. Konieczny, M. Kurc, R. Ratajczak, J. Siast, O. Stankiewicz, J. Stankowski, K. Wegner, „3D video compression by coding of disoccluded regions", 2012 IEEE International Conference on Image Processing (ICIP), Orlando, USA, 30 Sept. – 3 Oct. 2012.

[13] J. Stankowski, M. Domański, O. Stankiewicz, J. Konieczny, J. Siast, K. Wegner, „Extensions of the HEVC technology for efficient multiview video coding", 2012 IEEE International Conference on Image Processing (ICIP), Orlando, USA, 30 Sept, – 3 Oct., 2012.

[14] O. Stankiewicz, M. Domański, K. Wegner, „Stereoscopic depth refinement by mid-level hypothesis", IEEE International Conference on Multimedia & Expo, Singapore, July 2010.

[15] K. Wegner, O. Stankiewicz M. Domański, „Stereoscopic depth estimation using fuzzy segment matching", 28th Picture Coding Symposium (PCS2010), Nagoya, Japan, 8-10 Dec. 2010.

[16] O. Stankiewicz, K. Wegner, M. Domański, „Error concealment for MVC and 3D video coding", 28th Picture Coding Symposium (PCS2010), Nagoya, Japan, 8-10 Dec. 2010.

[17] K. Wegner, O. Stankiewicz, „Similiarity measures for depth estimation", 3DTV-Conference 2009 The True Vision Capture, Transmission and Display of 3D Video, Potsdam, Germany, 4-6 May 2009.

[18] O. Stankiewicz, K. Wegner, „A hybrid technique for stereoscopic depth estimation in video", International Conference on Signals and Electronic Systems (ICSES), Kraków, Sept. 2008.

[19] O. Stankiewicz , A. Łuczak, „Flexible processor architecture optimized for advanced coding algorithms", Picture Coding Symposium 2007, Lisboa, Portugal, 7-9 Nov. 2007.

[20] O. Stankiewicz, A. Łuczak, Antoni Roszak, „Temporal noise reduction for preprocessing of video streams in monitoring systems”, International Conference on Signals and Electronic Systems (ICSES'06), pages. 231-234, Łódź, Poland, 18-20 Sept. 2006,

Articles in polish journals

[21] O. Stankiewicz, K. Wegner, „System telewizji stereowizyjnej z wyznaczaniem mapy głębi”, Przegląd Telekomunikacyjny 4/2008, pages 213 - 216, April 2008.

[22] O. Stankiewicz, K. Wegner, „„Analiza dokładności syntezy obrazu w systemach telewizji wielowidokowej”, Przegląd Telekomunikacyjny 6/2009, pages 376 - 379, June 2009.

[23] K. Wegner, K. Klimaszewski, O. Stankiewicz, J. Stankowski, M. Domański, „Przygotowanie wielowidokowych sekwencji wizyjnych dla badań nad telewizją trójwymiarową”, Przegląd telekomunikacyjny 6/2010, pages 304-308, June 2010.

[24] M. Domański, T. Grajek, D. Karwowski, K. Klimaszewski, R. Ratajczak, J. Konieczny, M. Kurc, A. Łuczak, R. Ratajczak, J. Siast, O. Stankiewicz, J. Stankowski, K. Wegner, „Poznański kodek obrazów trójwymiarowych”, Przegląd Telekomunikacyjny, Vol. 82, No. 2-3, pages 81-83, 2013.

Articles published as documents of ISO/IEC Moving Pictures Experts group of International Organization for Standardization and ITU-T/ISO/IEC Joint Collaborative Team on 3D Video Coding Extension Development

[25] K. Wegner, O. Stankiewicz, M. Tanimoto, M. Domański, „Enhanced Depth Estimation Reference Software (DERS) for Free-viewpoint Television”, ISO/IEC JTC1/SC29/WG11 Doc. M31518, Geneva, Switzerland, October 2013.

[26] D. Rusanovskyy, O. Stankiewicz, D. Tian, J. Y. Lee, J. Lin, „JCT-3V AHG Report: 3D-AVC Software Integration (AHG4)", Joint Collaborative Team on 3D Video Coding Extensions of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, Doc. JCT3V-F0004, m31107, 6th Meeting, Geneva, Switzerland, 25 Oct. to 1 Nov. 2013.

[27] K. Wegner, O. Stankiewicz, M. Tanimoto, M. Domański, „Enhanced View Synthesis Reference Software (VSRS) for Free-viewpoint Television”, ISO/IEC JTC1/SC29/WG11 Doc. M31520 October 2013, Geneva, Switzerland.

[28] D. Rusanovskyy, O. Stankiewicz, D. Tian, J. Ying Lee, J. Lin, „JCT-3V AHG Report: 3D-AVC Software Integration (AHG4)", Joint Collaborative Team on 3D Video Coding Extensions of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, Doc. JCT3V-E0004, m30147, 5th Meeting, Vienna, Austria, 27 July – 2 Aug. 2013.

[29] O. Stankiewicz, K. Wegner, M. Domański, „Optimized QP/QD curve for 3D coding with half and full resolution depth maps”, Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 Doc. JTC3V-E0269, 5th Meeting: Wiedeń, Austria, 27 July – 2 August 2013.

[30] D. Rusanovskyy, O. Stankiewicz, D. Tian, J. Ying Lee, J. Lin, „JCT-3V AHG Report: 3D-AVC Software Integration (AHG4)", Joint Collaborative Team on 3D Video Coding Extensions of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, Doc. JCT3V-D0004, m29043, 4th Meeting, Incheon, South Korea, 20–26 April 2013.

[31] K. Wegner, O. Stankiewicz, M. Domański, „AHG14: Comparison of half resolution depth map coding versus full resolution depth map coding in 3D-ATM”, Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 Doc. JTC3V-D0080, 4th Meeting: Incheon, South Korea, 20–26 April 2013.

[32] O. Stankiewicz, K. Wegner, “3D-CE7.a Cross check of Improved Nonlinear Depth Representation by Poznan University of Technology”, Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 Doc. JCT3V-C0166, m27929, 3rd Meeting: Geneva, Switzerland Geneva, Switzerland, 17–23 Jan. 2013.

[33] O. Stankiewicz, K. Wegner, F.-C. Chen, C.-C. Lin, D. Rusanovskyy, „AHG12: Recommendation on MVC+D reference software”, Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 Doc. JTC3V-C0164, Geneva, Switzerland, 17–23 January 2013.

[34] D. Rusanovskyy, Y. Lee, J. Lin, D. Tian, O. Stankiewicz, „AHG4: 3D-AVC Software Integration (AHG4)", Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, Doc. JCT3V-C0004, m27931, 3rd Meeting, Geneva, Switzerland, 17–23 Jan. 2013.

[35] O. Stankiewicz, K. Wegner “AHG12: Report on results of cross-check on coding with JMVC+Depth software in interlace mode”, Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 Doc. JCT3V-C0036, m27759, 3rd Meeting: Geneva, Switzerland Geneva, Switzerland, 17–23 Jan. 2013.

[36] T. Grajek, O. Stankiewicz, K. Wegner, „AHG9: Correlation analysis between MOS data collected on stereoscopic and autostereoscopic displays”, Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 Doc. JTC3V-C0202, Geneva, Switzerland, 17–23 January 2013.

[37] S. Shimizu, O. Stankiewicz, S. Sugimoto, H. Kimata, K. Wegner, M. Domański, „3D-HEVC HLS on depth definition”, Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 Doc. JTC3V-B0164, 2nd Meeting: Shanghai, China, 13–19 October 2012.

[38] D. Rusanovskyy, Y. Lee, J. Lin, D.g Tian, O. Stankiewicz, „AHG4: 3D-AVC Software Integration (AHG4)", Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, Doc.JCT3V-B0004, m27128, 2nd Meeting, Geneva, Shanghai, China, 13–19 Oct. 2012.

[39] K. Wegner, O. Stankiewicz J. Siast, M. Domański, „Independent intra-period coding in HEVC”, Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 Doc. JTCVC-K0332 11th Meeting: Shanghai, China, 10–19 October 2012.

[40] O. Stankiewicz, K. Wegner, M. Domański, „3D-HEVC with reduced resolution of depth”, Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 Doc. JTC3V-B0183, 2nd Meeting: Shanghai, China, 13–19 October 2012.

[41] O. Stankiewicz, K. Wegner, „Cross-Check of Improved nonlinear depth representation (3D-ATM)”, Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 Doc. JTC3V-B0186, 2nd Meeting: Shanghai, China, 13–19 October 2012.

[42] O. Stankiewicz, K. Wegner, „AHG5 Cross-Check of MV-HEVC Software for HTM”, Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 Doc. JTC3V-B0181, 2nd Meeting: Shanghai, China, 13–19 October 2012.

[43] O. Stankiewicz, K. Wegner, M. Domański, „Impact of View Synthesis Optimization (VSO) on depth quality”, Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 Doc. JTC2-A0090, 1st Meeting: Stockholm, Sweden, 16–20 July 2012.

[44] K. Wegner, O. Stankiewicz, „Additional Cross Check of 3D-CE7.h Results on Global Depth and View Prediction of NICT by Poznan”, Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 Doc. JTC2-A0168, 1st Meeting: Stockholm, Sweden, 16–20 July 2012.

[45] J. Siast, O. Stankiewicz, K. Wegner, M. Domański, „Independent intra-period coding in 3D-HTM”, Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 Doc. JTC2-A0091, 1st Meeting: Stockholm, Sweden, 16–20 July 2012.

[46] O. Stankiewicz “CE2 summary report: depth representation and coding”, Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, Doc. JCT3V-A0089, m26045, 1st Meeting, Stockholm, Sweden, 16–20 July 2012.

[47] K. Wegner, O. Stankiewicz, J. Siast, „3D-CE1h results on Depth Map Disocclusion Coding by Poznan University of Technology”, ISO/IEC JTC1/SC29/WG11 Doc. M25014, Geneva, Switzerland, May 2012.

[48] K. Wegner, O. Stankiewicz, J. Siast, „3D-CE2a cross check of Samsung proposal on Adaptive Depth Quantization by Poznan University of Technology", ISO/IEC JTC1/SC29/WG11 Doc. M25018, Geneva, Switzerland, May 2012.

[49] K. Wegner, O. Stankiewicz, J. Siast, „3D-CE2a results on Nonlinear Depth Representation", ISO/IEC JTC1/SC29/WG11 Doc. M25019, Geneva, Switzerland, May 2012.

[50] O. Stankiewicz, „3D-CE2 summary report: Nonlinear Depth Representation & Coding", ISO/IEC JTC1/SC29/WG11 Doc. M25016, Geneva, Switzerland, May 2012.

[51] O. Stankiewicz, „3D-CE5a summary report: motion/mode parameter prediction", ISO/IEC JTC1/SC29/WG11 Doc. M25015, Geneva, Switzerland, May 2012.

[52] K. Wegner, O. Stankiewicz, J. Siast, „3D-CE2h results on Nonlinear Depth Representation", ISO/IEC JTC1/SC29/WG11 Doc. M25020, Geneva, Switzerland, May 2012.

[53] O. Stankiewicz "3D-CE2 summary report: Nonlinear Depth Representation & Coding", ISO/IEC JTC1/SC29/WG11 Doc. M25016., Geneva, Switzerland, May 2012.

[54] K. Wegner, O. Stankiewicz, J. Siast, „3D-CE2h cross check of Samsung proposal on Adaptive Depth Quantization by Poznan University of Technology", ISO/IEC JTC1/SC29/WG11 Doc. M25021, Geneva, Switzerland, May 2012.

[55] K. Wegner, O. Stankiewicz, J. Siast, „3D-CE2h results on Adaptive Depth Quantization combined with Nonlinear Depth Representation", ISO/IEC JTC1/SC29/WG11 Doc. M25022, Geneva, Switzerland, May 2012.

[56] K. Wegner, O. Stankiewicz, J. Siast, „3D-CE1h cross check of RWTH University proposal on Warping Based Prediction by Poznan University of Technology", ISO/IEC JTC1/SC29/WG11 Doc. M25187, Geneva, Switzerland, May 2012.

[57] K. Wegner, O. Stankiewicz, „Results of subjective evaluation in 3DV-CE2", ISO/IEC JTC1/SC29/WG11 Doc. M25244, Geneva, Switzerland, May 2012.

[58] O. Stankiewicz, K. Wegner, M. Kurc, „3D-AVC-CE3 results on Nonlinear Depth Representation and Coding", ISO/IEC JTC1/SC29/WG11, Doc. M23788, San Jose, USA, 6-10 Feb. 2012.

[59] K. Wegner, J. Siast, J. Konieczny, O. Stankiewicz, M. Domański, „Poznan University of Technology tools for 3DV coding integrated into 3D-HTM", ISO/IEC JTC1/SC29/WG11 Doc. M23783, San Jose, USA, February 2012.

[60] O. Stankiewicz, K. Wegner, M. Domański, „Nonlinear depth representation – extended results", ISO/IEC JTC1/SC29/WG11 Doc. M23791, San Jose, USA, February 2012.

[61] K. Wegner, J. Siast, J. Konieczny, O. Stankiewicz, M. Domański, „Poznan University of Technology tools for 3DV coding integrated into 3D-HTM", ISO/IEC JTC1/SC29/WG11 Doc. M23783, San Jose, USA, February 2012.

[62] O. Stankiewicz, K. Wegner, M. Kurc, „3D-AVC-CE3 cross-check of adaptive depth quantization", ISO/IEC JTC1/SC29/WG11 Doc. M23790, San Jose, USA, February 2012.

[63] O. Stankiewicz, K. Wegner, M. Kurc, „3D-AVC-CE3 results on Nonlinear Depth Representation & Coding", ISO/IEC JTC1/SC29/WG11 Doc. M23788, San Jose, USA, February 2012.

[64] O. Stankiewicz, „3D-AVC-CE3 summary report: Depth Representation & Coding", ISO/IEC JTC1/SC29/WG11 Doc. m23785, San Jose, USA, February 2012.

[65] O. Stankiewicz, „3D-AVC-CE2 summary report: Depth-based prediction", ISO/IEC JTC1/SC29/WG11 Doc. m23787, San Jose, USA, February 2012.

[66] M. Domański, T. Grajek, D. Karwowski, K. Klimaszewski, J. Konieczny, M. Kurc, A. Łuczak, R. Ratajczak, J. Siast, O. Stankiewicz, J. Stankowski, K. Wegner, „Technical Desciption of Poznan University of Technology proposal for Call on 3D Video Coding Technology", ISO/IEC JTC1/SC29/WG11, Doc. M22697, Geneva, Switzerland, November 2011.

[67] M. Domański, T. Grajek, D. Karwowski, K. Klimaszewski, J. Konieczny, M. Kurc, A. Łuczak, R. Ratajczak, J. Siast, O. Stankiewicz, J. Stankowski, K. Wegner, „Multiview HEVC – experimental results", Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, Doc. M22147, JCTVC-G582, Geneva, Switzerland, November 2011.

[68] K. Wegner, O. Stankiewicz, „Improved depth estimation with advanced occlusion handling", ISO/IEC JTC1/SC29/WG11 Doc. M21365, Torino, Italy, July 2011.

[69] K. Wegner, O. Stankiewicz, M. Domański, „3DV/FTV EE4 report on Poznan Street and Poznan CarPark sequences", ISO/IEC JTC1/SC29/WG11 Doc. M20192, Geneva, Switzerland, March 2011.

[70] K. Wegner, O. Stankiewicz, M. Domański "3DV/FTV EE4 report on Poznan Street and Poznan CarPark sequences", ISO/IEC JTC1/SC29/WG11 Doc. M20192, Geneva, Switzerland, March 2011.

[71] K. Wegner, O. Stankiewicz, M. Domański, „3DV/FTV EE1 report on Poznan Carpark sequence – improved depth for extended range of frames", ISO/IEC JTC1/SC29/WG11 Doc. M19391, Daegu, South Korea, January 2011.

[72] K. Klimaszewski, K. Wegner, O. Stankiewicz, M. Domański, „3DV/FTV EE4 report on Poznan Street sequence", ISO/IEC JTC1/SC29/WG11 Doc. M19392, Daegu, South Korea, January 2011.

[73] K. Wegner, O. Stankiewicz, M. Domański, „3DV/FTV EE1 report on Poznan Carpark sequence - improved depth for extended range of frames", ISO/IEC JTC1/SC29/WG11 Doc. M19391, Daegu, South Korea, January 2011.

[74] K. Wegner, O. Stankiewicz, „Frame range extension of Poznan Street and Poznan Carpark sequences (3DV/EE1)", ISO/IEC JTC1/SC29/WG11 Doc. M18506, Guangzhou, China, October 2010.

[75] K. Wegner, O. Stankiewicz, M. Domański, „Input on Call for Proposals on 3D Video Coding document", ISO/IEC JTC1/SC29/WG11 Doc. M17900, Geneva, Switzerland, July 2010.

[76] K. Wegner, O. Stankiewicz, K. Klimaszewski, M. Domański, „Comparison of multiview compression performance using MPEG-4 MVC and prospective HVC technology", ISO/IEC JTC1/SC29/WG11 Doc. M17913, Geneva, Switzerland, July 2010.

[77] K. Wegner, O. Stankiewicz, M. Domański, „Proposal of changes to Applications and Requirements on 3D Video Coding", ISO/IEC JTC1/SC29/WG11 Doc. M17899, Geneva, Switzerland, July 2010.

[78] O. Stankiewicz, K. Wegner, „Results of 3DV/FTV Exploration Experiment 1 (EE1) for Poznan sequences", ISO/IEC JTC1/SC29/WG11 Doc. M17613, Dresden, Germany, April 2010.

[79] O. Stankiewicz, K. Wegner, M. Domański, „Estimation of temporally consistent depth maps using noise removal from video", ISO/IEC JTC1/SC29/WG11 Doc. M17612, Dresden, Germany, April 2010.

[80] O. Stankiewicz, K. Wegner, K. Klimaszewski, „Newspaper sequence - Results of 3DV/FTV Exploration Experiments with depths and view synthesis", ISO/IEC JTC1/SC29/WG11 Doc. M17175, Kyoto, Japan, January 2010.

[81] J. Stankowski, K. Klimaszewski, O. Stankiewicz, K. Wegner, M. Domański, „Preprocessing methods used for Poznan 3D/FTV test sequences", ISO/IEC JTC1/SC29/WG11 Doc. M17174, Kyoto, Japan, January 2010.

[82] O. Stankiewicz, K. Wegner, M. Domański, „First version of depth maps for Poznan 3D/FTV test sequences", ISO/IEC JTC1/SC29/WG11 Doc. M17176, Kyoto, Japan, January 2010.

[83] O. Stankiewicz, K. Wegner, K. Klimaszewski, „Newspaper sequence - Results of 3DV/FTV Exploration Experiments with depths and view synthesis", ISO/IEC JTC1/SC29/WG11 Doc. M17051, Xian, China, October 2009.

[84] O. Stankiewicz, K. Wegner, M. Wildeboer, „A soft – segmentation matching in Depth Estimation Reference Software (DERS) 5.0", ISO/IEC JTC1/SC29/WG11 Doc. M17049, Xian, China, October 2009.

[85] M. Domański, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, K. Wegner, „Poznań Multiview Video Test Sequences and Camera Parameters", ISO/IEC JTC1/SC29/WG11 Doc. M17050, Xian, China, October 2009.

[86] K. Wegner, O. Stankiewicz, „An enhancement of depth estimation reference software with use of soft-segmentation", ISO/IEC JTC1/SC29/WG11 Doc. M16757, London, England, July 2009.

[87] O. Stankiewicz, K. Wegner, „Results of exploration experiments in 3D video coding, described in w10552, for 'newspaper' sequence", ISO/IEC JTC1/SC29/WG11 Doc. M16756, London, England, July 2009.

[88] O. Stankiewicz, K. Wegner, K. Klimaszewski, „Results of exploration experiments in 3D video voding, described in w10360, for alt moabit sequence”, ISO/IEC JTC1/SC29/WG11 Doc. M16328, Maui, USA, April 2009.

[89] O. Stankiewicz, K. Wegner, K. Klimaszewski, „Additional results of exploration experiments in 3D video coding, described in w10360, for alt moabit sequence”, ISO/IEC JTC1/SC29/WG11 Doc. M16460, Maui, USA, April 2009.

[90] K. Wegner, O. Stankiewicz, „Analysis of sub-pixel precision in depth estimation reference software and view synthesis reference software”, ISO/IEC JTC1/SC29/WG11 Doc. M16027, Lausanne, Switzerland, February 2009.

[91] O. Stankiewicz, K. Wegner, „Application of middle level hypothesis algorithm for improvement of depth maps produced by depth estimation reference software”, ISO/IEC JTC1/SC29/WG11 Doc. M16028, Lausanne, Switzerland, February 2009.

[92] O. Stankiewicz, K. Wegner, K. Klimaszewski, Results of 3DV/FTV exploration experiments, described in w10173, for alt moabit sequence”, ISO/IEC JTC1/SC29/WG11 Doc. M16026, Lausanne, Switzerland, February 2009.

[93] K. Wegner, O. Stankiewicz, „3DV/FTV EE1 and EE2 results on alt moabit sequence”, ISO/IEC JTC1/SC29/WG11 Doc. M15832, Busan, South Korea, Oct. 2008.

[94] O. Stankiewicz, K. Wegner, „Depth Map estimation software version 3”, ISO/IEC JTC1/SC29/WG11 Doc. M15540, Hannover, Germany, July 2008.

[95] O. Stankiewicz, K. Wegner, „Depth map estimation software version 2", ISO/IEC JTC1/SC29/WG11 Doc. M15338, Archamps, France, April 2008.

[96] O. Stankiewicz, K. Wegner, „Depth map estimation software", ISO/IEC JTC1/SC29/WG11 Doc. M15175, Antalya, Turkey, January 2008.

Submissions to U.S. Patent Office

[97] M. Domański, K. Klimaszewski, J. Konieczny, M. Kurc, R. Ratajczak, J. Siast, O. Stankiewicz, J. Stankowski, K. Wegner, "The image encoding method”, submission no. 13680652.

[98] M. Domański, T. Grajek, J. Konieczny, M. Kurc, A. Łuczak, J. Siast, O. Stankiewicz, J. Stankowski, K. Wegner, „Method for coding of stereoscopic depth”, submission no. 13680822.

[99] M. Domański, J. Konieczny, M. Kurc, R. Ratajczak, J. Siast, O. Stankiewicz, J. Stankowski, K. Wegner, „Method for predicting the shape of an encoded area based on depth map”, submission no. 13680740.

Submissions to Polish Patent Office

[100] O. Stankiewicz, K. Wegner, M. Domański, „Sposób wyznaczania modeli przestrzennych z wykorzystaniem redukcji szumów w wejściowych sekwencjach wizyjnych”, submission no. P-392496.

[101] M. Domański, J. Konieczny, M. Kurc, R. Ratajczak, J. Siast, O. Stankiewicz, J. Stankowski, K. Wegner, „Sposób przetwarzania obrazu zsyntezowanego”, submission no. P.397012.

[102] M. Domański, T. Grajek, J. Konieczny, M. Kurc, J. Siast, O. Stankiewicz, J. Stankowski, K. Wegner, „Sposób alokacji prędkości bitowej przy kodowaniu sekwencji wielowidokowych z informacją przestrzenną”, submission no. P.397014.

[103] M. Domański, T. Grajek, J. Konieczny, M. Kurc, A. Łuczak, J. Siast, O. Stankiewicz, J. Stankowski, K. Wegner, „Sposób wyznaczania parametrów kwantyzacji sterujących dekwantyzatorem sygnału resztkowego obrazu z wykorzystaniem mapy głębi odpowiadającej dekodowanemu obrazowi”, submission no. P.397013.

[104] M. Domański, J. Konieczny, M. Kurc, R. Ratajczak, J. Siast, O. Stankiewicz, J. Stankowski, K. Wegner, „Sposób predykcji kształtu obszaru kodowanego z wykorzystaniem map głębi”, submission no. P.397010.

[105] M. Domański, T. Grajek, J. Konieczny, M. Kurc, A. Łuczak, J. Siast, O. Stankiewicz, J. Stankowski, K. Wegner, „Sposób kodowania głębi stereoskopowej”, submission no. P.397016.

[106] M. Domański, T. Grajek, J. Konieczny, M. Kurc, J. Siast, O. Stankiewicz, J. Stankowski, K. Wegner, „Zastosowanie cech sygnału różnicowego mapy głębi do sterowania koderem sekwencji wizyjnych z informacją przestrzenną”, submission no. P.397015.

[107] M. Domański, T. Grajek, K. Klimaszewski, J. Konieczny, M. Kurc, J. Siast, O Stankiewicz, J. Stankowski, K. Wegner, „Sposób międzyobrazowej predykcji mapy odległości”, submission no. P.397011.

[108] M. Domański, T. Grajek, D. Karwowski, J. Konieczny, M. Kurc, J. Siast, O. Stankiewicz, J. Stankowski, K. Wegner, „Sposób kodowania parametrów kamer”, submission no. P.397009.

[109] M. Domański, T. Grajek, K. Klimaszewski, J. Konieczny, M. Kurc, J. Siast, O. Stankiewicz, J. Stankowski, K. Wegner, „Sposób przetwarzania map rozbieżności”, submission no. P.397017.

[110] M. Domański, K. Klimaszewski, J. Konieczny, M. Kurc, R. Ratajczak, J. Siast, O. Stankiewicz, J. Stankowski, K. Wegner, "Sposób kodowania obrazu”, submission no. P.397008.

Other references

(not co-authored by Olgierd Stankiewicz)

[111] ISO/IEC 14496-10, Int. Standard “Generic coding of audio-visual objects – Part 10: Advanced Video Coding”, 7th Ed., 2012, also: ITU-T Rec. H.264, Edition 7.0, 2012.

[112] Annex H “Multiview Video Coding” of ISO/IEC 14496-10, Int. Standard “Generic coding of audio-visual objects – Part 10: Advanced Video Coding”, 7th Ed., 2012, also: ITU-T Rec. H.264, Edition 8.0, 2013.

[113] A. Vetro, P. Pandit, H. Kimata, A. Smolic, and Y.-K. Wang, eds., „Joint Draft 8 of Multiview Video Coding”, Joint Video Team (JVT) Doc. JVTAB204, Hannover, Germany, July 2008.

[114] Annex I “Multiview and Depth video coding” of ISO/IEC 14496-10, Int. Standard “Generic coding of audio-visual objects – Part 10: Advanced Video Coding”, 8th Ed., 2013, also: ITU-T Rec. H.264, Edition 8.0, 2013.

[115] T. Suzuki, M. Hannuksela, Y. Chen, S. Hattori, “JCT-3 Draft Text of ISO/IEC 14496-10:2012/DAM 2 MVC extensions for inclusion of depth maps”, ISO/IEC JTC1/SC29/WG11, Doc. m26324, Stockholm, Sweden, 16–20 July 2012.

[116] T. Suzuki, M. M. Hannuksela, Y. Chen, G. Sullivan, S. Hattori, “Study Text of ISO/IEC 14496-10:2012/DAM2 MVC extension for inclusion of depth maps”, ISO/IEC JTC1/SC29/WG11, Doc. N13140, Shanghai, Oct. 2012.

[117] M. Hannuksela, Y. Chen, T. Suzuki, J.-R. Ohm, G. Sullivan, „Text of ISO/IEC 14496-10:2012/PDAM3 AVC compatible video-plus-depth extension”, ISO/IEC JTC1/SC29/WG11, Doc.N13142, Shanghai, Oct. 2012.

[118] M. M. Hannuksela, Y. Chen, T. Suzuki, J.-R. Ohm, G. Sullivan, “JCT-3V 3D-AVC Draft Text 8”, ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, F1002 Doc. JCT3V-m31703 , 6th Meeting, Geneva, Switzerland, 25 Oct. – 1 Nov. 2013.

[119] "3D-AVC Draft Text 6”, JCT-3V of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, Doc. JCT3V-D1002, Incheon, South Korea, 2013.

[120] D. Rusanovskyy, F. C. Chen, L. Zhang, T. Suzuki, „3D-AVC Test Model 5”, ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11,Doc. JCT3V-C1003, Geneva, Switzerland, Jan. 2013.

[121] ISO/IEC 23008-2:2013, Int. Standard, “MPEG H - Part 2: High Efficiency Video Coding”, also: ITU-T Rec. H.264, Edition 1.0, 2013.

[122] G. Tech, K. Wegner, Y. Chen, M. Hannuksela, J. Boyce, „MV-HEVC Draft Text 3”, JCT3V Document, JCT3V-C1004, Geneva, Switzerland, Jan. 2013.

[123] G. Tech, K. Wegner, Y. Chen, S.Yea, "3D-HEVC Test Model 4", JCT-3V of ITU-T SG 16 WP 3 and ISO/IEC JTC1/SC 29/WG 11, Doc. JCT3V-D1005, Incheon, South Korea, 2013.

[124] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, Y. Mori, "Reference softwares for depth estimation and view synthesis", ISO/IEC JTC1/SC29/WG11, Archamps, France, Tech. Rep. M15377, Apr. 2008.

[125] "View synthesis algorithm in view synthesis reference software 3.0 (VSRS3.0)", ISO/IEC JTC1/SC29/WG11 Doc. M16090 , Feb.2009.

[126] M. Tanimoto, T. Fujii, K. Suzuki, „Video Depth Estimation Reference Software (DERS) with Image Segmentation and Block Matching", ISO/IEC MPEG M16092, Lausanne, Switzerland, Feb. 2009.

[127] G. Bjontegaard, "Calculation of Average PSNR Differences between RD-curves", ITU-T SG16, Doc. VCEG-M33, April 2001.

[128] "ITU-R BT.500-12 Recommendation: Methodology for the subjective assessment of the quality of television pictures", approved Jan 2012.

[129] "Call for Proposals on 3D Video Coding Technology", ISO/IEC JTC1/SC29/WG11 Doc. N12036, Geneva, Switzerland, March 2011.

[130] A. Vetro, T. Wiegand, G.J. Sullivan, "Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC Standard", Proceedings of the IEEE, Vol. 99, No. 4,2011, pages 626-642.

[131] K. Müller, P. Merkle, T. Wiegand "3-D video representation using depth maps", Proceedings of the IEEE, Vol. 99, No. 4, pages 643-656, April 2011.

[132] F. Shao, G. Jiang, M. Yu, K. Chen, Y.-S. Ho, "Asymmetric Coding of Multi-View Video Plus Depth Based 3-D Video for View Rendering", IEEE Transactions on Multimedia, Vol. 14, pages 157- 167, 2012.

[133] P. Merkle, Y. Morvan, A. Smolic, et al., "The effects of multiview depth video compression on multiview rendering", Signal Processing: Image Communication, Vol. 24, No. 1+2, pages 73-88, January 2009.

[134] P. Merkle, A. Smolic, K. Müller, T. Wiegand, "Multi-view video plus depth representation and coding", Proceedings of IEEE International Conference on Image Processing (ICIP'07), TX, USA, Vol. I, pages 201-204, 2007.

[135] "Call for 3D Test Material: Depth Maps & Supplementary Information", ISO/IEC JTC1/SC29/WG11 MPEG/N10359, Laussane, Switzerland, Feb. 2009.

[136] T. Senoh, K. Yamamoto, R. Oi, Y. Ichihashi, T. Kurita, „Depth Estimation Experiment with Poznan Hall", ISO/IEC JTC1/SC29/WG11, Doc. M18221, Guangzhou, China, October 2010.

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

[138] D. Tian, P.-L. Lai, P. Lopez, C. Gomila, "View Synthesis Techniques for 3D Video", Apagesof Digital Image Proc. XXXII. Proc, of the SPIE, Volume 7443, 2009.

[139] A. Smolic, K. Müller, T. Wiegand, et al., "Intermediate View Interpolation Based on Multiview Video Plus Depth for Advanced 3D Video Systems", IEEE Int. Conf. on Image Processing, ICIP2008, San Diego, CA, USA, October 2008.

[140] C. Fehn, R. de la Barr´e, S. Pastoor, "Interactive 3-DTV – concepts and key technologies", Proc. IEEE, Vol. 94, No. 3, pages 524–538, Mar. 2006.

[141] C. Fehn, "Depth-image-based rendering (dibr), compression and transmission for a new approach on 3d-tv", Prceedings of SPIE, Vol. 5291A, pages 93–104, 2004.

[142] D. Scharstein, R. Szeliski, "Middlebury Stereo Vision Page", webpage http://vision.middlebury.edu/stereo/ - online 1st Dec 2013.

[143] D. Scharstein, R. Szeliski, „A taxonomy and evaluation of dense two-frame stereo correspondence algorithms", International Journal of Computer Vision 2002.

[144] Tanimoto Laboratory, Dept.of Info., Elec. Eng., Nagoya University, Japan, "MPEG-FTV web-page", http://www.tanimoto.nuee.nagoya-u.ac.jp/ - online 1st Dec 2013.

[145] S. Kemp, "BBC Hits Pause on 3D TV Ambitions", article, http://www.hollywoodreporter.com/behind-screen/bbc-hits-pause-3d-tv-581091, online 1st Dec 2013..

[146] "Applications and Requirements on 3D Video Coding", ISO/IEC JTC1/SC29/WG11, Doc. N11829, Geneva, Switzerland, March 2011.

[147] P. Kauff, K. Müller, A. Smolic, et al., "Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability", Signal Processing: Image Communication, Vol. 22, No. 2, 2007.

[148] K. Klimaszewski, K. Wegner, „Joint Intra Coding of Video and Depth Maps", IEEE Int. Conf. Signals and Electronic Systems – ICSES 2010, Gliwice, Poland, Sept. 2010.

[149] K. Klimaszewski, „Algorytmy kompresji sekwencji wielowidokowych", PhD Dissertation at Poznan University of Technology, Faculty of Electronics and Telecommunications, Supervisor: prof. dr hab. inż. M. Domański, Reviewers: dr hab. inż. A. Przelaskowski, prof. dr hab. inż. R. Stasiński, 2012.

[150] L. Ma, Y.Q. Chen, K.L. Moore, "Analytical piecewise radial distortion model for precision camera calibration", Vision, Image and Signal Processing, IEEE Proceedings, Vol.:153, Issue: 4, pages 468 – 474, August 2006.

[151] F. Devernay, Olivier Faugeras, "Straight lines have to be straight,", Machine Vision and Application, Vol. 13, No. 1, pages 14-24, 2001.

[152] Z. Zhang, „Determining the epipolar geometry and its uncertainty:A review", International Journal of Computer Vision, 27(2):161-1195, 1998.

[153] C. Loop, Z. Zhang, „Computing rectifying homographies for stereo vision", Proceedings of the 1999 Conference on Computer Vision and Pattern Recognition (CVPR '99), 1999.

[154] C. Doutre, P. Nasiopoulos, "Color correction preprocessing for multiview video coding,", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 19, No. 9, pages 1400-1405, Sep. 2009.

[155] O. Faugeras, "Three-Dimensional Computer Vision: A Geometric Viewpoint", MIT Press, ISBN: 9780262061582, 1993.

[156] R. Hartley A. Zisserman, "Multiple View Geometry in Computer Vision", Second Edition, Cambridge University Press, ISBN: 0521540518, 2004.

[157] Jie Zhao, Dong-Ming Yan, Guo-Zun Men, Ying-Kang Zhang, „A Method of Calibrating the Intrinsic and Extrinsic Camera Parameters Separately for Multi-Camera Systems", International Conference on Machine Learning and Cybernetics, Vol.:3, pages 1548 - 1553, Hong Kong, China, 19-22 Aug. 2007.

[158] Z. Zhang "A flexible camera calibration by viewing a plane from unknown orientations", Proceedings of the 7th International Conference on Computer Vision, pages 666-673, Corfu, Greece, 1999.

[159] K. Müller, P. Merkle, T. Wiegand, "3-D video representation using depth maps", Proceedings of the IEEE, Vol. 99, No. 4, pages 643-656, April 2011.

[160] David G. Lowe, "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, Vol. 60, No. 2, pages 90-110, 2004.

[161] I. Tošic, B. A. Olshausen, B. J. Culpepper, "Learning Sparse Representations of Depth", IEEE Journal on Selected Topics in Signal Processing, Vol. 5, No. 5, pages 941-952, September 2011.

[162] K. Yongtae, K. Jiyoung, S.-K.-H. Wanghoon, "Fast Disparity and Motion Estimation for Multi-view Video Coding", IEEE Transactions on Consumer Electronics, Vol. 53, No. 2, pages 712-719, 2007.

[163] Martin A. Fischler, R. C. Bolles, „Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography", Communications of the Association for Computing Machinery (ACM), Vo.l. 24, Issue 6, pages 381–395, June 1981.

[164] Liang Cheng, Hao Hu, Yecheng Wang, Manchun Li, „A new method for remote sensing image matching by integrating affine invariant feature extraction and RANSAC", 2010 3rd International Congress on Image and Signal Processing (CISP), Vol.: 4, Pages 1605 - 1609, Yantai, 2010.

[165] Nikos Georgis, Josef Kittler, M. Bober, „Accurate Recovery of Dense Depth Map for 3D Motion Based Coding.", European Transactions on Telecommunications, Vol. 11, No.2, March-April 2000.

[166] Di Wu, Huyin Zhang, Xin Li, Long Qian, „Depth Map Generation Algorithm for Multiview Video", 2013 Fifth International Conference on Computational and Information Sciences (ICCIS), Shiyang, China, 2013.

[167] T.W. Ryan, R.T. Gray, B.R. Hunt, "Prediction of Correlation Errors in Stereo-Pair Images", Optical Eng., Nol. 19, No. 3, pages 312-322, 1980.

[168] D. Marr, T.A. Poggio, "Cooperative Computation of Stereo Disparity", Science, Vol. 194, No. 4262, pages 283-287, 1976.

[169] M. Agrawal, L.S. Davis, "Window-based, discontinuity preserving stereo", Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition (CVPR 04), Washington, DC, USA, 2004.

[170] A. Fusiello, V. R.o, E. Trucco, "Efficient stereo with multiple windowing", Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), pages 858, Washington, DC, USA, 1997.

[171] D. Rzeszotarski, P. Skulimowski, P. Strumiłło, "A method for verification of dense disparity maps computed from the matching algorithm implemented in the stereovision system", Signal Processing Symposium SPS 2007, Jachranka Village May 24-26, 2007.

[172] Lee Sang Hwa Lee, S. Sharma, „Real-time disparity estimation algorithm for stereo camera systems", Consumer Electronics, IEEE Transactions on, Vol.57, Issue: 3, pages 1018-1026, August 2011.

[173] R. Zabih, J. Woodfill, "Non-Parametric Local Transforms for Computing Visual Correspondence", Proceedings of European Conference on Computer Vision, 1994.

[174] Jędrzej Kowalczuk, Eric T. Psota, Lance C. Pérez, „Real-time Temporal Stereo Matching using Iterative Adaptive Support Weights", 2013 IEEE International Conference on Electro/Information Technology (EIT), pages 1-6, Rapid City, SD, USA, 9-11 May 2013.

[175] A. Hosni, M. Bleyer, M. Gelautz, „Near Real-Time Stereo With Adaptive Support Weight Approaches", International Symposium 3D Data Processing, Visualization and Transmission (3DPVT) 2010, Paris, France, pages 1-8, 2010.

[176] K. He, J. Sun, X. Tang "Guided image filtering", Proceedings of European Conference on Computer Vision (ECCV 2010), 2010.

[177] A. Hosni, M. Bleyer, C. Rhemann, M. Gelautz "Real-time local stereo matching using guided image filtering", 2011 IEEE International Conference on Multimedia and Expo (ICME), pages 1-6, Barcelona, Spain, 11-15 July 2011.

[178] Q. Yang, P. Ji, D. Li, S. Yao, M. Zhang. Near real-time stereo matching using adaptive guided filtering. Submitted to Image and Vision Computing 2013.

[179] Huiyan Han, Xie Han, Fengbao Yang, „An improved gradient-based dense stereo correspondence algorithm using guided filter", International Journal for Light and Electron OpticsVolume 125, Issue 1 , pages 115-120, accepted for publication, 2014.

[180] T. Kanade, M. Okutomi "A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16, No. 9, pages 920-932, Sept. 1994.

[181] Y. Boykov, O. Veksler, R. Zabih "A variable window approach to early vision", IEEE Trans. Pattern Anal. Mach. Intell., Vol. 20, No. 12, pages 1283–1294, Dec. 1998.

[182] P. Skulimowski, P. Strumiłło, "Refinement of depth from stereo camera ego-motion parameters", IEEE Electronics Letters, Volume 44, Issue 12, pages 729-730, June 2008.

[183] D. Sun et al.,, „Secrets of Optical Flow Estimation and Their Principles", in IEEE International Conference on Computer Vision & Pattern Recognition, 2010.

[184] J. Sun, Y. Li, S. Kang, H. Shum "Symmetric stereo matching for occlusion handling", Proc. IEEE Conf. Computer Vision and Pattern Recognition, pages 399–406, 2005.

[185] Woo-Seok Jang, Yo-Sung Ho, „Efficient disparity map estimation using occlusion handling for various 3D multimedia applications", IEEE Transactions on Consumer Electronics, Vol. 57, No. 4, pages 1937-1943, Nov. 2011.

[186] Chao Liang, Liang Wang, Hongyun Liu, „Stereo matching with cross-based region, hierarchical belief propagation and occlusion handling", International Conference on Mechatronics and Automation (ICMA), pages.1999-2003, 7-10 Aug. 2011.

[187] R. Ben-Ari, N. Sochen "Stereo matching with Mumford-Shah regularization and occlusion handling", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.:32, Issue: 11, pages 2071-2084, Nov. 2010.

[188] W. Chen, M. Zhang, Z. Xiong "Segmentation-based stereo matching with occlusion handling via region border constrains", Computer Vision and Image Understanding, CVIU, 2009.

[189] A. Arranz, A. Sanchez-Miralles, A. de la Escalera, M. Alvar, J. Boal, "Genetic algorithm for stereo correspondence with a novel fitness function and occlusion handling", 8th International Conference on Computer Vision Theory and Applications VISAPP 2013, Barcelona, Spain, 21-24 February 2013.

[190] D.M. Greig, B.T. Porteous, A.H. Seheult, "Exact maximum a posteriori estimation for binary images", Journal of the Royal Statistical Society Series B, 51, pages 271–279, 1989.

[191] Y. Boykov, O. Veksler, R. Zabih, „Markov Random Fields with Efficient Approximations", International Conference on Computer Vision and Pattern Recognition (CVPR), 1998.

[192] Y. Boykov, O. Veksler, R. Zabih, „Fast approximate energy minimisation via graph cuts", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, pages 1222-1239, 2001.

[193] L.R. Ford Jr., D.R Fulkerson, "Maximal flow through a network:", Canad. J. Math, Vol. 8, pages 399-404. doi:10.4153/CJM-1956-045-5, 1956.

[194] P. Elias, A. Feinstein, C.E. Shannon "Note on Maximum Flow Through a Network", IRE Trans. on Information Theory, IT. 2, No. 4, pages 117-119, 1956.

[195] Yuri Boykov, Vladimir Kolmogorov "An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision", IEEE Transactions on Pattern Anal. Mach. Intell. 26(9): 1124-1137, 2004.

[196] Christos H. Papadimitriou, Kenneth Steiglitz, „The Max-Flow, Min-Cut Theorem", Combinatorial Optimization: Algorithms and Complexity. Dover. pages 120–128. ISBN 0-486-40258-4, 1998.

[197] M.F. Tappen, W.T. Freeman "Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters", IEEE International Conference on Computer Vision, 2003.

[198] S. Geman, G. Geman "Stochastic Relaxation, Gibbs Distribuition and the Bayesian Restoration of Images", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 6, pages 721-741, 1984.

[199] Li Hong, G. Chen "Segment-based stereo matching using graph cuts", Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004, Vol.:1,pages 74-82, 27 June-2 July 2004.

[200] A. Zureiki, M. Devy, R. Chatila, „Stereo Matching using Reduced-Graph Cuts", IEEE International Conference on Image Processing, 2007. ICIP 2007. Vol:1, pages 237-240, San Antonio, TX, USA, Sept. 16 2007-Oct. 19 2007.

[201] A. Klaus, M. Sormann, K. Karner "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure", International Conference on Pattern Recognition 2006, Hong Kong, China, August 2006.

[202] M.I. Jordan "Learning in Graphical Models", MIT Press, ISBN: 9780262600323, January 1999.

[203] I.J. Cox, S.L. Hingorani, S.B. Rao, B.M. Maggs "A Maximum-Likelihood Stereo Algorithm", Computer Vision and Image Understanding, Vol. 63, No. 3, pages 542-567, 1996.

[204] Jian Sun, Nan-Ning Zheng, Heung-Yeung Shum, „Stereo Matching Using Belief Propagation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, Issue: 7, pages 787 - 800, 2003.

[205] L. Cheng, T. Caelli "Bayesian Stereo Matching", Proc. Conf. Computer Vision and Pattern Recognition Workshop, pages 192-192, 2004.

[206] D. Min, K. Sohn "Cost aggregation and occlusion handling with WLS in stereo matching", IEEE Tansactions on Image Processing, Vol. 17, No. 8, pages 1431–1442, 2008.

[207] K. Tanaka, J. Inoue, D.M. Titterington, „Loopy belief propagation and probabilistic image processing", 2003 IEEE 13th Workshop on Neural Networks for Signal Processing (NNSP'03), pages 29-338, 17-19 Sept. 2003.

[208] L. E. Baum, T. Petrie, „Statistical Inference for Probabilistic Functions of Finite State Markov Chains". The Annals of Mathematical Statistics 37, Vol. 6, pages 1554–1563. doi:10.1214/aoms/1177699147, 1966.

[209] Y.Weiss, W.T. Freeman "On the optimality of solutions of the max-product belief propagation algorithm in arbitrary graphs", IEEE Transactions on Information Theory, 47(2), pages 723–735, 2001.

[210] J. Pearl, „Reverend Bayes on inference engines: A distributed hierarchical approach". Proceedings of the Second National Conference on Artificial Intelligence (AAAI-82), AAAI Press, Pittsburgh, PA. Menlo Park, California, pages 133–136, 1982.

[211] J. Pearl, „Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference", Morgan Kaufmann Press, San Francisco, CA, USA, 1988.

[212] P.F. Felzenszwalb, S.P. Huttenlocher "Efficient belief propagation for early vision", Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)., Vol.:1, pages 261-268, 27 June-2 July 2004.

[213] T. Montserrat, J. Civit, O.D. Escoda, J.-L. Landabaso, "Depth estimation based on multiview matching with depth/color segmentation and memory efficient Belief Propagation", 2009 16th IEEE International Conference on Image Processing (ICIP), pages 2353 – 2356, Cairo, Egypt, 7-10 Nov. 2009.

[214] J.M. Perez, P. Sanchez, M. Martinez, „High memory throughput FPGA architecture for High-Definition Belief-Propagation Stereo Matching", 3rd International Conference on Signals, Circuits and Systems (SCS), pages 1-6, Medenine, Tunisia, 6-8 Nov. 2009.

[215] S.Z. Li. "Markov Random Field Modeling in Image Analysis", Springer, 2009.

[216] Li Zhang, S.M. Seitz, „Estimating Optimal Parameters for MRF Stereo from a Single Image Pair", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.:29 , Issue: 2, pages 331 - 342, 2007.

[217] P. Viola, W. Wells, "Alignment by Maximization of Mutual Information", 1995. Proceedings of Fifth International Conference on Computer Vision, pages 16-23, Cambridge, MA , USA, 20-23 Jun 1995.

[218] Seon-Yeong , Jong-Hun Kim, Jeong-Ho Kim, Dae-Woo Lee, "3D depth estimation for target region using optical flow and mean-shift algorithm", International Conference on Control, Automation and Systems (ICCAS 2008), pages 34 – 39, Seoul, South Korea, 14-17 Oct. 2008.

[219] A. Banno, K. Ikeuchi, “Disparity map refinement and 3D surface smoothing via directed anisotropic diffusion”, The 2009 IEEE International Workshop on 3-D Digital Imaging and Modeling, Kyoto, Japan, October 2009.

[220] T. Pock, C. Zach, H. Bischof, “Mumford-Shah Meets Stereo: Integration of Weak Depth Hypotheses”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR '07), Minneapolis, USA, 2007.

[221] E. Scott Larsen et al., “Temporally Consistent Reconstruction from Multiple Video Streams Using Enhanced Belief Propagation”, International Conference on Computer Vision (ICCV 2007), 2007.

[222] F.F. Pedro, P. H. Daniel, “Efficient Belief Propagation for Early Vision”, International Journal of Computer Vision Vol. 70, No.1, Oct. 2006.

[223] H. Tao, H.S. Sawhney, R. Kumar, “Dynamic Depth Recovery from Multiple Synchronized Video Streams”, International Conference on Image Processing, 2003.

[224] C. Cigla, A.A.Alatan, „Temporally consistent dense depth map estimation via Belief Propagation", 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, pages 1-4, 4-6 May 2009.

[225] S. V. Vaseghi, “Advanced Digital Signal Processing and Noise Reduction (Third Edition)”, John Wiley & Sons 2006, ISBN: 978-0-470-09495-2, 2006.

[226] W.C. Van Etten, “Introduction to Random Signals and Noise”, John Wiley & Sons 2006. ISBN 0-470-02411-9, 2006.

[227] R. Dugad, N. Ahuja, “Video denoising by combining Kalman and Wiener estimates”, Proceedings of International Conference on Image Processing, pages 152-156, 1999.

[228] K. Arakawa “Median filter based on fuzzy rules and its application to image restoration”, Fuzzy Sets Syst., 1996.

[229] M. Lang, H. Guo, J.E. Odegard; C.S. Burrus, R.O. Wells, “Noise reduction using an undecimated discrete wavelet transform”, Signal Processing Letters, IEEE Volume 3, Issue 1, Jan. 1996.

[230] P. R. Schrater, D. C. Knill, E. P. Simoncelli, „Mechanisms of visual motion Detection”, Nature America, 2000.

[231] T. Wittkop, V. Hohmann, B. Kollmeier, “Noise reduction strategies employing interaural parameters”, Journal of The Acoustical Society of America - Vol.105, No. 2, 1999.

[232] D. Kersten, “Statistical efficiency for the detection of visual noise.”, Vision Research, Volume 27, Issue 6, Pages 1029–1040, 1987.

[233] S. V. Vaseghi, “Advanced Digital Signal Processing and Noise Reduction (Third Edition)”, John Wiley & Sons 2006, ISBN: 978-0-470-09495-2, 2006.

[234] W.C. Van Etten, “Introduction to Random Signals and Noise”, John Wiley & Sons 2006, ISBN 0-470-02411-9, 2006.

[235] R. Dugad, N. Ahuja “Video denoising by combining Kalman and Wiener estimates”, Proceedings of International Conference on Image Processing, pages 152-156, 1999.

[236] “Common Test Conditions for AVC and HEVC-based 3DV”, ISO/IEC JTC1/SC29/WG11 Doc. N12560, San Jose, USA, February 2012.

[237] “Description of Exploration Experiments in 3D Video”, ISO/IEC MPEG N9596, Antalya, Turkey, January 2008.

[238] “Description of Core Experiments in 3DVideo Coding", ISO/IEC JTC1/SC29/WG11 N12561, San Jose, USA, February 2012.

[239] Yo-Sung Ho, E.-K. Lee, C. Lee “Video Test Sequence and Camera Parameters”, ISO/IEC MPEG M15419, Archamps, France, April 2008.

[240] I. Feldmann, A. Smolic, T. Wiegand et.al., „HHI Test Material for 3D Video”, ISO/IEC JTC1/SC29/WG11, Doc M15413, Archamps, France, April 2008.

[241] Gi-M. Um, G. Bang, N. Hur, J. Kim and Yo-S. Ho, „Video Test Material of Outdoor Scene", ISO/IEC JTC1/SC29/WG11, MPEG/ M15371, Archamps, France, April 2008.

[242] M. Tanimoto, T. Fujii, N. Fukushima, “1D Parallel Test Sequences for MPEG-FTV”, MPEG M15378, Archamps, France, April 2008.

[243] J. Zhang, Ri Li, H. Li, D. Rusanovskyy, M.M. Hannuksela, “Ghost Town Fly 3DV sequence for purposes of 3DV standardization”, ISO/IEC JTC1/SC29/WG11, Doc. M20027, Geneva, Switzerland, March 2011.

[244] D. Rusanovskyy, P. Aflaki, M. M. Hannuksela, “Undo Dancer 3DV sequence for purposes of 3DV standardization”, ISO/IEC JTC1/SC29/WG11, Doc. M20028, Geneva, Switzerland, March 2011.

[245] M. Fizick, A. Balakhnin, T. Schniede “Mv-tools web-page”, http://avisynth.org.ru/mvtools/mvtools.html - online 1 Dec 2013.

[246] R. Berg, K. Post, W. Dijkhof, E. Peche, B. Rudiak-Gould “Avi-Synth webpage”, www.avisynth.org - online 1 Dec 2013.

[247] D. Wood, “The truth about stereoscopic television”, International Broadcasting Convention IBC 2009, Conference Publication, Amsterdam, 2009.

[248] Y. Sugihara, M. Tanimoto, T. Fujii, T. Senoh, T. Itoh, Y. Miyamoto, H. Nakamura, A. Ishikawa, M. Kitahara, H. Kimata, „Requirements for FTV and 3DTV to Multi-view video coding", IEC JTC 1/SC 29/WG 11 Doc. M13169, Montreux, Switzerland, April 2006.

[249] M. Tanimoto, T. Fujii, H. Kimata, S. Sakazawa, „Proposal on Requirements for FTV", ISO/IEC JTC1/SC29/WG11 Doc. M14417, San Jose, USA, April 2007.

[250] M. Tanimoto, M. P. Tehrani, T. Fujii, T. Yendo "Free-viewpoint TV", IEEE Signal Processing Magazine, Vol. 28, pages 67-76, January 2011.

[251] M. Tanimoto, „FTV: Free-viewpoint Television", Signal Processing: Image Communication, Vol. 27, Issue 6, pages 555–570, July 2012.

[252] K. Müller, P. Merkle, T. Wiegand ,"3-D video representation using depth maps", Proceedings of the IEEE, Vol. 99, No. 4, pages 643-656, April 2011.

[253] G. J. Sullivan, J. R. Ohm "Recent Developments in Standardization of High-Efficiency Video Coding (HEVC)", Proc. SPIE, Vol. 7798, Aug. 2010.

[254] F. Kossentini, "Jasper: A software-based JPEG-2000 codec implementation", IEEE Int. Conf. on Image Proc., 2, pages 53–56, October 2000.

[255] R. Krishnamurthy, B. Chai, H. Tao, S. Sethuraman, "Compression and transmission of depth maps for image-based rendering", Proceedings of International Conference on Image Processing, Vol. 3. IEEE, pages 828–831, 2001.

[256] I. Daribo, C. Tillier, B. Pesquet-Popescu,"Adaptive wavelet coding of the depth map for stereoscopic view synthesis", 2008 IEEE 10th Workshop on Multimedia Signal Processing, Proc. IEEE, 413–417, Oct. 2008.

[257] M. Maitre, M.N. Do, "Joint encoding of the depth image based representation using shape-adaptive wavelets", Int. Conf. Image Proc., ICIP 2008, Proc. IEEE, pages 1768–1771, Oct. 2008.

[258] A. Sanchez, G. Shen, A. Ortega, "Edge-preserving depth-map coding using tree-based wavelets", Asilomar Conference on Signals and Systems, Nov. 2009.

[259] W.-S. Kim, A. Ortega, P. Lai, D. Tian, C. Gomila, "Depth map distortion analysis for view rendering and depth coding", Int. Conf. Image Proc., ICIP 2009, Proc. IEEE, Nov. 2009.

[260] R. M. Willett, R. D. Nowak, "Platelets: a multiscale approach for recovering edges and surfaces in photon-limited medical imaging", IEEE Transactions on Medical Imaging 22, pages 332–350, March 2003.

[261] Y. Morvan, D. Farin, P. H. N. de With, "Novel coding technique for depth images using quadtree decomposition and plane approximation", Visual Communications and Image Proc., July 2005.

[262] Y. Morvan, D. Farin, P. H. N. With, "Platelet-based coding of depth maps for the transmission of multiview images", Proceedings of SPIE, Stereoscopic Displays and Applications, Vol. 6055, San Jose, USA, January 2006.

[263] Y. Morvan, D. Farin , P. H. N. de With, "Depth-Image Compression based on an R-D Optimized Quadtree Decomposition for the Transmission of Multiview Images", IEEE International Conference on Image Processing 2007, San Antonio, USA, Sept. 2007.

[264] D. Donoho, "Wedgelets: nearly minimax estimation of edges", Annals of Statistics 27, pages 859–897, March 1999.

[265] P. Merkle, C. Bartnik, K. Muller, D.Marpe, T. Wiegand, "3D video: Depth coding based on inter-component prediction of block partitions", Picture Coding Symposium (PCS), 2012 , Pages 149 - 152 , Kraków, Poland, 7-9 May 2012.

[266] K. Muller, P. Merkle, G. Tech, T. Wiegand, "3D video coding with depth modeling modes and view synthesis optimization", Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), USA, p.1-4, 3-6 Dec. 2012.

[267] F. Jaeger, "Contour-based segmentation and coding fordepth map compression", Proceedings of IEEE Visual Communications and Image Processing, pages 1–4, Tainan, Taiwan, 6-9 Nov. 2011.

[268] I. Feldmann, O. Schreer, P. Kauff, "Nonlinear depth scaling for video applications", Proceedings of 4th Workshop on Digital Media Processing for Multimedia Interactive Services, E. Izquierdo (ed.), World Scientific, pages 433-438, London, UK,2003.

[269] T. Senoh, K. Yamamoto, R. Oi, Y. Ichihashi, T. Kurita, "Proposal on non-linear normalization of depth maps to 8 bits", ISO/IEC JTC1/SC29/WG11, MPEG Doc. M21189, Torino, Italy, July 2011.

[270] F. Jaeger, "Simplified depth map intra coding with an optional depth lookup table", International Conference on 3D Imaging (IC3D), pages 1-4, Liège, Belgium, 3-6 Dec 2012.

[271] Filip Lewandowski, F. Paluszkiewicz, T. Grajek, K. Wegner, "Subjective Quality Assessment Methodology for 3D Video Compression Technology", IEEE International Conference on Signals and Electronic Systems – ICSES 2012, Wrocław, Poland, September 2012.

[272] Filip Lewandowski, F. Paluszkiewicz, T. Grajek, K. Wegner, „Methodology for 3D Video Subjective Quality Evaluation", International Journal of Electronics and Telecommunications, Vol. 59, No. 1. pages 25-32, 2013.

[273] I. Lim, H. Wey, D. Park, „3D-CE7.a Improved Nonlinear Depth Representation", ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, Doc. JCT3V-C0094, Geneva, Switzerland, Jan. 2013.

[274] B. Tae Oh, "Cross check of 3D-CE2.a results of Poznan Univ.", ISO/IEC JTC1/SC29/WG11, MPEG Doc. m25056, Geneva, Switzerland, May 2012.

[275] K. Müller, A. Vetro, Vittorio Baroncini, „Report of Subjective Test Results from the Call for Proposals on 3D Video Coding Technology", ISO/IEC JTC1/SC29/WG11, Doc. N12347, Geneva, Switzerland, November 2011.

[276] K. Müller, A. Vetro, V. Baroncini, „AHG Report on 3D Video Coding", ISO/IEC JTC1/SC29/WG11, MPEG11/m21469, Geneva, Switzerland, November 2011.

[277] D. Rusanovskyy, M. M. Hannuksela, "Video Description of 3D Video Coding Technology Proposal by Nokia”, ISO/IEC JTC1/SC29/WG11, Doc. m22552, Geneva, Switzerland, November 2011.

[278] H. Schwarz, P. Merkle, K. Müller, G. Tech, T. Wiegand, et Al., „Description of 3D Video Coding Technology Proposal by Fraunhofer HHI (HEVC compatible, configuration A)”, ISO/IEC JTC1/SC29/WG11, Doc. m22570, Geneva, Switzerland, November 2011.

[279] H. Schwarz, P. Merkle, K. Müller, G. Tech, T. Wiegand, et Al., „Description of 3D Video Coding Technology Proposal by Fraunhofer HHI (HEVC compatible, configuration B)”, ISO/IEC JTC1/SC29/WG11, Doc. m22571, Geneva, Switzerland, November 2011.

[280] N. Stefanoski, A. Smolic, K. Müller, H. Schwarz, T. Wiegand, et Al., „Description of 3D Video Coding Technology Proposal by Disney Research Zurich and Fraunhofer HHI", ISO/IEC JTC1/SC29/WG11, Doc. M22668, Geneva, Switzerland, Nov. 2011.

[281] H. Schwarz, K. Wegner, T. Rusert, “Overview of 3DV coding tools proposed in the CfP”, ISO/IEC JTC1/SC29/WG11 Doc. N12348, Geneva, Switzerland, Dec. 2011.

[282] J. Konieczny, M. Domański, “Extended inter-view direct mode for Multiview Video Coding”, IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2011, pages 845-848, Prague, Czech Republic, May 2011.

[283] J. Konieczny, M. Domański, „Depth-based inter-view motion data prediction for HEVC-based multiview video coding”, Picture Coding Symposium, PCS 2012, Kraków, Poland, pages 33-36, May 2012.

[284] J. Konieczny, M. Domański, „Depth-based inter-view motion data prediction for HEVC-based multiview video coding”, Picture Coding Symposium, PCS 2012, Kraków, pages 33-36, May 2012.

[285] J.A. Martins, „Low bit rate LPC vocoders using vector quantization and interpolation”, International Conference on Acoustics, Speech, and Signal Processing, 1991, Vol. 1, pages 597- 600, Toronto, Canada, 1991.

Appendix

Linear correlation coefficient for noise values

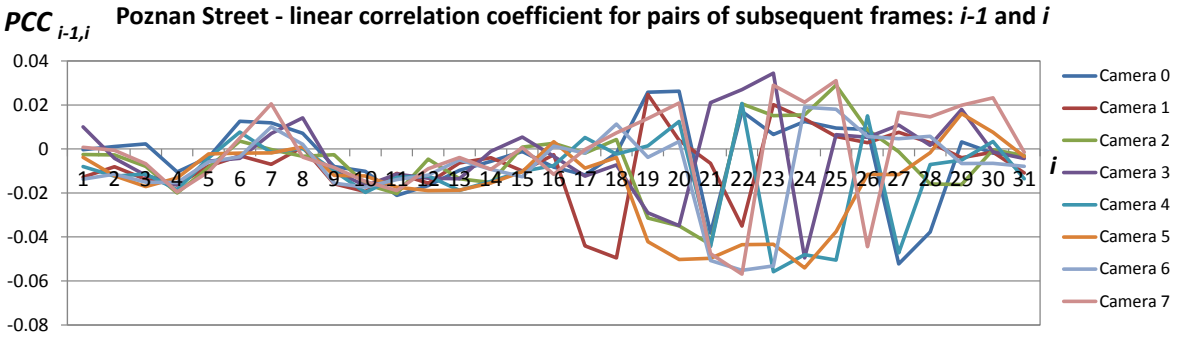


Fig. 81. Linear correlation coefficient measured for noise values in two subsequent frames, index $i-1$ and index i, for Poznan Street sequence.

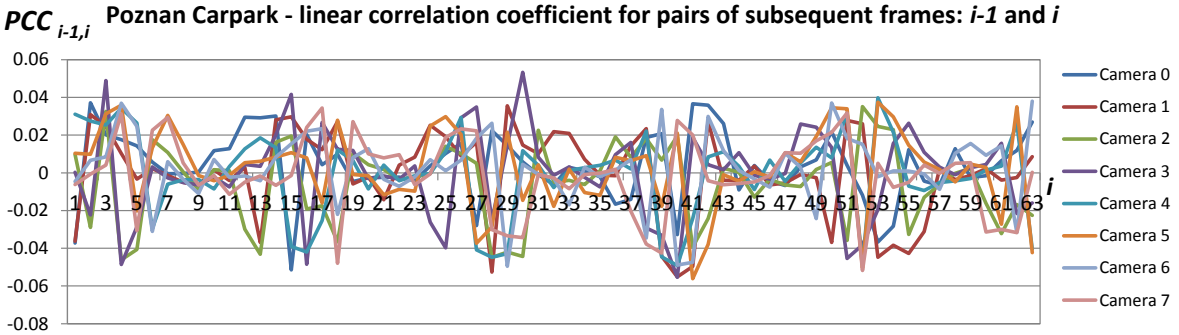


Fig. 82. Linear correlation coefficient measured for noise values in two subsequent frames, index $i-1$ and index i, for Poznan Carpark sequence.

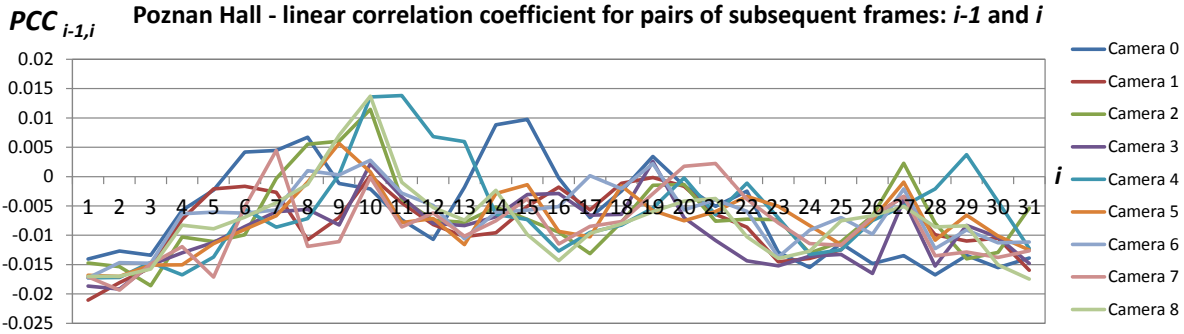


Fig. 83. Linear correlation coefficient measured for noise values in two subsequent frames, index $i-1$ and index i, for Poznan Hall sequence.

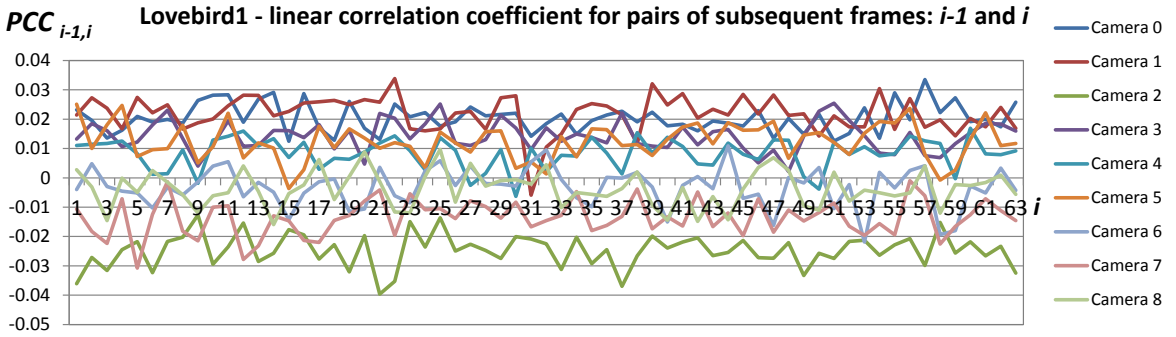


Fig. 84. Linear correlation coefficient measured for noise values in two subsequent frames, index $i-1$ and index i, for Lovebird1 sequence.

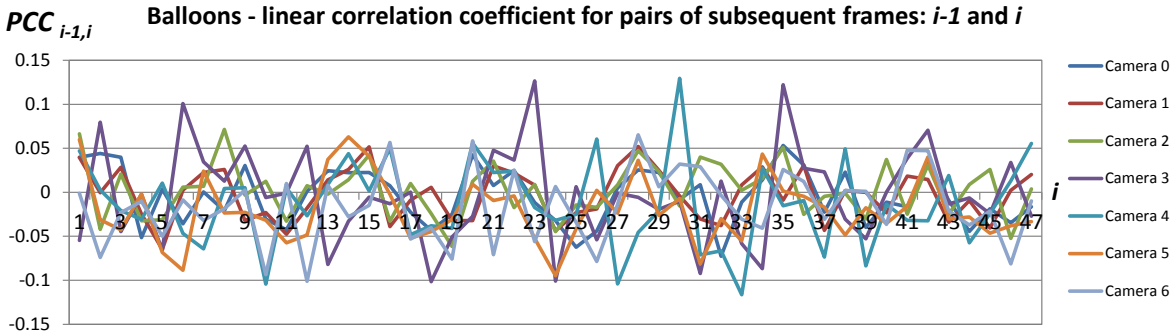


Fig. 85. Linear correlation coefficient measured for noise values in two subsequent frames, index $i-1$ and index i, for Balloons sequence.

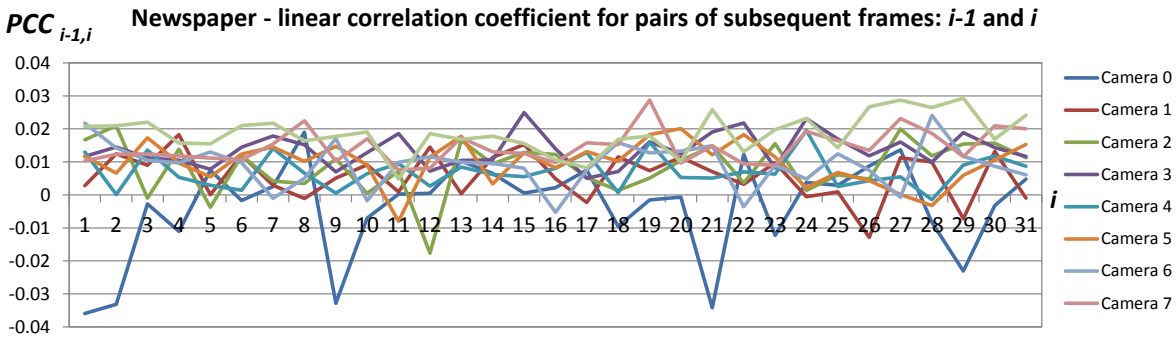


Fig. 86. Linear correlation coefficient measured for noise values in two subsequent frames, index $i-1$ and index i, for Newspaper sequence.

Probability distributions of noise values

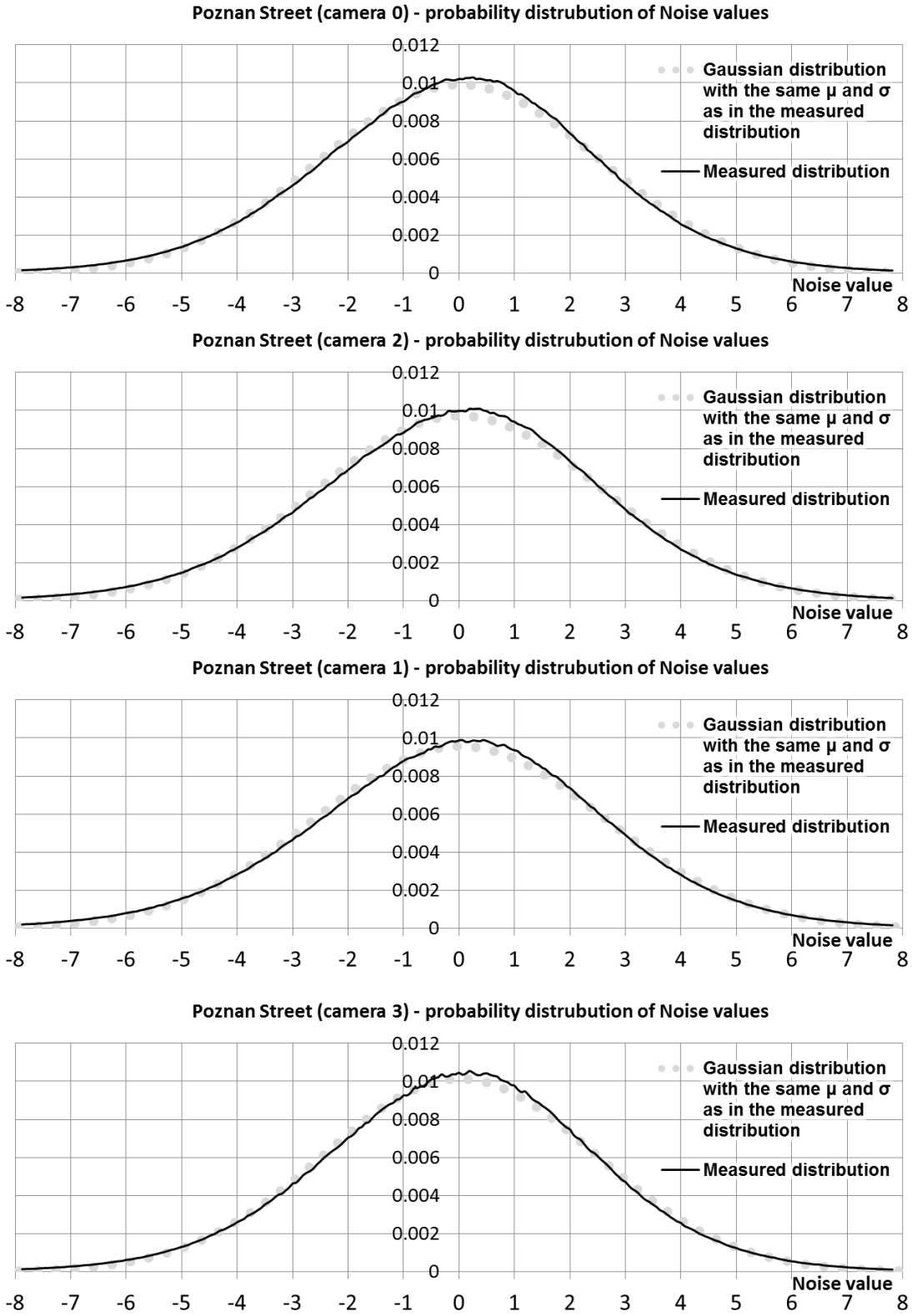


Fig. 87. Measured probability distributions of noise values in Poznan Street sequence (cameras 0…3), estimated with histogram bin size of $^1/_{16}$.
See (38) on page 57 for more details.

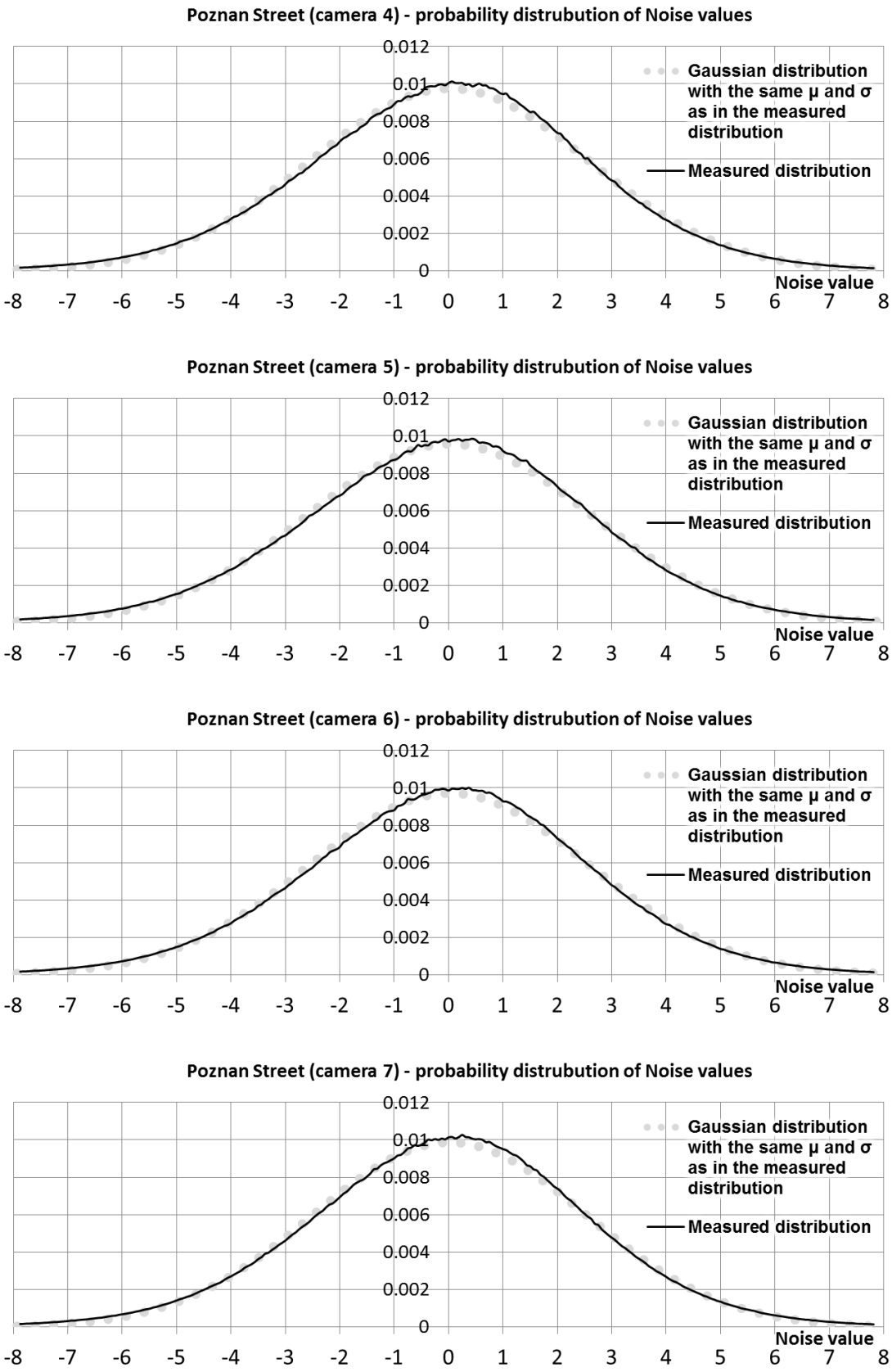


Fig. 88. Measured probability distributions of noise values in Poznan Street sequence (cameras 4…7), estimated with histogram bin size of $1/16$.
See (38) on page 57 for more details.

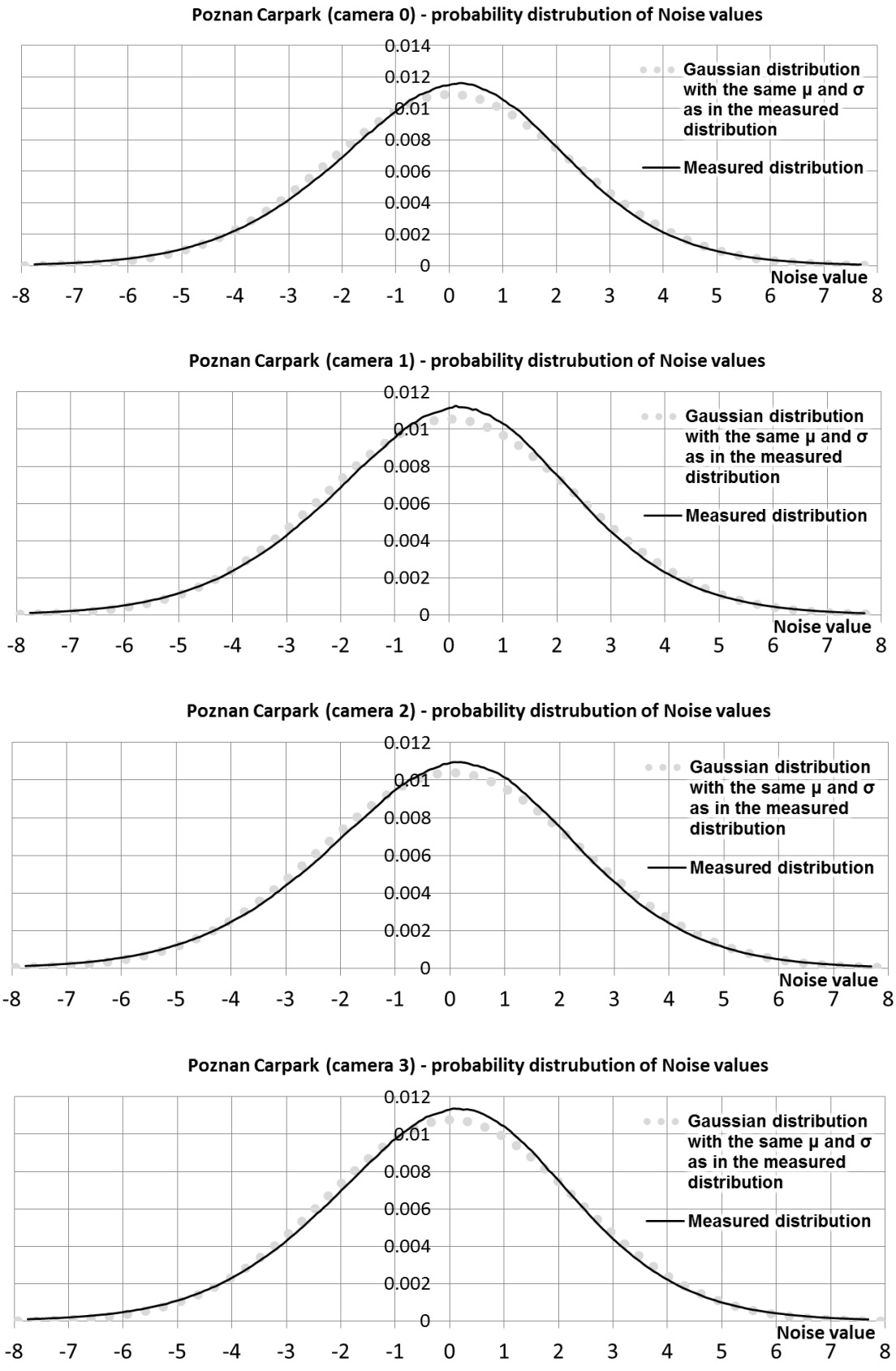


Fig. 89. Measured probability distributions of noise values in Poznan Carpark sequence (cameras 0…3), estimated with histogram bin size of $1/16$.
See (38) on page 57 for more details.

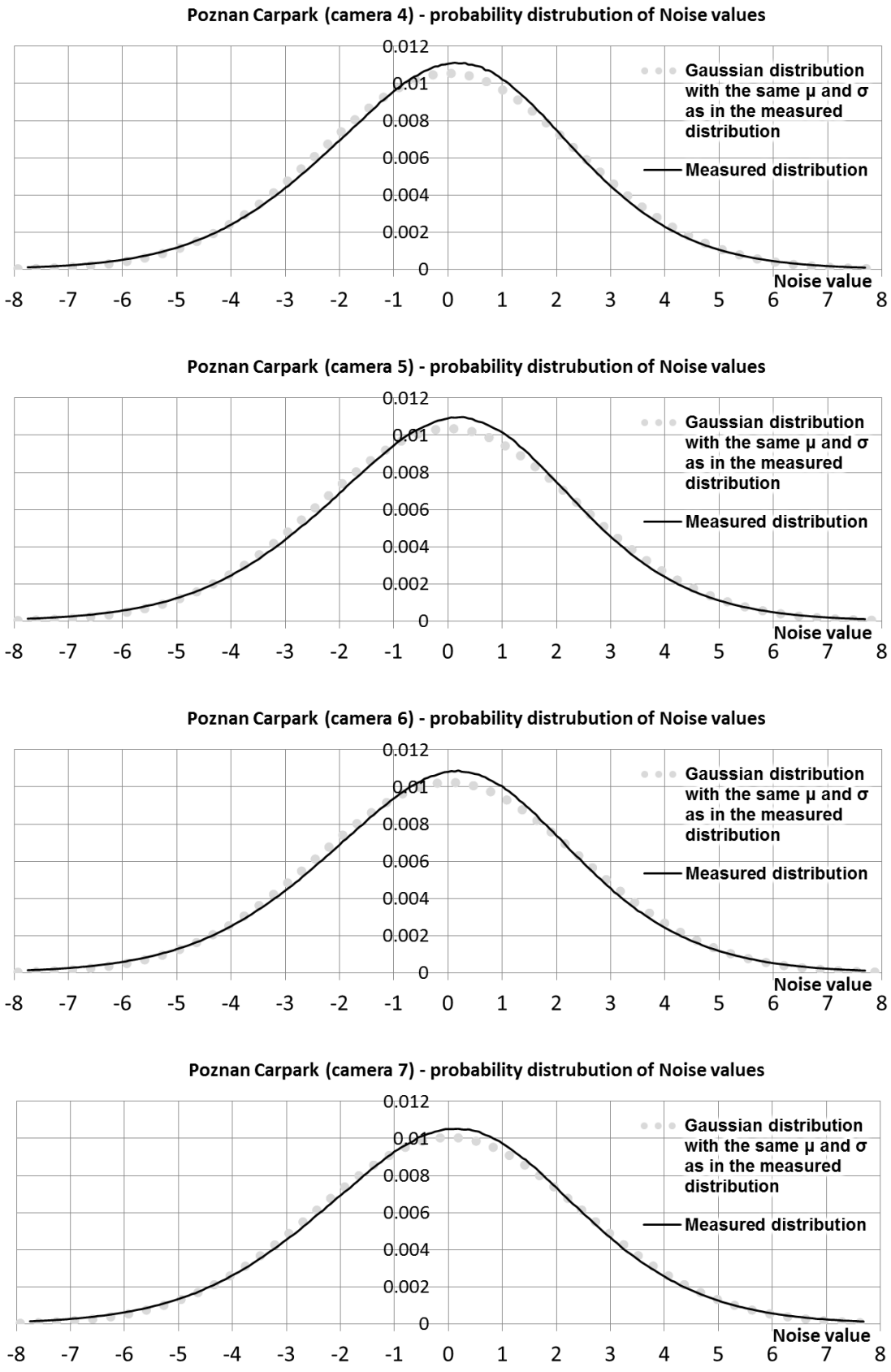


Fig. 90. Measured probability distributions of noise values in Poznan Carpark sequence (cameras 4…7), estimated with histogram bin size of $1/16$.
See (38) on page 57 for more details.

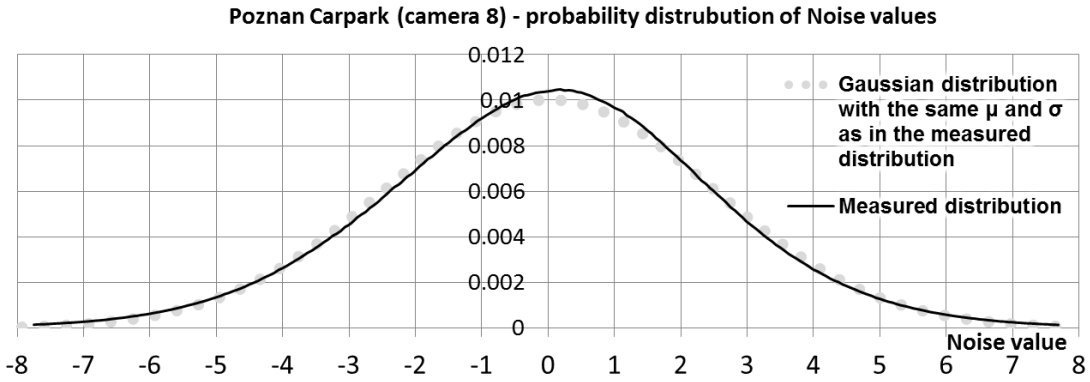


Fig. 91. Measured probability distributions of noise values in Poznan Carpark sequence (camera 8), estimated with histogram bin size of $^1/_{16}$.
See (38) on page 57 for more details.

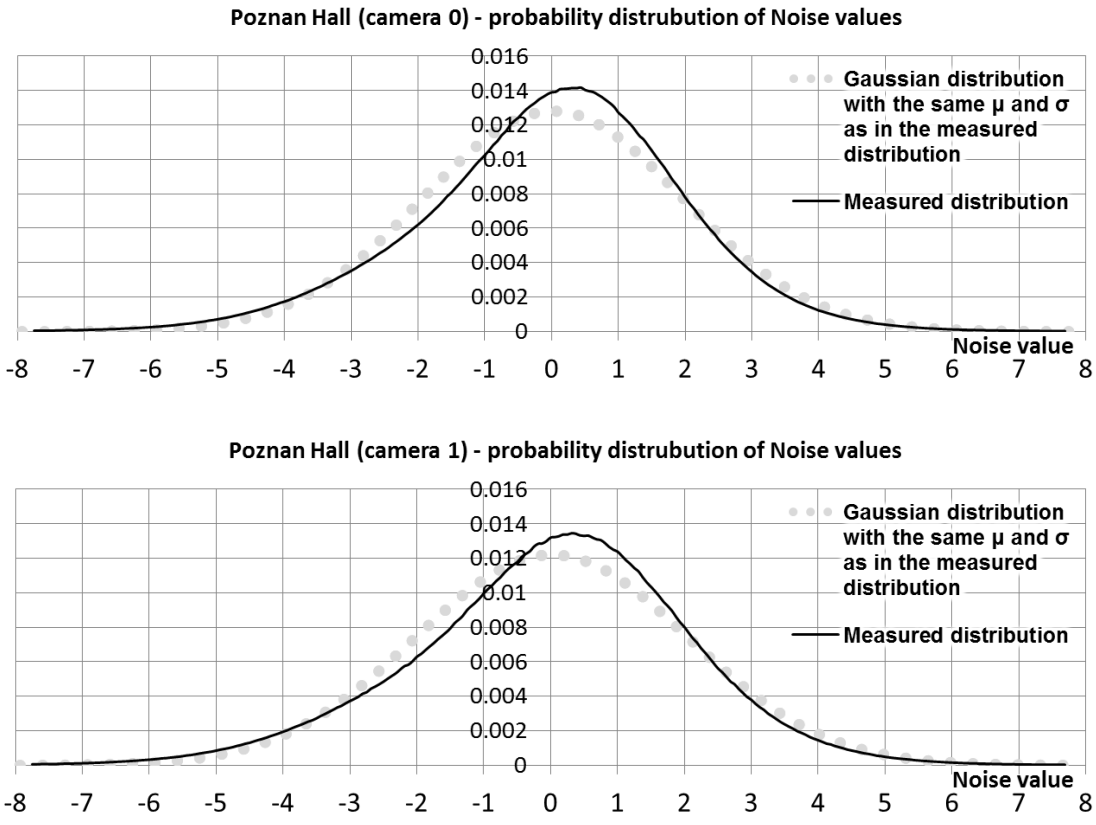


Fig. 92. Measured probability distributions of noise values in Poznan Hall sequence (cameras 0…1), estimated with histogram bin size of $^1/_{16}$.
See (38) on page 57 for more details.).

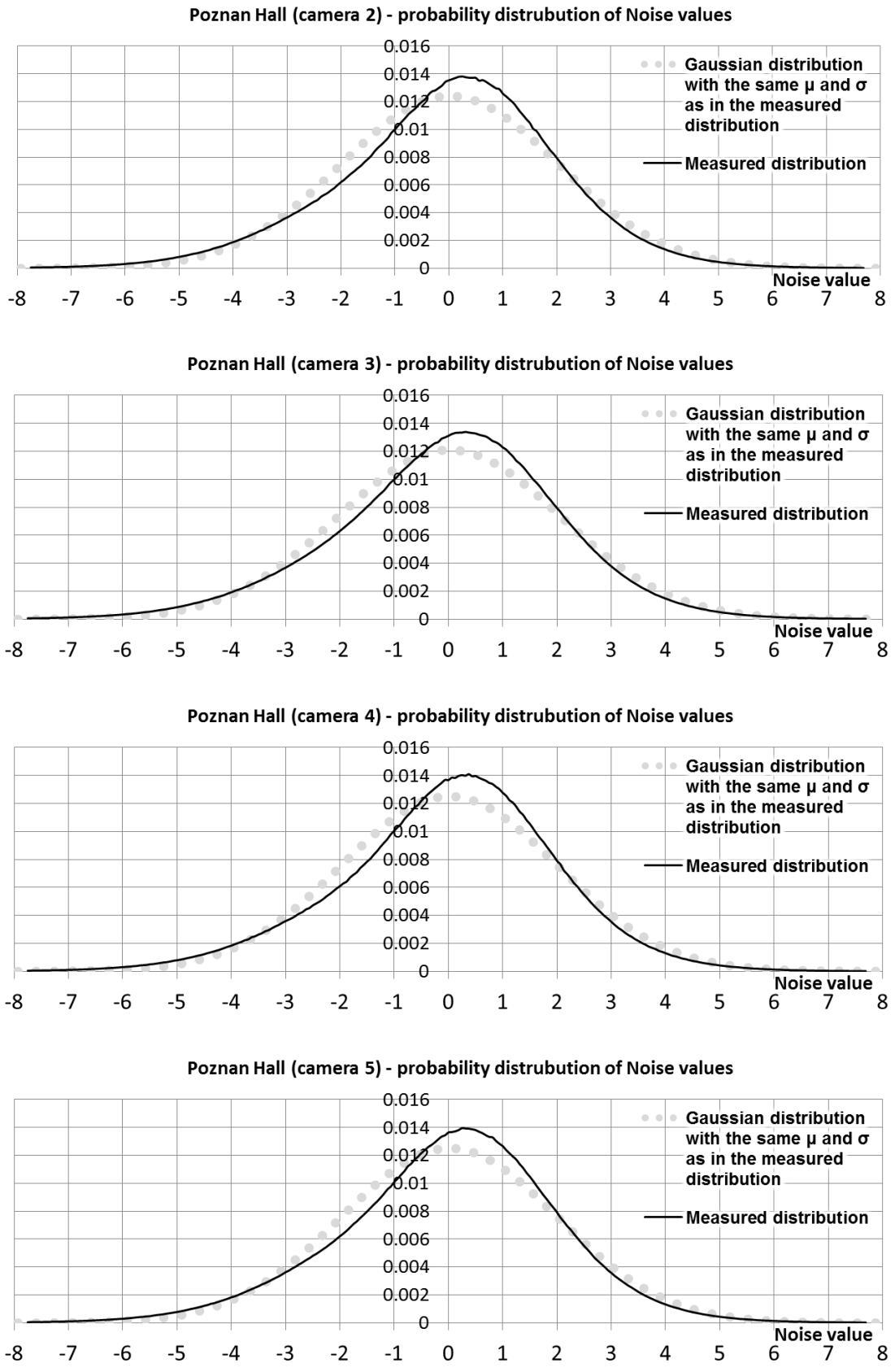


Fig. 93. P Measured probability distributions of noise values in Poznan Hall sequence (cameras 2…5), estimated with histogram bin size of $1/16$.
See (38) on page 57 for more details.

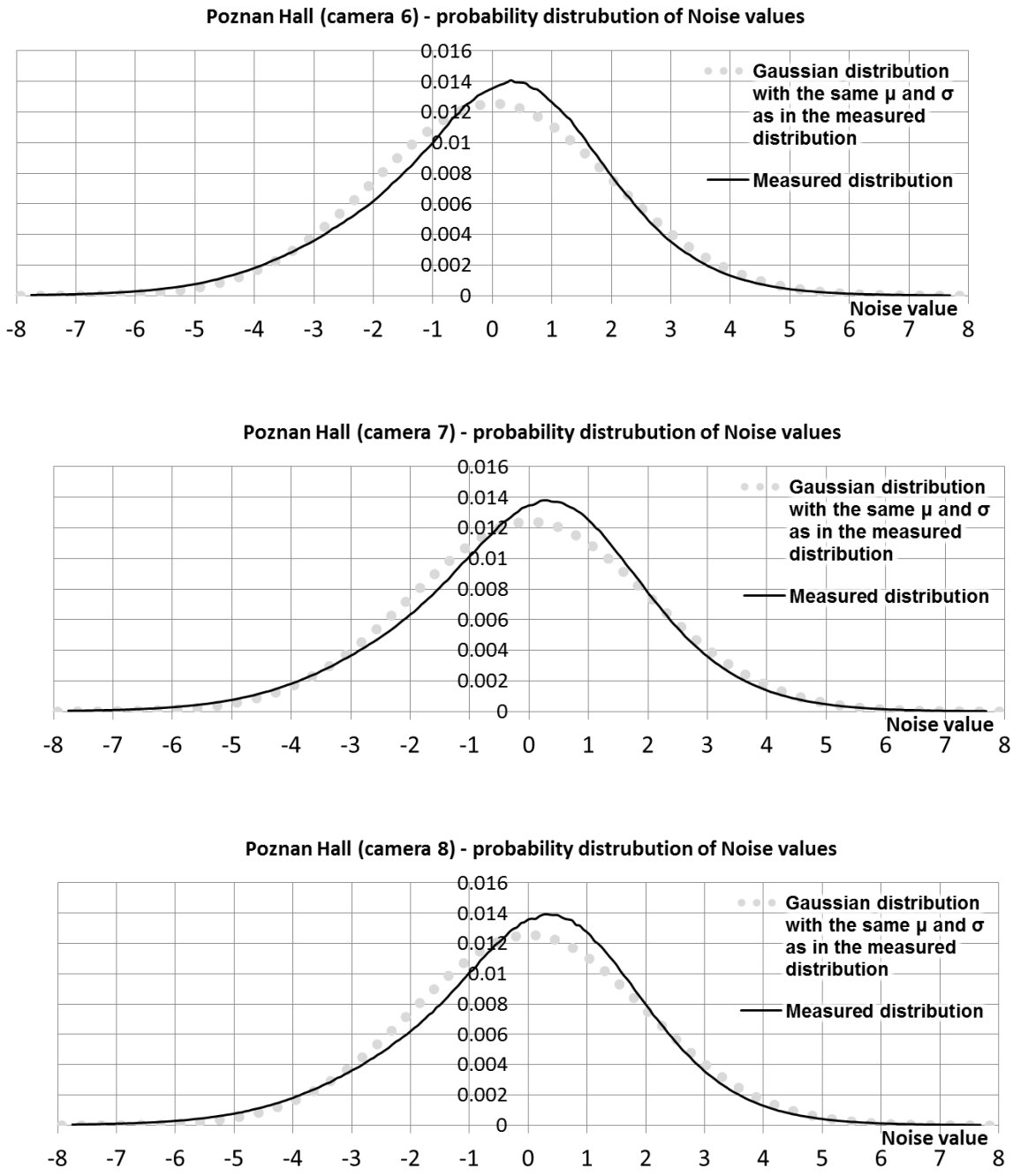


Fig. 94. Measured probability distributions of noise values in Poznan Hall sequence (cameras 6…8), estimated with histogram bin size of $1/16$.
See (38) on page 57 for more details.

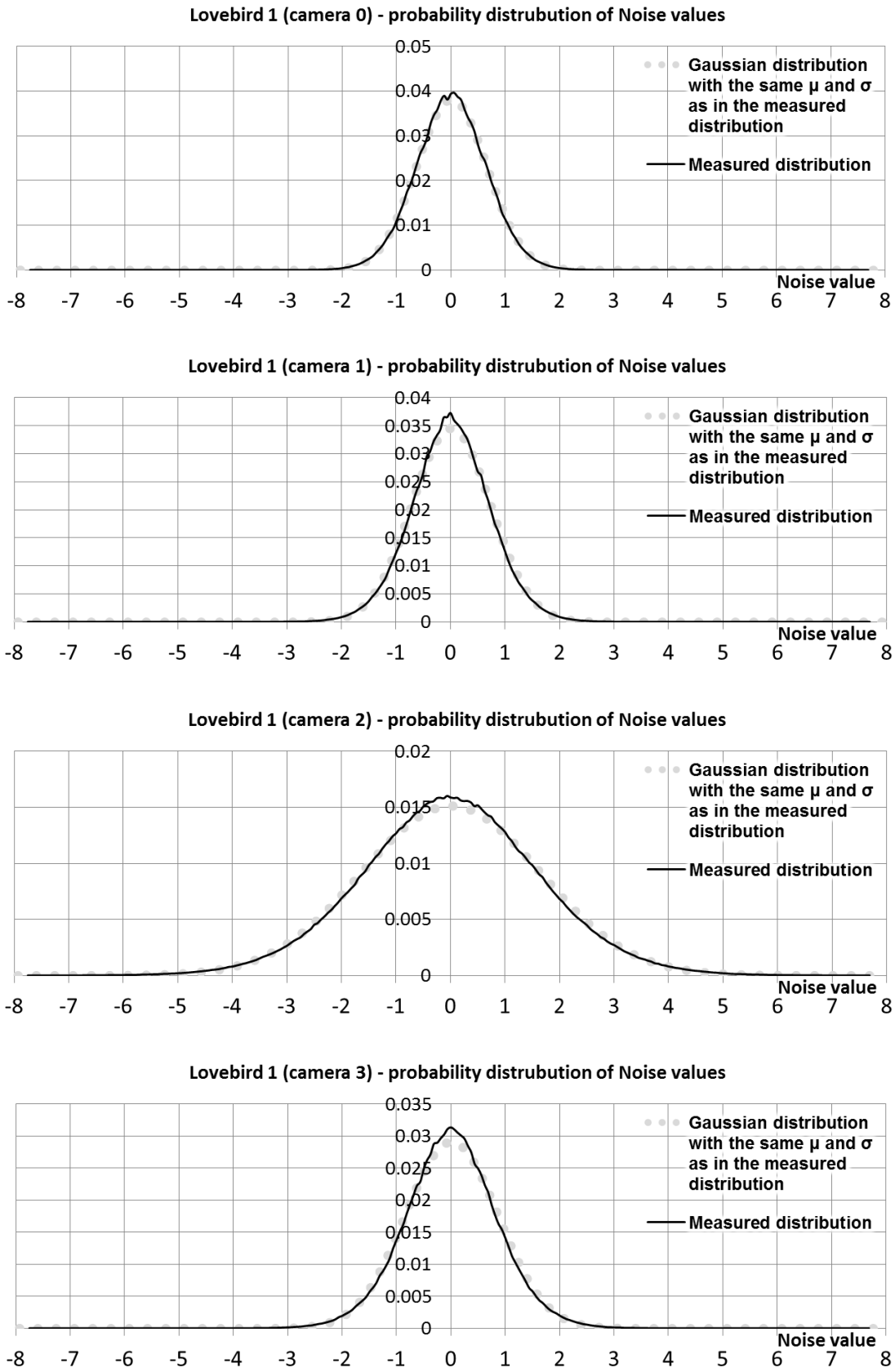


Fig. 95. Measured probability distributions of noise values in Lovebird1 sequence (cameras 0…3), estimated with histogram bin size of $1/16$.
See (38) on page 57 for more details.

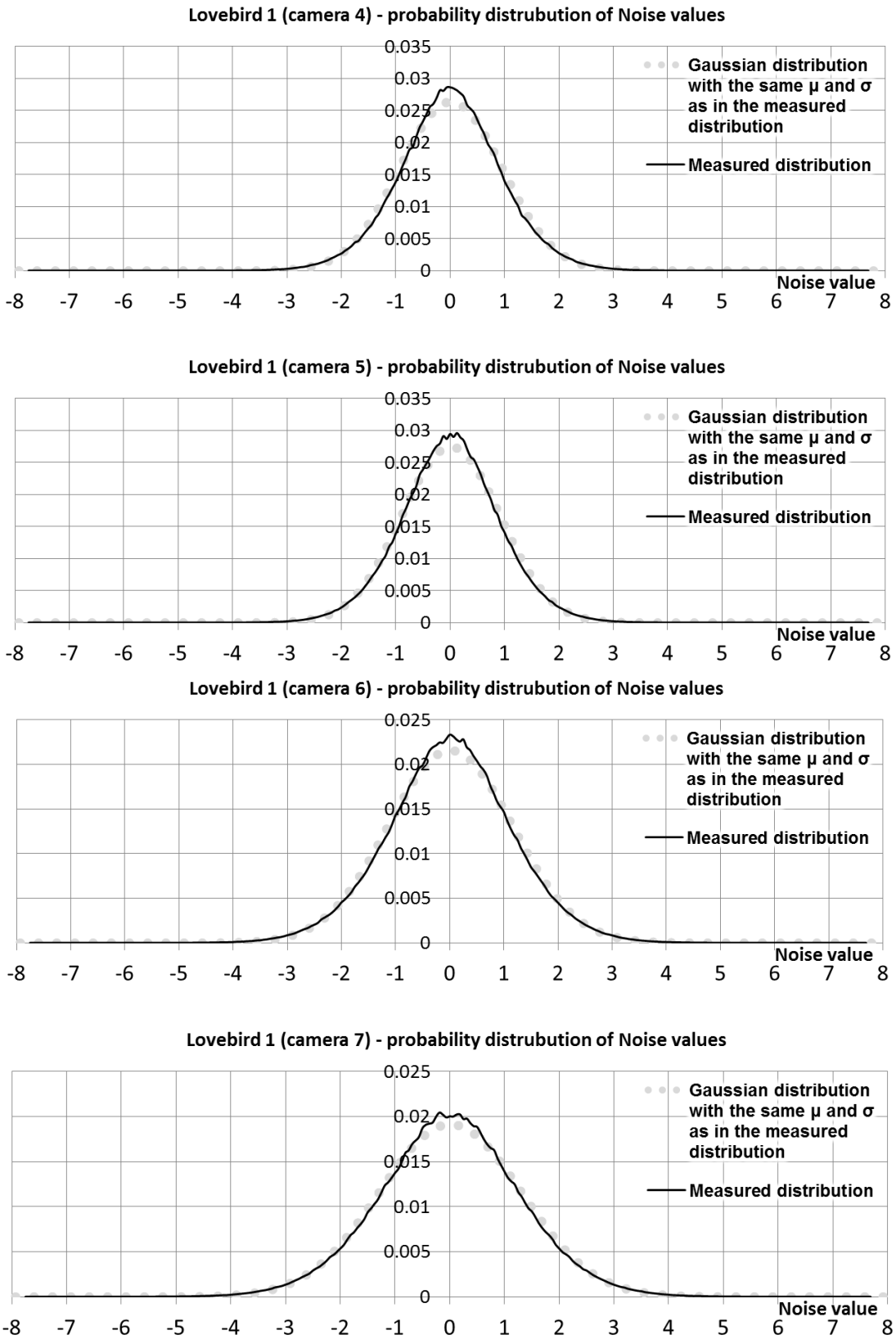


Fig. 96. Measured probability distributions of noise values in Lovebird1 sequence (cameras 4…7), estimated with histogram bin size of $1/16$.
See (38) on page 57 for more details.

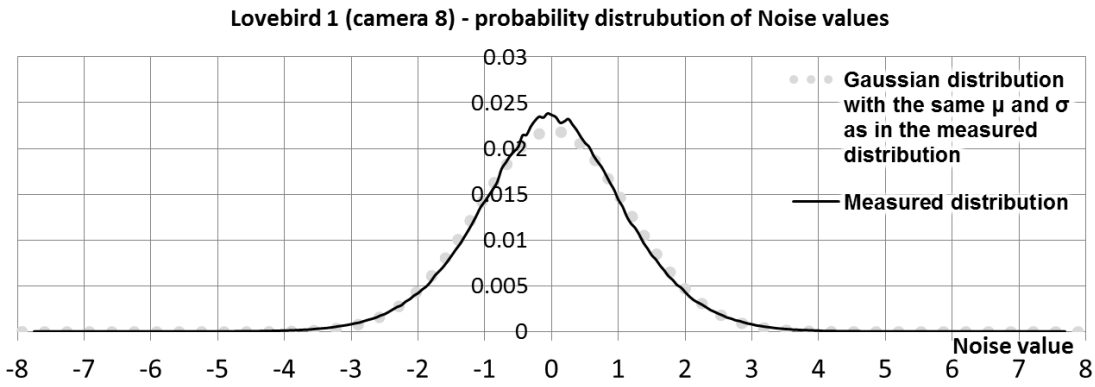


Fig. 97. Measured probability distribution of noise values in Lovebird1 sequence (camera 8), estimated with histogram bin size of $1/16$.
See (38) on page 57 for more details.).

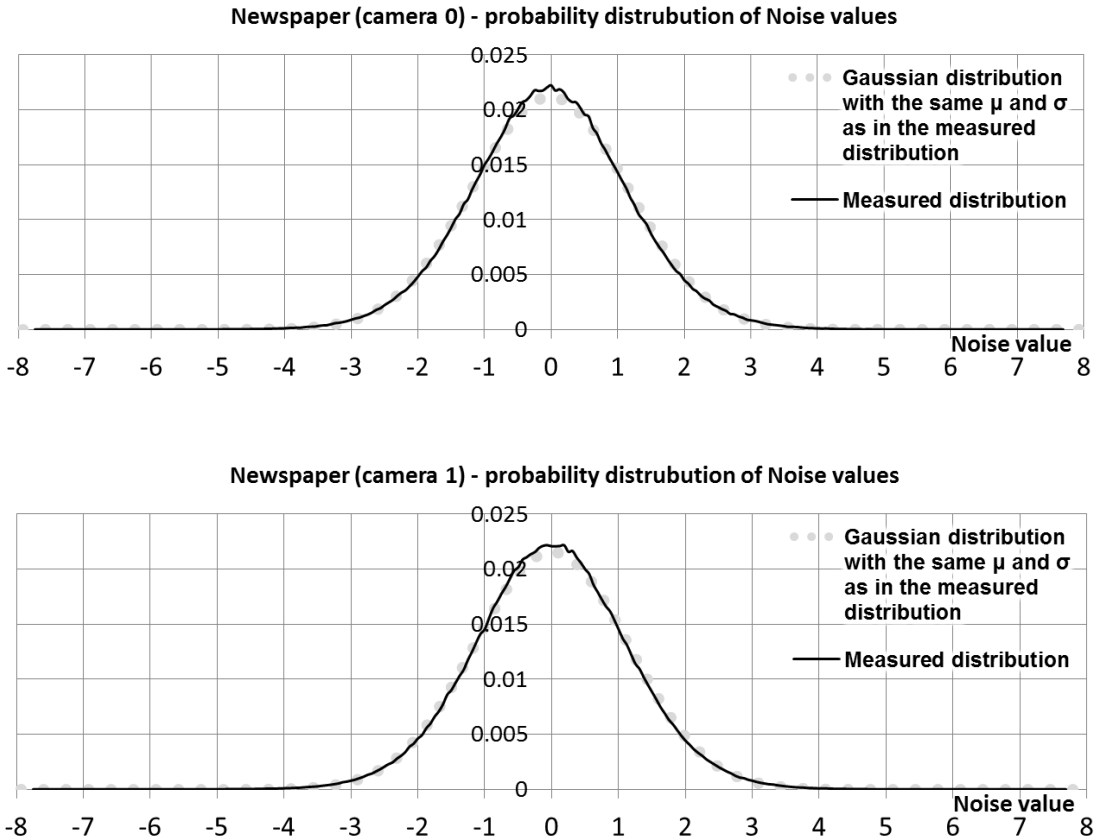


Fig. 98. Measured probability distributions of noise values in Newspaper sequence (cameras 0…1), estimated with histogram bin size of $1/16$.
See (38) on page 57 for more details.

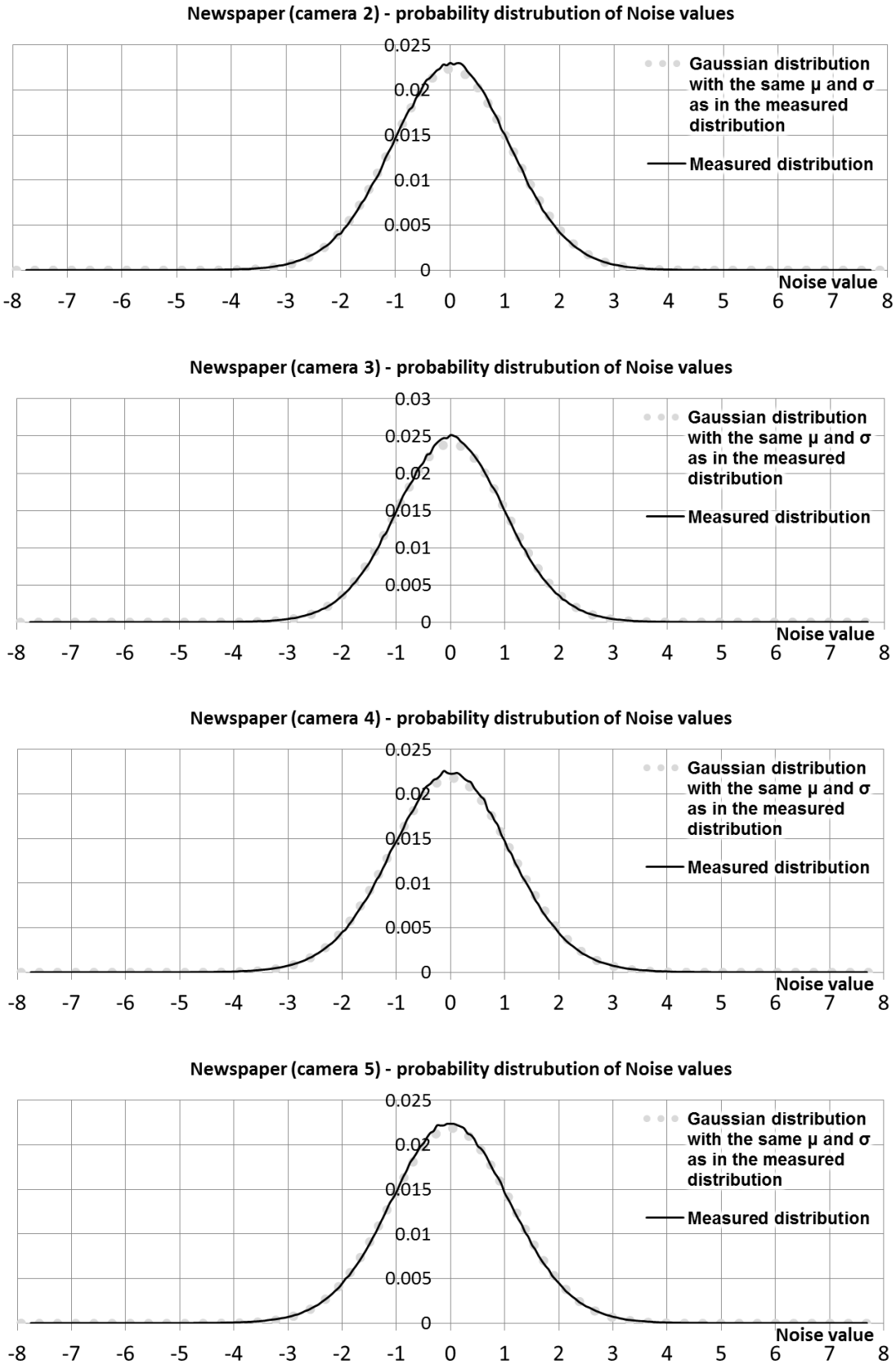
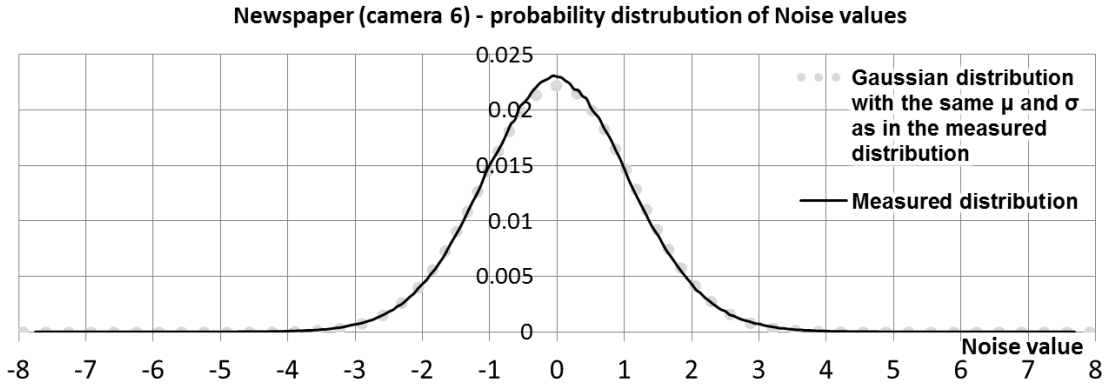
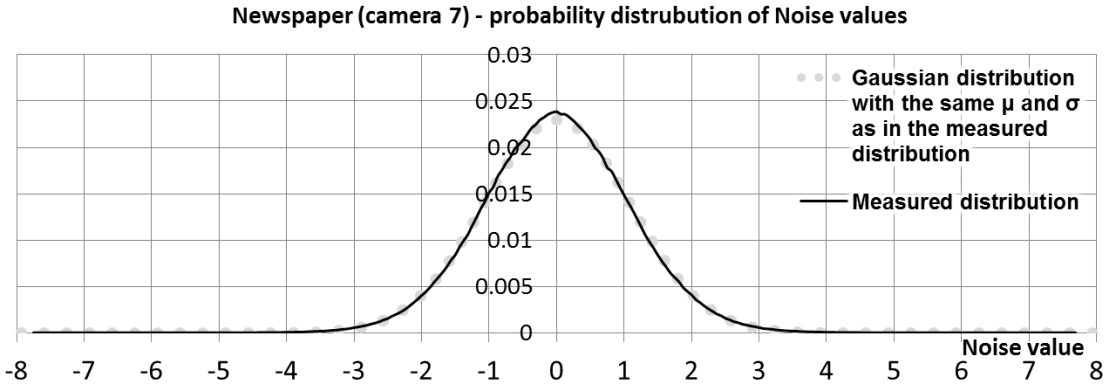


Fig. 99. Measured probability distributions of noise values in Newspaper sequence (cameras 2…5), estimated with histogram bin size of $1/16$.
See (38) on page 57 for more details.





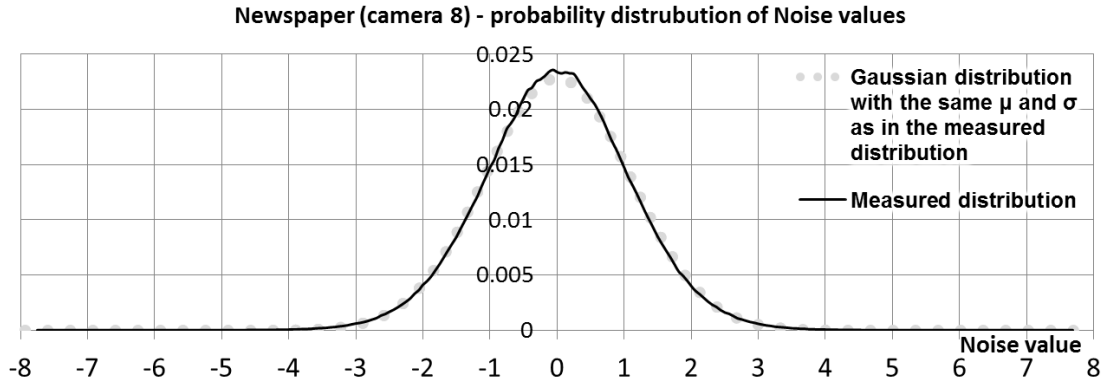


Fig. 100. Measured probability distributions of noise values in Newspaper sequence (cameras 6…8), estimated with histogram bin size of $1/16$.
See (38) on page 57 for more details.

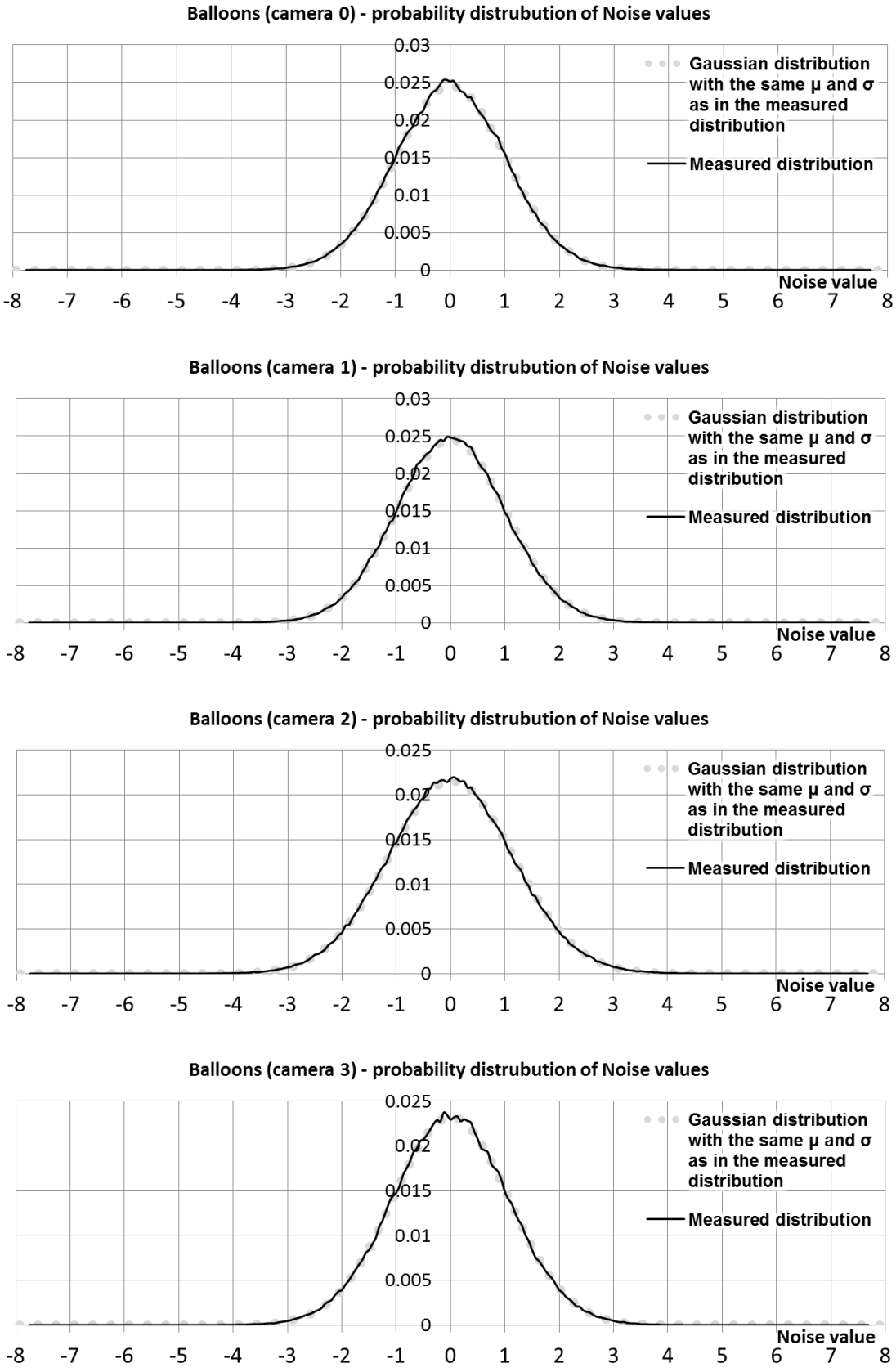


Fig. 101. Measured probability distributions of noise values in Balloons sequence (cameras 0…3), estimated with histogram bin size of $1/16$.
See (38) on page 57 for more details..

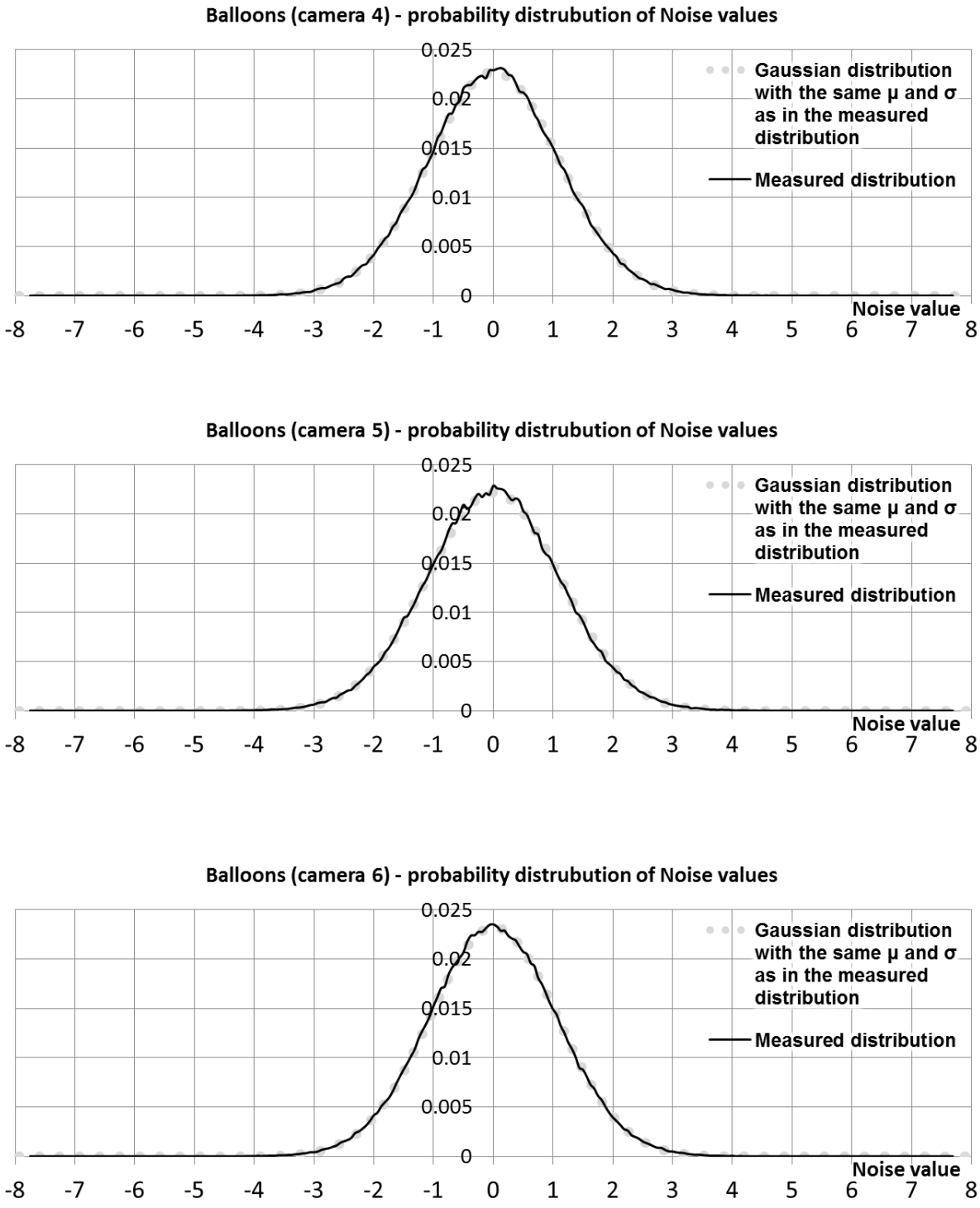
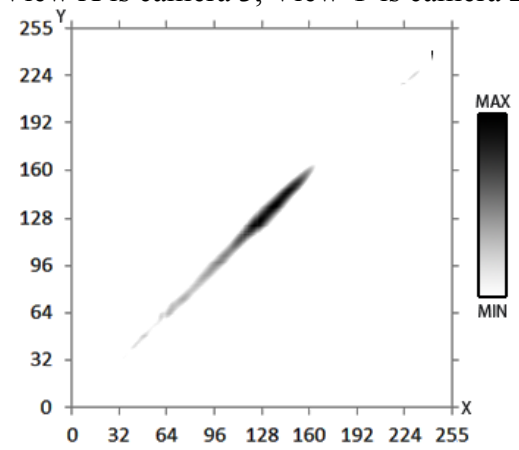


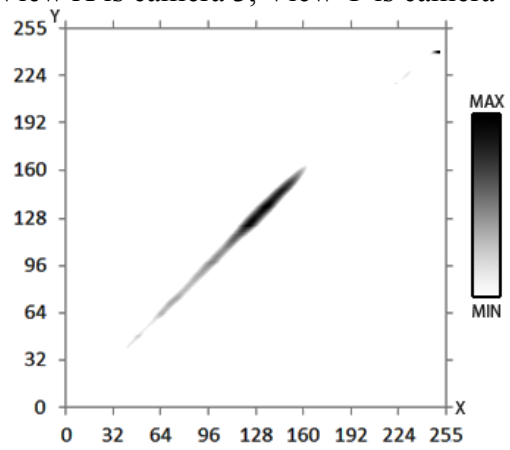
Fig. 102. Measured probability distributions of noise values in Balloons sequence (cameras 4…6), estimated with histogram bin size of $1/16$.
See (38) on page 57 for more details.

Raw 2-D histograms of luminance values in 2 neighboring views

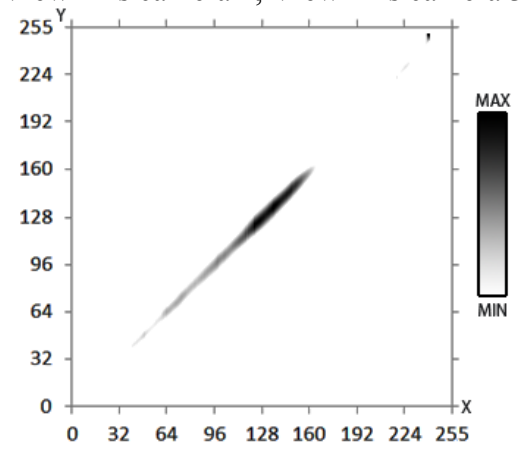
Poznan Street sequence.
View X is camera 3, View Y is camera 2.



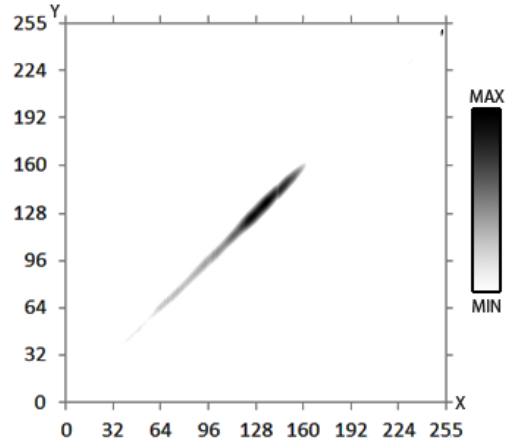
Poznan Street sequence.
View X is camera 3, View Y is camera 4.



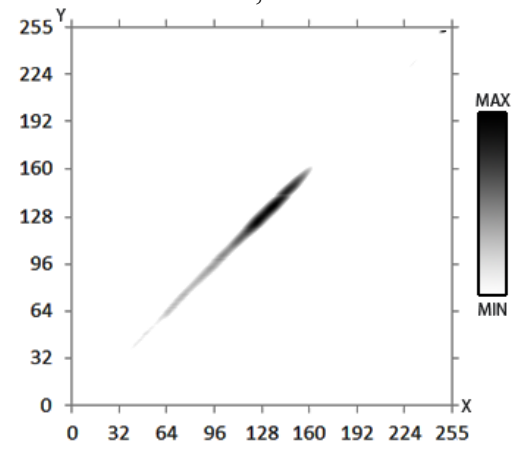
Poznan Street sequence.
View X is camera 4, View Y is camera 3.



Poznan Street sequence.
View X is camera 4, View Y is camera 5.



Poznan Street sequence.
View X is camera 5, View Y is camera 4.



Poznan Street sequence.
View X is camera 5, View Y is camera 6.

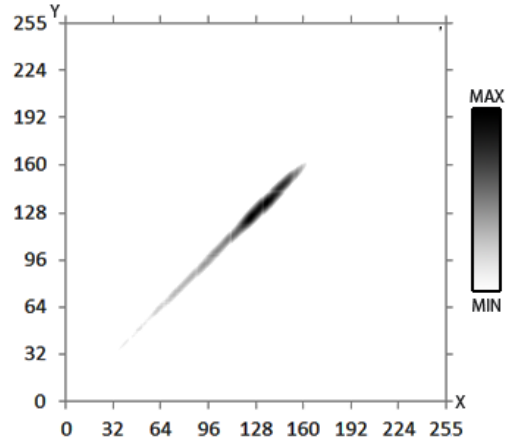
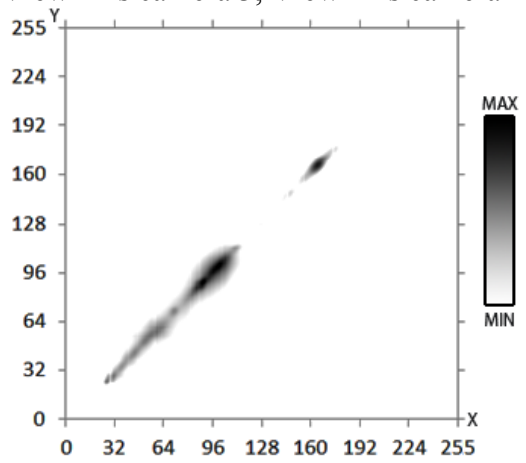
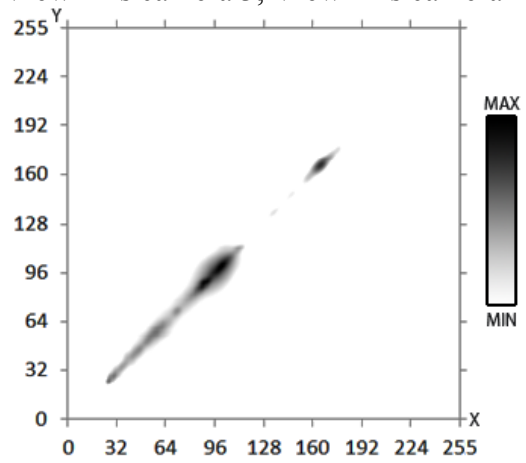


Fig. 103. Graphs of 2-dimentional histograms of luminance values (in logarithmic gray-level scale) of corresponding pixels in the views: (X=3 Y=2,4) (X=4, Y=3,5) (X=5, Y=4,6) of Poznan Street sequence.

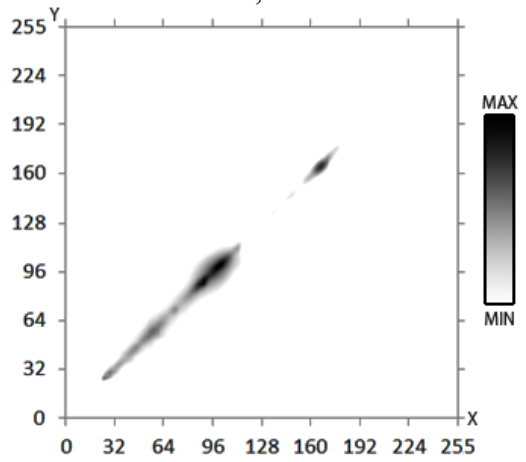
Poznan Carpark sequence.
View X is camera 3, View Y is camera 2.



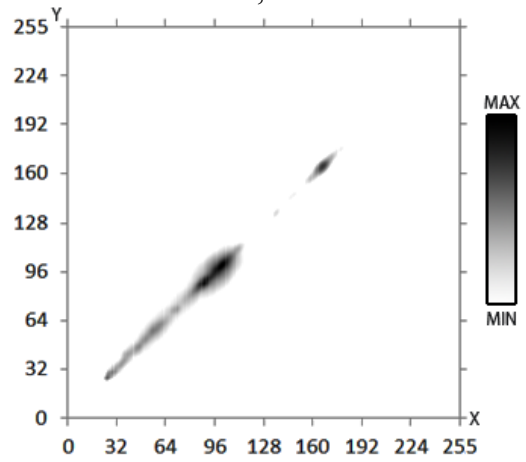
Poznan Carpark sequence.
View X is camera 3, View Y is camera 4.



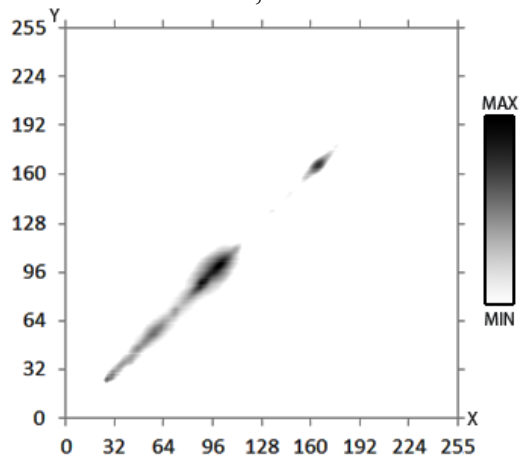
Poznan Carpark sequence.
View X is camera 4, View Y is camera 3.



Poznan Carpark sequence.
View X is camera 4, View Y is camera 5.



Poznan Carpark sequence.
View X is camera 5, View Y is camera 4.



Poznan Carpark sequence.
View X is camera 5, View Y is camera 6.

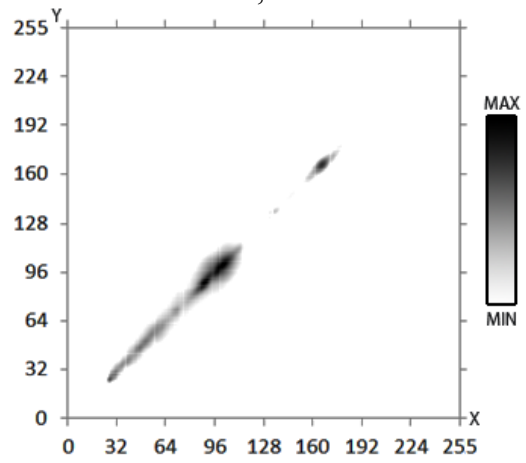
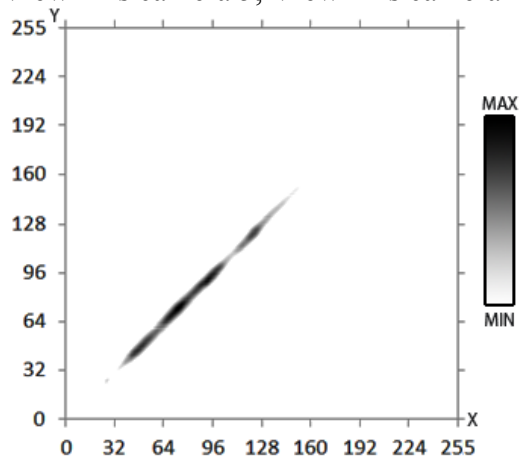
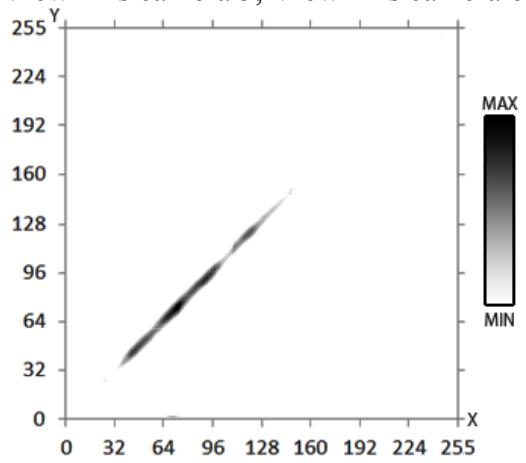


Fig. 104. Graphs of 2-dimentional histograms of luminance values (in logarithmic gray-level scale) of corresponding pixels in the views: (X=3 Y=2,4) (X=4, Y=3,5) (X=5, Y=4,6) of Poznan Carpark sequence.

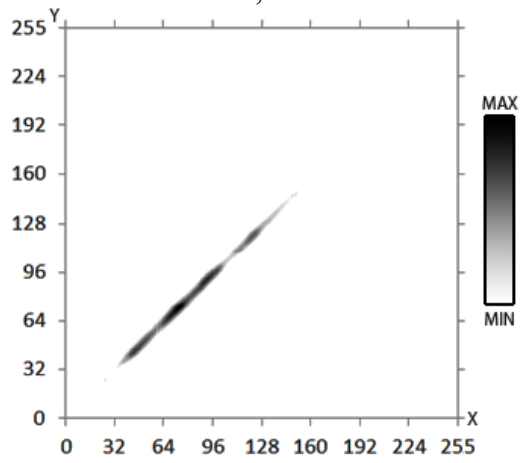
Poznan Hall2 sequence.
View X is camera 5, View Y is camera 4.



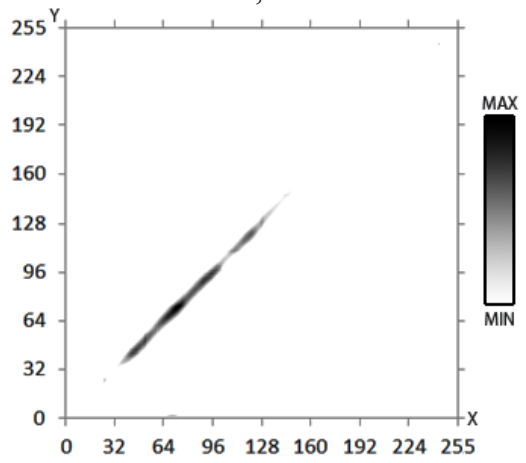
Poznan Hall2 sequence.
View X is camera 5, View Y is camera 6.



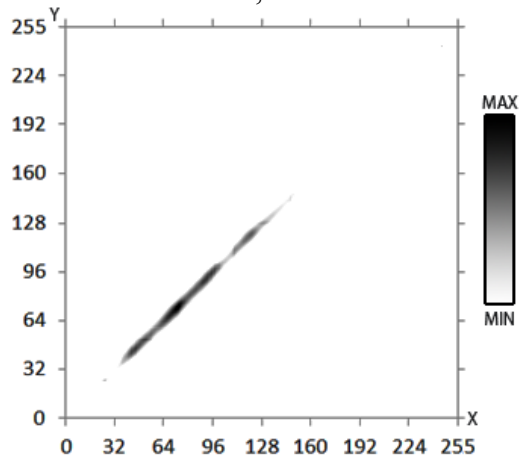
Poznan Hall2 sequence.
View X is camera 6, View Y is camera 5.



Poznan Hall2 sequence.
View X is camera 6, View Y is camera 7.



Poznan Hall2 sequence.
View X is camera 7, View Y is camera 6.



Poznan Hall2 sequence.
View X is camera 7, View Y is camera 8.

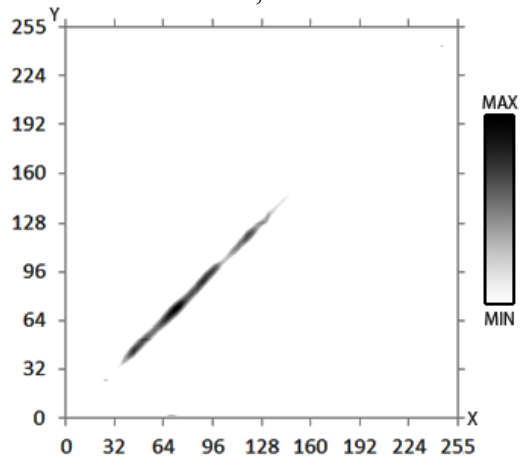
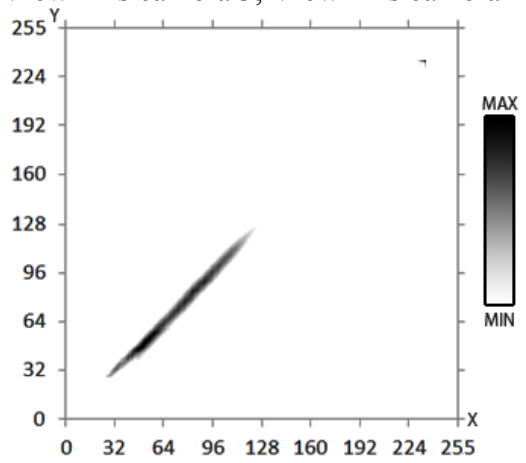
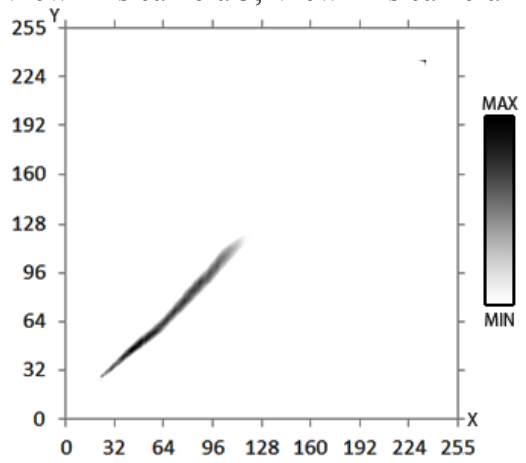


Fig. 105. Graphs of 2-dimentional histograms of luminance values (in logarithmic gray-level scale) of corresponding pixels in the views: (X=5 Y=4,6) (X=6, Y=5,7) (X=7, Y=6,8) of Poznan Hall2 sequence.

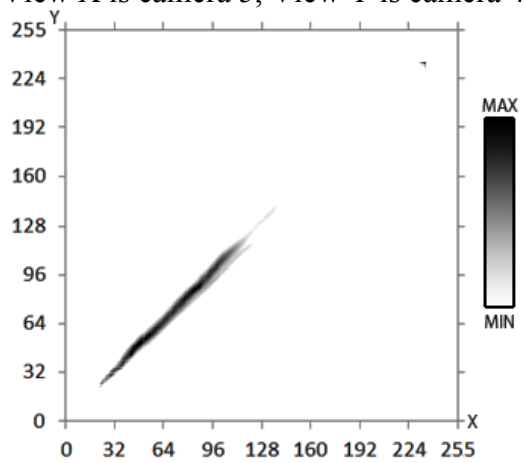
Lovebird1 sequence.
View X is camera 3, View Y is camera 2.



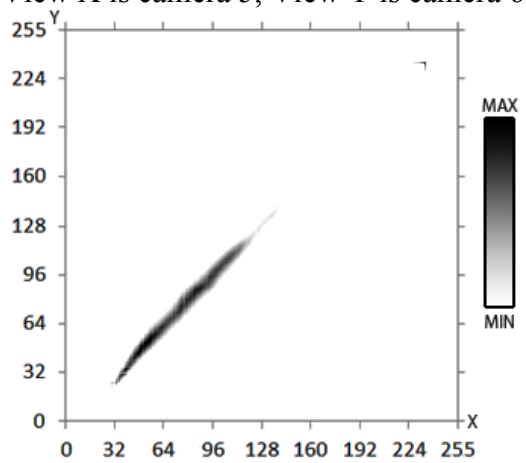
Lovebird1 sequence.
View X is camera 3, View Y is camera 4.



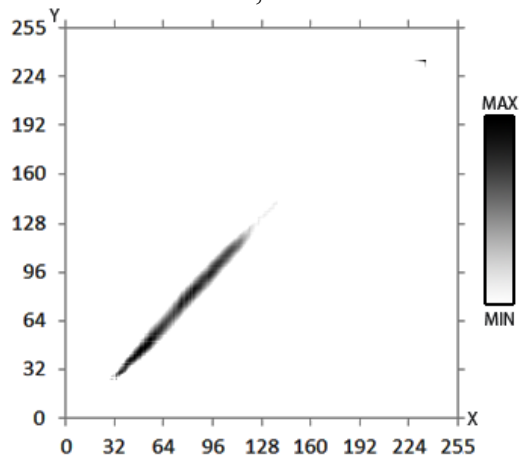
Lovebird1 sequence.
View X is camera 5, View Y is camera 4.



Lovebird1 sequence.
View X is camera 5, View Y is camera 6.



Lovebird1 sequence.
View X is camera 7, View Y is camera 6.



Lovebird1 sequence.
View X is camera 7, View Y is camera 8.

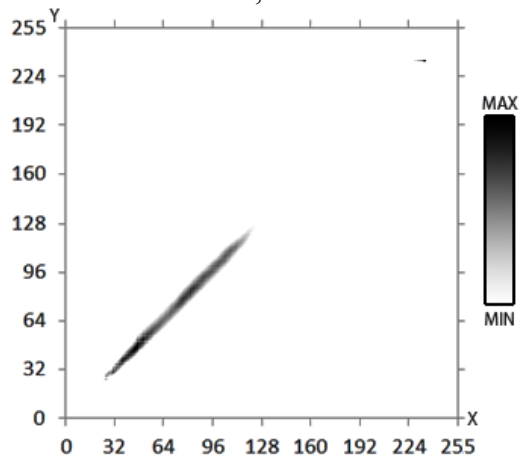
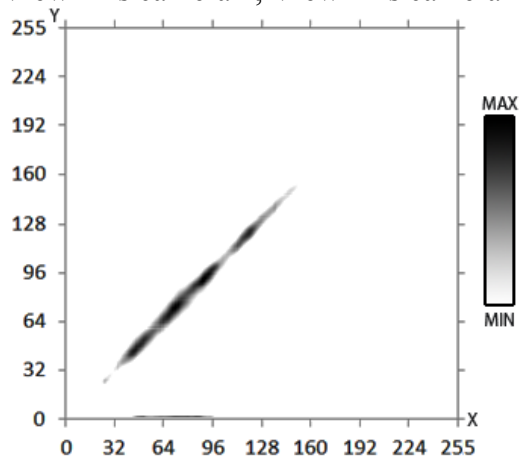
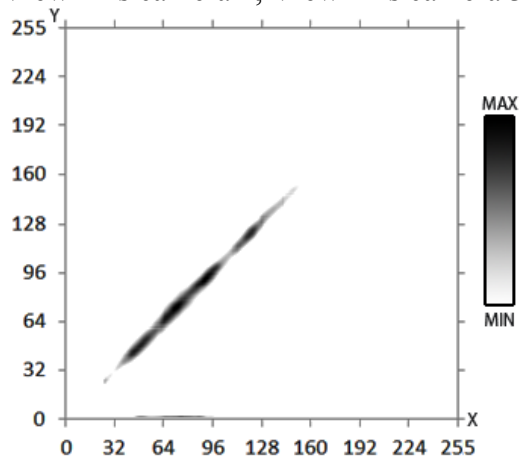


Fig. 106. Graphs of 2-dimentional histograms of luminance values (in logarithmic gray-level scale) of corresponding pixels in the views: (X=4 Y=3,5) (X=6, Y=5,7) (X=8, Y=7,9) of Lovebird1 sequence.

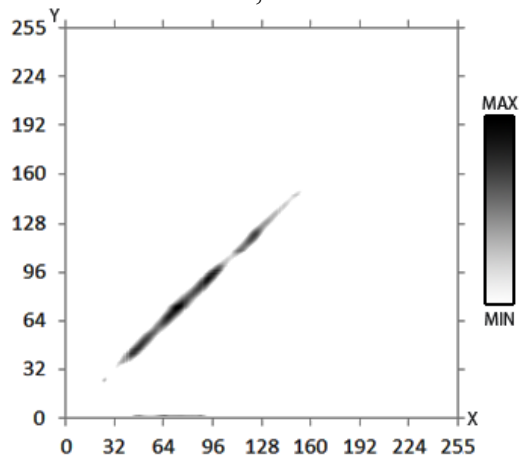
Newspaper sequence.
View X is camera 2, View Y is camera 1.



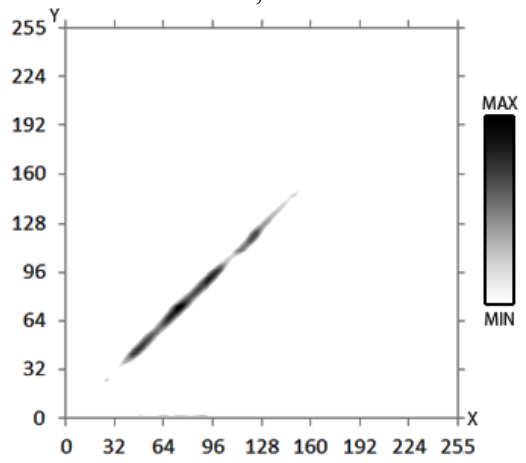
Newspaper sequence.
View X is camera 2, View Y is camera 3.



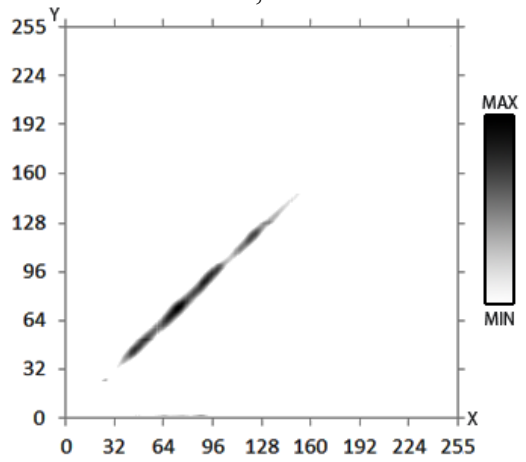
Newspaper sequence.
View X is camera 4, View Y is camera 3.



Newspaper sequence.
View X is camera 4, View Y is camera 5.



Newspaper sequence.
View X is camera 6, View Y is camera 5.



Newspaper sequence.
View X is camera 6, View Y is camera 7.

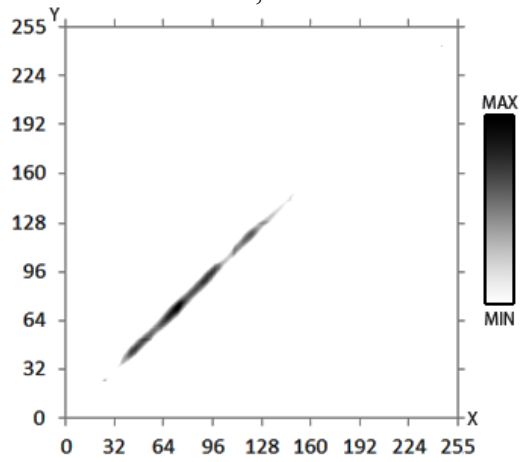
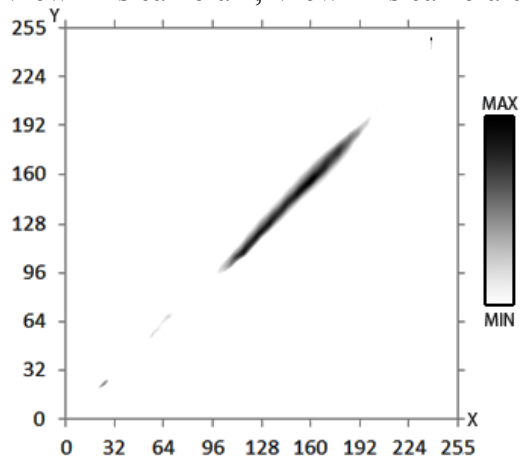
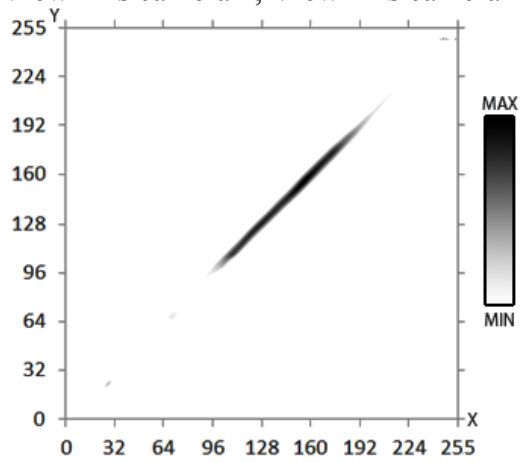


Fig. 107. Graphs of 2-dimentional histograms of luminance values (in logarithmic gray-level scale) of corresponding pixels in the views: (X=2 Y=1,3) (X=4, Y=3,5) (X=6, Y=5,8) of Newspaper sequence.

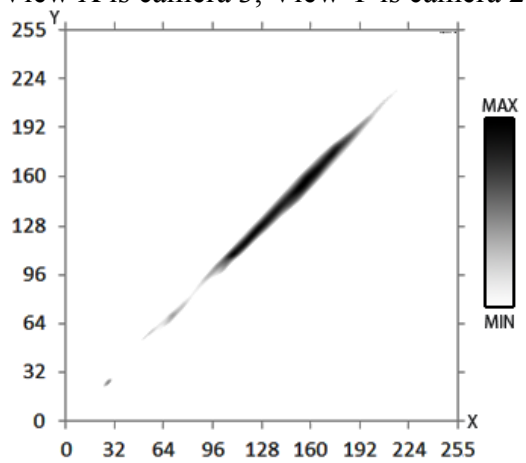
Balloons sequence.
View X is camera 1, View Y is camera 0.



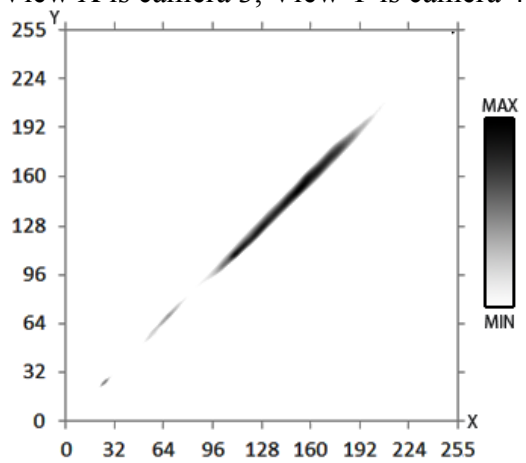
Balloons sequence.
View X is camera 1, View Y is camera 2.



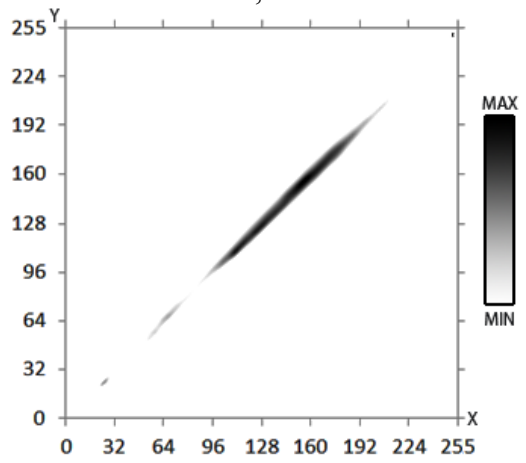
Balloons sequence.
View X is camera 3, View Y is camera 2.



Balloons sequence.
View X is camera 3, View Y is camera 4.



Balloons sequence.
View X is camera 5, View Y is camera 4.



Balloons sequence.
View X is camera 5, View Y is camera 6.

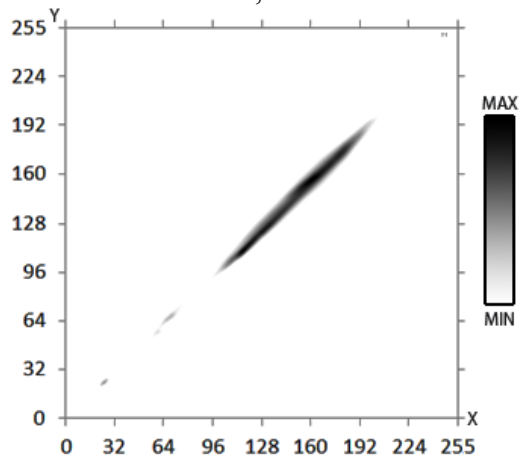
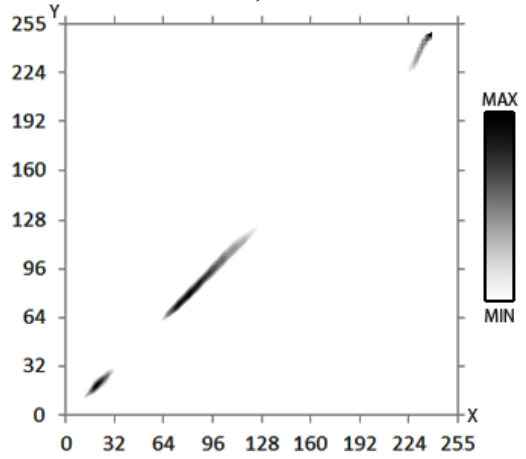
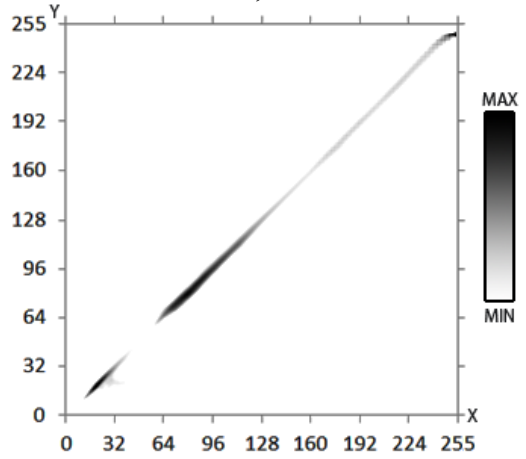


Fig. 108. Graphs of 2-dimentional histograms of luminance values (in logarithmic gray-level scale) of corresponding pixels in the views: (X=1 Y=0,2) (X=3, Y=2,4) (X=5, Y=4,6) of Balloons sequence.

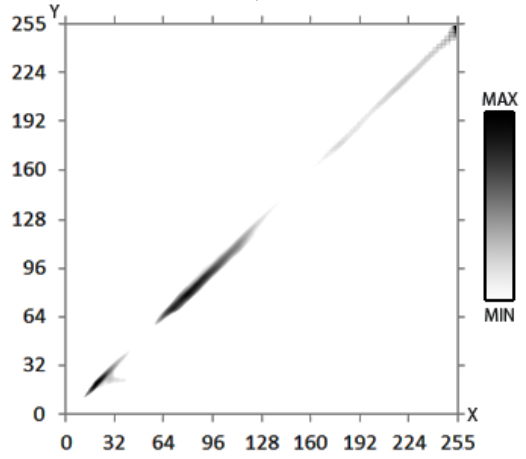
Kendo sequence.
View X is camera 1, View Y is camera 0.



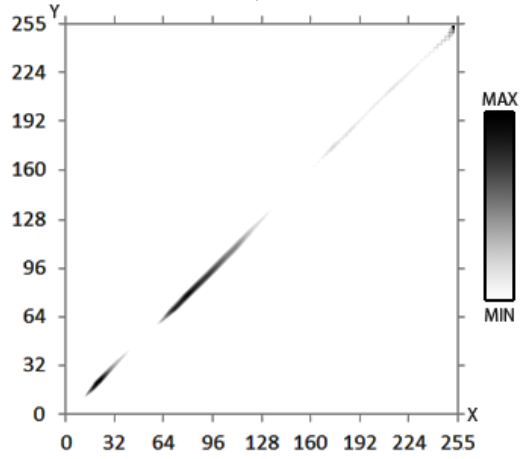
Kendo sequence.
View X is camera 1, View Y is camera 2.



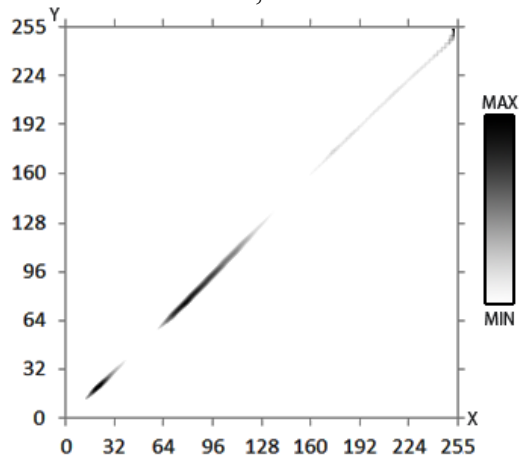
Kendo sequence.
View X is camera 3, View Y is camera 2.



Kendo sequence.
View X is camera 3, View View Y is camera 4.



Kendo sequence.
View X is camera 5, View Y is camera 4.



Kendo sequence.
View X is camera 5, View Y is camera 6.

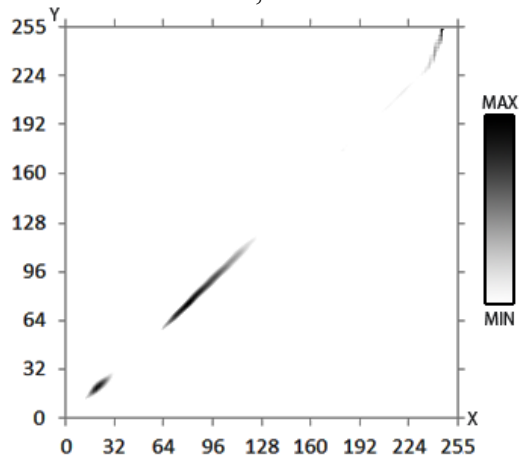
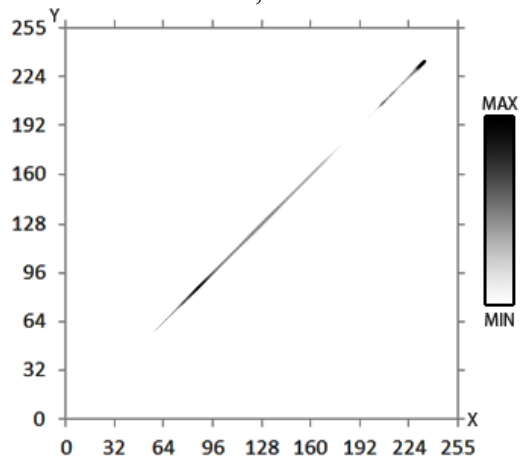
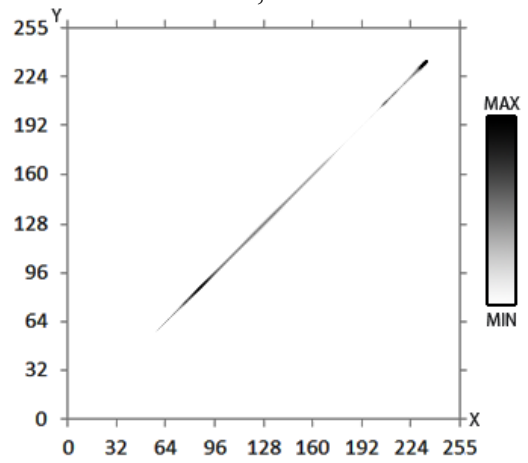


Fig. 109. Graphs of 2-dimentional histograms of luminance values (in logarithmic gray-level scale) of corresponding pixels in the views: (X=1 Y=0,2) (X=3, Y=2,4) (X=5, Y=4,6) of Kendo sequence.

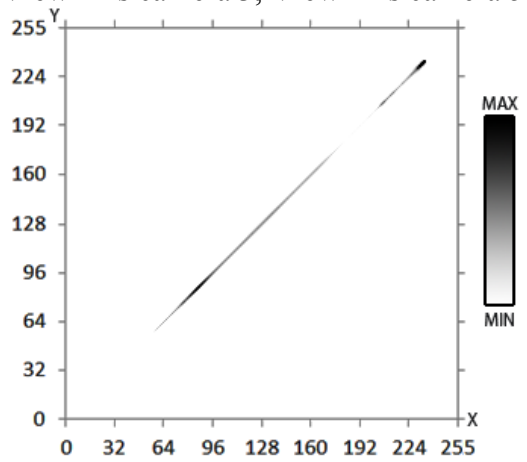
Computer-generated Undo Dancer sequence. View X is camera 3, View Y is camera 1.



Computer-generated Undo Dancer sequence. View X is camera 3, View Y is camera 2.



Computer-generated Undo Dancer sequence. View X is camera 3, View Y is camera 5.



Computer-generated Undo Dancer sequence. View X is camera 3, View Y is camera 9.

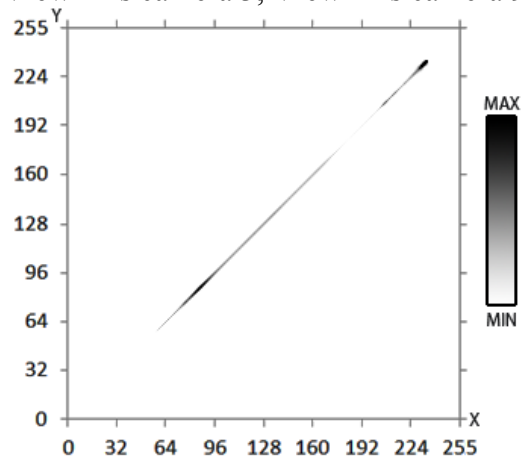
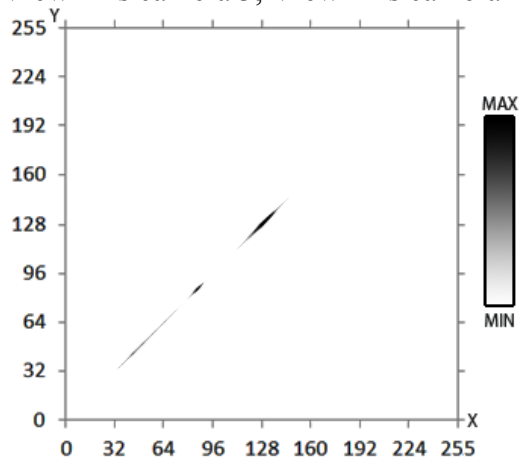
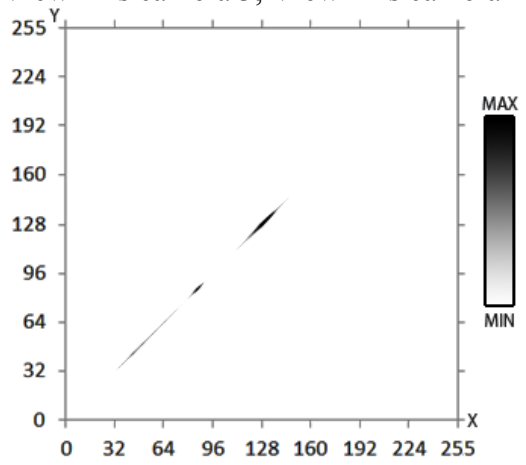


Fig. 110. Graphs of 2-dimentional histograms of luminance values (in logarithmic gray-level scale) of corresponding pixels in the views X=3 Y=1,2,3,5,9 of Undo Dancer sequence. Undo Dancer is a computer-generated sequence and in its case a perfect 2-dimensional histogram can be observed between the views – therefore there is no noise, color profiles are strictly compatible and model of reflectance is Lambertian.

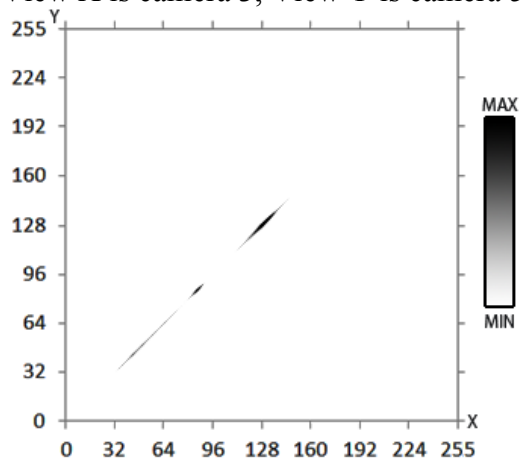
Computer-generated GT Fly sequence.
View X is camera 3, View Y is camera 1.



Computer-generated GT Fly sequence.
View X is camera 3, View Y is camera 2.



Computer-generated GT Fly sequence.
View X is camera 3, View Y is camera 5.



Computer-generated GT Fly sequence.
View X is camera 3, View Y is camera 9

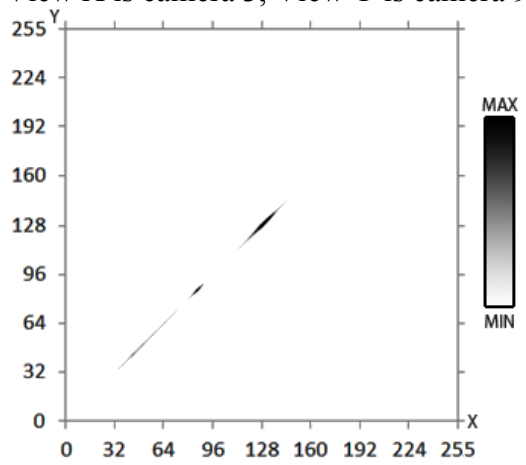


Fig. 111. Graphs of 2-dimentional histograms of luminance values (in logarithmic gray-level scale) of corresponding pixels in the views X=3 Y=1,2,3,5,9 of GT Fly sequence.
GT Fly is a computer-generated sequence and in its case a perfect 2-dimensional histogram can be observed between the views – therefore there is no noise, color profiles are strictly compatible and model of reflectance is Lambertian.

Histograms of luminance values

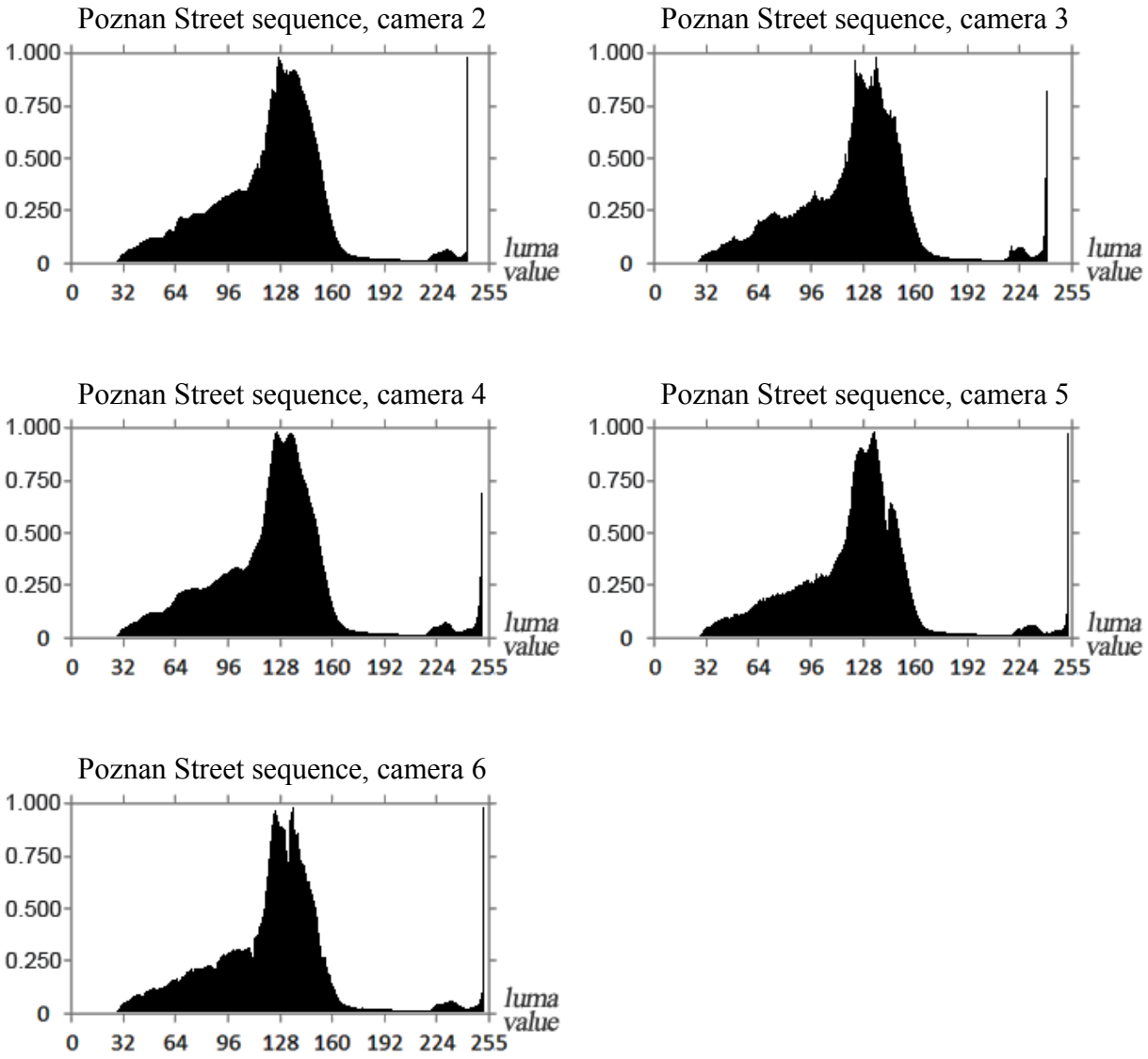


Fig. 112. Histograms of luminance values of pixels in the views X=2,3,4,5,6 of Poznan Street sequence. The graphs have been normalized to range [0;1].

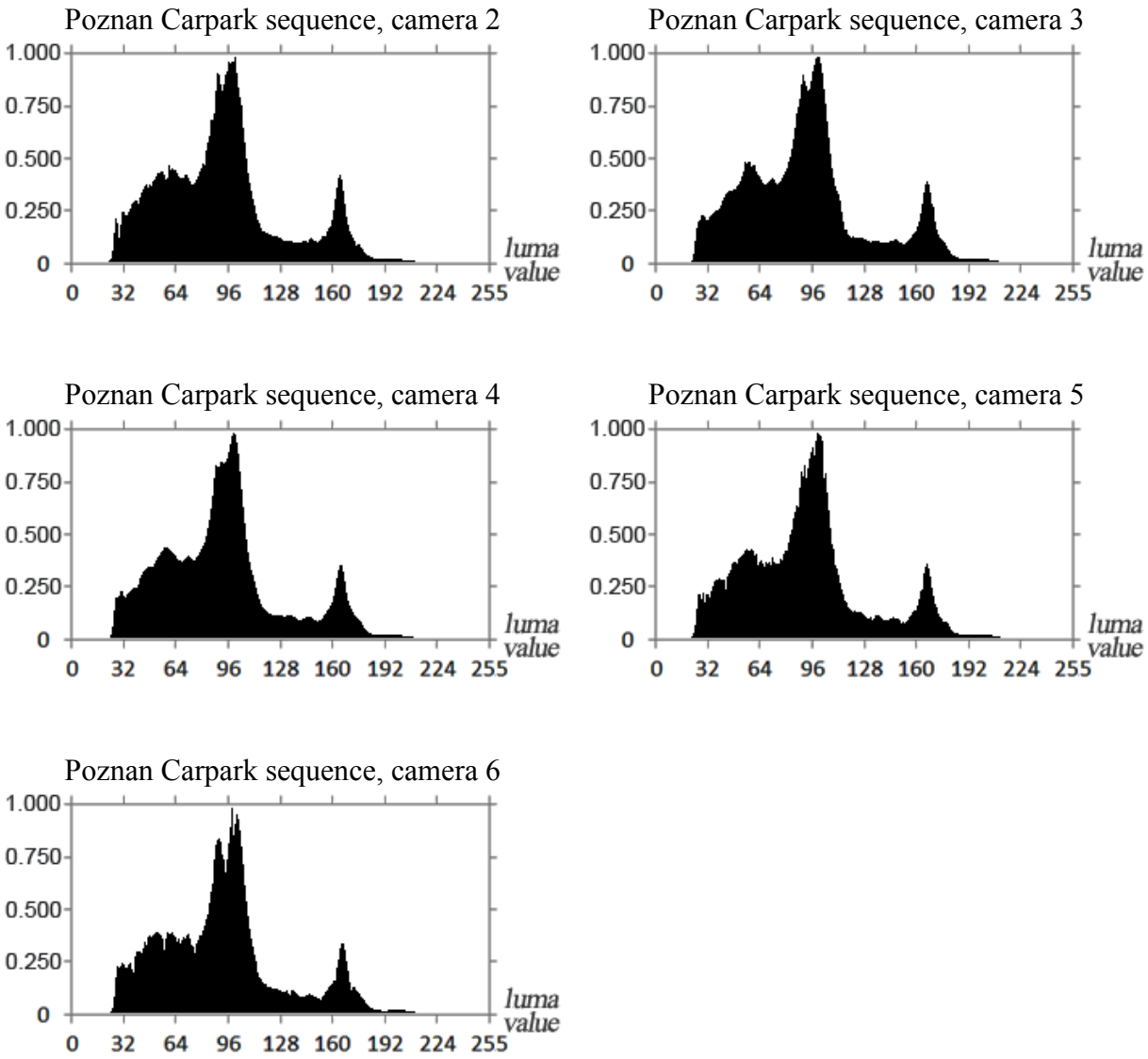


Fig. 113. Histograms of luminance values of pixels in the views X=2,3,4,5,6 of Poznan Carpark sequence. The graphs have been normalized to range [0;1].

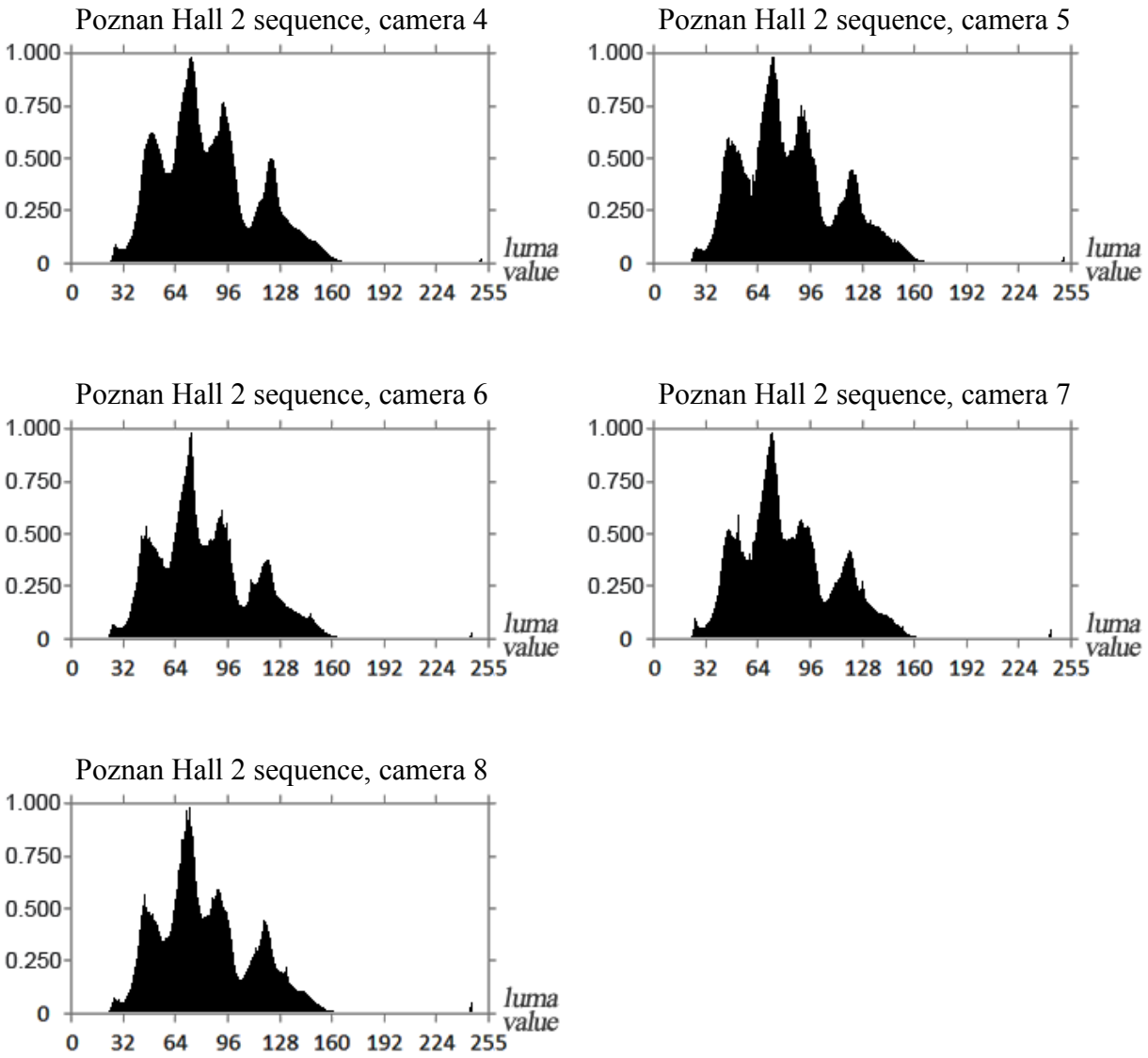


Fig. 114. Histograms of luminance values of pixels in the views X=4,5,6,7,8 of Poznan Hall 2 sequence. The graphs have been normalized to range [0;1].

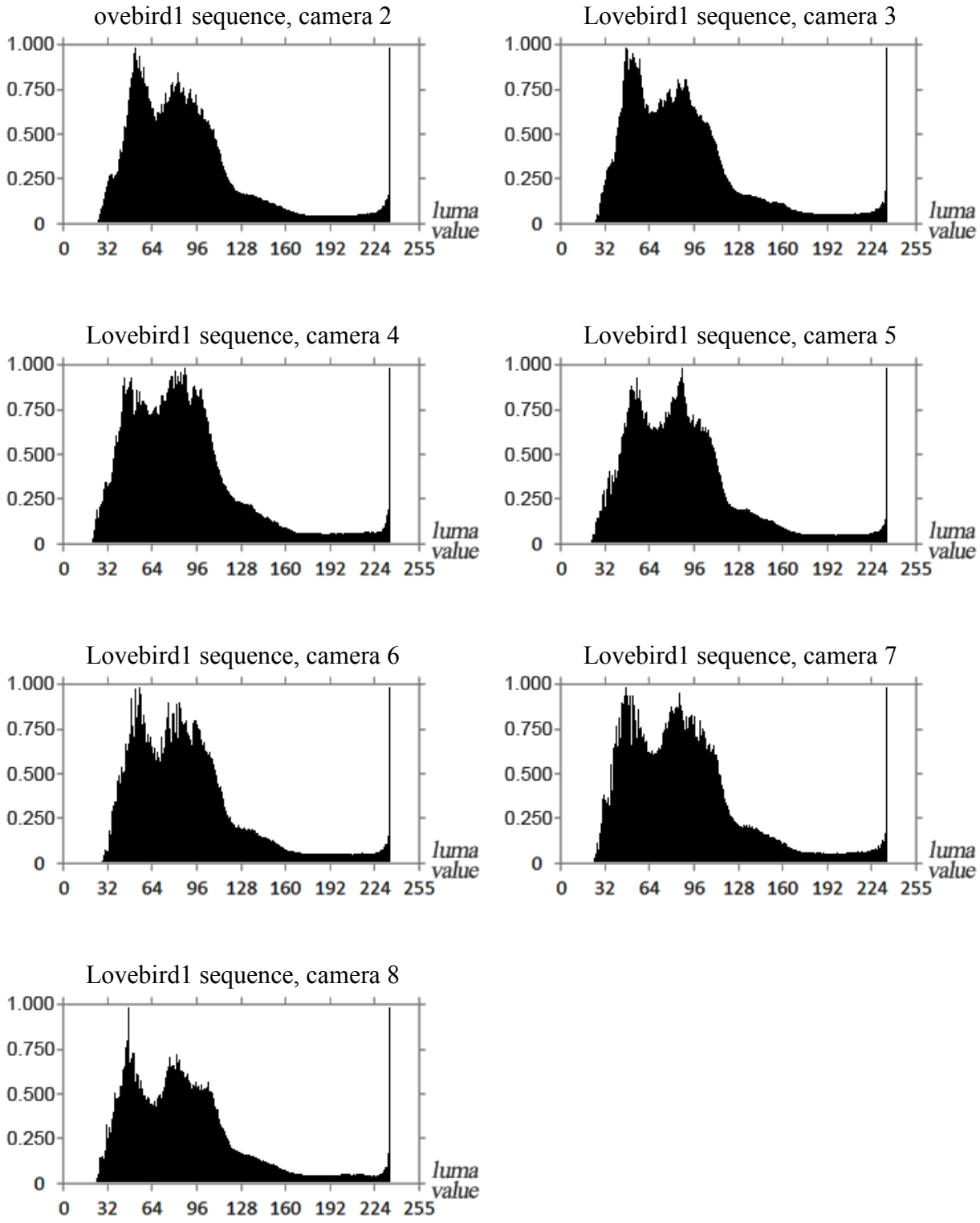


Fig. 115. Histograms of luminance values of pixels in the views X=3,4,5,6,7,8,9 of Lovebird1 sequence. The graphs have been normalized to range [0;1].

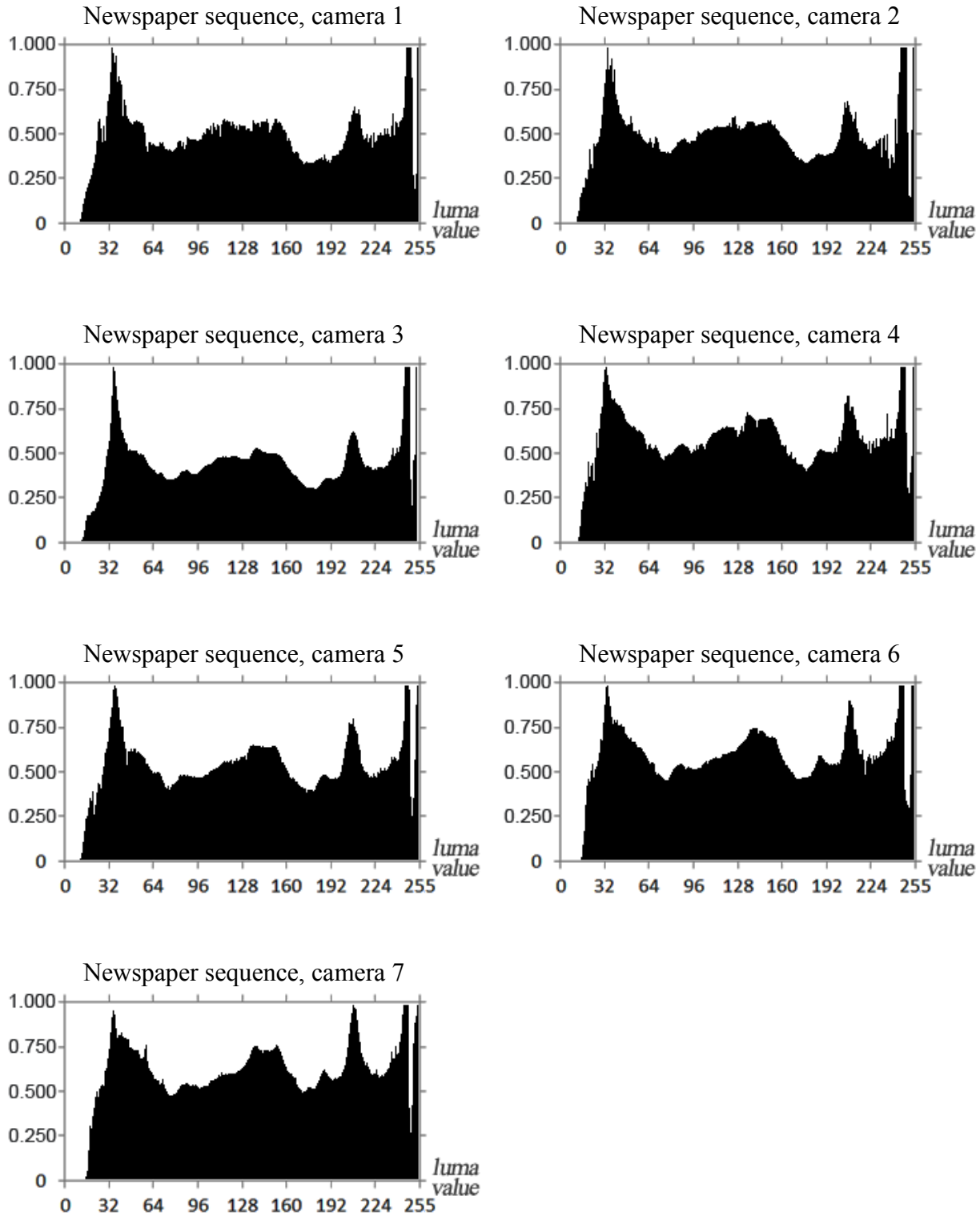


Fig. 116. Histograms of luminance values of pixels in the views X=1,2,3,4,5,6,7 of Newspaper sequence. The graphs have been normalized to range [0;1].

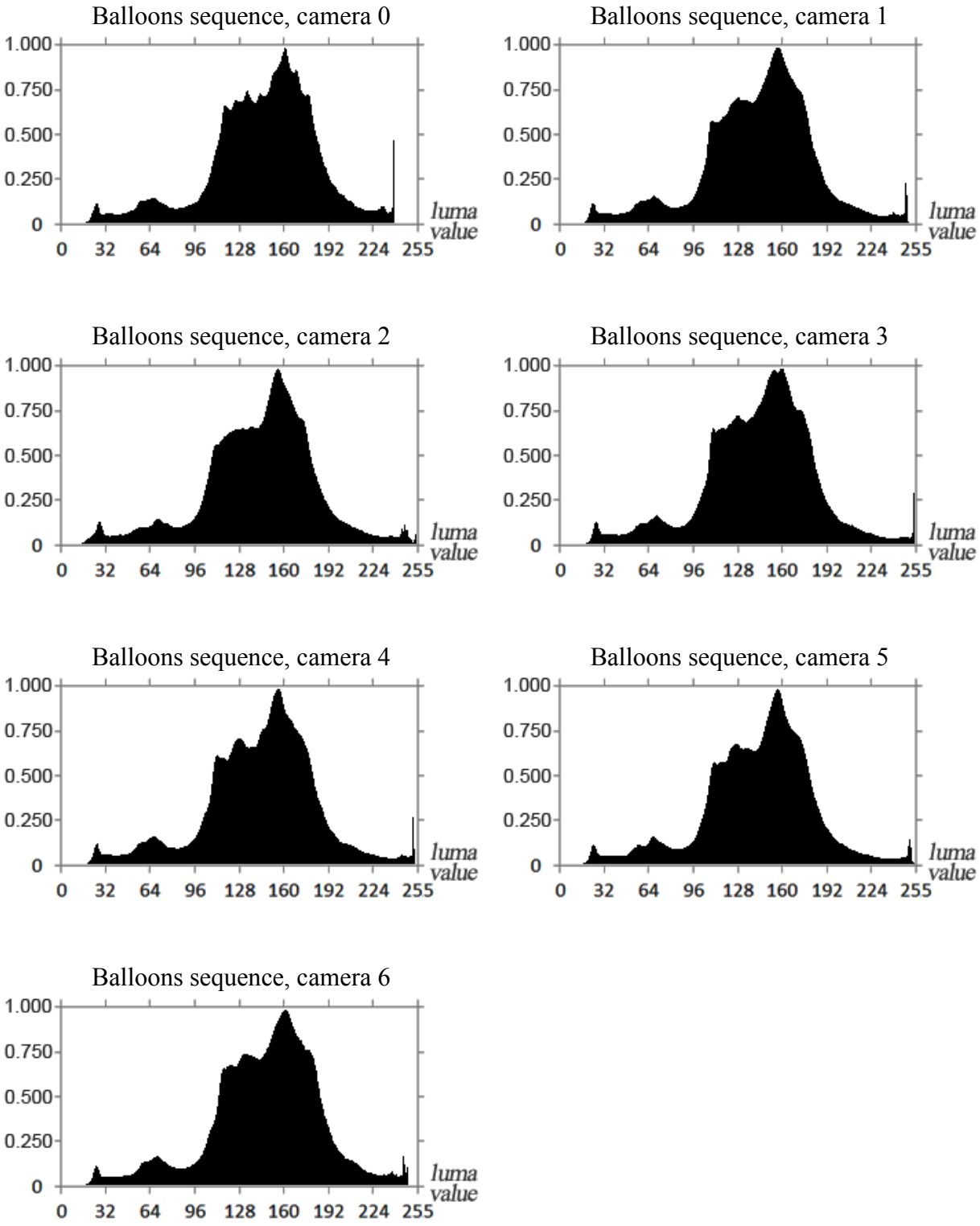


Fig. 117. Histograms of luminance values of pixels in the views X=0,1,2,3,4,5,6 of Balloons sequence. The graphs have been normalized to range [0;1].

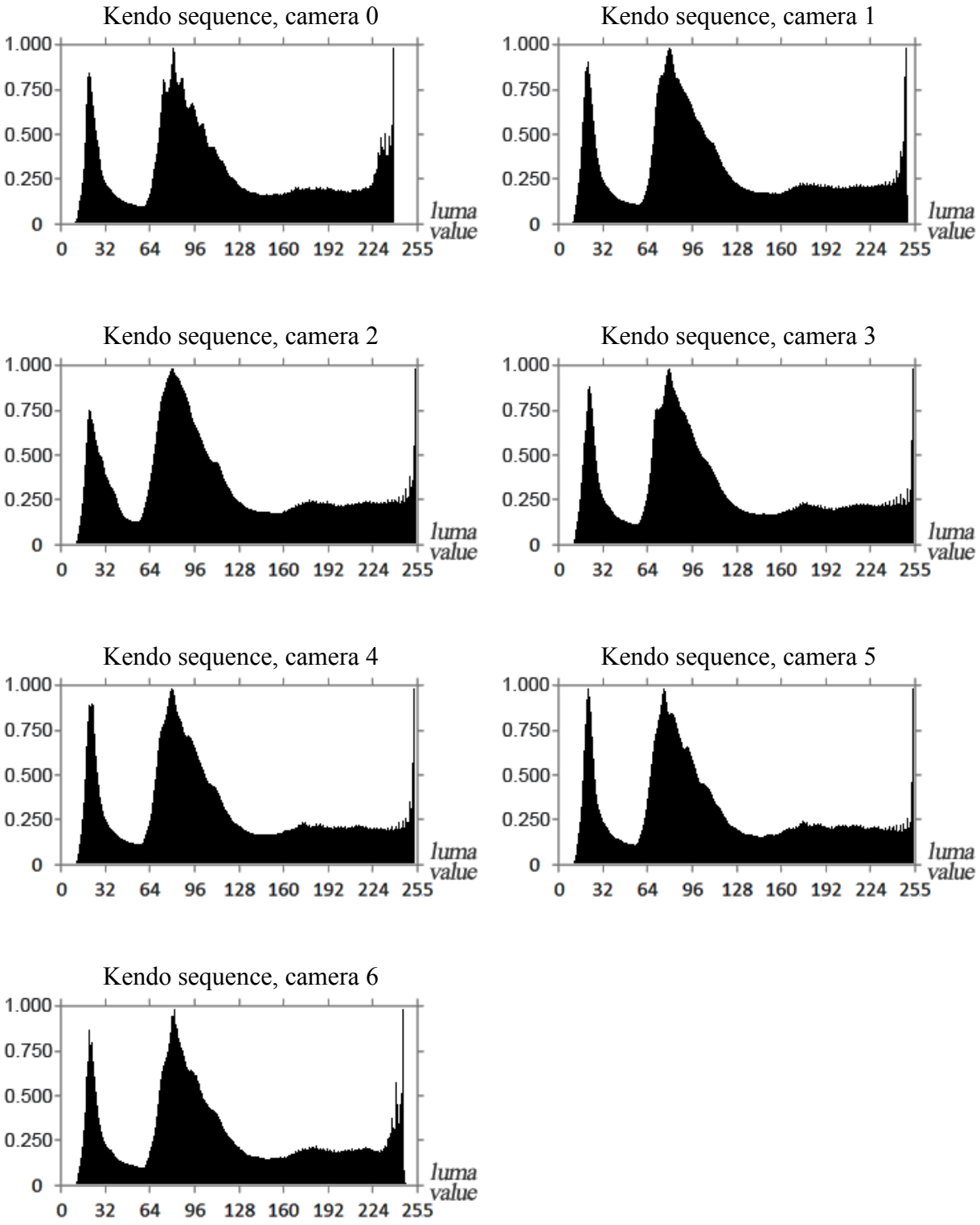


Fig. 118. Histograms of luminance values of pixels in the views X=0,1,2,3,4,5,6 of Kendo sequence. The graphs have been normalized to range [0;1].

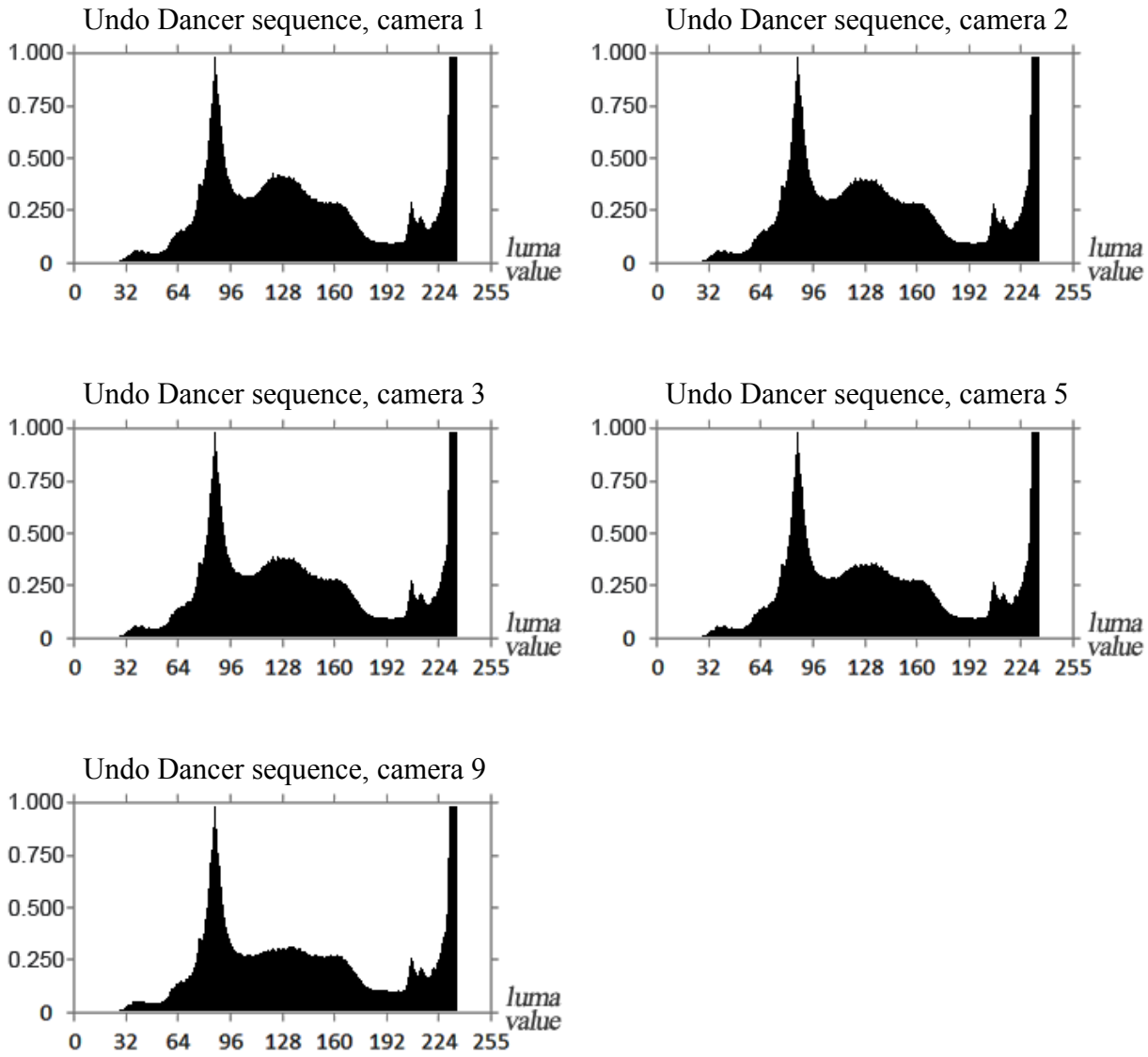


Fig. 119. Histograms of luminance values of pixels in the views X=1,2,3,5,9 of Undo Dancer sequence. The graphs have been normalized to range [0;1].

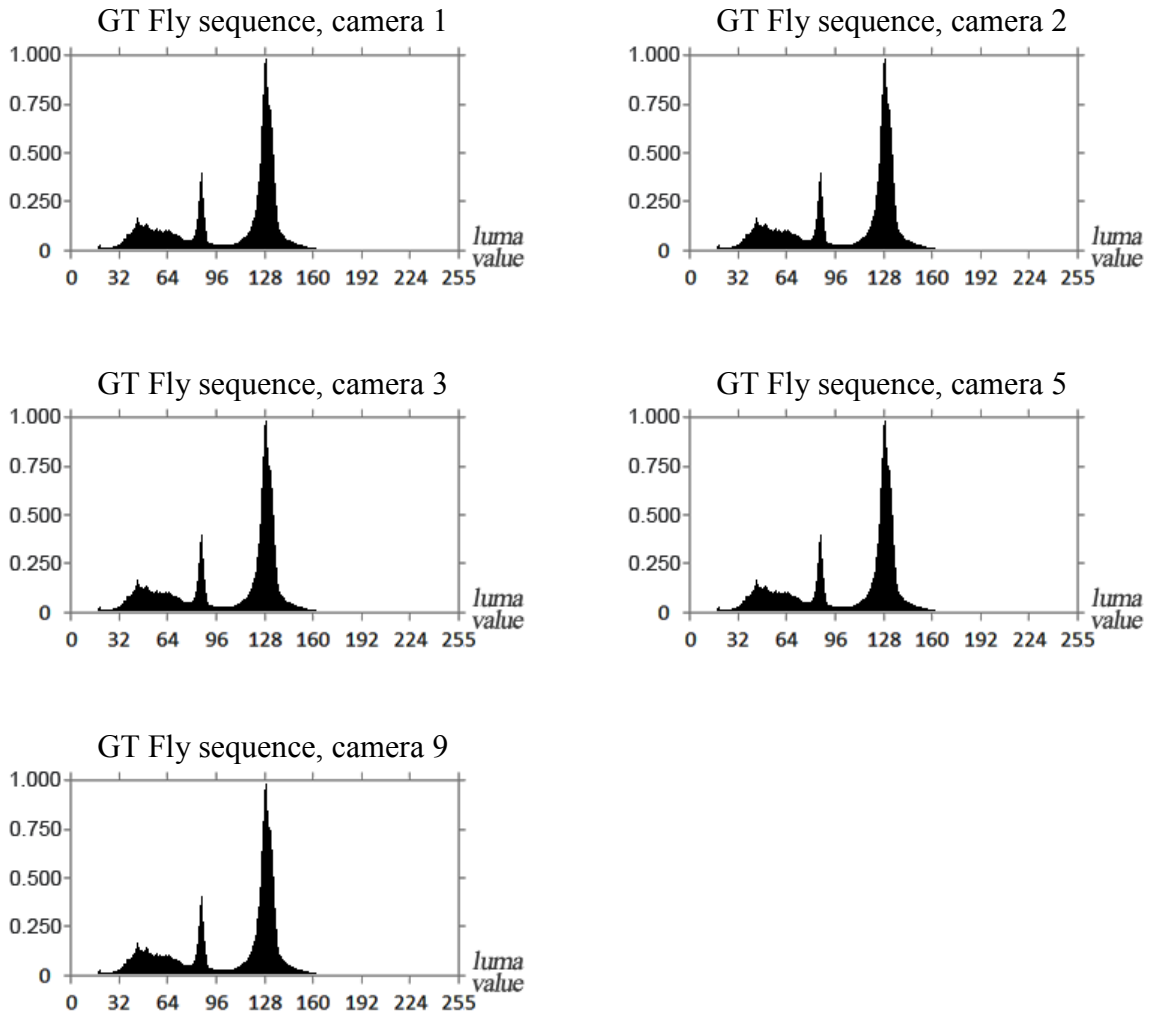
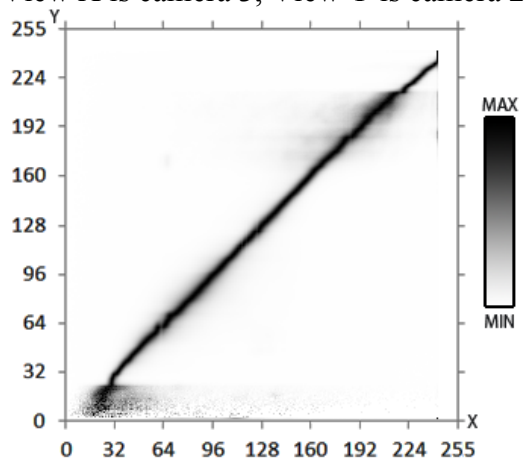


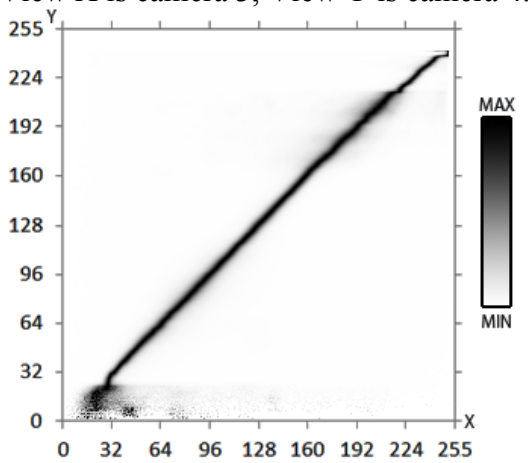
Fig. 120. Histograms of luminance values of pixels in the views X=1,2,3,5,9 of GT Fly sequence. The graphs have been normalized to range [0;1].

Normalized 2-D histograms of luminance values in neighboring views

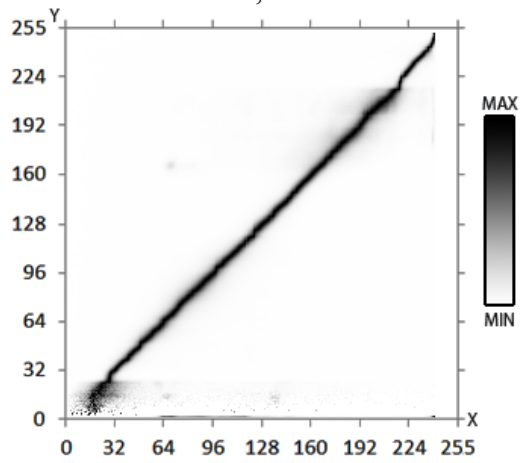
Poznan Street sequence.
View X is camera 3, View Y is camera 2.



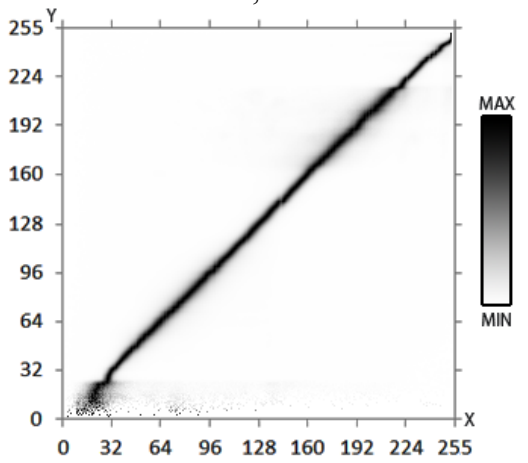
Poznan Street sequence.
View X is camera 3, View Y is camera 4.



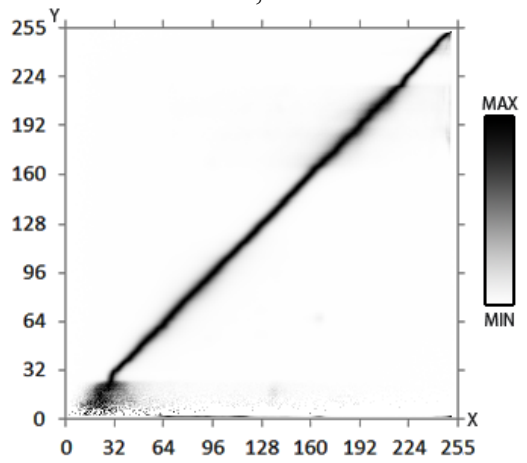
Poznan Street sequence.
View X is camera 4, View Y is camera 3.



Poznan Street sequence.
View X is camera 4, View Y is camera 5.



Poznan Street sequence.
View X is camera 5, View Y is camera 4.



Poznan Street sequence.
View X is camera 5, View Y is camera 6.

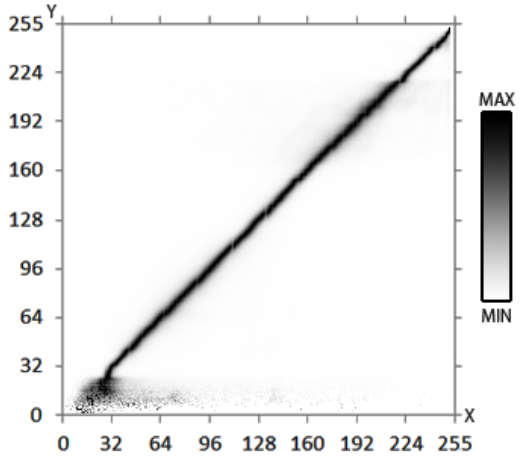
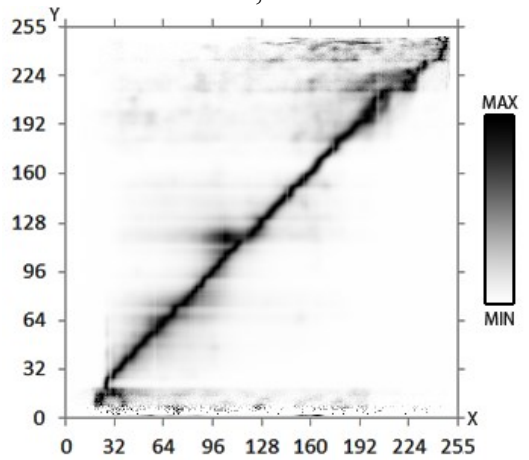
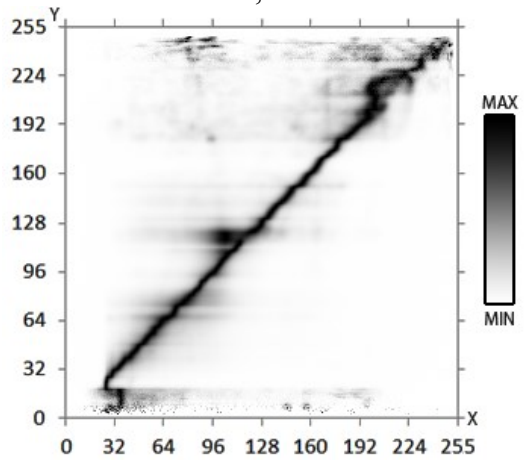


Fig. 121. Graphs of 2-dimentional histograms of luminance values of corresponding pixels in the views: (X=3 Y=2,4) (X=4, Y=3,5) (X=5, Y=4,6) of Poznan Street sequence, normalized with respect to 1-dimensional histogram of luminance values in view X.

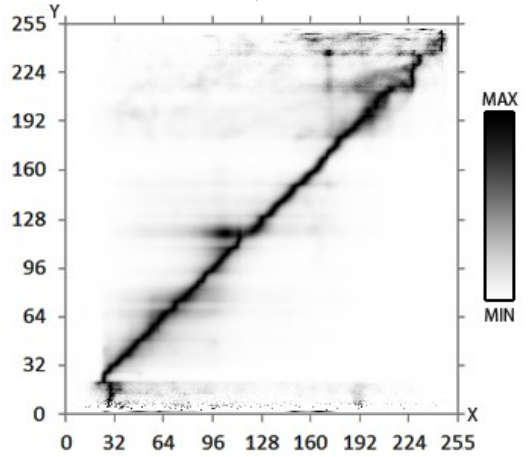
Poznan Carpark sequence.
View X is camera 3, View Y is camera 2.



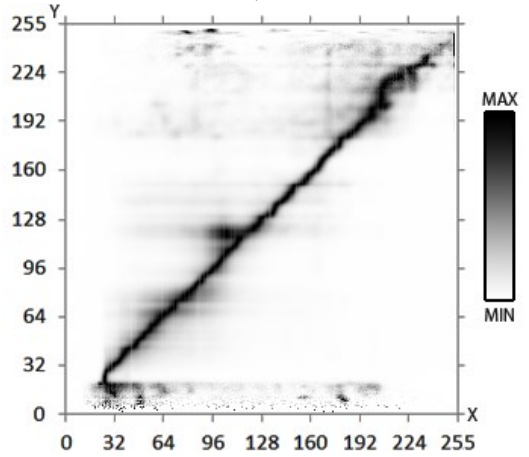
Poznan Carpark sequence.
View X is camera 3, View Y is camera 4.



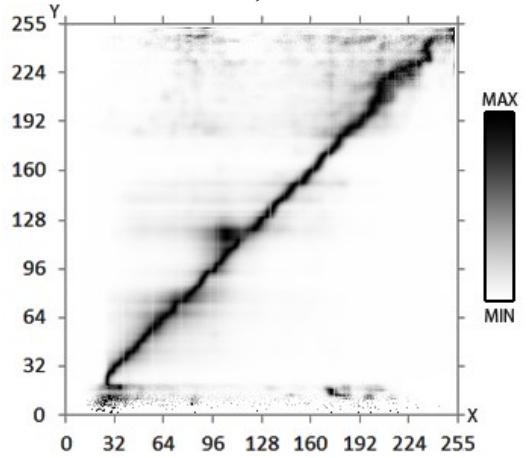
Poznan Carpark sequence.
View X is camera 4, View Y is camera 3.



Poznan Carpark sequence.
View X is camera 4, View Y is camera 5.



Poznan Carpark sequence.
View X is camera 5, View Y is camera 4.



Poznan Carpark sequence.
View X is camera 5, View Y is camera 6.

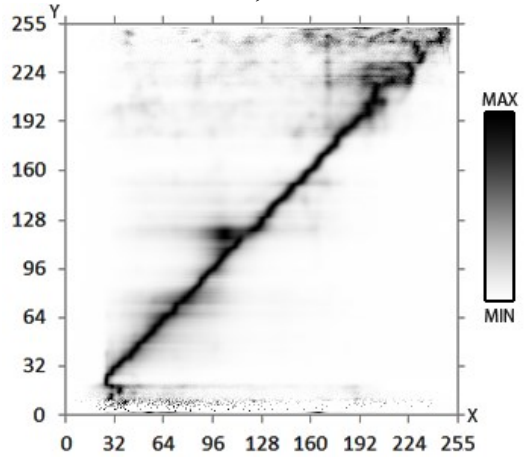
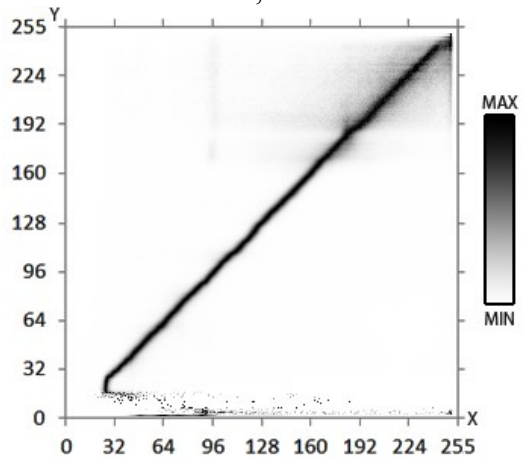
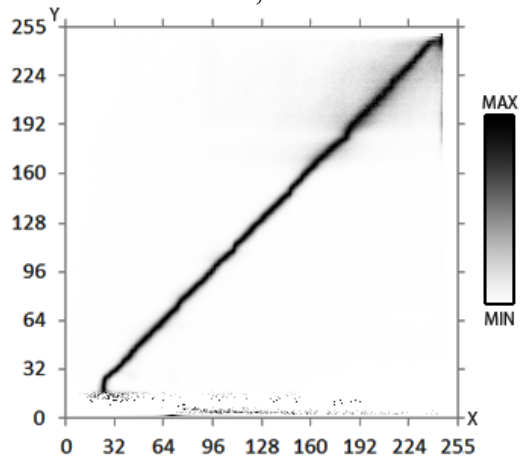


Fig. 122. Graphs of 2-dimentional histograms of luminance values of corresponding pixels in the views: (X=3 Y=2,4) (X=4, Y=3,5) (X=5, Y=4,6) of Poznan Carpark sequence, normalized with respect to 1-dimensional histogram of luminance values in view X.

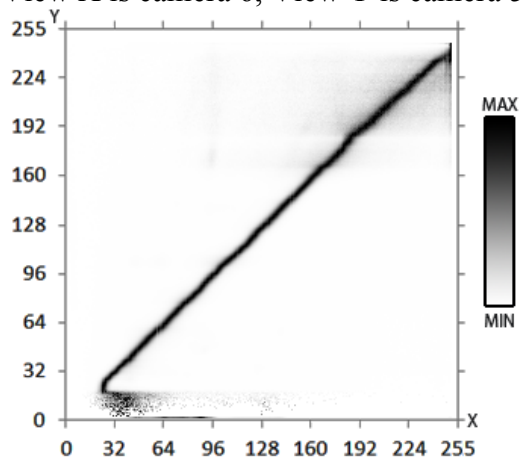
Poznan Hall2 sequence.
View X is camera 5, View Y is camera 4.



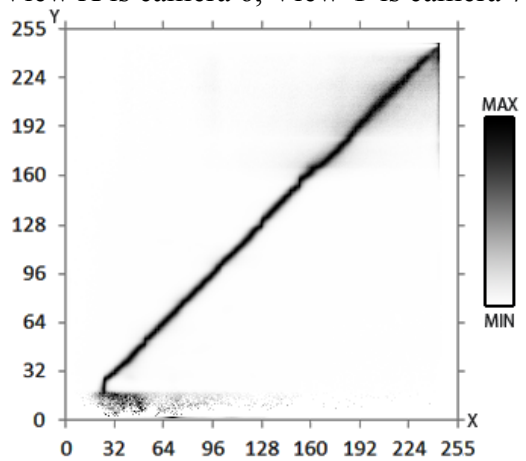
Poznan Hall2 sequence.
View X is camera 5, View Y is camera 6.



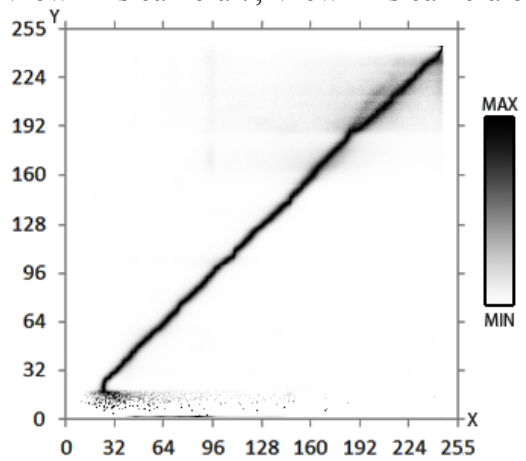
Poznan Hall2 sequence.
View X is camera 6, View Y is camera 5.



Poznan Hall2 sequence.
View X is camera 6, View Y is camera 7.



Poznan Hall2 sequence.
View X is camera 7, View Y is camera 6.



Poznan Hall2 sequence.
View X is camera 7, View Y is camera 8.

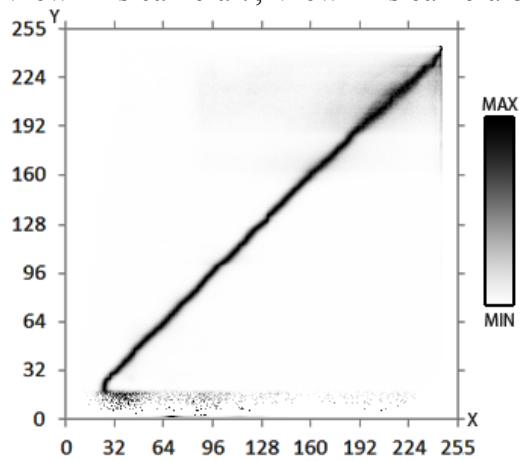
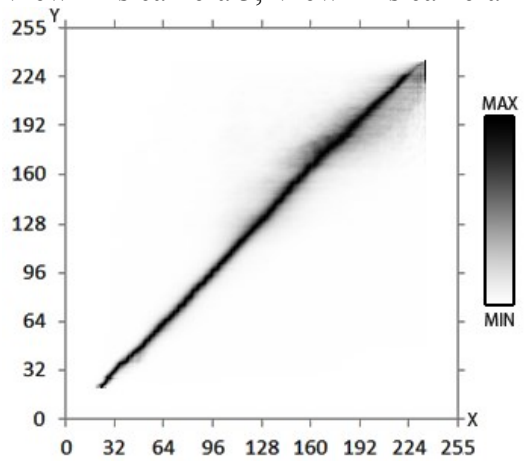
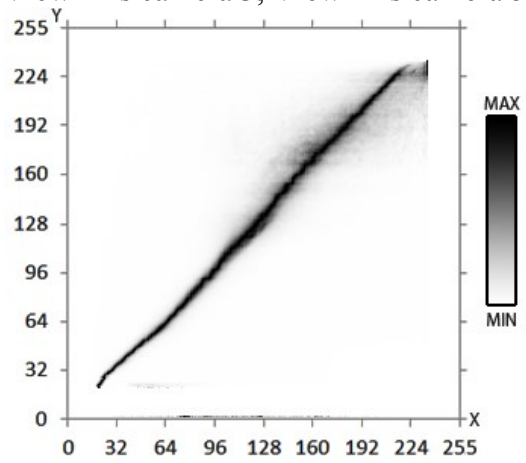


Fig. 123. Graphs of 2-dimentional histograms of luminance of corresponding pixels in the views: (X=5 Y=4,6) (X=6, Y=5,7) (X=7, Y=6,8) of Poznan Hall2 sequence, normalized with respect to 1-dimensional histogram of luminance values in view X.

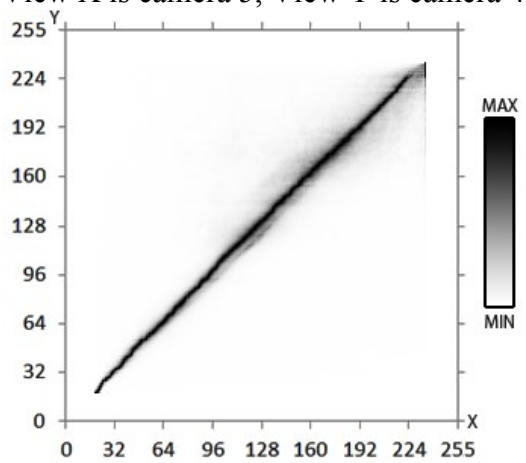
Lovebird1 sequence.
View X is camera 3, View Y is camera 2.



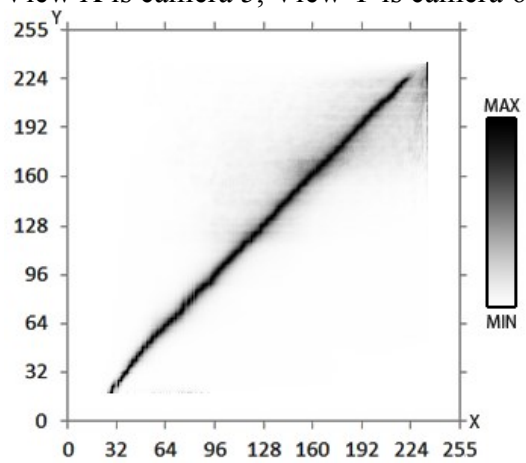
Lovebird1 sequence.
View X is camera 3, View Y is camera 5.



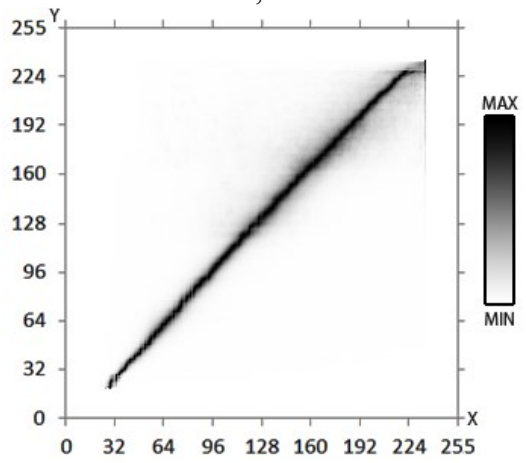
Lovebird1 sequence.
View X is camera 5, View Y is camera 4.



Lovebird1 sequence.
View X is camera 5, View Y is camera 6.



Lovebird1 sequence.
View X is camera 7, View Y is camera 6.



Lovebird1 sequence.
View X is camera 7, View Y is camera 8.

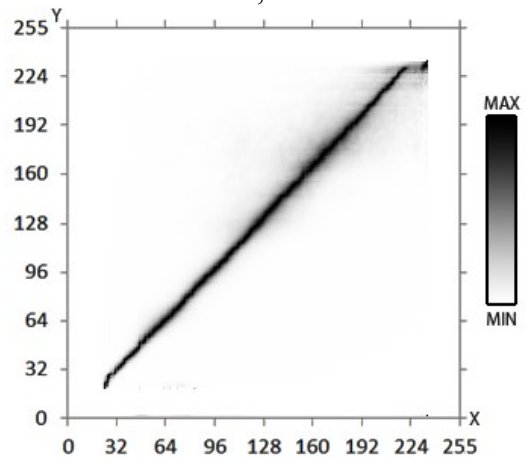
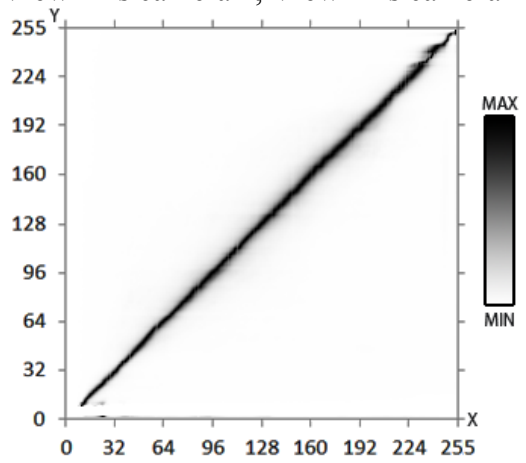
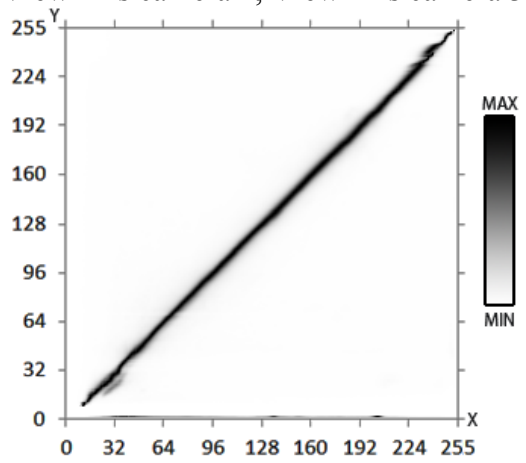


Fig. 124. Graphs of 2-dimentional histograms of luminance of corresponding pixels in the views: (X=4 Y=3,5) (X=6, Y=5,7) (X=8, Y=7,9) of Lovebird1 sequence, normalized with respect to 1-dimensional histogram of luminance values in view X.

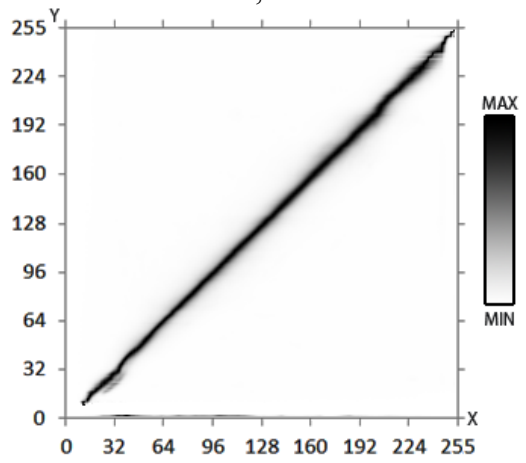
Newspaper sequence.
View X is camera 2, View Y is camera 1.



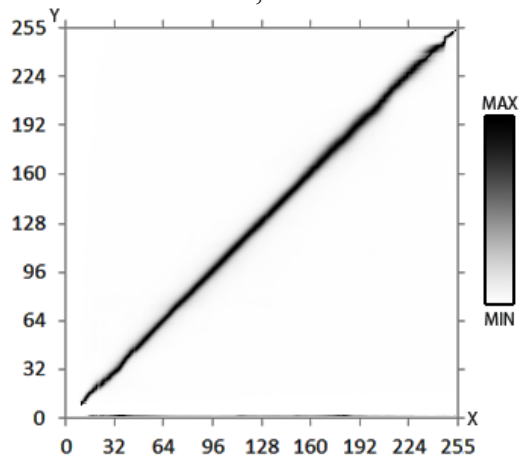
Newspaper sequence.
View X is camera 2, View Y is camera 3.



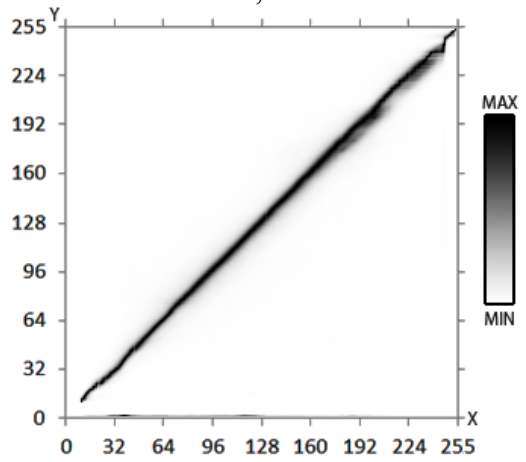
Newspaper sequence.
View X is camera 4, View Y is camera 3.



Newspaper sequence.
View X is camera 4, View Y is camera 5.



Newspaper sequence.
View X is camera 6, View Y is camera 5.



Newspaper sequence.
View X is camera 6, View Y is camera 7.

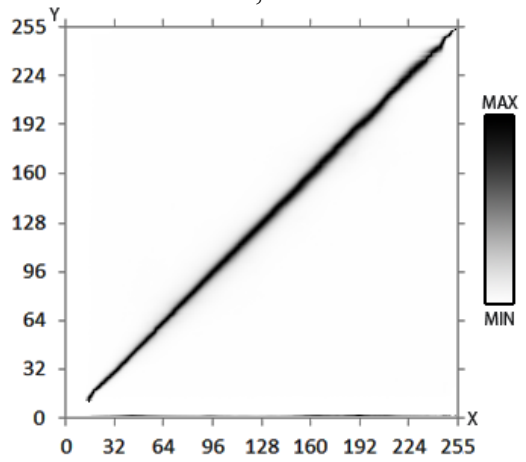
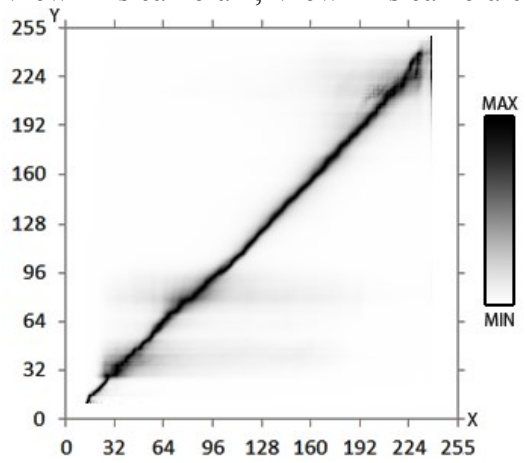
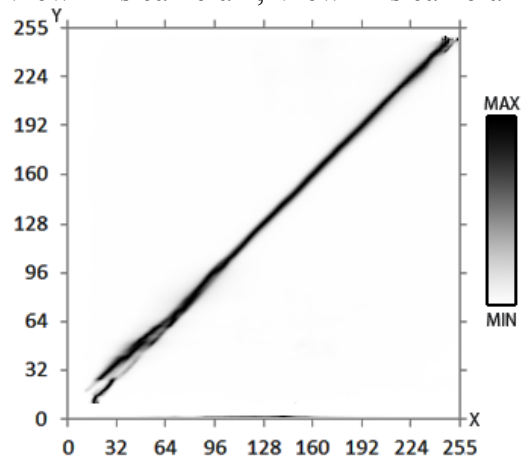


Fig. 125. Graphs of 2-dimentional histograms of luminance values of corresponding pixels in the views: (X=2 Y=1,3) (X=4, Y=3,5) (X=6, Y=5,8) of Newspaper sequence, normalized with respect to 1-dimensional histogram of luminance values in view X.

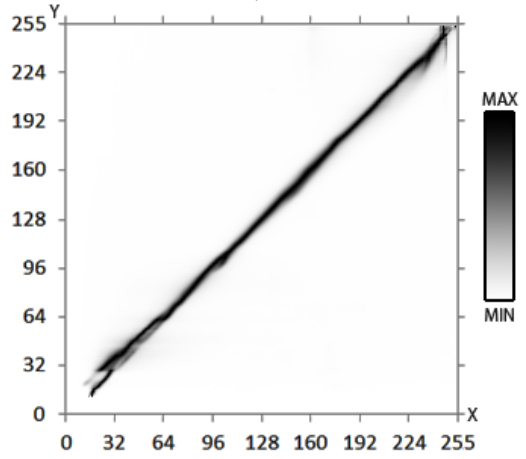
Balloons sequence.
View X is camera 1, View Y is camera 0.



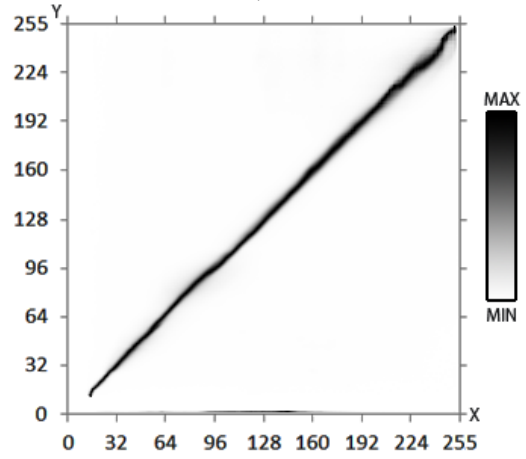
Balloons sequence.
View X is camera 1, View Y is camera 2.



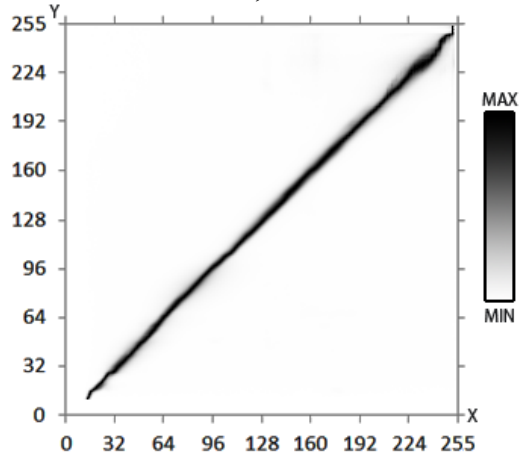
Balloons sequence.
View X is camera 3, View Y is camera 2.



Balloons sequence.
View X is camera 3, View Y is camera 4.



Balloons sequence.
View X is camera 5, View Y is camera 4.



Balloons sequence.
View X is camera 5, View Y is camera 6.

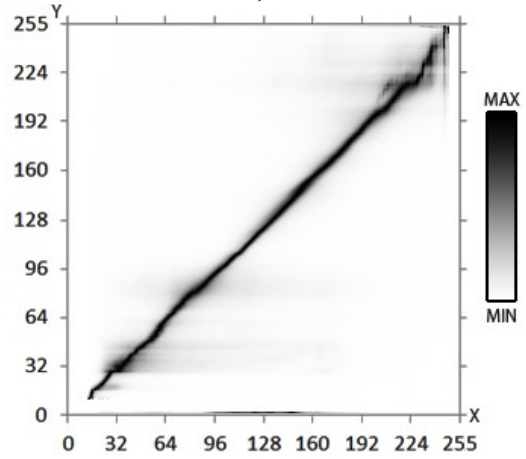
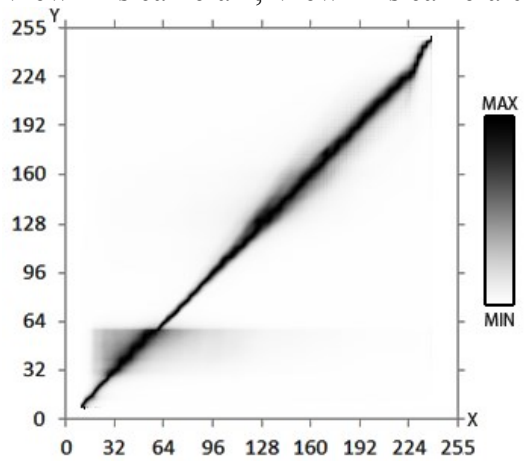
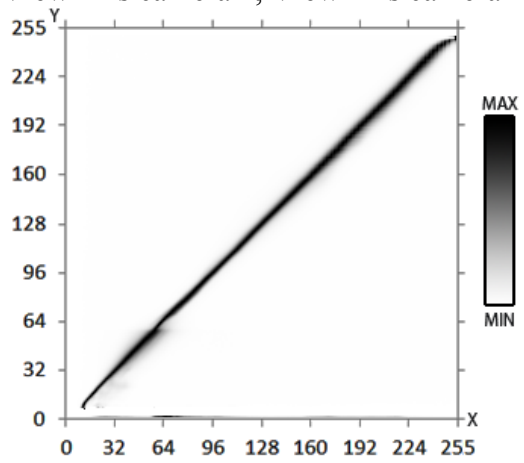


Fig. 126. Graphs of 2-dimentional histograms of luminance values of corresponding pixels in the views: (X=1 Y=0,2) (X=3, Y=2,4) (X=5, Y=4,6) of Balloons sequence, normalized with respect to 1-dimensional histogram of luminance values in view X.

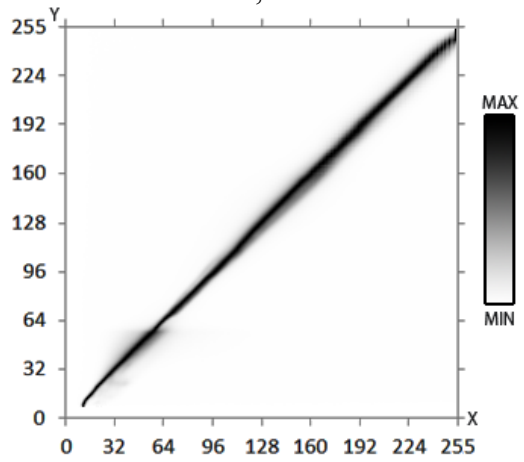
Kendo sequence.
View X is camera 1, View Y is camera 0.



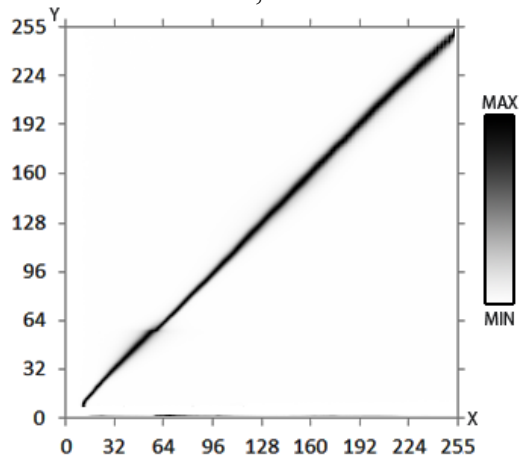
Kendo sequence.
View X is camera 1, View Y is camera 2.



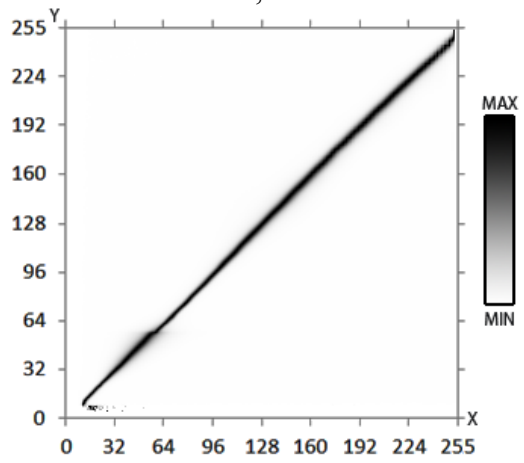
Kendo sequence.
View X is camera 3, View Y is camera 2.



Kendo sequence.
View X is camera 3, View Y is camera 4.



Kendo sequence.
View X is camera 5, View Y is camera 4.



Kendo sequence.
View X is camera 5, View Y is camera 6.

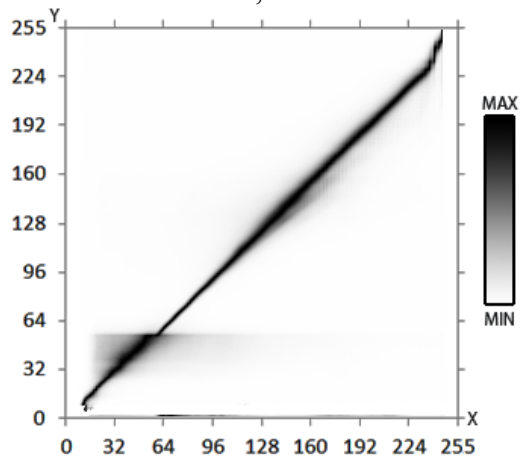
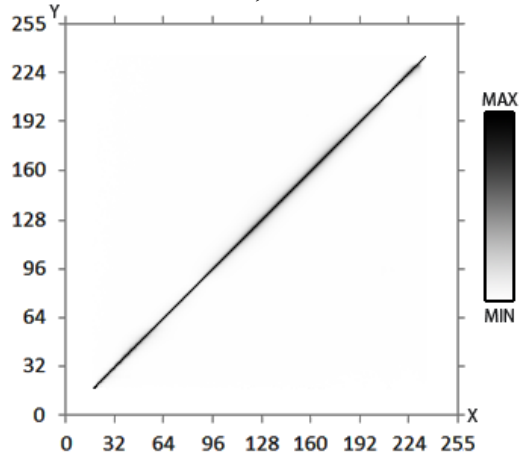
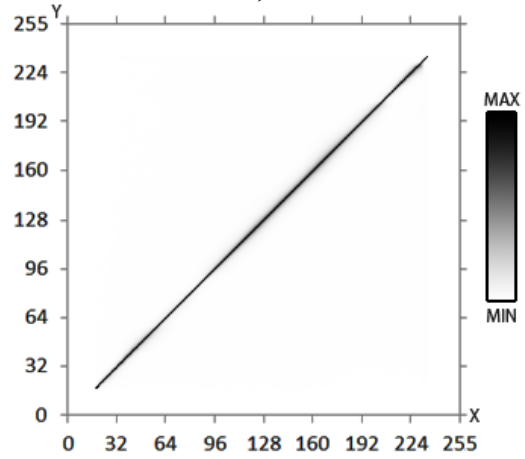


Fig. 127. Graphs of 2-dimentional histograms of luminance of corresponding pixels in the views: (X=1 Y=0,2) (X=3, Y=2,4) (X=5, Y=4,6) of Kendo sequence, normalized with respect to 1-dimensional histogram of luminance values in view X.

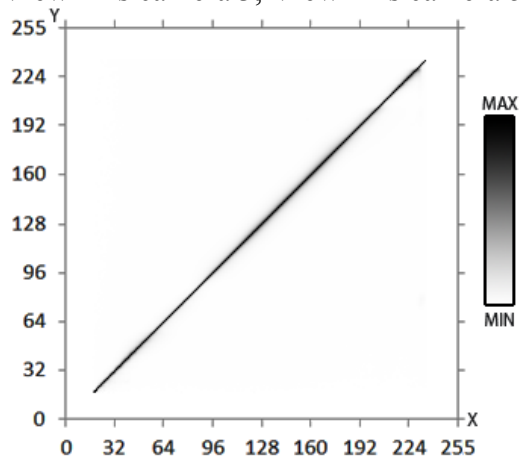
Computer-generated Undo Dancer sequence. View X is camera 3, View Y is camera 1.



Computer-generated Undo Dancer sequence. View X is camera 3, View Y is camera 2.



Computer-generated Undo Dancer sequence. View X is camera 3, View Y is camera 5.



Computer-generated Undo Dancer sequence. View X is camera 3, View Y is camera 9.

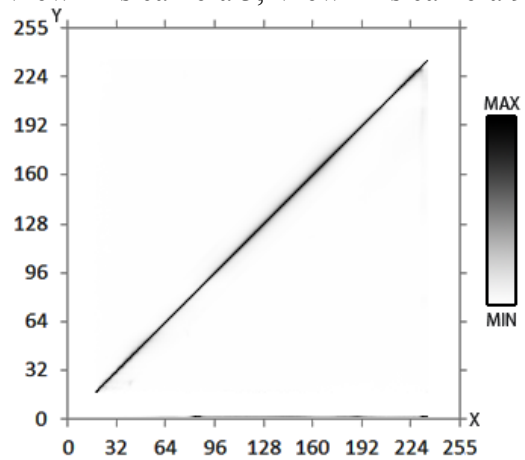
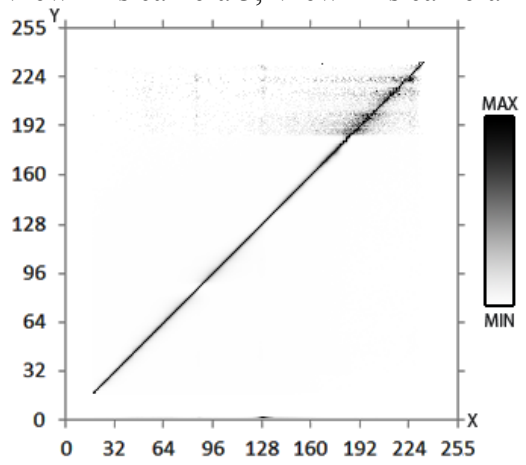
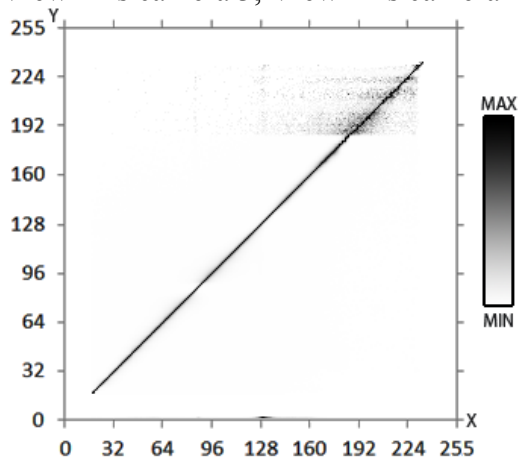


Fig. 128. Graphs of 2-dimentional histograms of luminance values of corresponding pixels in the views: X=3 Y=1,2,3,5,9 of Undo Dancer sequence, normalized with respect to 1-dimensional histogram of luminance values in view X. Undo Dancer is a computer-generated sequence and in its case a perfect 2-dimensional histogram can be observed between the views – therefore there is no noise, color profiles are strictly compatible and model of reflectance is Lambertian.

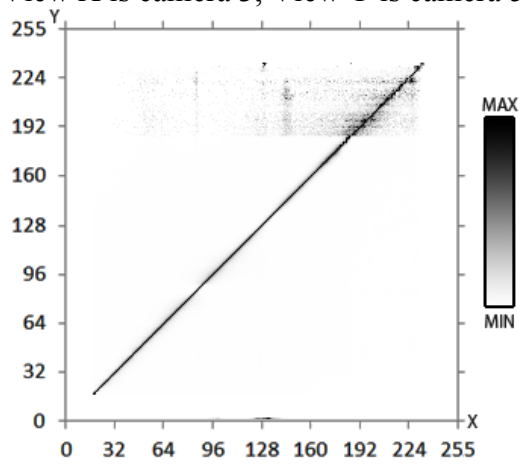
Computer-generated GT Fly sequence.
View X is camera 3, View Y is camera 1.



Computer-generated GT Fly sequence.
View X is camera 3, View Y is camera 2.



Computer-generated GT Fly sequence.
View X is camera 3, View Y is camera 5.



Computer-generated GT Fly sequence.
View X is camera 3, View Y is camera 9.

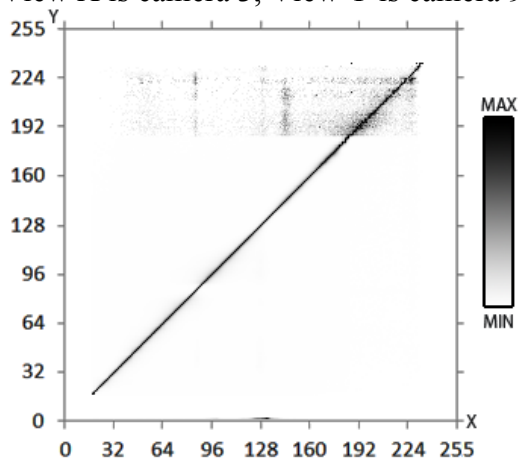


Fig. 129. Graphs of 2-dimentional histograms of luminance values of corresponding pixels in the views: X=3 Y=1,2,3,5,9 of GT Fly sequence, normalized with respect to 1-dimensional histogram of luminance values in view X. GT Fly is a computer-generated sequence and in its case a perfect 2-dimensional histogram can be observed between the views – therefore there is no noise, color profiles are strictly compatible and model of reflectance is Lambertian.

Histograms of normalized disparity values

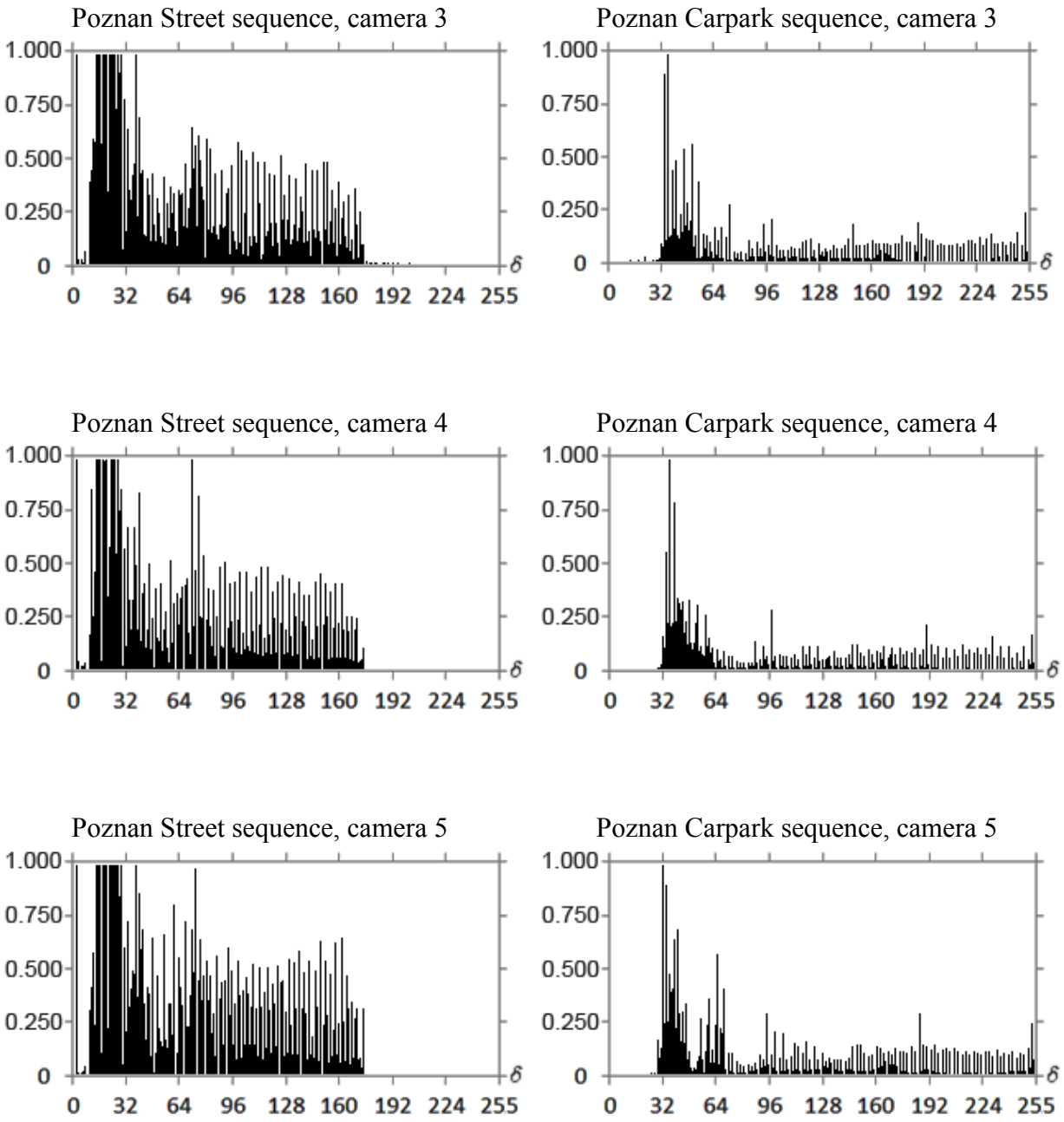


Fig. 130. Histograms of normalized disparity values of pixels in disparity maps in the views 3,4,5 of Poznan Street and Poznan Carpark sequences. The graphs have been normalized to range [0;1]. Please note that the histograms are sparse, as only some of the normalized disparity values are present in the disparity maps.

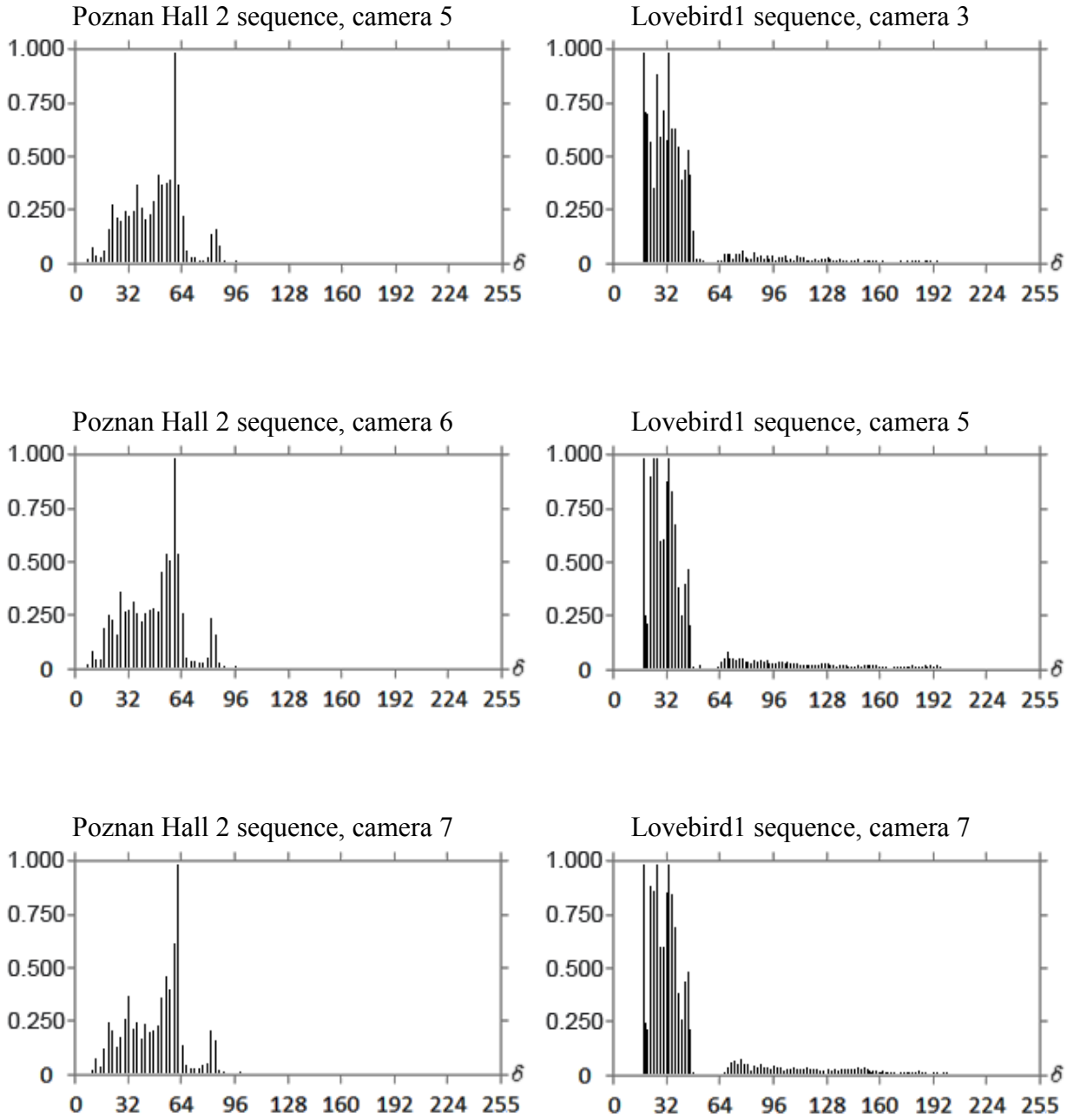


Fig. 131. Histograms of normalized disparity values of pixels in disparity maps in the views 5,6,7 of Poznan Hall 2 sequence and in the views 3,5,7 of Lovebird1 sequence. The graphs have been normalized to range [0;1]. Please note that the histograms are sparse, as only some of the normalized disparity values are present in the disparity maps.

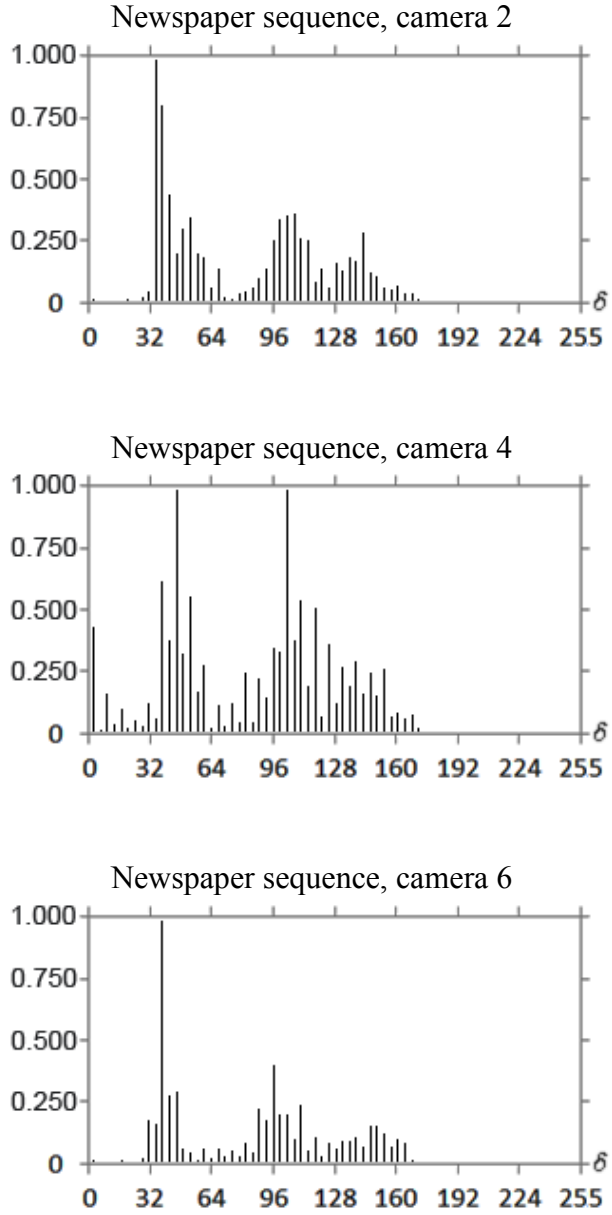


Fig. 132. Histograms of normalized disparity values of pixels in disparity maps in the views 2,4,6 of Newspaper sequence. The graphs have been normalized to range [0;1]. Please note that the histograms are sparse, as only some of the normalized disparity values are present in the disparity maps.

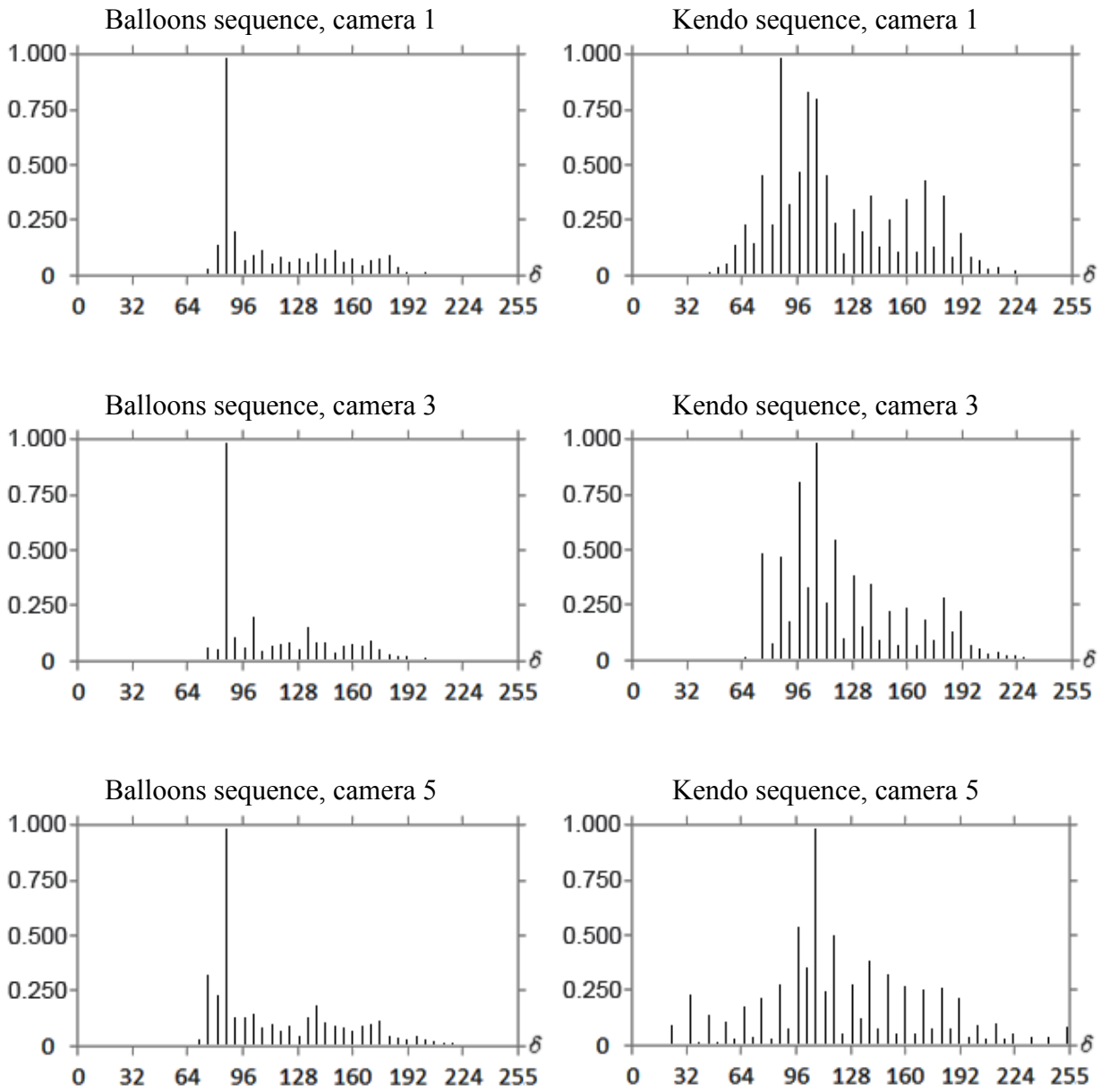


Fig. 133. Histograms of normalized disparity values of pixels in disparity maps in the views 1,3,5 of Ballons and Kendo sequences. The graphs have been normalized to range [0;1]. Please note that the histograms are sparse, as only some of the normalized disparity values are present in the disparity maps.

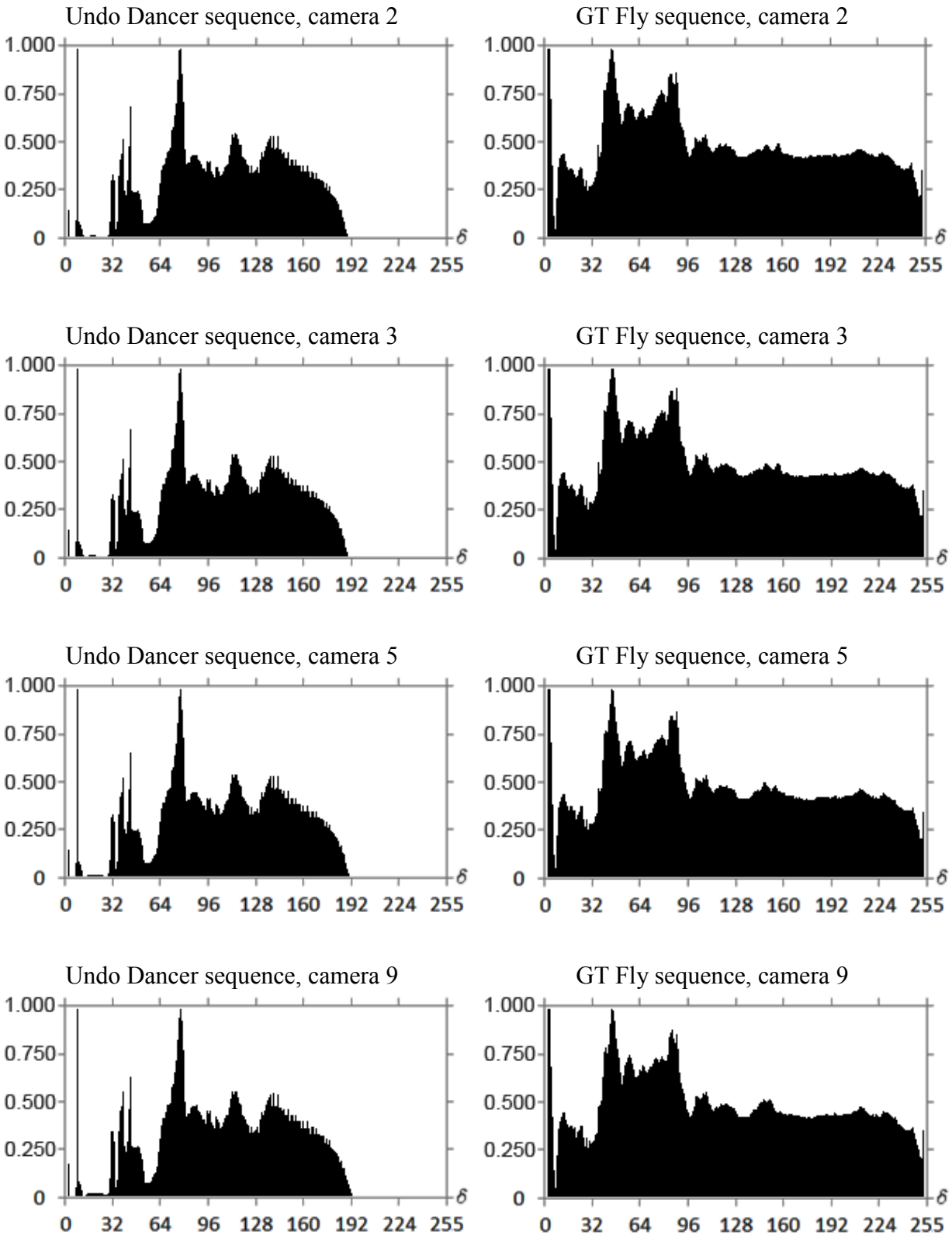
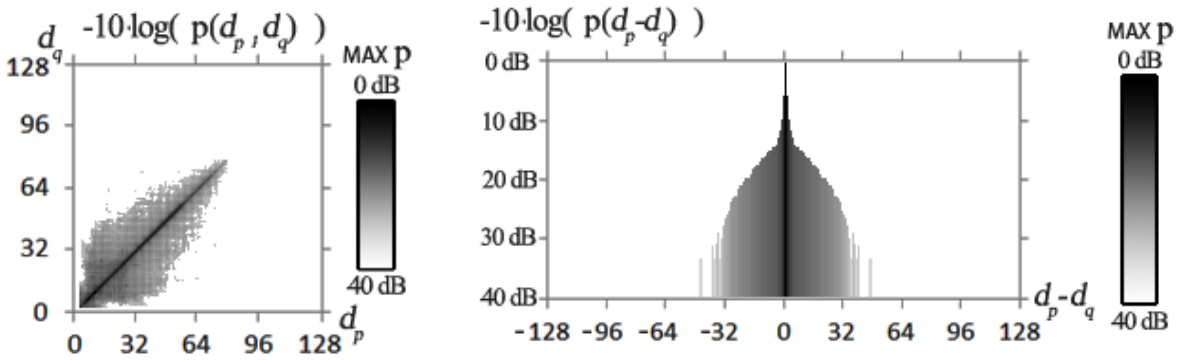


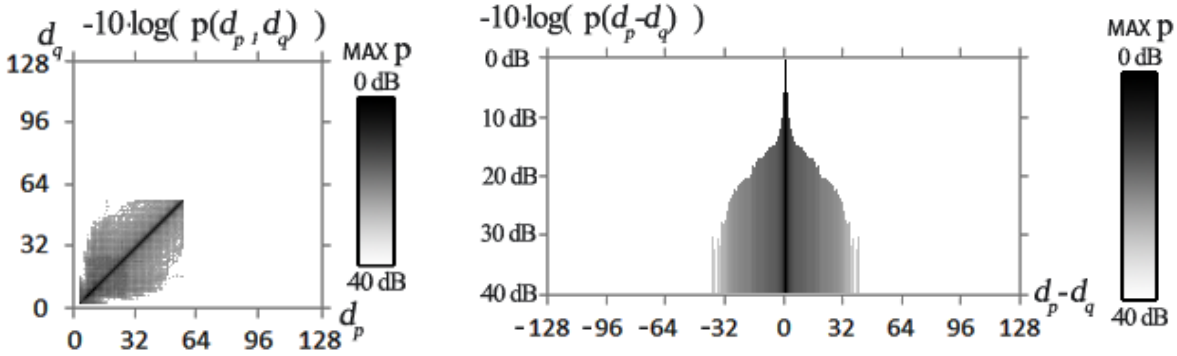
Fig. 134. Histograms of normalized disparity values of pixels in normalized disparity maps in the views 2,3,5,9 of Undo Dancer and GT Fly sequences. The graphs have been normalized to range [0;1]. Please note, that despite other sequences, GT Fly is a computer-generated sequence and the normalized disparity histogram is dense.

Histograms of disparity values of neighboring pixels

Poznan Street, camera 3



Poznan Street, camera 4



Poznan Street, camera 5

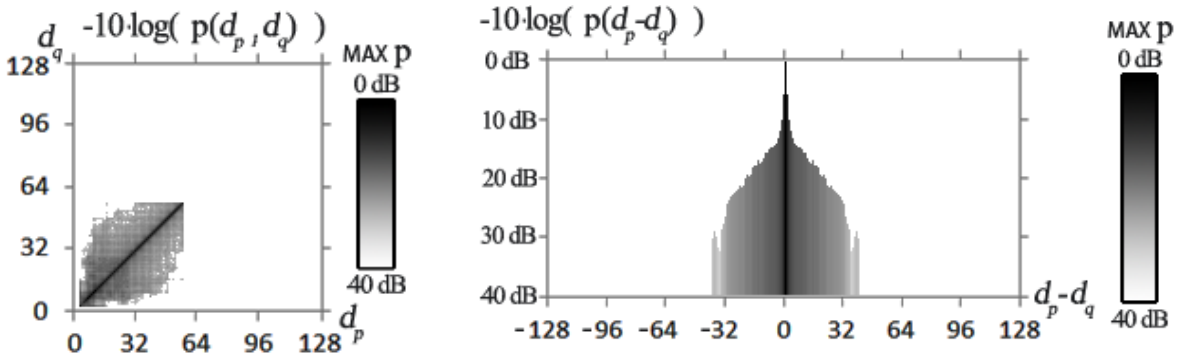
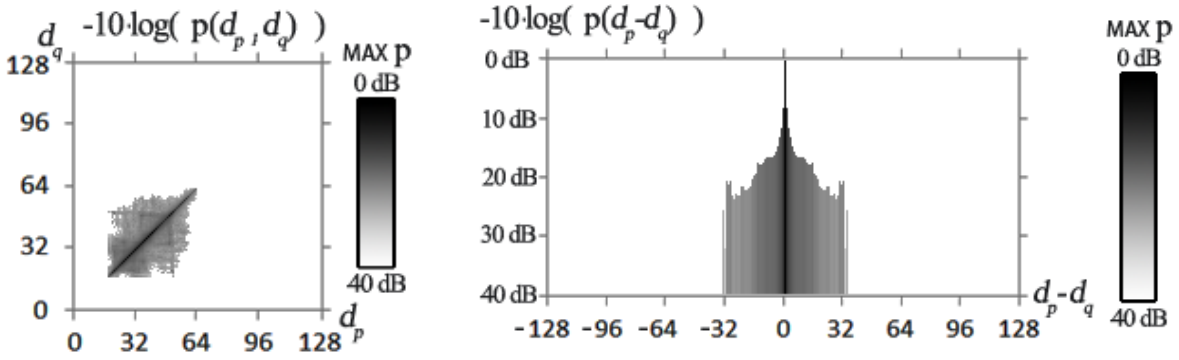
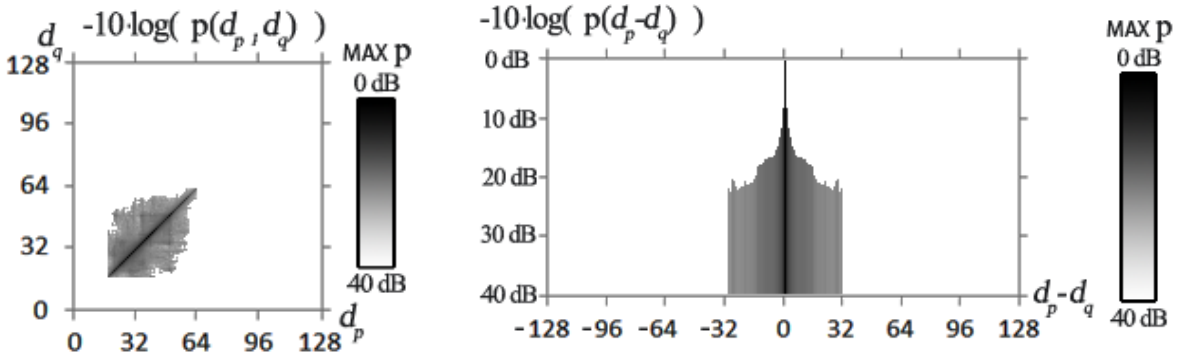


Fig. 135. Histogram of disparity values d_p and d_q of neighboring pixels p and q, in ground truth disparity maps for Poznan Street sequence for views 3,4,5. The histograms have been visualized as 2D plots (left) and histograms in domain of $d_p - d_q$ disparity difference (right). Both types of plots are presented in logarithmic scale and in the same shading. See Fig. 36 for explanation.

Poznan Hall 2, camera 5



Poznan Hall 2, camera 6



Poznan Hall 2, camera 7

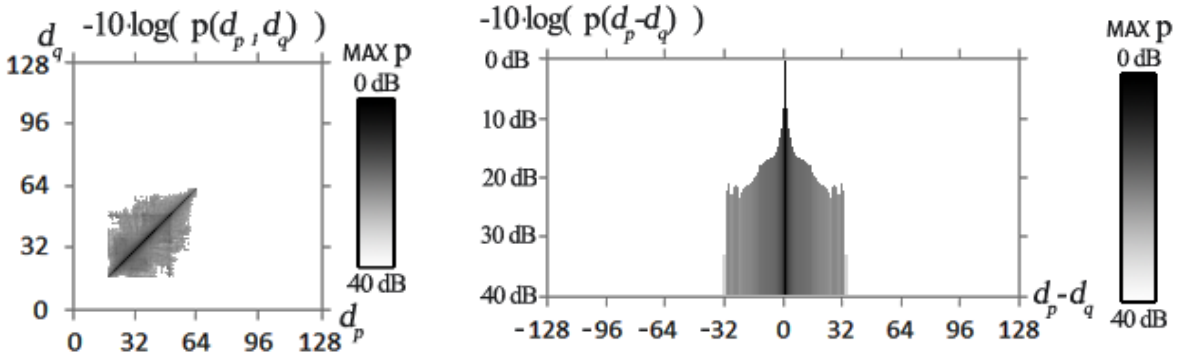
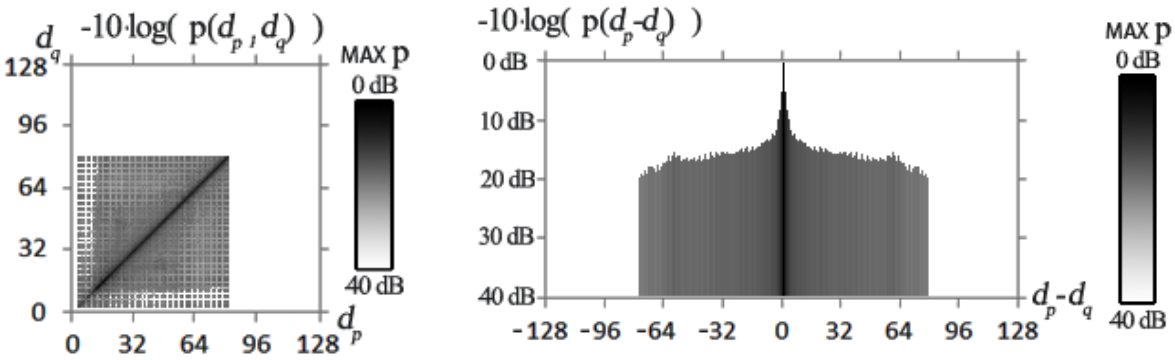
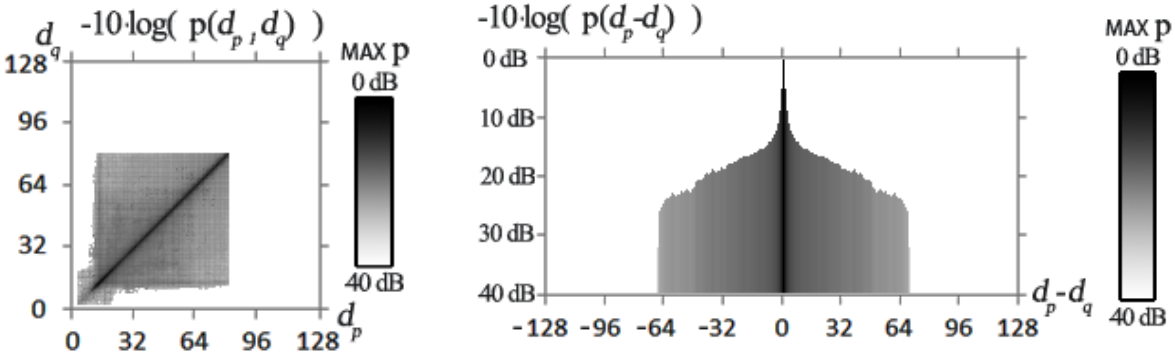


Fig. 136. Histogram of disparity values d_p and d_q of neighboring pixels p and q, in ground truth disparity maps for Poznan Hall 2 sequence for views 5,6,7. The histograms have been visualized as 2D plots (left) and histograms in domain of $d_p - d_q$ disparity difference (right). Both types of plots are presented in logarithmic scale and in the same shading. See Fig. 36 for explanation.

Poznan Carpark, camera 3



Poznan Carpark, camera 4



Poznan Carpark, camera 5

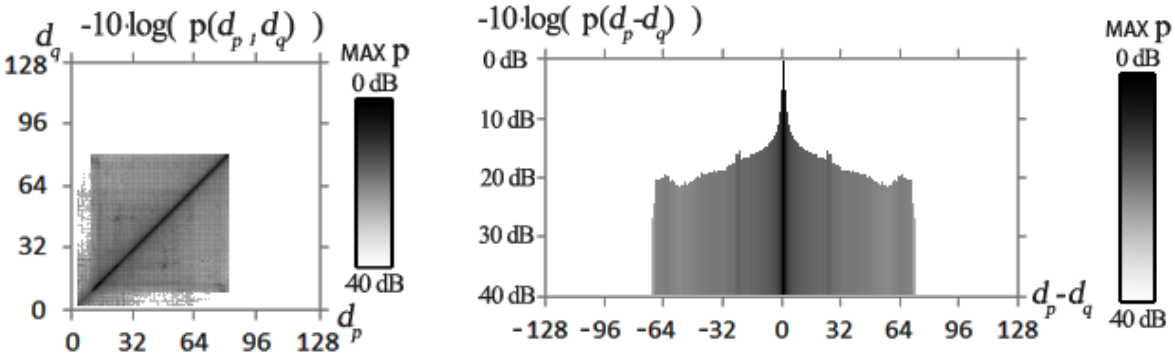
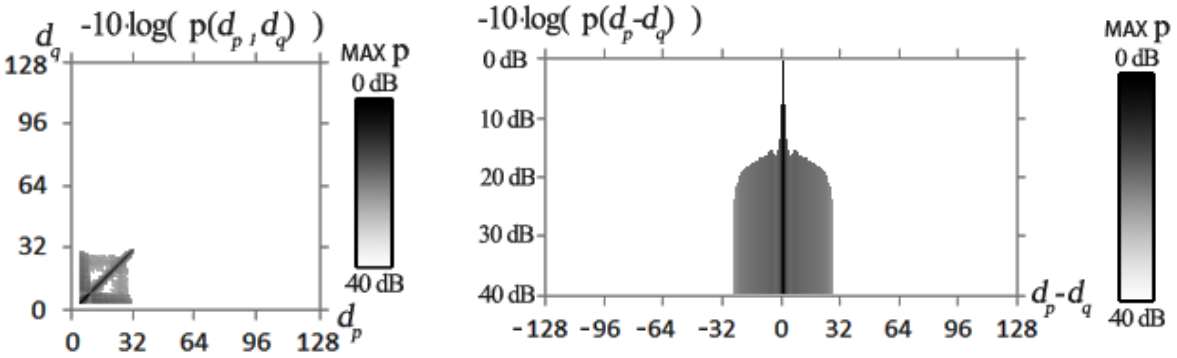
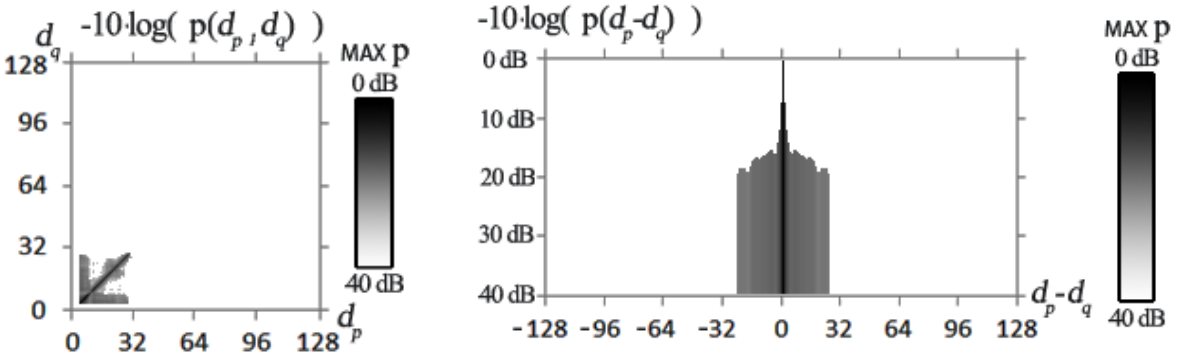


Fig. 137. Histogram of disparity values d_p and d_q of neighboring pixels p and q, in ground truth disparity maps for Poznan Carpark sequence for views 3,4,5. The histograms haev been visualized as 2D plots (left) and histograms in domain of $d_p - d_q$ disparity difference (right). Both types of plots are presented in logarithmic scale and in the same shading. See Fig. 36 for explanation.

Lovebird1, camera 3



Lovebird1, camera 5



Lovebird1, camera 7

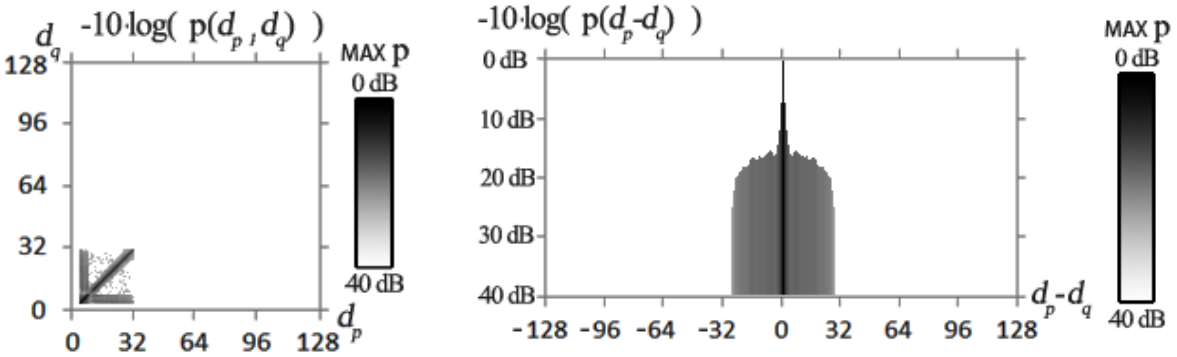
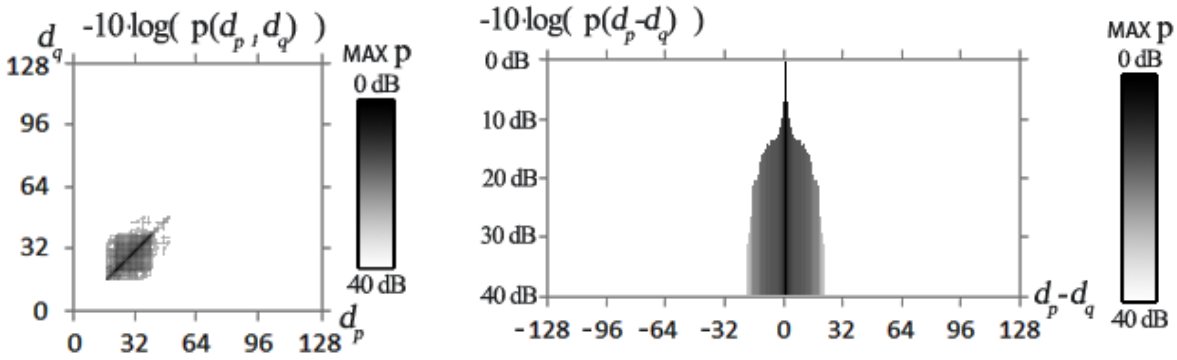
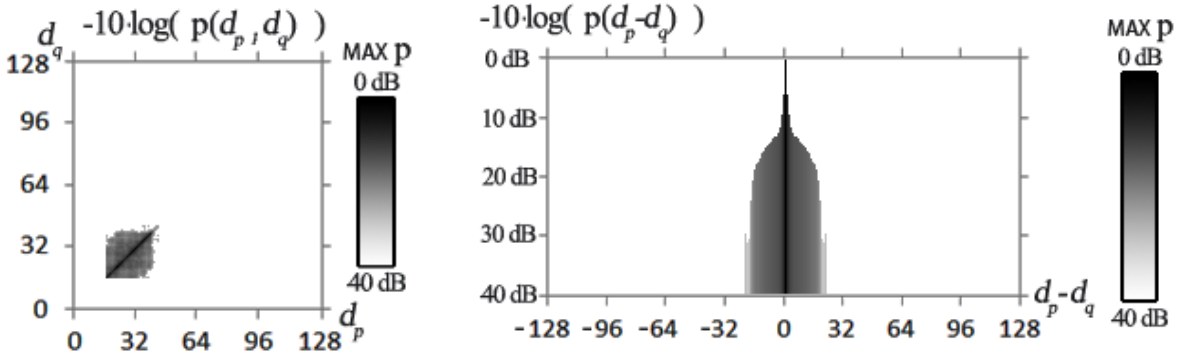


Fig. 138. Histogram of disparity values d_p and d_q of neighboring pixels p and q, in ground truth disparity maps for Lovebird1 sequence for views 3,5,7. The histograms have been visualized as 2D plots (left) and histograms in domain of $d_p - d_q$ disparity difference (right). Both types of plots are presented in logarithmic scale and in the same shading. See Fig. 36 for explanation.

Newspaper, camera 2



Newspaper, camera 4



Newspaper, camera 6

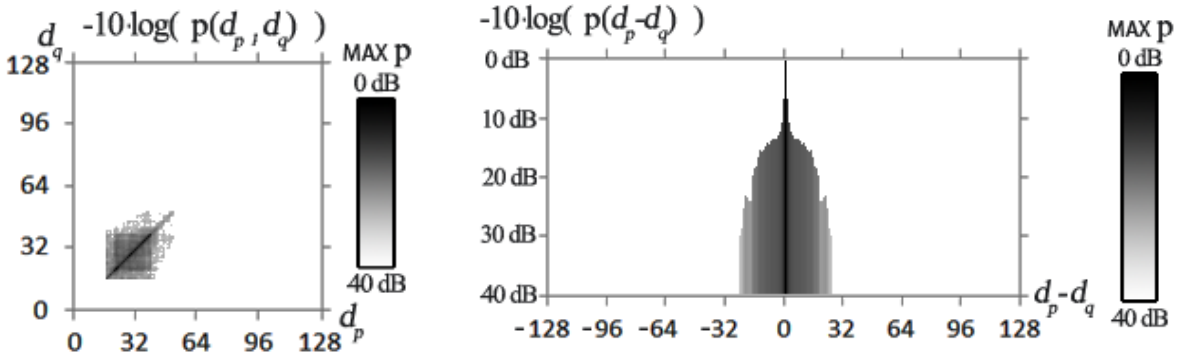
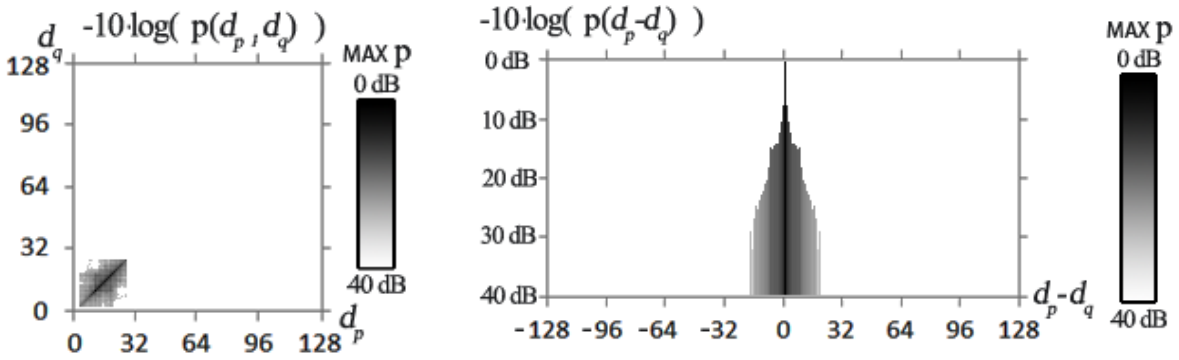
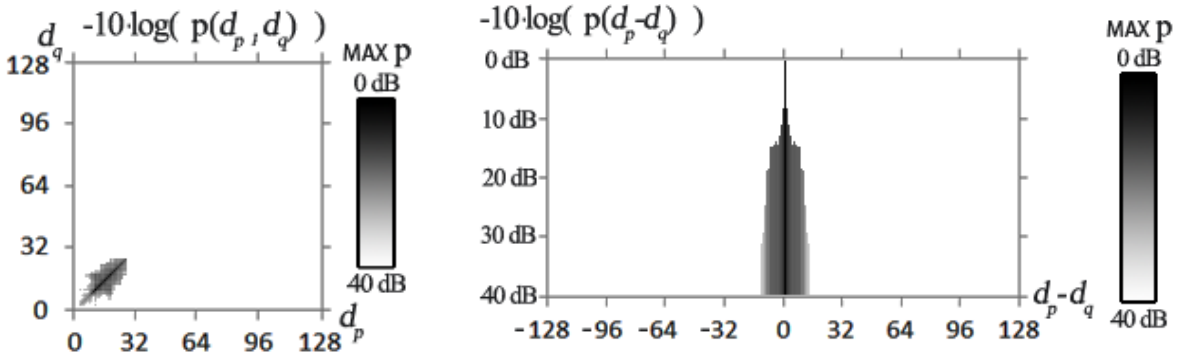


Fig. 139. Histogram of disparity values d_p and d_q of neighboring pixels p and q, in ground truth disparity maps for Newspaper sequence for views 2,4,6. The histograms have been visualized as 2D plots (left) and histograms in domain of $d_p - d_q$ disparity difference (right). Both types of plots are presented in logarithmic scale and in the same shading. See Fig. 36 for explanation.

Balloons, camera 1



Balloons, camera 3



Balloons, camera 5

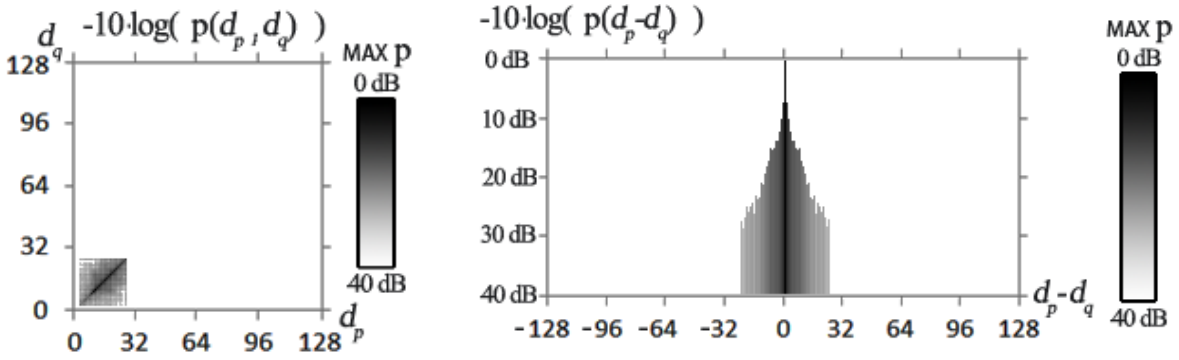
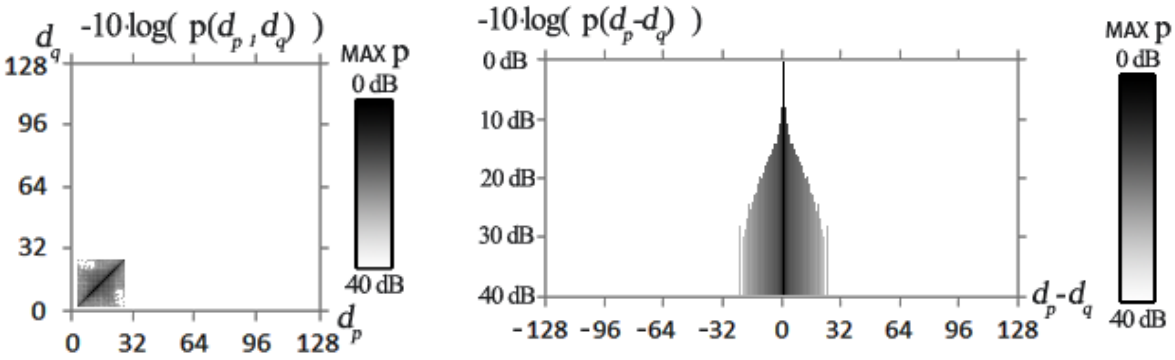
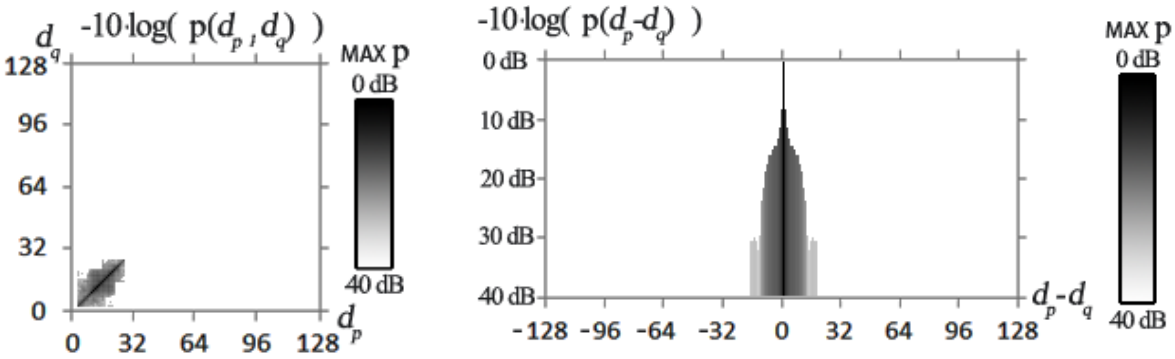


Fig. 140. Histogram of disparity values d_p and d_q of neighboring pixels p and q, in ground truth disparity maps for Balloons sequence for views 1,3,5. The histograms have been visualized as 2D plots (left) and histograms in domain of $d_p - d_q$ disparity difference (right). Both types of plots are presented in logarithmic scale and in the same shading. See Fig. 36 for explanation.

Kendo, camera 1



Kendo, camera 3



Kendo, camera 5

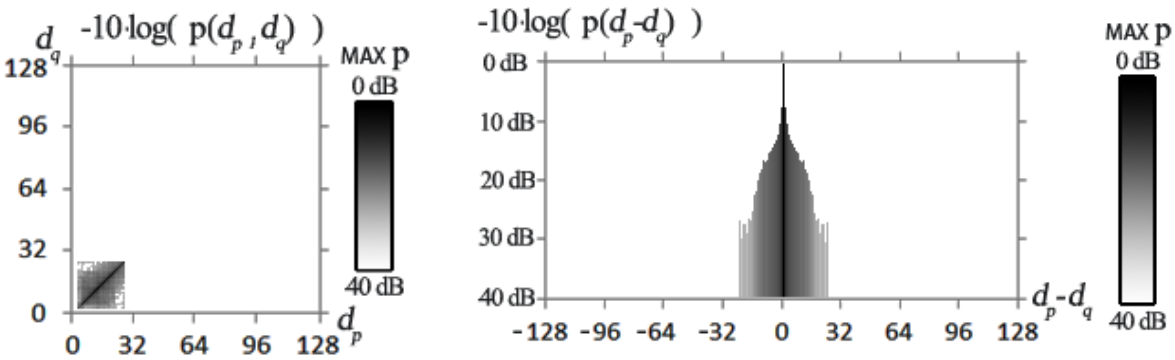
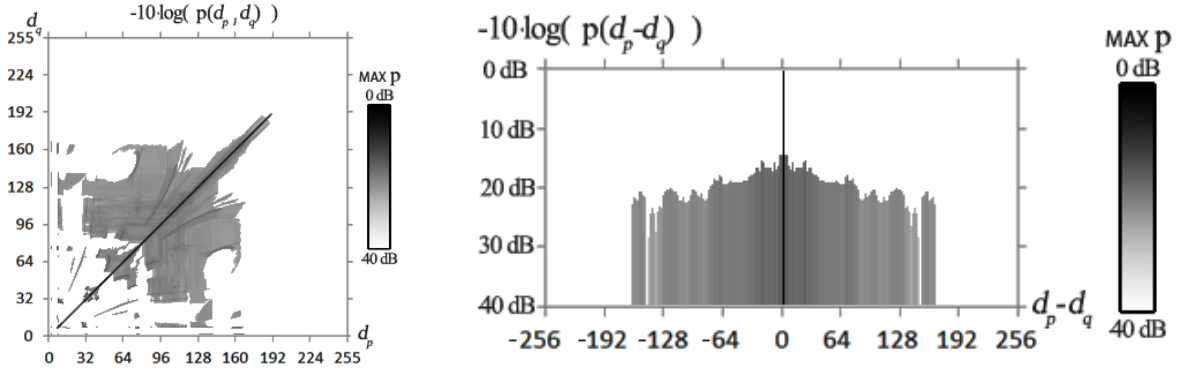
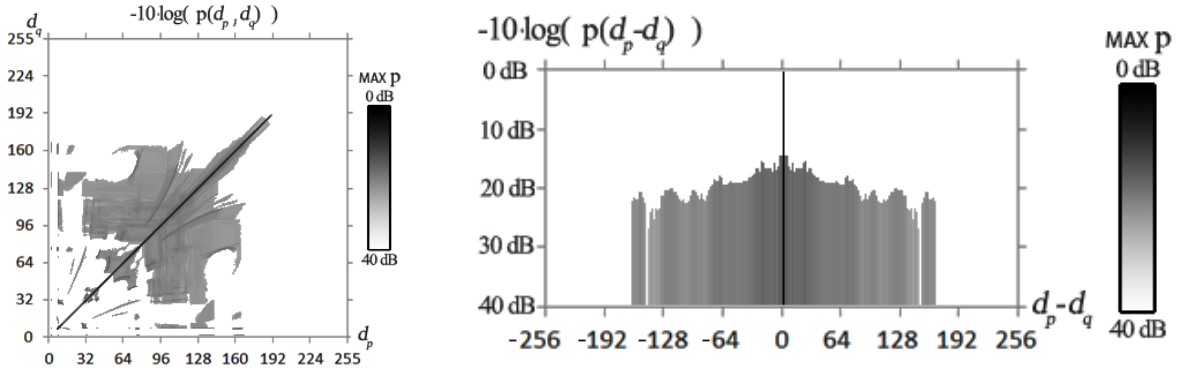


Fig. 141. Histogram of disparity values d_p and d_q of neighboring pixels p and q, in ground truth disparity maps for Kendo sequence for views 1,3,5. The histograms have been visualized as 2D plots (left) and histograms in domain of $d_p - d_q$ disparity difference (right). Both types of plots are presented in logarithmic scale and in the same shading.
See Fig. 36 for explanation.

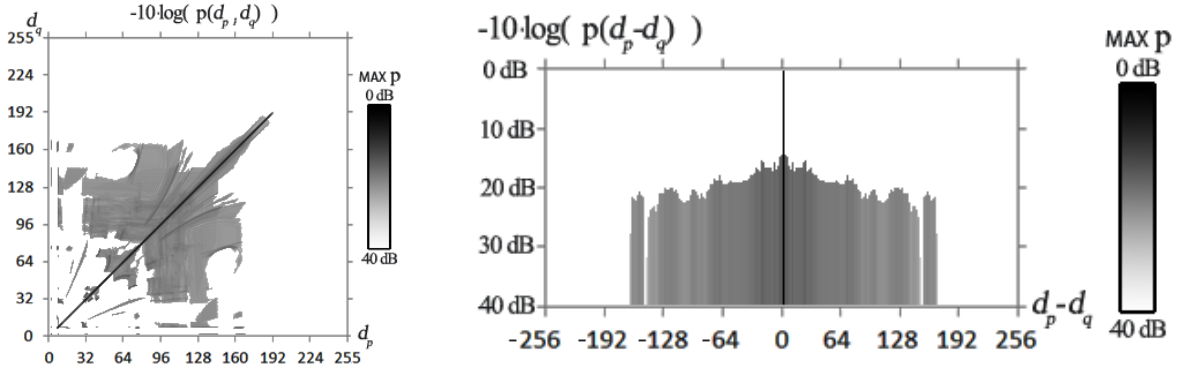
Undo Dancer, camera 2



Undo Dancer, camera 3



Undo Dancer, camera 5



Undo Dancer, camera 9

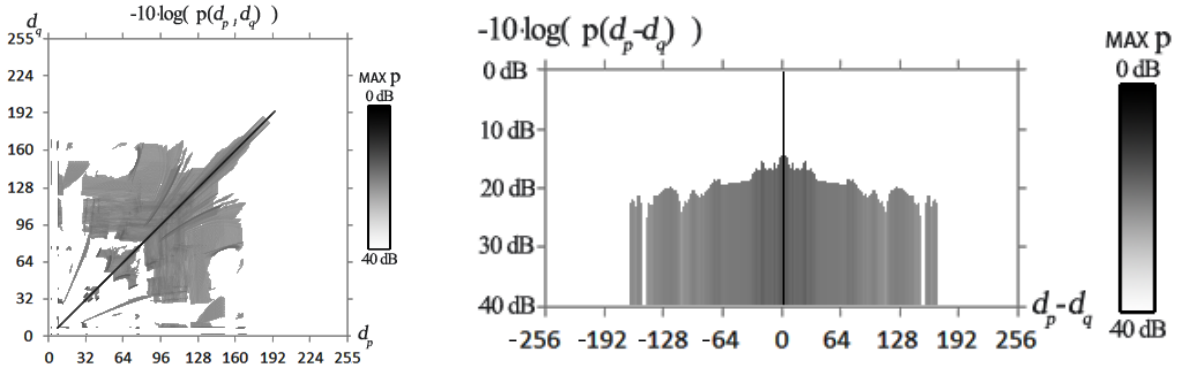


Fig. 142. Histogram of disparities d_p and d_q of neighboring pixels p and q, in ground truth disparity maps for Undo Dancer sequence (views 2,3,5,9). The histograms have been visualized as 2D plots (left) and histograms in domain of $d_p - d_q$ disparity difference (right). Both types of plots are presented in logarithmic scale and in the same shading (See: Fig. 36).

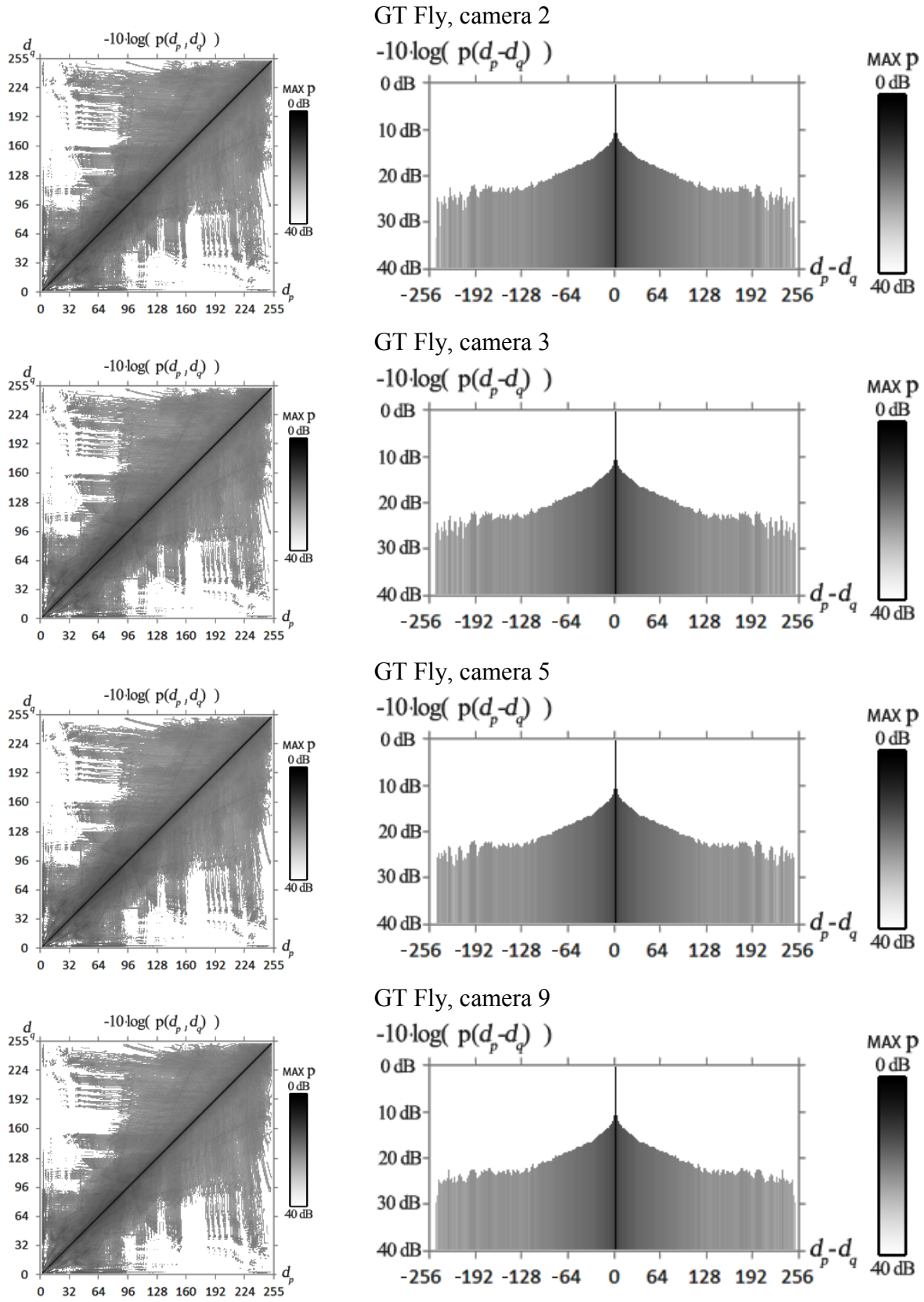
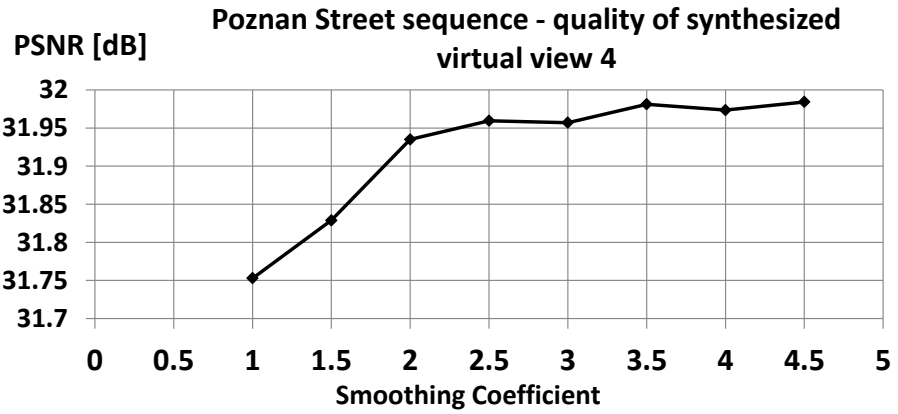
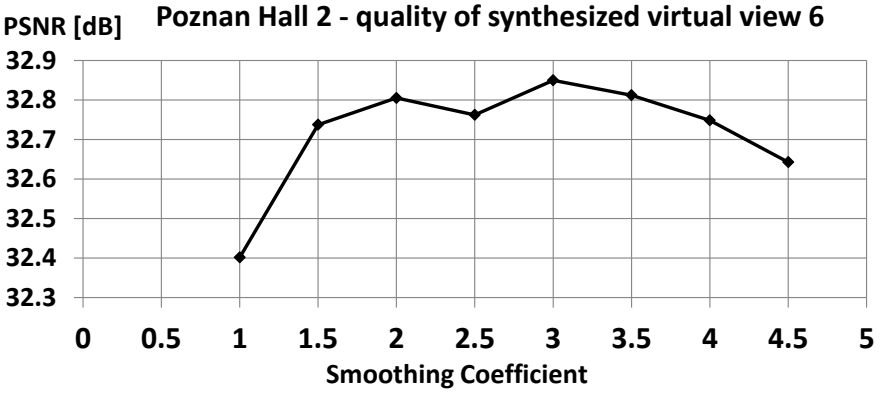


Fig. 143. Histogram of disparities d_p and d_q of of neighboring pixels p and q, in ground truth disparity maps for GT Fly sequence for views 2,3,5,9. The histograms have been visualized as 2D plots (left) and histograms in domain of $d_p - d_q$ disparity difference (right). Both types of plots are presented in logarithmic scale and in the same shading. See Fig. 36 for explanation.

Detailed results of experiments in area of depth estimation





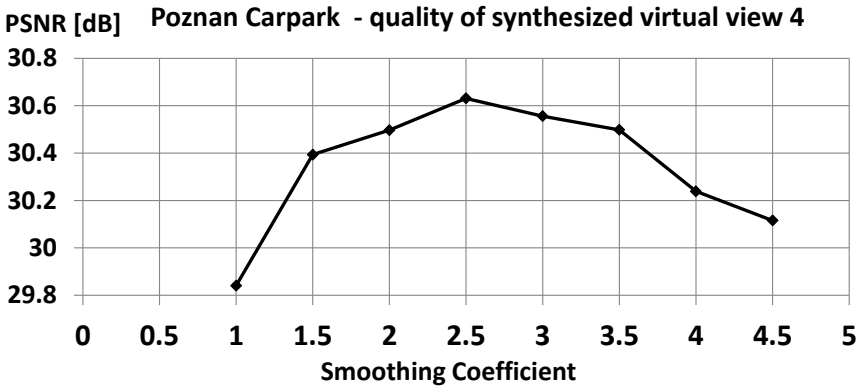


Fig. 144. PSNR of a virtual view, synthesized with use of depth maps estimated with original (unmodified DERS), related to the original view in function of Smoothing Coefficient.

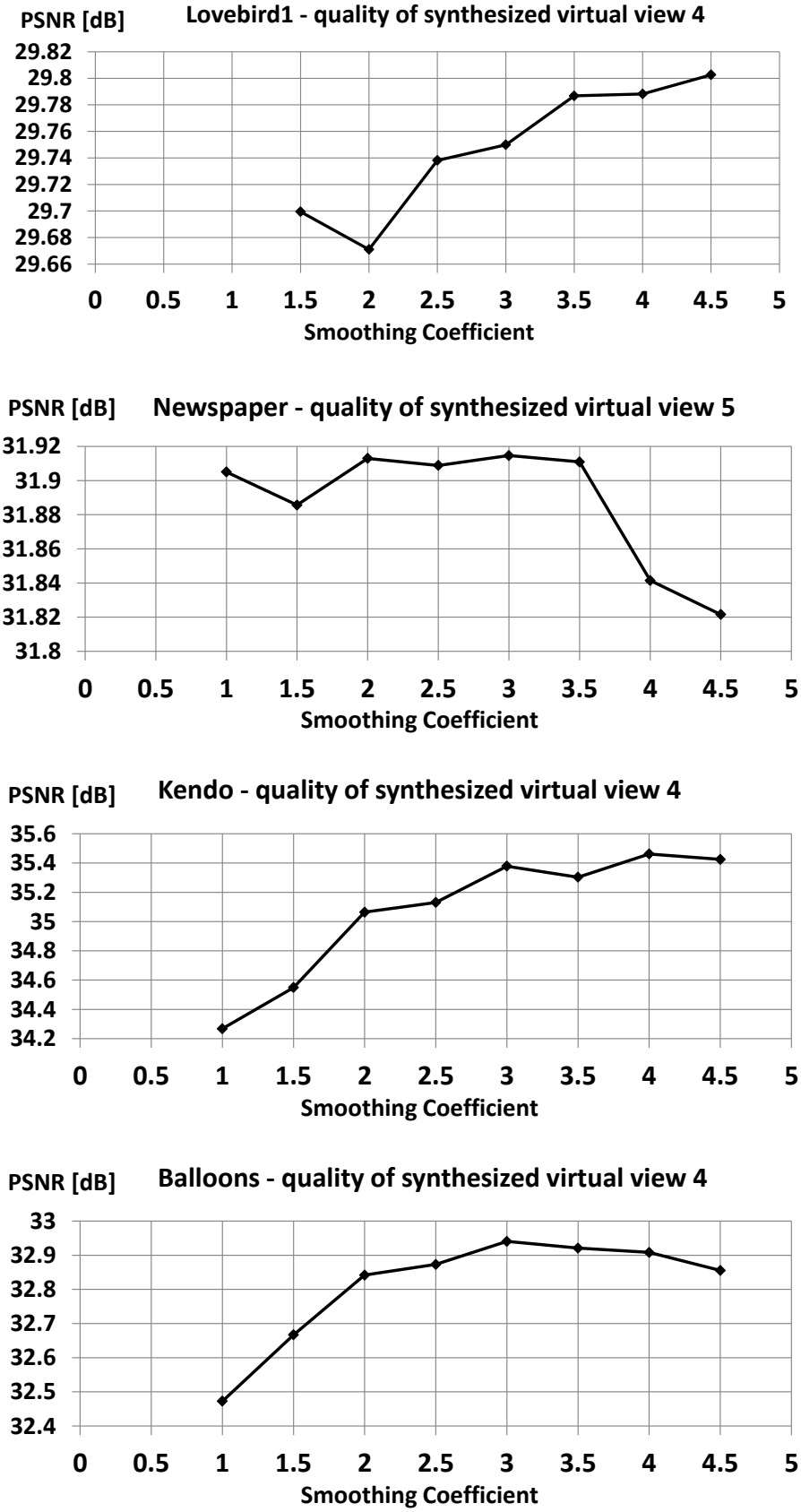


Fig. 145. PSNR of a virtual view, synthesized with use of depth maps estimated with original (unmodified DERS), related to the original view in function of Smoothing Coefficient.