TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Péter Tamás Kovács

# Methods for Light Field Display Profiling and Scalable Super-Multiview Video Coding

Julkaisu 1598 • Publication 1598

Tampere 2018

Tampereen teknillinen yliopisto. Julkaisu 1598
Tampere University of Technology. Publication 1598

Péter Tamás Kovács

# Methods for Light Field Display Profiling and Scalable Super-Multiview Video Coding

Thesis for the degree of Doctor of Science in Technology to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB104, at Tampere University of Technology, on the 23rd of November 2018, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2018

Doctoral candidate: Péter Tamás Kovács
3D Media Research Group
Faculty of Computing and Electrical Engineering
Tampere University of Technology
Finland

Supervisor: Prof. Atanas Gotchev
3D Media Research Group
Faculty of Computing and Electrical Engineering
Tampere University of Technology
Finland

Pre-examiners: Prof. Patrick Le Callet
IRCCyN lab
University of Nantes
France

Prof. András Kemény
Virtual Reality and Immersive Simulation Center
Renault Group / Arts et Métiers ParisTech
France

Opponent: Dr. Sebastian Schwarz
Volumetric Video Coding Research and Standardisation Team
Nokia
Finland

ISBN 978-952-15-4261-9 (printed)
ISBN 978-952-15-4274-9 (PDF)
ISSN 1459-2045

# Abstract

Light field 3D displays reproduce the light field of real or synthetic scenes, as observed by multiple viewers, without the necessity of wearing 3D glasses. Reproducing light fields is a technically challenging task in terms of optical setup, content creation, distributed rendering, among others; however, the impressive visual quality of hologram-like scenes, in full color, with real-time frame rates, and over a very wide field of view justifies the complexity involved. Seeing objects popping far out from the screen plane without glasses impresses even those viewers who have experienced other 3D displays before.

Content for these displays can either be synthetic or real. The creation of synthetic (rendered) content is relatively well understood and used in practice. Depending on the technique used, rendering has its own complexities, quite similar to the complexity of rendering techniques for 2D displays. While rendering can be used in many use-cases, the holy grail of all 3D display technologies is to become the future 3DTVs, ending up in each living room and showing realistic 3D content without glasses. Capturing, transmitting, and rendering live scenes as light fields is extremely challenging, and it is necessary if we are about to experience light field 3D television showing real people and natural scenes, or realistic 3D video conferencing with real eye-contact.

In order to provide the required realism, light field displays aim to provide a wide field of view (up to 180°), while reproducing up to ~80 MPixels nowadays. Building gigapixel light field displays is realistic in the next few years. Likewise, capturing live light fields involves using many synchronized cameras that cover the same display wide field of view and provide the same high pixel count. Therefore, light field capture and content creation has to be well optimized with respect to the targeted display technologies. Two major challenges in this process are addressed in this dissertation.

The first challenge is how to characterize the display in terms of its capabilities to create light fields, that is how to *profile* the display in question. In clearer terms this boils down to finding the equivalent spatial resolution, which is similar to the screen resolution of 2D displays, and angular resolution, which describes the smallest angle, the color of which the display can control individually. Light field is formalized as 4D approximation of the plenoptic function in terms of geometrical optics through spatially-localized and angularly-directed light rays in the so-called ray space. Plenoptic Sam-

pling Theory provides the required conditions to sample and reconstruct light fields. Subsequently, light field displays can be characterized in the Fourier domain by the effective display bandwidth they support. In the thesis, a methodology for display-specific light field analysis is proposed. It regards the display as a signal processing channel and analyses it as such in spectral domain. As a result, one is able to derive the display throughput (i.e. the display bandwidth) and, subsequently, the optimal camera configuration to efficiently capture and filter light fields before displaying them.

While the geometrical topology of optical light sources in projection-based light field displays can be used to theoretically derive display bandwidth, and its spatial and angular resolution, in many cases this topology is not available to the user. Furthermore, there are many implementation details which cause the display to deviate from its theoretical model. In such cases, profiling light field displays in terms of spatial and angular resolution has to be done by measurements. Measurement methods that involve the display showing specific test patterns, which are then captured by a single static or moving camera, are proposed in the thesis. Determining the effective spatial and angular resolution of a light field display is then based on an automated analysis of the captured images, as they are reproduced by the display, in the frequency domain. The analysis reveals the empirical limits of the display in terms of pass-band both in the spatial and angular dimension. Furthermore, the spatial resolution measurements are validated by subjective tests confirming that the results are in line with the smallest features human observers can perceive on the same display. The resolution values obtained can be used to design the optimal capture setup for the display in question.

The second challenge is related with the massive number of views and pixels captured that have to be transmitted to the display. It clearly requires effective and efficient *compression techniques* to fit in the bandwidth available, as an uncompressed representation of such a super-multiview video could easily consume ~20 gigabits per second with today's displays. Due to the high number of light rays to be captured, transmitted and rendered, distributed systems are necessary for both capturing and rendering the light field. During the first attempts to implement real-time light field capturing, transmission and rendering using a brute force approach, limitations became apparent. Still, due to the best possible image quality achievable with dense multi-camera light field capturing and light ray interpolation, this approach was chosen as the basis of further work, despite the massive amount of bandwidth needed. Decompression of all camera images in all rendering nodes, however, is prohibitively time consuming and is not scalable. After analyzing the light field interpolation process and the data-access patterns typical in a distributed light field rendering system, an approach to reduce the amount of data required in the rendering nodes has been proposed. This approach, on the other hand, requires rectangular parts (typically vertical bars in case of a Horizontal

Parallax Only light field display) of the captured images to be available in the rendering nodes, which might be exploited to reduce the time spent with decompression of video streams. However, partial decoding is not readily supported by common image / video codecs. In the thesis, approaches aimed at achieving partial decoding are proposed for H.264, HEVC, JPEG and JPEG2000 and the results are compared.

The results of the thesis on display profiling facilitate the design of optimal camera set-ups for capturing scenes to be reproduced on 3D light field displays. The developed super-multiview content encoding also facilitates light field rendering in real-time. This makes live light field transmission and real-time teleconferencing possible in a scalable way, using any number of cameras, and at the spatial and angular resolution the display actually needs for achieving a compelling visual experience.

# Preface

The research problems addressed in this thesis were identified during my work at Holografika (2006-2016). The research for this thesis has been performed at Tampere University of Technology (2013-2018) and at Holografika. Having worked on successive generations of light-field 3D displays for a decade provided enough insight into the exciting world of 3D displays. Working on cutting-edge display technology, which I consider the best in the 3D world, brought many interesting challenges to solve - challenges that have scientific value and at the same time are rooted from practical industrial needs to advance the technology. I was a member of the PROLIGHT Industry-Academia Partnerships and Pathways program, during which I had the opportunity to join TUT to perform research at the University, solving issues identified in the industry. I consider myself very lucky for the possibility of working and collaborating with the sharpest minds in both the industrial and research world.

I would like to express my deepest gratitude and appreciation to Tibor Balogh, founder of Holografika and inventor of HoloVizio light-field display technology. I am grateful for the opportunity to work with him and learn from him. His optimism, wisdom, and practical approach to solving very complex problems serve as a good example to be followed. I am very grateful to my supervisor Prof. Atanas Gotchev for initiating the PROLIGHT project, for offering me the opportunity to become a visiting researcher at Tampere University of Technology and start my PhD studies during my time there. I am especially grateful for his guidance in the academic world, discussing and reviewing my work, and continuous support and inspiration. Organizing the 3DTV-Conference 2014 and the preceding Summer School together was also an exciting experience, which eventually resulted in being invited to MPEG.

I highly appreciate the comments and feedback of Prof. Patrick Le Callet and Prof. András Kemény who were per-examiners of my thesis. I am grateful to Dr. Sebastian Schwarz for being my opponent during the public defense of this thesis.

I am grateful to all colleagues and co-authors of my papers for the possibility to work together. From Tampere, I would like to especially thank Dr Atanas Boev, who is a great friend, tutor, who had a major role in helping me get started at TUT. Special thanks to Dr Robert Bregovic for his continuous support and merciless comments on everything I have ever written during my PhD studies. I am also thankful to Alireza Zare for supporting my research work, Olli Suominen, Evgeny Belyaev, Suren Vagharshakyan, Aleksandra Chuchvara and Dr Satu Jumisko-Pyykkö. I am grateful to Ulla Siltaloppi and Elina Orava for swiftly arranging all things that led to my graduation.

From Holografika, I would like to thank Attila Barsi, Kristof Lackner, Vamsi Kiran Adhikarla, Dr Zsolt Nagy, Dr Zoltán Megyesi, Zoltán Gaál, Ákos Balázs, Dave Singhal, Ádám Fekete, Alexander Ouazan, Gáspár Balogh, László Bordács and Zsuzsa Dobrányi. I would like to thank my co-authors and collaborators from outside Holografika and TUT, especially Frederik Zilly, Antoine Dricot, Joel Jung, Gauthier Lafruit, Marek Domanski, Krzystof Wegner, Adrian Muntenau, Pawel Wozniak, Peter Kara, Maria Martini, Caroline Conti, Lajos Hanzo, Paulo Nunes and Amar Aggoun. I am grateful to Masayuki Tanimoto for giving me the opportunity to be involved in MPEG activities.

I am grateful to the Department of Signal Processing, the 3D Media Research Group and later Centre for Immersive Visual Technologies for providing me the environment to perform research undisturbed, and other departments of Tampere University of Technology for the interesting and useful courses taught during my post-graduate studies.

My warmest thanks go to my family for their continuous support and love. I thank my parents for supporting me during my graduate studies, and urging me to finish my PhD studies. I thank my wife Zsuzsi for always supporting me during all my work and travels, and for using her excellent organizational skills to help me with the local organization of the 3DTV Conference in 2014. I thank my sons Barnabás and Gergely for being nice and patient while I was working on my research.

Péter Tamás Kovács

Tampere, 07.11.2018

# Contents

x

# List of Figures

# List of Abbreviations

2D …… Two Dimensional

3D …… Three Dimensional

3D-HEVC … 3D High Efficiency Video Coding

3DTV …….. Three Dimensional Television

AC …….. Alternating Current (used as the diff. from the average value in JPEG)

CABAC ….. Context-Adaptive Binary Arithmetic Coding

CAVLC ….. Context-adaptive variable-length coding

CDF …… Cohen–Daubechies–Feauveau (wavelet)

CG ……. Computer Graphics

CGI …….. Computer-Generated Imagery

CPU ……. Central Processing Unit

DC …… Direct Current (used as the average value in JPEG)

DCT ……. Discrete Cosine Transform

EBCOT …… Embedded Block Coding with Optimized Truncation

EPI …….. Epipolar Plane Image

FFT ……. Fast Fourier Transform

FOV ……. Field Of View

FTV ……. Free-Viewpoint Television

GOP …… Group Of Pictures

GPU …… Graphics Processing Unit

H.264 ……. MPEG-4 Part 10, Advanced Video Coding (MPEG-4 AVC)

HD …….. High Definition

HDR …… High Dynamic Range

HEVC ……. High Efficiency Video Coding

HMD …… Head Mounted Display

HPO …… Horizontal Parallax Only

HVS …… Human Visual System

ISO ……. International Organization for Standardization

IEC ……. International Electrotechnical Commission

JPEG …….. Joint Photographic Experts Group

LCD ……. Liquid-Crystal Display

LED ……. Light-Emitting Diode

LF ……. Light Field

MCU …… Minimum Coded Unit (JPEG)

MPEG ……. Moving Picture Experts Group

MV-HEVC .. Multi-View High Efficiency Video Coding

MVC …… Multiview Video Coding

OLED …….. Organic Light-Emitting Diode

QPI …….. Quantum Photonic Imager

RDMA ……. Remote Direct Memory Access

RG ……. Ray Generator

RGB ……. Red Green Blue

SFM ……. Structure From Motion

SVC ……. Scalable Video Coding

TOF ……. Time Of Flight

USB ……. Universal Serial Bus

VGA ……. Video Graphics Array

YUV ……. luminance (Y), chrominance (UV)

# List of Publications

The thesis consists of a summary and the following original publications. Author's contribution to, and the importance of each publication are discussed in Section 5.2.

I. P. T. Kovács, T. Balogh, "3D Visual Experience", in High-Quality Visual Experience: Creation, Processing and Interactivity of High-Resolution and High-Dimensional Video Signals (eds M. Mrak, M. Grgic, M. Kunt), Springer, Signals and Communication Technology, 2010, DOI: 10.1007/978-3-642-12802-8

II. P. T. Kovács, A. Boev, R. Bregović, A. Gotchev, "Quality Measurements of 3D Light-Field Displays", in Proc. Eighth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM 2014), January 31, 2014, Chandler, Arizona, USA

III. P. T. Kovács, K. Lackner, A. Barsi, Á. Balázs, A. Boev, R. Bregović, A. Gotchev, "Measurement of Perceived Spatial Resolution in 3D Light field Displays", in Proc. IEEE International Conference on Image Processing (ICIP), 2014, DOI:10.1109/ICIP.2014.7025154

IV. R. Bregović, P. T. Kovács, A. Gotchev, "Optimization of light field display-camera configuration based on display properties in spectral domain," Optics Express, vol. 24, no. 3, pp. 3067-3088, 2016

V. P. T. Kovács, R. Bregović, A. Boev, A. Barsi, A. Gotchev, "Quantifying Spatial and Angular Resolution of Light-Field 3-D Displays", in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 7, pp. 1213-1222, Oct. 2017. DOI: 10.1109/JSTSP.2017.273860

VI. T. Balogh, P. T. Kovács, "Real-time 3D light field transmission", in Proc. SPIE 7724, Real-Time Image and Video Processing 2010, 772406, 2010, DOI:10.1117/12.854571

VII. P. T. Kovacs, F. Zilly, "3D capturing using multi-camera rigs, real-time depth estimation and depth-based content creation for multi-view and light field autostereoscopic displays", in ACM SIGGRAPH 2012 Emerging Technologies (SIGGRAPH '12), DOI: 10.1145/2343456.2343457

VIII. P. T. Kovács, K. Lackner, A. Barsi, V. K. Adhikarla, R. Bregović, A. Gotchev, "Analysis and Optimization of Pixel Usage of Light field Conversion from Multi-Camera Setups to 3D Light field Displays", in Proc. IEEE International Conference on Image Processing (ICIP), 2014, DOI: 10.1109/ICIP.2014.7025016

IX. P. T. Kovács, Z. Nagy, A. Barsi, V. K. Adhikarla and R. Bregović, "Overview of the applicability of H.264/MVC for real-time light field applications", 2014 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), Budapest, 2014, pp. 1-4. DOI: 10.1109/3DTV.2014.6874744

X. P. T. Kovács, A. Zare, T. Balogh, R. Bregovic, A. Gotchev, “Architectures and codecs for real-time light field streaming”, in Journal of Imaging Science and Technology, 61(1), [010403], 2017, DOI:10.2352/J.ImagingSci.Technol.2017.61.1.010403

# 1 Introduction

Displays are the primary means for human-machine, and increasingly remote interpersonal communication. While other channels that address human senses such as audio, tactile, or olfactory are also available [1], visual information is the richest in terms of information content, and also the one that requires the widest bandwidth when stored or transmitted. This makes displays and associated visual technologies highly relevant for businesses as well as in everyday life.

Most displays used today show a 2D projection of real or virtual scenes. As human observers are exposed to different kinds of 2D representations of the real 3D world at an early age, the association and transfer of 3D to 2D and vice versa are learnt at early ages [2]. However this association is not natural, which is especially apparent when understanding complex spatial scenarios [3], interacting with the content [4], or when using displays as a means to facilitate interpersonal communication [5]. The transformation involved while capturing 3D scenes with a camera discards one dimension by projecting all 3D points to their 2D representation on the screen.

3D displays aim to represent visual scenes in their natural three dimensions; scenes appear as popping out from a screen, behind a screen, or both [6]. This is possible if the display can address at least the stereopsis [7], i.e. the ability of the human visual system to fuse two different-perspective images, causing retinal disparity, aka binocular visual cue. There are several display techniques for showing 3D imagery, with different levels of implemenation complexity and information content, aimed at reproducing the binocular visual cue. The most well-known types of 3D displays (see [I] for an overview) include stereoscopic displays [8] and multiview autostereoscopic displays [9]. However, these displays fail to reproduce other visual cues such as focus and continuous parallax. Therefore, more advanced 3D displays, such as volumetric displays [10], light field displays [6] and holographic displays [11][12][13] have been attempted. Among these [14], light field displays are perhaps the most interesting ones as they aim to reproduce the light as it naturally is, i.e. in terms of dense bunches of rays with different locations and directions. As such, they provide impressive image quality over a large field of view and depth of field without 3D glasses [7], and are also available commercially [15].

## 1.1 Objectives of Research

The topics discussed in this thesis target enabling light field displays to show *real-live* 3D content, captured with cameras, reproducing the diversity of real scenes, including realistic, live people; surfaces with real specular reflections, anisotropic effects, transparency, atmospheric effects, subsurface scattering, and other phenomena that occur in the real world, but not so much in synthetic, rendered scenes. The long term goal of this research is to enable light field displays to be used as the future 3DTVs, and to enable high-end use cases like real 3D videoconferencing [16][17]. These use cases can only be served if live content can be captured, transmitted, and rendered.

Advancing light field displays goes through the formalization of the light field as a multi-dimensional function which describes the light formation, propagation and perception. For the displays under question, light is modeled in terms of geometrical optics. That is, each light ray is parameterized by its location and direction. A thorough **model** is needed to understand how light field displays generate fields of light. Such a model should facilitate the design and optimization of such displays and formalize the content creation for them. It would be instrumental, if such a model is developed in signal processing terms, that is, regarding the display as a signal processing channel, which gets light generators as an input and generates continuous light field at the output.

Capturing real light fields generally requires multi-camera rigs [18] that capture the scene from the necessary number of viewpoints, covering the necessary viewing angle, with the resolution, field of view, and frame rate sufficient for the targeted display. In formal terms, such camera rigs are regarded as samplers of the continuous light field function, which then has to be processed digitally for proper driving of the target display. However, in case of many 3D displays, projection-based light-field displays included, the number of necessary viewpoints, the angle between adjacent views, and even the equivalent resolution are not known. This is because the screen does not have an explicit pixel structure [19]. Instead, it is formed by a set of light generators originating from many projectors in a complex optical setup, which includes a special screen which recombines these light generators in continuously superimposed light beams, eventually forming the desired continuous-parallax light field [6]. Multi-camera setups for capturing content for 3D displays are generally designed based on rule of thumb, or physical constraints of the cameras used [18][20][21]. To design a capture setup optimal for a given 3D display, the display first needs to be **profiled** in terms of **spatial** and **angular resolution** and **field of view**, or more generally, in terms of the **bandwidth** of the light field function it generates. Profiling projection-based light field displays is therefore one objective of the research, followed by the design of optimal camera setups for a known display.

Transmitting the captured light fields is challenging due to the sheer amount of information present in such a light field video stream. As an example, the multi-view test sequences [22] provided by

Nagoya University for experimentation in MPEG consist of 80 video streams, each with 1280x960 pixel resolution, 30 frames per second, YUV4:2:0 color format, which account to 98 Mpixel for each frame, or 5.89 Gigabytes of uncompressed image information per second. In comparison, a 4K Ultra HD image in 2D, which is generally considered as state of the art at the time of writing, consists of 8 Mpixels. While specialized hardware encoders are available to support video streams of widely used formats, these generally target maximum 4K resolution, and 2D video streams. While MPEG developed standards for multi-view video coding (H.264/MVC [23], 3D-HEVC [24], MV-HEVC [24]), these generally consider one centralized encoder and decoder, and a single bitstream containing the full video stream. A single centralized encoder or decoder for such high resolution / many views is clearly out of scope on today's hardware when real-time applications are targeted. Therefore the objective is to find a suitable solution to **encode** / **decode multiview video streams** of many views (referred to as **super-multiview**) on today's hardware, which can serve the light field rendering process with live data.

## 1.2 Research Questions

This thesis addresses the following research questions:

- How can a light field display be modeled as a signal processing channel assuming underlying geometrical optics and multi-dimensional light field parameterization?
- How can light field displays be profiled in terms of display bandwidth, or equivalently, in terms of spatial and angular resolution?
- How can an optimal capture setup be designed for a light field display with known parameters?
- How can real-time encoding and decoding of captured light fields be supported on hardware available today?
- How can the coding method feed the light field rendering process with data in real time?

The question of display profiling is addressed by considering the camera-display pair as a signal processing channel, and analyzing the light field sampling by cameras and reproduction by the display in the frequency domain.

The optimal camera setup for capturing light fields for display purposes builds on the parameters either known at display design time, or acquired during the display profiling process. The question of optimizing the camera setup is addressed from the perspective of ray space analysis, as well as from analyzing the light field rendering process.

Runtime efficient encoding and decoding for high pixel count light fields consisting of many views is approached by analyzing the peculiarities of the underlying light field function, its capturing and

reconstruction, and how the contemporary codecs designed for encoding 3D content perform in the targeted use case. Starting from the analysis of bottlenecks of an initial light field rendering and capturing system, followed by a rendering system based on depth estimation and image+depth based rendering, leads to the use of a light field interpolation system *without* estimated depth maps. Then encoding methods designed to encode 3D video content are analyzed for suitability. Based on this analysis, recommendations about how to use these codecs for the case under consideration, and modifications to the codecs are proposed and implemented.

Feeding the light field rendering process with data in real time is supported by the decoder, and thus very strongly connected to the previous question. The data flow in a typical light field rendering algorithm is analyzed, which allows exploiting the locality of data access. Based on this observation, two different architectures are proposed for runtime efficient light field decoding systems, both of which are feasible on hardware available today.

## 1.3 Structure of the Dissertation

This dissertation consists of two parts. The first part introduces the problems, presents the state of the art, and summarizes the approaches to solving these problems. The second part consists of publications in their original form, which describe the solutions to the described problems in full detail and present the obtained scientific results.

Section 2 introduces light fields, 3D displays, highlighting some implementations, as well as a historical retrospection on 3D displays. Light field displays, the choice of display technology for the presented work are introduced in detail, followed by a discussion about the limitations of 3D displays. 3D video representations and codecs are introduced, and the connection between display quantification and content creation is established.

Section 3 describes methods for light field display profiling. It first presents state of the art in display quantification, models light field display's behavior in the frequency domain and derives an analytical evaluation of camera-display relation, followed by the novel objective and subjective measurement methods for spatial and angular resolution measurement.

Section 4 describes the aspects of 3D video compression and light-field processing for light field displays, followed by the challenges and proposed solutions for the compression / decompression methods for light field video storage and transmission.

Section 5 concludes the dissertation by explaining the contribution of the papers in the second part of the dissertation and identifies topics for further research.

# 2 Preliminaries

## 2.1 Basics of Light Field

### 2.1.1 The Plenoptic Function

The intensity of light rays in 3D space can be described by the plenoptic function [25]. It is a 7D function in the general form L = P(x, y, z, θ, φ, t, λ), where (x, y, z) is a ray location in 3D space, (θ, φ) describe its direction, t is time, and λ is the wavelength of light (see Figure 1). While this continuous function provides a full description of an arbitrary, time varying light field in 3D space, it is difficult to maintain in its seven dimensions. Therefore, it is usually simplified to a smaller number of dimensions and discretized for practicality.

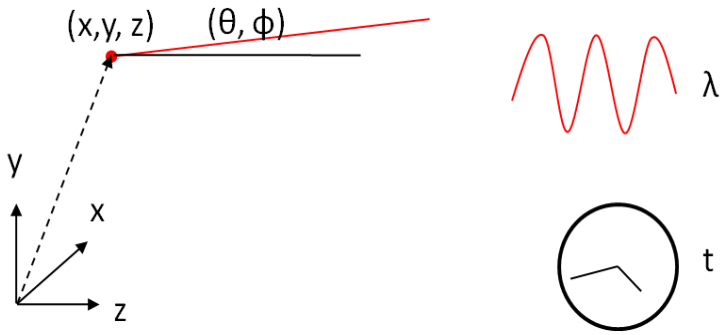



Figure 1. Parameters of the plenoptic function

### 2.1.2 Two-Plane Parameterization

One common simplification is the 4D parameterization which describes a static, grayscale light field between two parallel planes. This can be described by a two-plane parameterization (x, y, s, t), where (x, y) describe the coordinates of a hit point on one plane, and (s ,t) describe the hit point on a parallel plane (see Figure 2, left). The other common description with just one plane is (x, y, θ, φ),

where (x, y) describe the coordinates of a hit point on the plane, while (θ, φ) describe the direction of the light ray (see Figure 2, right).

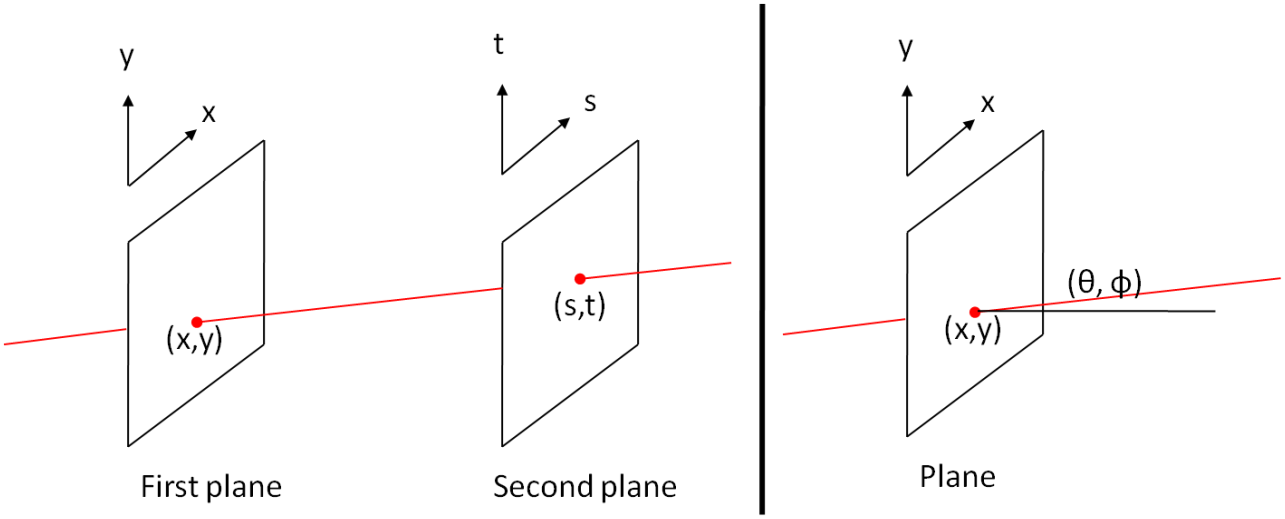



Figure 2. Left: two-plane parameterization. Right: hit point + angle parameterization

When working with displays which provide horizontal parallax only (HPO) one can omit one parameter from the 4D parameterization, so that the vertical direction of light rays is not taken into account. On the other hand, the displays emit colored RGB images, which can be described by three single-channel light fields. Also, as the display screen is updated with video frame rates (for example, 30 frames per second), discretized time is needed to describe an animated, colored HPO light field.

The position of two parallel planes can be chosen depending on the application. Two such positions, where the distances between parameterizing planes is taken as a unit, are given in Figure 3. According to the figure, the propagation of light rays through space can be mathematically expressed as [26][19]

$$L_2\left(\begin{bmatrix}x_2\\ s_2\end{bmatrix}\right) = L_1\left(\begin{bmatrix}x_1\\ s_1\end{bmatrix}\right) = L_1\left(\begin{bmatrix}1 & -d\\ 0 & 1\end{bmatrix}\begin{bmatrix}x_2\\ s_2\end{bmatrix}\right) \quad (1)$$

$$L_2\left(\begin{bmatrix}x_2\\ \varphi_2\end{bmatrix}\right) = L_1\left(\begin{bmatrix}x_1\\ \varphi_1\end{bmatrix}\right) = L_1\left(\begin{bmatrix}x_2 - d\tan\varphi_2\\ \varphi_2\end{bmatrix}\right) \quad (2)$$

with $L_1$ and $L_2$ referring to LFs on plane position 1 and plane position 2, respectively, and d being the distance between the plane positions along the z axis. As can be seen from Eq. (2), when considering propagation of light rays in plane and direction representation, the relation between parameters on both planes is not strictly linear. However, for small angles, this nonlinearity can be ignored.

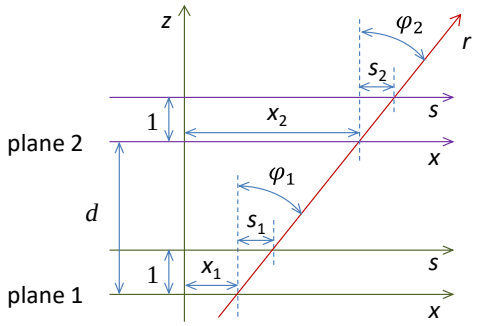



Figure 3. Light ray (r) propagation through space (representation on two planes)

### 2.1.3 Light Field Capture

To support capturing, processing, and the subsequent reconstruction of light fields using digital computers, the LF function needs to be sampled in all dimensions. However, sampling individual light rays in 3D space is impractical. In practice, spatial and directional sampling happens when capturing the light field with cameras of finite number and resolution. Considering a camera arrangement of equidistant cameras laid out on a row, with parallel camera axes, the following can be observed. For simplicity, a pinhole camera model [27] is used. The captured light rays cross the camera plane at a low number of positions – equal to the number of cameras used. No samples are taken between the cameras. As for the angles, each camera takes samples on a grid of horizontal and vertical samples determined by the optics in front of the image sensor. The number of samples captured by each camera is determined by the number of pixels captured by the sensor (for simplicity, the approximation of wavelength by RGB color representation [28], as well as the effects of Bayer coding [29] for capturing color images are disregarded). The arrangement of equidistant cameras can be regarded as a practical implementation of the two-plane LF parameterization, where camera sensors are placed on one of the planes and are focused on the other, thus forming *camera* and *image* planes. Figure 4 illustrates the camera arrangement for the case of HPO.

The number of times each light ray is sampled depends on the frame rate of the cameras. To ensure that samples represent the state of the light field at the same time, or in other words, are taken at the same time is ensured in two ways. To ensure samples captured by one camera are taken at the same time, a camera with a global shutter [30] (as opposed to a rolling shutter) should be used. To ensure that cameras constituting the camera array capture the same time, cameras with a trigger input and a synchronized trigger signal [30] should be used. Due to mechanical imprecision, a multi-camera array mounted on a rig does not represent precisely equidistant and parallel cameras. Moreover, camera lenses, even of the same type, do not have the exact same geometry, or may suffer from other manufacturing tolerances (i.e. off-center, tilted).

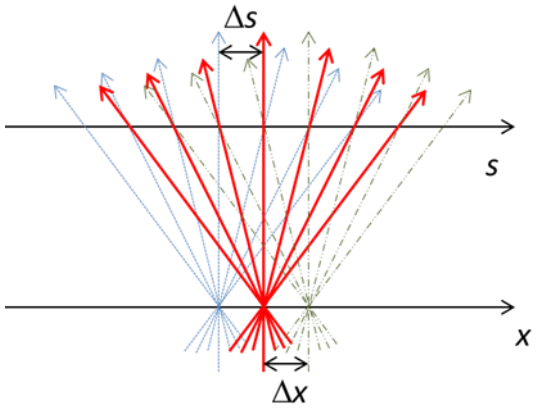



Figure 4. Equidistant cameras along plane (line) x focused on plane (line s). Sampling is introduced by the distance between cameras (baseline) and camera pixel resolution.

Individual cameras are typically described by intrinsic camera parameters [31][32], which consist of principal point, focal length, and parameters of a suitable lens model. Commonly used lens models are the OpenCV lens model [33] for narrow angle cameras and the OCAM lens model for wide angle cameras [34]. The position and orientation of cameras with respect to each other (or a designated reference camera) is described by the extrinsic camera parameters, which is typically described by a translation vector and rotation matrix for each camera. Intrinsic and extrinsic camera parameters of real cameras are typically determined by camera calibration algorithms [31][32][34] that estimate these parameters based on capturing patterns of known geometry.

Cameras also have differences in terms of capturing colors faithfully; small differences in terms of intensity / color may exist between the images showing the exact same scene by two different cameras. The color reproduction of cameras can also be calibrated [35], and the calibration information used to compensate for the differences before or during the rendering process.

Some better known camera arrays for light field capturing include the Stanford camera array [18], the Nagoya camera array [20], and the successive Light Stage rigs [36][37].

Light fields may also be constructed based on images captured by an array of cameras arranged on a different shape than a line or 2D array. Cameras may also be arranged on an arc around the scene, or - in case of small, static scenes - can be captured using a turntable and a single camera [39].

### 2.1.4 Epipolar Images

Epipolar geometry [40] describes the relationship between two images showing the same scene, as captured from two different viewpoints. An epipolar line of one camera consists of a set of points directly in line with the camera's optical center – as a consequence, all these points are projected

to a single pixel in the image captured by this camera. The other camera on the other hand captures this line as a line in 2D image space.

Epipolar Plane Images (EPIs) of a scene captured by a horizontal multiview camera can be constructed by stacking 2D images representing the same scene at the same time to form a 3D volume (see Figure 5), and re-slicing the volume so that one axis represents the horizontal (X) coordinate of the image, while the other axis represents the horizontal position of the camera (see Figure 6).

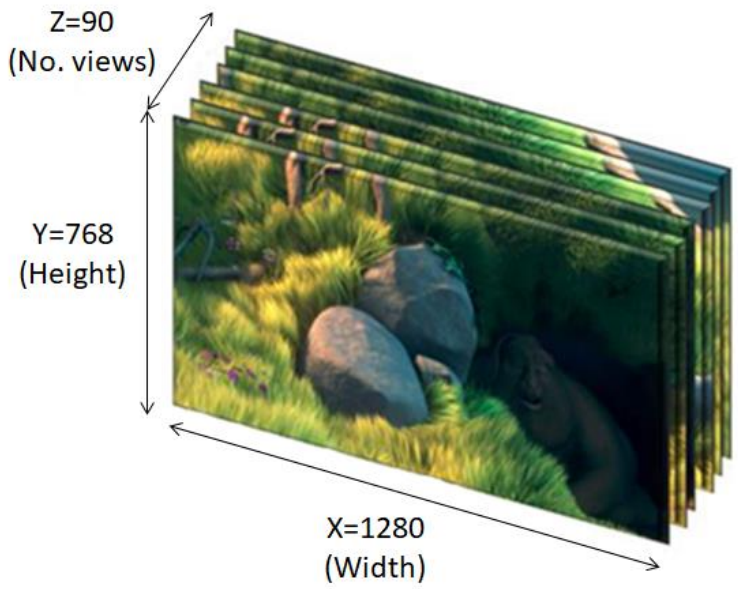



Figure 5. Image stack representing the same scene from multiple viewpoints (EPI volume)

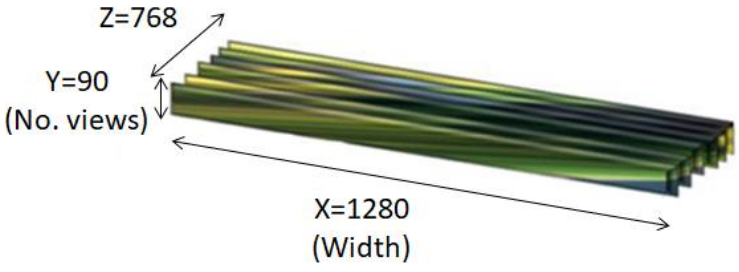



Figure 6. Epipolar images after reslicing the epipolar volume along the number of views.

The slanted lines typically visible on EPI represent the same object point, as it moves sideways on the image when captured from different viewpoints. The angle of the slanted line thus corresponds to the distance of the point from the camera.

EPIs are instrumental for analyzing and processing light fields as they represent the information carried by directional rays in the structured forms of slanted (sheared) lines and wider stripes, representing objects at different depths. EPIs also have a direct interpretation in terms of the LF two-plane parameterization, as they can be regarded as rearranging the light rays on the x-s coordinate system.

### 2.1.5 Use Cases of Light Fields

While this thesis considers light fields as an input to 3D displays, light fields have been used for many different purposes, including refocusing, relighting, or synthetic aperture imaging.

Refocusing involves changing the focal plane of the captured image, after the image was captured [41]. In terms of LF representation this would mean changing the position of the focus plane which is equivalent to shearing the EPI changing the slopes of the corresponding epipolar lines.

Relighting means rendering images showing a real object as it was lit by a set of arbitrary light sources, based on image samples that capture the object illuminated by known and controlled light probes [37]. Relighting would involve geometrical and CG models along the LF primitives.

Synthetic aperture imaging / focusing means generating an image that shows a part of the scene that is partially occluded from all captured views [38].

Camera arrays are also used for artistic reasons to implement time slice / bullet time effects, where a scene is shown from adjacent viewpoints while time is “frozen”.

## 2.2 3D Perception in the Human Visual System

The human visual system (HVS) uses several mechanisms to understand the spatial relations of real world objects, resulting in the perception of depth. The most important visual depth cues [42][43] used by the HVS are binocular disparity, motion parallax, vergence and accommodation, and pictorial cues [44][45]. The brain area called anterior intraparietal cortex (AIP), integrates all visual cues into a consistent perception of depth.

Binocular disparity is the difference of the position of object points when projected onto the left and right retinas. The amount of horizontal difference depends on the distance of the object, which is then reconstructed by the HVS by matching the feature points and estimating their distance. Please note that depth estimation algorithms based on stereoscopic cameras attempt to perform the same matching of features and estimation of their distance based on their disparity [46]. This visual cue is considered the most important one, thus stereoscopic display systems aim to reproduce this visual cue by showing two different images captured from two different positions.

Motion parallax is the effect when objects closer to the viewer appear to move faster than objects further away, when the viewer is moving. This is a monocular cue, that is, it can be observed with just one eye, as it can be observed with many animals not having stereo vision due to the positioning of their eyes [47]. Motion parallax is typically observed due to larger movements of the body and the head, though it has been reported that humans also use unconscious micro head movements to repeatedly check the consistency of the mental 3D model [48]. The computer vision technique called Structure From Motion [49] (SFM) attempts to mimic this mechanism of the HVS.

Vergence and accommodation are two corresponding oculomotor functions related to depth perception. Vergence is the simultaneous rotation of both eyes in opposite directions, so that the object of interest is projected to the center of the retina in both eyes. When looking at close objects, the eyes rotate towards each other (called convergence); when looking at an object further away, the eyes rotate away from each other (called divergence). Accommodation is the focusing of the eye's lens to the object of interest, so that is appears sharp on the retina. Retinal blur is the primary effect that controls the eye's accommodation [14]. Vergence and accommodation typically work together to provide sharp and high-resolution images, as well as depth information to the HVS. However, when using artificial means to reproduce depth perception (such as 3D glasses), the synchronization between vergence and accommodation is usually broken due to the screen not being in the depth where the represented object appears to be [50].

Pictorial cues [51], such as occlusion, perspective, relative size, depth from defocus, pattern scaling, shadows, and atmospheric effects are monocular depth cues that are often connected to learned experiences. The advantage of pictorial cues is that most of them work for any distance, thus the HVS typically relies on them when the primary depth cues do not work sufficiently due to the large distance.

As both binocular disparity and motion parallax occur mostly in horizontal direction - due to the eye's horizontal displacement and the primarily horizontal movement of people – most 3D displays do not even attempt to reproduce vertical parallax. In the author's experience, most viewers do not notice missing vertical parallax unless the content provokes them to move vertically.

## 2.3 3D Displays

The common purpose of 3D displays is to visualize real world or virtual spatial objects or scenes as they would appear in reality. All 3D displays aim to reproduce the plenoptic function as precisely as possible. The quality of approximation heavily depends on the 3D display technology used [52]. As humans observe the real world via two eyes, the only way to achieve such an illusion is to show different perspectives to the two eyes, as well as to reproduce other visual cues that make up the

full 3D illusion such as focus cues (being able to focus on objects ‘flying’ in space) and continuous parallax (being able to see the scene as it changes from different perspectives).

Showing different images to the two eyes can either be achieved by showing two different images directly to the eyes, or by creating a surface that has different appearance when observed from different directions (such as parallax barrier based, lenticular lens based, light field or holographic displays), or by presenting an object that really has a volume, and is able to show pixels (or voxels) at different depths (so called volumetric displays).

A detailed overview of different 3D display technologies can be found in [I].

### 2.3.1 First Historical Attempts for Creating 3D Displays

As there have been several early attempts to create 3D displays that ultimately led to today’s technologies, we present some of the early ones to show when and how they started, and how they contributed to the 3D displays we know today.

The first attempt to create stereoscopic 3D displays date back to 1838 when Wheatstone created a stereoscope [53] utilizing a pair of mirrors and two drawings depicting the same object from the left eye’s and right eye’s perspective. This simple approach proved the feasibility of tricking the brain into seeing 3D objects by presenting two matching images of the same object. The Brewster Stereoscope from 1849 (although not invented by Brewster himself) was a more compact unit that allowed the creation of compact hand-held devices, and was demonstrated to a wide audience during the Great Exhibition in 1851. The device was improved by Duboscq, and some 250,000 stereoscopes were manufactured in a short time. Between 1860 and 1930 stereoscopic photographs were used extensively. The stereoscope also appeared as the method for showing 3D movies. In Friese-Greene’s patent a stereoscope was used to fuse the images played back from two films in the late 1890s. A more practical approach for 3D movies was to use anaglyph, which was first demonstrated in 1915 by Porter [54]. The first publicly shown anaglyph movie was presented in 1922 by Fairall [55]. The anaglyph method has been widely used since then, mostly in low-cost use cases. Also in 1922 Hammond demonstrated the Teleview system [56] (see Figure 7), which used alternating left-right frames projected from a pair of projectors, and an individual shutter device for each viewer which was mounted on an adjustable gooseneck on each seat. Stereoscopic displays are still in use today in Virtual and Augmented Reality glasses, 3D cinemas and 3DTVs.

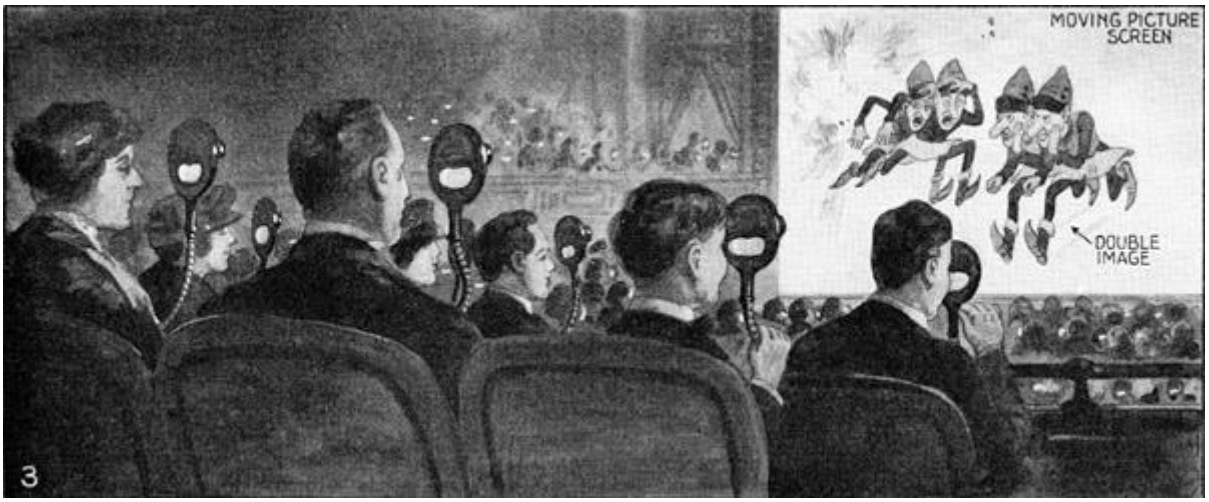



Figure 7. An illustration from a 1922 article about Televiewer

Parallax barrier was the first feasible enabling technology to implement autosteroscopic imagery. The left and right (or multiple) images are interleaved in a striped image, positioned behind a very dense set of equidistant vertical barriers. The first known implementation (on paintings, see Figure 8) dates back to 1692 by G. A. Bois-Clair. The first known photograph-based parallax barrier method appeared in 1903, when Frederick E. Ives demonstrated the Parallax Stereogram [57].

Figure 8. Gaspar Antoine de Bois-Clair: Double Portrait of King Frederik IV and Queen Louise of Mecklenburg-Güstow of Denmark

Lenticular lens-based displays originate from early attempts to achieve 3D imaging using integral lenses. The first well known implementation is from 1908 by Lippmann [58], who used small spherical lenses (fly-eye lens array) to capture and reproduce imagery including both horizontal and vertical parallax. The integral lens method was later simplified to a lenticular lens array in the 1920's that can be used to reproduce images with horizontal parallax only. This method was used

mainly for advertising purposes from the 1960's. The technology for capturing / creating content and manufacturing lenticular lens-based 3D pictures developed rapidly, and is still used today. This is all, however, for static, printed 3D content only. Lenticular lenses can also be used in combination with flat 2D screens to generate animated 3D imagery, which was first published in 1996 by van Berkel [59]. These displays are still in development and use today.

The origins of holography come from the work of Gabor from 1947, who worked on the improvement of the resolution of electron microscopes [60]. While the resolution improvement of electron microscopes was also achieved using different methods, the idea of holography triggered follow-up research by others. Gabor was awarded the Nobel Prize in physics later in 1971. Initially the depth of holograms was limited by the properties of the mercury vapour lamp technology used, which was later superseded by lasers in the 1960s. Leith and Upatnieks presented the first laser-based transmission holograms in 1964 [61]. The mass production of holograms was made possible by Dr. Stephen A. Benton, who invented holography with white light transmission (rainbow hologram), which enabled holograms to be seen in white light. While the development of static holograms is an interesting topic in itself, researchers like Benton and Lloyd pushed holographic techniques further, seeing moving 3D imagery (holographic TV / cinema, or interactive holograms) as the end goal. To that end, Benton, Hilaire and Lucente implemented successive generations of dynamic holographic systems based on electronic / computational holography, which calculate and generate the hologram patterns by computational means [62]. Electro-holographic displays are still an active area of research [63][64].

### 2.3.2 Stereoscopic, Volumetric and Autostereoscopic Displays

Stereoscopic displays [65] represent the most well-known technique for making 3D displays. In this technique, two images with horizontal disparities between corresponding object points are shown to the two eyes. Images with disparities, when projected on the eyes, generate retinal disparities, which evoke the stereopsis visual cues and trick the brain that objects are at different depths. The separation of two images shown at the same time to the right and left eye correspondingly can be achieved in a multitude of ways. The most widespread one is based on polarized light, and separation by polarization filters, placed in the glasses [8]. When used in cinemas, typically two projectors are used, equipped with different polarizers [8]. When this technique is used in televisions, polarization is performed on the screen surface, applying different polarization on alternating rows of pixels [66]. A different approach involves showing images for the left and right eyes alternating rapidly, and covering the eye which is not supposed to see the image being presented, using active glasses [66]. Two displays, a beamsplitter, and stereo glasses can also be used to implement desktop 3D displays [67][68].

In the case of a single user, it is possible to use two displays in front of the eyes, mounted on some kind of headwear, so that the displays move together with the viewer's eyes [69]. Such displays are known as Head Mounted Displays (HMD). HMDs either occlude the vision of the viewer completely

[70], or superimpose a virtual image on top of the real image [71][72]. HMDs track head movements, which enables virtual- or augmented reality applications [73]. Recently mobile phone displays are used with extra lenses as HMDs [74].

Volumetric displays [10][75] reproduce 3D imagery by making up a volume inside which voxels emit light. This can be achieved by having multiple light emitting layers. A typical implementation involves a rotating screen onto which a rotating projector projects images corresponding to the angle [10]. Instead of using moving layers, one can use multiple layers [76]. Another typical implementation includes a matrix of light emitting devices which rapidly rotate or oscillate in a given space [77]. A solution that does not involve time multiplexing is made up of several physical layers, each layer being a display panel with light emitting pixels [78].

Displays that have a flat screen, yet able to produce an image with 3D appearance without the user wearing any kind of apparatus in front of their eyes are referred to as autostereoscopic displays [79]. These displays can create a 3D image by means of direction selective light emission, which means that each pixel can have a different appearance based on the direction they are observed from. Autostereoscopic displays mainly reproduce stereopsis by generating a number of perspective views. In these displays continuous parallax is rather limited due to the limited number of views. Parallax barriers [80] are one example of autostereoscopic displays. These can be implemented by a fixed set of slits, as well as by means of a dual layer LCD, the upper layer of which forms the barrier when activated [81]. The latter solution allows the barrier to be active or inactive, enabling the same display to be used as 2D display. Parallax barriers are typically used to create two views, dominantly in mobile devices. Lenticular lenses used in conjunction with a flat display enable the creation of 3D pixels, by each lenslet covering multiple pixels of the underlying display [9]. Such a lenslet allows the viewer to see the color of different underlying pixels, based on the viewing direction. Lenticular lens based displays typically support more than two views to enable motion parallax. This is the technique used in most desktop autostereoscopic displays.

### 2.3.3 Light Field Displays

Light field displays show a very dense set of different directional light rays. Projection-based light field displays [6][82][83] generate these light rays by means of multi-projection. A dense horizontal array of projection modules project light towards a common screen, to the same area (see Figure 9), but from different directions, typically from behind (so called back-projected displays). The projector plane and the screen plane can be well described by the two-plane parameterization of light fields, though only one angle is reproduced, as projection-based light field displays typically reproduce the horizontal parallax only (HPO). As such, these displays reproduce (x, y) positions on the screen plane, one angle in the horizontal direction (θ), time (t), and color (λ) with an RGB approximation.

The screen enables the projected light rays to pass through without changing their direction, applying only minor horizontal diffusion. This diffusion can be considered as a discrete-to-continuous transformation of the light rays emitted by a finite number of light sources, transforming them into a continuous and homogeneous image. Using this setup, each point on the screen can emit different intensity or color to the different directions. Viewers on the other side of the screen can observe a subset of the emitted light rays by each eye, making up a stereoscopic view. Viewers moving sideways will see motion parallax, as they will see different light rays emitted from the same screen positions. The light rays emitted from the screen can be described by the hit point + angle parametrization. The light field display can be considered as a digital to analog converter that converts the sampled representation of the light field into an approximation of a continuous light field that was originally captured.

An alternative configuration is front-projected, in which case the projection modules are on the same side as the viewers, and the holographic screen is replaced with a reflective holographic screen.

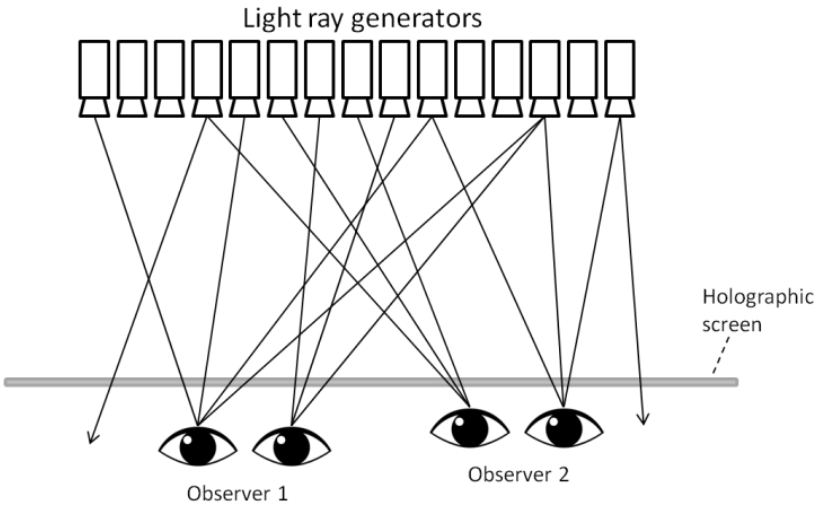



Figure 9. Light field display architecture

The main advantage of light field displays is scalability in terms of pixel count, screen size, field of view and angular resolution, thanks to the distributed nature of the projection modules that serve as the source of light rays. Due to this arrangement it is possible to create 3D displays with 100+ MPixels (million light rays), close to 180 degrees Field Of View, or sub-degree angular resolution, which are all difficult to achieve by means of flat panel based autostereoscopic displays. The number of projection modules is typically 24 to 80, have VGA to HD resolution, and updated with up to 60 frames per second. Those light rays emitted by projection modules on the sides of the setup, which could otherwise not reach the screen can be reused by means of side mirrors, which turn them back towards the screen. This technique also increases the Field of View of the display.

Along with binocularity, light field displays can potentially recreate the motion parallax cue and the focus cue, as a sufficiently dense set of rays can be rendered.

It is important to note that light field displays do not track viewers to update the image according to their position, as the whole light field is projected all the time. As a consequence, no latency is involved when viewers observe different parts of the image, but this also means that the whole light field has to be rendered every frame.

The complexity of light field displays and especially the number of light rays controlled simultaneously necessitate a distributed control and rendering system. This is typically implemented by using a number of GPU-equipped computers. In such a setup, each GPU output is responsible for a projection module, rendering one slice (as opposed to one view) of the complete light field. The computers (rendering nodes) are connected using a high-speed network. The display is typically controlled from a user computer that feeds the display with rendering commands. The software running on the rendering cluster allows running arbitrary rendering algorithms and synchronized update of the light field.

As the projection modules and the screen may not be precisely positioned according to the optimal placement, slight misalignments may occur. These are compensated by means of display calibration [84], which determines the precise mapping between the pixels of the projection modules and the light rays emitted from the screen. A separate color / intensity calibration step measures the slight differences between projection modules. Geometry and color calibration data are used during the rendering process to ensure the precise reproduction of the desired light field. Further details on projection-based light field displays are described in [6].

Light field displays reproduce all depth cues to some extent. Binocular disparity is reproduced for objects in the depth range. Motion parallax is reproduced in the horizontal direction, within the FOV of the display, with a given angular resolution. The vergence-accommodation conflict observed in stereoscopic displays is much reduced with light-field displays [85]. Pictorial cues depend on the way content is captured or rendered, and generally can be well reproduced.

### 2.3.4 Other 3D Display Technologies that may Benefit from the Presented Results

Today many companies and research labs work towards solving the challenge of creating the perfect 3D experience. While these efforts follow different approaches for implementing the display, the challenges associated with showing live 3D imagery on these displays are inherently the same.

Most display technologies (except stereoscopic ones) can benefit from the work related to spatial and angular resolution measurement presented in this dissertation for both content creation and to facilitate objective comparison of the capabilities of different displays. Those with high pixel count can benefit from distributed rendering, and utilize the results related to 3D video encoding and decoding to support the transmission and visualization of live 3D imagery.

Voxiebox / Voxon VX1 [86] is a volumetric display that shows 500 million points per second into an enclosed 3D volume using high-speed projection technology and a reciprocating screen. A volumetric display can be modelled in ray space as multiple (time multiplexed) emitting planes, which emit light to all directions. As such, considering the plenoptic function $P(x, y, z, \theta, \varphi, t, \lambda)$, the parameters (x, y) are positions on the reciprocating screen, z is the position of the screen at a given time instant t (considering the high frequency of the moving screen), light is emitted to all directions ($\theta$, $\varphi$), t (considering the progression of time when showing animated scenes) represents time, and $\lambda$ is constant due to projecting with a single color. The Voxon VX1 thus creates a four-parameter approximation of the plenoptic function $P'(x, y, z, t)$. Transmission and processing of live 3D data is definitely an issue for such displays, as explicitly discussed on Voxon's blog [87]: "The biggest constraint is really the huge volume of volumetric data (full 360 degrees) in a full game of soccer. If someone capturing that 'volumetric video' can crunch that data and create a stream that could be displayed in real-time, then I have no doubt we could display it on a volumetric display." Also, there is no known work that addresses the measurement of effective resolution of a volumetric display – all we know is the total pixel count, like in the case of most 3D displays. Therefore both issues addressed in this work are also relevant for state-of-the art volumetric displays.

Ostendo Technologies proposed the Quantum Photonic Imager (QPI) [88], which, thanks to its high brightness, power efficiency and compact size, can be used to create light field images. To create larger format displays, many QPIs need to be tiled together, as demonstrated by the company using a 4x2 array. This indicates that processing high number of pixels in a distributed way will be necessary, once the technology matures, and will be used for showing live imagery. Also, being a light field display, there is no methodology to quantify spatial or angular resolution.

The near-eye light field display [89] from Nvidia and Stanford provides light field displaying capabilities using microlenses over two OLED displays in front of viewer's eye's. As this display effectively works as an integral imaging / light field display, both the resolution measurement and video content compression techniques apply. As the authors note, "practical applications will necessitate manufacturing larger microdisplays with smaller pixel pitches, enabling wide fields of view and high resolutions, respectively", indicating that to produce production-grade near-eye light field displays, the pixel count involved is expected to increase dramatically.

Compressive light field displays [90] generalize the idea behind parallax barrier, but instead of one fixed barrier, they use multiple LCD layers and a directional backlight to emit light. The image is then formulated by optimizing the image layers using nonnegative tensor factorization in a way that the resulting light field is as close to the desired light field as possible. These displays effectively do compression of the content in optics, and thus the range of images that can be displayed is limited. Measuring the effective spatial and angular resolution of these displays is especially interesting, as it does not only depend on the physical properties of the display elements, but also the factorization algorithm used (approximate or exact, and the quality of the approximation). On the

content compression side, one may think that the optical compression inherent to compressive displays is efficient enough in order not to need any further compression. However, the compression that happens inside these displays is highly display dependant. In a future content distribution system a display independent compression solution needs to be used, with subsequent conversion to display specific representations. Super-multiview video is a good candidate for fulfilling all these requirements.

Both near-eye and compressive light field displays reproduce (x, y) positions from the plenoptic function, as well as two directions (horizontal and vertical: θ, φ), time (t), and color (λ) with an RGB approximation.

### 2.3.5 Choice of Display Technology

The display technology of choice for the following discussion is projection-based light field type. Light field displays provide one of the highest quality 3D imagery without glasses. Due to the inherent scalability of the technology due to projection, displays of arbitrary size can be implemented, which has already been proven in displays with screen diagonals ranging from 10” to 140”, while image resolution, field of view, angular resolution, frame rate are also scalable depending on the capabilities and the number of imaging components used. Projection-based light field displays are the clearest form for light field displays. Though they can reproduce the plenoptic function up to five parameters (x, y, θ, t, λ), thus ignoring vertical parallax, all the other parameters can be reproduced with almost arbitrary granularity. Missing vertical parallax is typically not recognized by viewers, due to the horizontal displacement of human eyes, as well as the typical motion of viewers is horizontal, thus horizontal motion parallax is more important for subjective quality [91].

With these properties in mind, projection-based light field displays are expected to be one of the most important technologies in the field of high-end 3D visualization, and as such, research targeting the enhancement of these displays is beneficial on the long run. On the other hand, contrary to very well understood technologies like stereoscopic 3D based on glasses, there are several challenges to be solved, partly due to the unusual image formation, partly due to the number of pixels / light-rays making up the 3D image, which necessitates a distributed rendering system in most cases. The work presented targets solving some of these challenges.

## 2.4 Limitations of 3D Displays

The amount of visual information available in the real world is unlimited, made up of light rays propagating from continuous locations to an infinite number of directions, changing during all time instants, and of different wavelengths on a continuous scale. This infinite set of light rays is best described by the plenoptic function, as introduced in Section 2.1.

All displays target some approximation of the plenoptic function up to a given number of dimensions, extent and discretization. In general, the more precise the approximation is, the displays is deemed to provide a more faithful representation of reality.

2D displays generally reproduce light rays for a range of discretized positions in space (modeled by a regular 2D matrix of pixels), with a subset of wavelengths visible to the human eye (modeled by several color channels and discretized intensities on each channel) with discretized time (represented by the refresh rate of the display), and with no reproduction of directional light (see Figure 10).

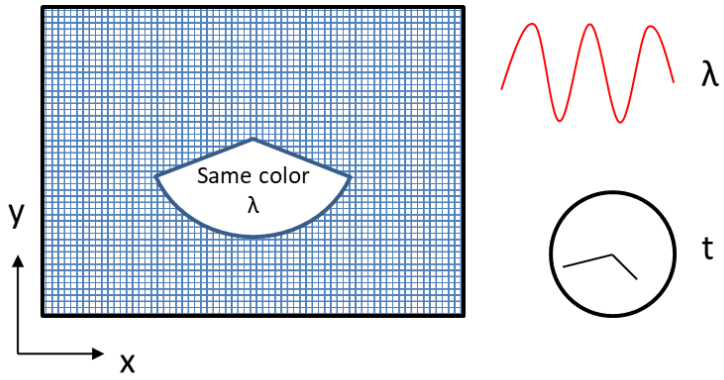



Figure 10. A 2D display reproduces pixels along (x,y), with color (λ) along time (t)

Autostereoscopic 3D displays reproduce binocular parallax by mimicking the different appearance of the same scene when observed from different directions, that is, by means of direction selective light emission, up to a certain viewing angle and a certain discretization of directions (characterized by angular resolution), see Figure 11.

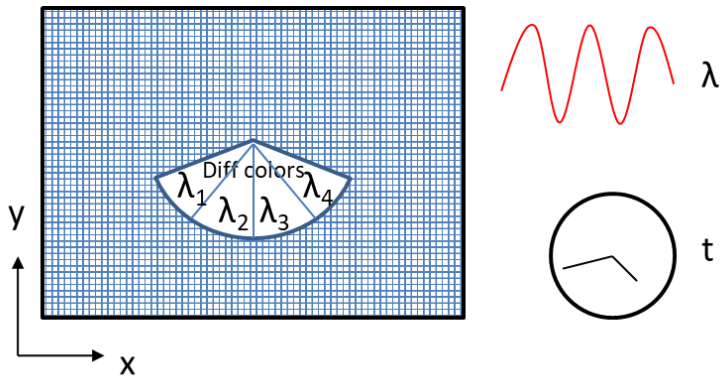



Figure 11. A multiview autostereoscopic display reproduces pixels along (x,y), with view dependent color (λ) along time (t)

Assuming that we have a display device capable of emitting a given number of pixels / light rays with a given frequency, those light rays can be exploited in a multitude of ways (setting aside color

reproduction for the moment). The simplest trade-off is between horizontal (X) and vertical (Y) resolution (number of pixels) along a flat display surface having X*Y pixels in total. A more complex trade-off is when N pixels can be grouped together to form a 3D pixel (that is, a pixel that has different appearance when observed from different directions), in which case the number of 3D pixels we may reproduce will be the number of 2D pixels / N. These N directional (sub)pixels can then be used for reproducing horizontal directions, vertical directions or both. Assuming that our display reproduces only horizontal directions, we may choose to represent N directions over a wide Field Of View resulting in low angular resolution between views, or over a narrow Field Of View resulting in higher angular resolution but narrower viewing angle (see Figure 12).

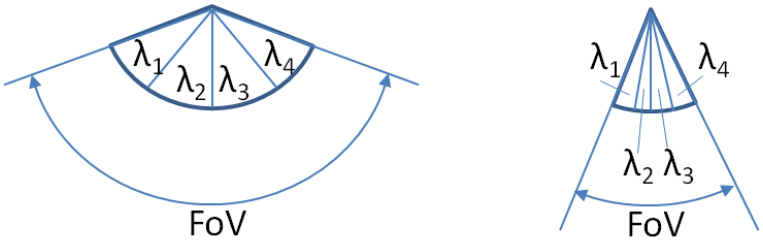



Figure 12. Left: wide field of view with coarse angular resolution. Right: narrow field of view with fine angular resolution, using the same number of views

It's also possible to have both: repeating the N directions several times over a wide Field Of View is another possible trade-off, which on one hand results in high angular resolution and wide field of view, but on the other hand the field of view will contain repeating visual information, resulting in invalid zones in the viewing zone. In case we have pixels that can be controlled with very high speed, the display may resort to time multiplexing, in which case one pixel is reused several times to create (seemingly) different pixels, without compromising image quality in a major way. Going further, one may trade off color depth or the number of color channels for refresh rate, in case the total bandwidth of some component of the system is a concern.

As the above examples show, the properties of the visible image cannot be described by a single metric, and it is also quite difficult to find a group of metrics that can be equally well used for all the different kinds of 3D displays. The metrics proposed in this thesis enable comparing 3D displays using different technologies in terms of image reproduction capabilities.

The total number of controlled pixels with a refresh rate that is appropriate for the human eye may be considered to be in line with the expected image quality, but the way these pixels are distributed over the 3D image can have a major effect on the perceived quality and usability of the display.

On top of the theoretical limitations due to the number of pixels, there are other, practical factors that affect image quality. The contrast ratio, crosstalk between adjacent pixels, temporal crosstalk, brightness, color gamut, color uniformity all affect perceived image quality in some way. The metrics currently used for the quantification of 2D and 3D displays [92] attempt to capture and quantify

these properties, and relate them to the subjective quality of the imagery as it appears on the screen.

## 2.5 3D Video Representations

Some 3D scene representations, which are generally considered to be suitable for rendering multiview or light field images are introduced in the following subsections. This list is not exhaustive, as it targets the specific use case of representing real scenes, and targeting projection-based light field displays.

### 2.5.1 Light Field Display-Specific Representation

The representation that is closest to the display hardware requires one image to be projected by each projection module. This representation is highly display-specific, as one needs to take into account display geometry (including projection modules, screen, and mirrors) to generate it. When side mirrors are used to extend the display field of view, a projector image contains multiple sub-images. It is possible to compress the images / videos corresponding to projector images directly, which makes real-time playback of pre-processed light field videos possible.

A direct method to generate projector images directly is ray tracing, when a computer renders exactly those light rays that are needed for the display [93]. This results in the highest possible quality, with no ray interpolation taking place anywhere in the image generation process (provided that display calibration data is taken into account in the rendering process). The downsides of this method are: it is only applicable for synthetic scenes, as the necessary light rays cannot be captured by practical cameras available today; the generated images are highly display dependant; no accelerated rasterization-based rendering techniques can be used for the synthetic content.

### 2.5.2 Image-Based Representations

In order to transmit and store content in a format that is generic to light field displays (as well as other 3D displays), and to allow practical capture of both real and synthetic content, the representation typically used is a multi-view representation that shows the scene from many different viewpoints. The viewing points may be arranged on a line, with cameras looking parallel, arranged on a line with camera targets pointing to the center of the scene, or the cameras may form an arc around the scene (see Figure 13). The number of cameras is typically matched with the number of projection modules (plus virtual projection modules), and range between 32 and 128 in practical cases.

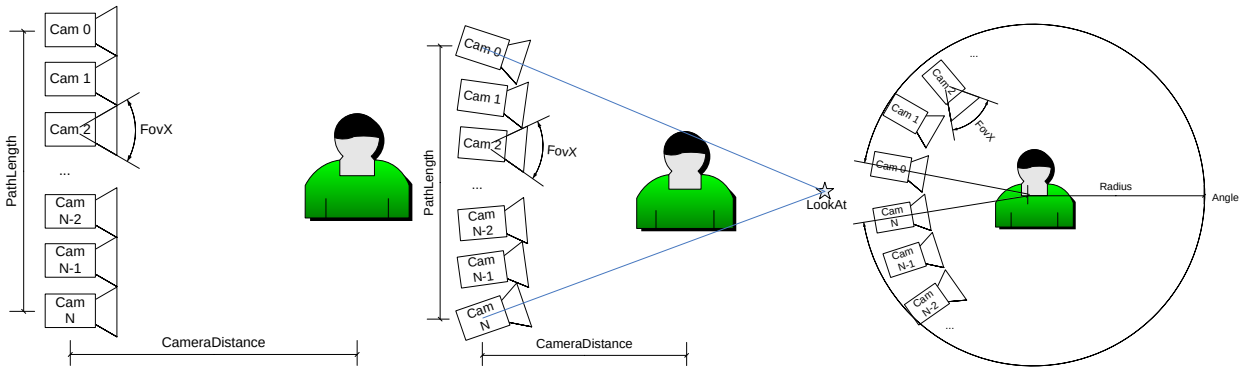



Figure 13. Left: Linear, parallel camera setup. Middle: Linear, converging camera setup. Right: circular / arc camera setup

Image sets with the above configurations can either be rendered (any rendering tool is capable of rendering such image sets with the right automation), or captured using an array of cameras. Once the images are rendered / captured, using the correspondence between the capture cameras and the parameters of the LF display, the matching between captured and reproduced light rays is calculated, and light rays emitted by the display are interpolated from the captured light rays [39]. Compensation with calibration data on both the capture and display side (both geometry and color) are necessary to reach high-quality reproduction of scenes. While the number of views to be stored / transmitted is high, resulting in significant bandwidth requirements, this dense light field representation is still the preferred one, due to its display independence, and the high visual fidelity that it enables.

### 2.5.3 Image Plus Depth-Based Representations

Representations that store a 3D scene using image + depth representation are popular in the research community, in the hope that having depth information helps synthesize new views quickly, removing the burden of transmitting many views. While image+depth based representations combined with high-quality view synthesis can provide good results [94][95][96] (see also Figure 14), pure image-based representations have a wider range of applicability, and do not suffer from some of the typical artifacts and limitations of image+depth based representations.

According to plenoptic sampling theory [98][99] there is a trade-off between the number of images and depth map precision when targeting a constant reconstruction quality. That is, the same quality can be achieved with a large number of images and pure ray interpolation, or a small number of images and precise depth maps. This suggests that even a depth map with a few bits precision can greatly enhance the reconstruction quality (see Figure 15), as it guides ray interpolation: depth information suggests where the light rays shall cross each other, instead of assuming a single depth plane [100].

Figure 14. Left: four images with estimated depth maps. Right: four-camera rig from project MUSCADE [97]. © Springer, reproduced with permission.

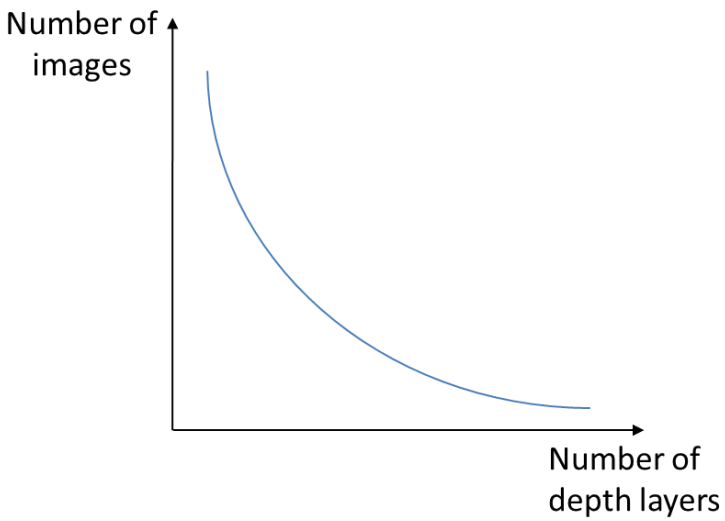



Figure 15. In joint image and geometry space, there is a tradeoff between image samples and depth layers for a given rendering quality

There are other approaches to synthetic view generation then “blind” ray interpolation, which use pixel shifting or warping [101], back and forward projections [102], layered processing [103], or other similar means to generate novel views by directly manipulating the pixels of the source views based on depth information of the pixels, as opposed to considering the input images as a set of known light rays.

The naïve approach assumes that the depth map is a perfect representation of the scene’s depth at the given pixel locations, which is typically not the case. In reality, depth maps suffer from several issues, regardless of the means they have been captured / generated. Please note that we use depth and disparity interchangeably (also later, where applicable), but this does not limit the discussion due to the possibility to convert between the two. Estimated depth / disparity maps based on stereo matching [104], depth maps measured with Time Of Flight (TOF) sensors [105],

structured light [106], or the combination of these [107] all exhibit some form of artifacts, which have a negative effect on all algorithms that rely on them as perfect input data. The only exceptions are synthetic depth maps generated by a rendering tool during the rendering process of the images.

This is the reason why all advanced view synthesis algorithms start with simple view synthesis, but then have to use heavy pre- and post-processing of the input data as well as the resulting views, in an attempt to mitigate the most apparent artifacts that are caused by the errors contained in the depth maps [108][109][110].

Some of the typical depth map errors include: depth incongruity on diffuse surfaces; unreliable boundaries; temporal instability; limited depth resolution. Depth incongruity means that pixels that are on the same depth do not have the same depth value. This typically manifests itself as depth gradients on an object that has the same depth over its surface. It also happens between images, that is, the same object is assigned a substantially different depth value in two adjacent views. Unreliable boundaries result in flickering edges at depth discontinuities, even with objects that are standing still. Averaged depth values between the foreground and the background depth layer can also be observed in some cases, which is obviously not correct for any of them. Temporal instability means that the same still object is changing its depth value in subsequent frames, sometimes even dramatically.

If we consider the original (that is, as it was originally defined and used in computer graphics) meaning of depth maps: Depth map is an image or image channel that contains information relating to the distance of the surfaces of scene objects from a viewpoint. Analyzing typical depth map artifacts from this point of view, it is clear that typical depth map artifacts represent situations impossible in physical reality. Typical depth incongruities would mean walls and floors have huge bumps in them, or are slanted to one or more directions. Unreliable boundaries would mean that objects do not have well defined edges, skin of people as well as objects would be toothed. Depth values that are averaged between a foreground person and background wall would mean that some parts of the body would be floating somewhere behind the person, or people would melt into the walls. Temporal instability would mean that walls and objects are constantly moving forward and backward, as well as changing their shape with no reason. People would move forward meters and then back in fractions of seconds. Input data that represents such impossible situations clearly causes problems during view synthesis.

Even though the quality of depth maps as well as the visual quality of synthesized views based on depth maps is getting better and better every year, the visual quality of most depth-map based view synthesis algorithms so far is clearly insufficient for everyday use for non-professional viewers. Some researchers argue that depth-map estimation is an ill-posed problem and should be avoided altogether [111].

Proponents of disparity-based pixel shifting and related methods claim that erroneous disparity values can only occur if the algorithm that generated the disparity values found a correspondence with another pixel of the same color on the same object, and therefore will only be used during the rendering process for rendering that same object and same color, not causing any issues in the output. The assumption here is that depth / disparity values are used only for view interpolation, and not for view extrapolation. Wide-FOV light field displays on the other hand typically require heavy view extrapolation when fed with typical narrow-baseline image+depth content [94], and in such cases, the assumption breaks, resulting in disturbing view synthesis artifacts.

View extrapolation over the leftmost or rightmost camera poses another challenge, as unknown parts of the scene will be revealed behind occluding objects. While this may occur with image interpolation in a small scale, image extrapolation makes this appear on massive scale. Image inpainting targets filling the unknown image areas with plausible content. Inpainting algorithms [112][113][114] however achieve varying levels of success in filling the missing regions, and also very complex computationally.

Image+depth based approaches also have difficulties representing scenes where depth values are ambiguous or otherwise not well defined. Consider semi-transparent surfaces or mirroring surfaces, non-transparent gases, fire or fog in a scene. While using multiple depth layers may solve some of these challenges, they are clearly not generic enough to represent all kinds of natural scenes that occur in the real world.

### 2.5.4 Geometry-Based Representations

Due to massive advances in computer graphics in recent years, scene representations that contain geometry, texture, and other material information, light sources, animation, and other supporting information are perfectly suited for computer-generated scenes [115].

There are efforts towards using the same or similar representations for live scenes [116][117], arguing that images rendered by today's computer graphics algorithms can be practically indistinguishable from real images. While this might be true, it is often overlooked that the issue of capturing and converting real scenes to geometry-based representations is lagging far behind in terms of visual realism [118][119]. Those synthetic scenes that are indistinguishable from real images are created by computer graphics artists, using massive amount of manual labor and parameter tuning.

## 2.6 Image and Video Codecs

Image codecs are used to store or transmit images in a compact digital form, in order to reduce the storage space / bandwidth necessary. Many image codecs have been used in the history of com-

puters, both lossless (which reproduce the original image exactly) and lossy (which reproduce an approximate image, typically targeting perceptual similarity).

### 2.6.1 JPEG and JPEG2000

In this work, JPEG and JPEG2000 image codecs are used, which have been both developed by JPEG, the Joint Photographic Experts Group (ISO/IEC JTC 1/SC 29/WG 1). JPEG [120] is based on block-based encoding, approximating the contents of each block (MCU, Minimum Coded Unit) using perceptually similar content that can be represented in a compact digital form.

The image is first converted to YUV color space, to separate luminance from chrominance data. The chrominance channels are subsampled (typically by 2), as the human visual system is less sensitive to differences in color than intensity [121]. The color channels are then divided into 8x8 pixel MCUs. MCUs are first transformed to frequency domain using Discrete Cosine Transform (DCT), followed by quantization of the resulting coefficients using a quantization table recommended by the JPEG standard. The quantization table is different for the luminance and chrominance channels. The DC coefficients of each successive MCU are encoded separately using differential encoding. AC coefficients are ordered using a zigzag scan pattern to maximize the number of successive zero coefficients. Both DC and AC data is then encoded using Huffman encoding.

While JPEG is a still image codec, it is also used for encoding motion pictures (Motion-JPEG or MJPEG), in which case individual frames are encoded as JPEG images.

JPEG2000 [122] uses wavelet transform to compress images more efficiently then JPEG. After converting the image to YUV color space, the chrominance channels are usually subsampled. Depending on the codec's configuration, the whole image, or tiles are transformed using wavelet transform. The wavelet transform is chosen from Cohen–Daubechies–Feauveau (CDF) 9/7 wavelet (in case of lossy coding) or CDF 5/3 wavelet in case of lossless (reversible) coding. The wavelet coefficients are quantized according to the quality / bitrate trade-off defined for the codec. The coefficients represent sub-bands, which are further subdivided into rectangular regions in the wavelet domain. The Embedded Block Coding with Optimal Truncation (EBCOT) scheme then orders coefficient bit planes into three passes, ordered by significance. The bits resulting from this encoding pass are subsequently encoded by an arithmetic encoder.

JPEG2000 is also used for encoding moving images, with frame-by-frame encoding. Both JPEG and JPEG2000 support lossy and lossless encoding.

### 2.6.2 H.264 and HEVC

Due to the large amount of information contained in moving images (typically consisting of 25 to 30 images per second), there are several methods and standards for performing video compression, which are used in practically all areas where moving pictures are stored or transmitted in a digital

format. The most well-known video compression standards originate from MPEG, the Moving Picture Experts Group, which is an ISO/IEC working group tasked with the coding of moving pictures and audio (ISO/IEC JTC 1/SC 29/WG 11). The successive video compression standards by MPEG are all based on encoding blocks using approximate, but perceptually lossless techniques, and block-based motion compensation. The successive standards improved the tools to enable more and more compact representation of videos, at the price of more complexity during encoding and decoding.

H.264 [123] first groups subsequent video frames into Groups Of Pictures (GOPs). Depending on the GOP structure, some frames are encoded with no reference to other frames in the GOP (so called I-frames), while other frames will be encoded based on reference frames and motion compensation (so called P-frames when referencing one frame, or B-frames when referencing two frames). In all cases, images are transformed to YUV color space, where luma channels are subsampled. Images are subdivided into Macroblocks of 16x16 pixels. Macroblocks in I-frames are estimated by several pre-defined prediction modes. Macroblocks in B and P frames are estimated based on similarity with image areas in the reference image(s). This mechanism basically exploits the fact that objects moving across the image have a similar appearance, thus image areas can be reused in subsequent frames. The residuals, which represent the difference between the image to be encoded and the approximation provided by all the prediction / motion compensation schemes are transformed with DCT and quantized. The resulting coefficients are then ordered in a zigzag pattern and encoded by using the Context-adaptive variable-length coding (CAVLC) entropy encoder.

HEVC [124], while follows a similar method as H.264, extended the coding tools in H.264 in several aspects. Coding Tree Units are variable sized blocks (4x4 to 64x64 pixels) used to subdivide the image. The entropy encoder is context-adaptive binary arithmetic coding (CABAC), while prediction schemes were also extended. Motion vectors can also be predicted in HEVC.

MPEG also created extensions on top of the standardized codecs to enable sophisticated use cases of the basic codecs. To encode stereoscopic and multi-view 3D video, MPEG developed H.264/MVC, and later MV-HEVC and 3D-HEVC. How these match the use case of light field video encoding is discussed in [IX].

# 3 Quantification of Light Field Displays

The most apparent property of a display is screen size, typically characterized by horizontal and vertical screen size, or by screen diagonal and aspect ratio. The second most significant property of raster displays is their resolution, the number of pixels making up the whole image on screen. Metrics related to the spatial distribution of pixels like luminance step response, or resolution from contrast modulation (which is the closest to our proposed method) also apply for 3D displays. These are applied on the screen plane on 2D displays, and also applicable for 3D displays on the screen plane, resulting in the effective 2D resolution of the display. This, however, is not equal with the total number of pixels making up the 3D image; one typically cannot see all the pixels at the same time on the screen plane, being spread across the field of view of the display. While most 3D display manufacturers report the total number of pixels, there is no easy way to check where all these pixels are emitted by the display. It would make perfect sense to measure spatial resolution at different depth layers, too, which requires different tools than a light meter. Those measurements that assume rectangular pixels however cannot be applied to all 3D displays, which may not have a regular pixel structure. The spatial resolution measurement method presented later generalizes spatial resolution measurement in a way that can be applied to any 3D display that is not head-mounted (including those with an irregular pixel structure), and can measure resolution at different depth layers, too.

Intensity and color related 2D display metrics such as black level, maximum brightness, contrast ratio, linearity of gray scale, color fidelity, color gamut, brightness uniformity, color uniformity and contrast uniformity also apply for all 3D displays. These can be measured on the screen plane using light and color meters, provided the display has a screen surface and shows 2D content.

One can also measure viewing angle related metrics on a 2D display. These, however, measure the changes in the perceived image when observed from different viewing angles in terms of how the brightness, contrast and colors change. These metrics attempt to measure whether a viewer who observes the screen from a different angle will be able to conveniently see the content like a

viewer observing the screen from the center of the viewing zone. Examples include viewing angle luminance change ratio, viewing-angle color variation and viewing-angle color inversion. The assumption behind is that the same image is supposed to be shown to all directions, while in the case of autostereoscopic 3D displays, the opposite is true when showing 3D content. An autostereoscopic display must show different content to different viewing angles in order to reproduce a 3D image. This means that viewing angle-related measurements, as defined for 2D screens can only be used when the 3D display shows a 2D image, and will quantify how well the 3D screen operates as a 2D screen in terms of viewing angle. What is more interesting in case of a 3D screen however, is how well it can reproduce 3D content. This primarily depends on its angular resolution, which also determines the depth range of the display. Angular resolution means how small is the angle that the display can control independently from the light emitted to adjacent angles. The angular resolution measurement method proposed measures angular resolution over the field of view of the display. In cases where the angular resolution is non-uniform, the method can be extended to determine the angular resolution over different areas of the field-of view. Angular resolution can also be defined and measured for horizontal and vertical directions, as these might be different. All the 2D display measurements referenced above are described in the International Display Measurement Standard [92].

## 3.1 3D display's Passband Estimation in the Fourier Domain

Spatial and angular resolutions describe what a display is capable to reproduce in terms of the frequency of the content without excessive distortions. From a signal processing point of view, they describe the bandwidth or passband of a light-field generator device in the Fourier domain. The passband of a projection-based light-field display with a known internal structure can be estimated based on the geometrical parameters of the projection and screen setup [IV]. The approach is briefly described next.

As illustrated in Figure 16, a typical projection-based light field display consists of $N_p$ projection engines uniformly distributed on the ray generators (RG) plane ($p$ - plane) over distance $d_p$ thereby making the distance between engines $x_p = d_p / (N_p – 1)$. Each projection engine generates $N_x$ rays over its field of view $FOV_p$. We assume that the rays hit a certain plane (screen plane, $s$ - plane) parallel to the RG plane at equidistant points. Without loss of generality, we can assume that this results in an angular resolution at the RG plane as $\alpha_p = FOV_p / N_x$.

The 'trajectory' of each ray is uniquely defined by its origin $x_0^{(r)}$ at the RG plane and its direction determined by angle $\varphi^{(r)}$. The position of the ray at distance $z$ from the display is given as

$$\begin{bmatrix} x_z^{(r)} \\ \varphi_z^{(r)} \end{bmatrix} = \begin{bmatrix} x_0^{(r)} + z\tan(\varphi^{(r)}) \\ \varphi^{(r)} \end{bmatrix}.$$

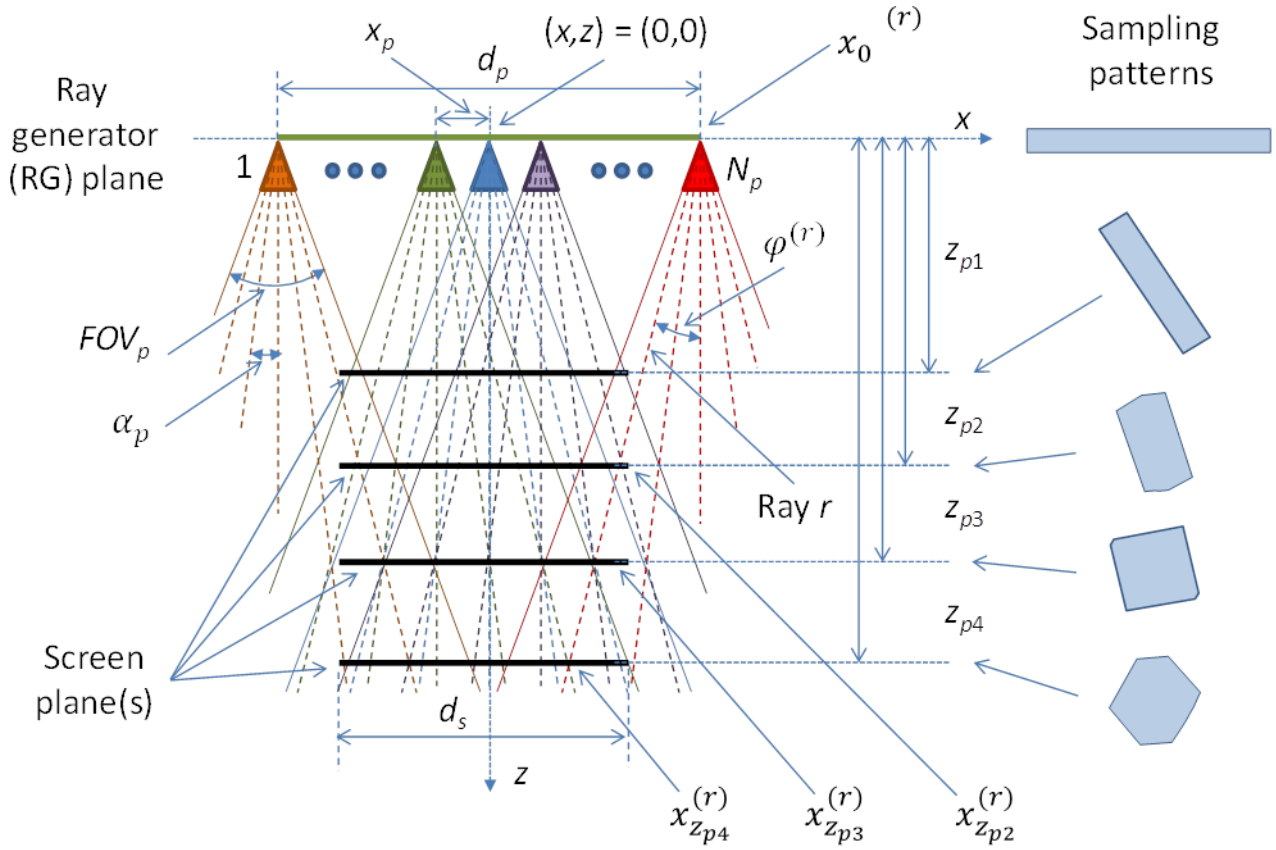



Figure 16. Ray propagation in a light field display – different sampling patterns are illustrated for different positions of the screen plane.

Each of the rays propagates to the screen plane (several positions for the screen are illustrated in Figure 16 with thick black lines). Depending on the distance between the RG and screen plane, each ray will contribute to a different part of the screen, and consequently, to different part of reconstructed light field.

For the need of frequency analysis, each ray is considered as a sample, positioned in the 2D $(x, \varphi)$ ray-space plane for fixed *z* (in the case of full parallax, this turns into a 4D plane). This is visualized for several distances in Figure 17. One can observe that the sampling pattern changes with distance and that for every distance, the sampling pattern is regular although not rectangular. The fact that the sampling patterns are regular, enables us to utilize multi-dimensional sampling theory [126][127].

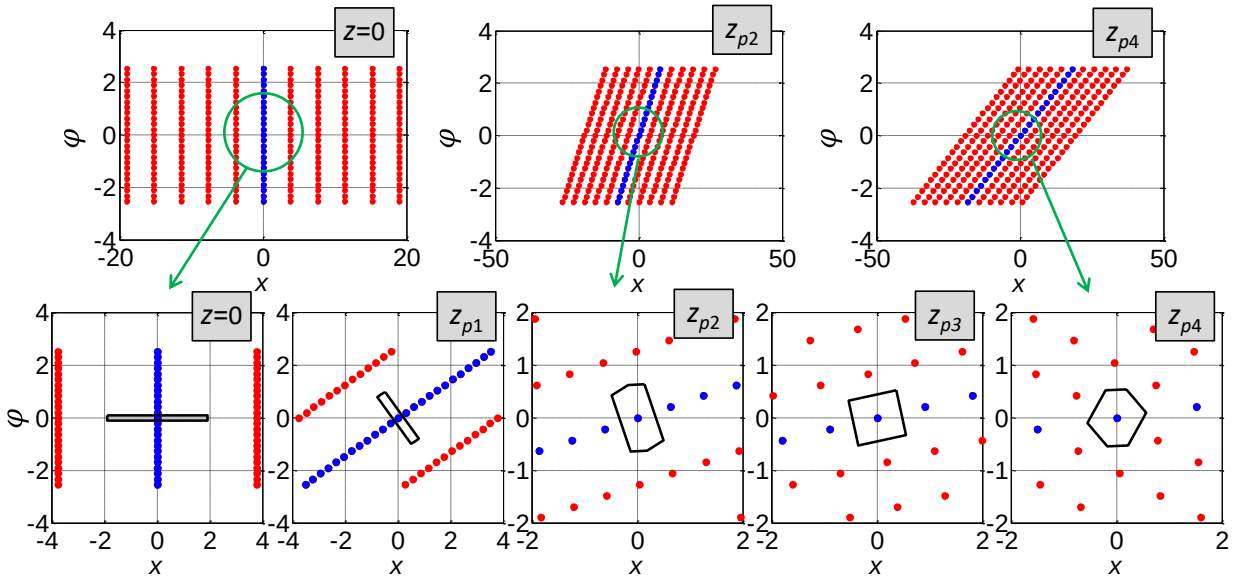



Figure 17. Light field display – ray space spatial sampling patterns at different distances from the RG plane

Samples of any regular 2D pattern can be described through the notion of a sampling lattice $\Lambda$,

$$\Lambda(\boldsymbol{V}) = \{n_1\boldsymbol{v}_1 + n_2\boldsymbol{v}_2 \mid n_1, n_2 \in \mathbb{Z}\}$$

with $\boldsymbol{v}_k = \left[v_k^{(x)} \;\; v_k^{(\varphi)}\right]^{\mathrm{T}}$ for $k = 1{,}2$ referred to as basis vectors and $^{\mathrm{T}}$ being the transpose operator. The vectors building the lattice can be expressed in matrix form as

$$\boldsymbol{V}(\boldsymbol{v}_1, \boldsymbol{v}_2) = [\boldsymbol{v}_1 \;\; \boldsymbol{v}_2] = \begin{bmatrix} v_1^{(x)} & v_2^{(x)} \\ v_1^{(\varphi)} & v_2^{(\varphi)} \end{bmatrix}.$$

Furthermore, for a regular grid described with a lattice $\Lambda$, one can also define a unit cell *P* that is a set in $\mathbb{R}^2$ such that the union of all cells centered on each lattice point covers the whole sampling space without overlapping or leaving empty space. The shape of the unit cell depends on the sampling pattern – see Figure 17 for examples of unit cells.

Having the sampling matrix describing a sampling pattern, the corresponding frequency domain representation can be evaluated as [19][126]

$$\Lambda^*(\boldsymbol{V}) = \Lambda((\boldsymbol{V}^T)^{-1}).$$

The passband of the display corresponds to any unit cell of the given lattice $\Lambda^*$. Each possible unit cell describes a set of bandlimited functions that can be represented by the sampling pattern and can be reconstructed from a given discrete representation assuming that the reconstruction filter has the shape of the selected unit cell.

The most compact (isotropic) unit cell for a given sampling pattern is a Voronoi cell. The Voronoi cell, denoted by *P**, is a set in $\mathbb{R}^2$ such that all elements of the set are closer, based on Euclidean distance, to the one lattice point that is inside the cell than to any other lattice point. The importance of this unit cell is twofold. First, it represents frequency support that treats equally both directions (spatial and angular direction in ray space representation) – this is beneficial from the HVS viewpoint. Second, the screen in the display that performs the D/C conversion has for practical reasons a 'low-pass' type characteristics (typically it is rectangular with Gaussian type weights [6]) that has to be matched to available ray distribution or vice versa. As such, the Voronoi cell is the most convenient unit cell to match the screen reconstruction filter. The display bandwidth, in terms of spatial and angular resolution, is directly given by the support of the Voronoi cell.

To illustrate the discussion in this section, following the notations as given in Figure 16, we can estimate the bandwidth of a hypothetical display with $(x_p, \alpha_p, z_p) = (30\text{mm}, 0.0365°, 1570\text{mm})$. Here, $x_p$ is the distance between adjacent ray sources on the ray generator plane, $\alpha_p$ is the angular resolution at the ray generator plane, and $z_p$ is the distance between the ray generator and screen plane. The parameters of the display have been selected so that they correspond to a realistic setup, e.g. , the selected angular resolution $\alpha_p = 70°/1920$ that would correspond to a ray source having 1920 px in horizontal direction over a 70 degree FoV. The sampling pattern in spatial-angular domain on the ray generator plane of such setup is given in Figure 18, left. After performing the analysis outlined earlier in this section, the resulting spatial-angular sampling at the screen plane is shown in Figure 18, right. The corresponding support in the frequency domain, marked as blue square in Figure 19, represents the estimated display bandwidth.

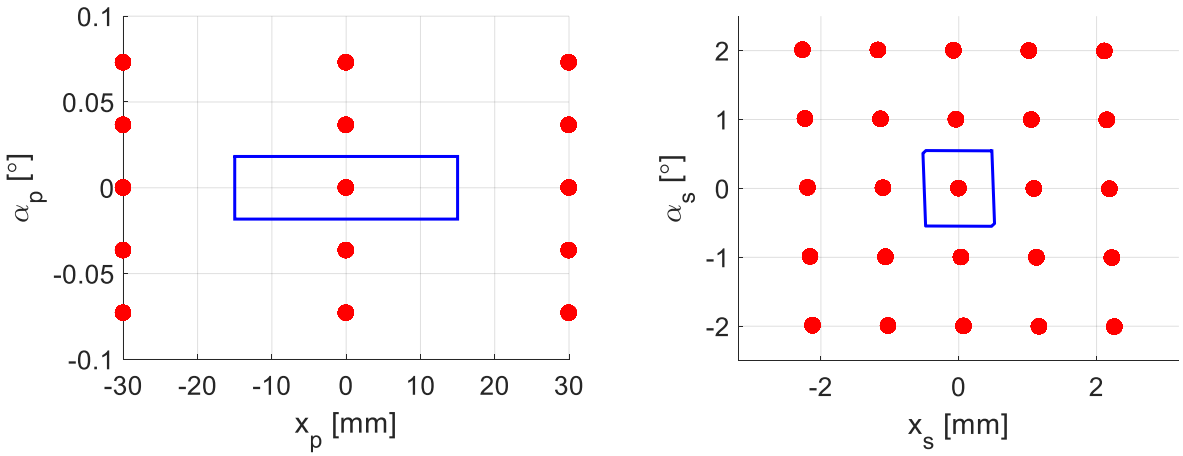



Figure 18. Left: sampling pattern in spatial-angular domain at the ray generator. Right: sampling pattern in spatial-angular domain at the screen plane

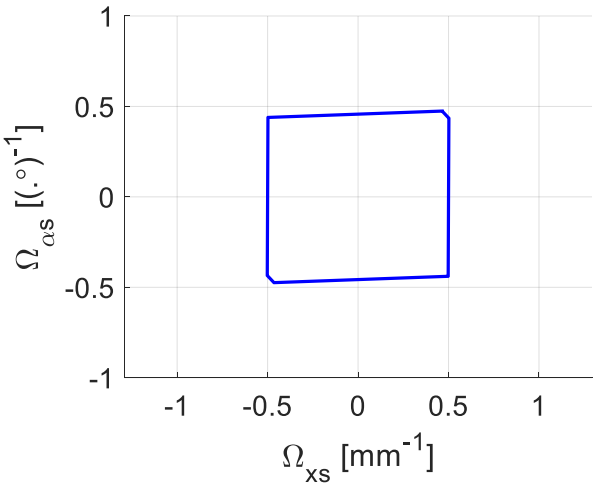



Figure 19. Estimated display bandwidth in the frequency domain at the screen plane

The same approach for estimating the display bandwidth can also be used to determine the optimal parameters of the display for a desired performance. Furthermore, one can also establish an optimal relation between cameras (light field sampling) and display (light field reconstruction), that is, create optimal camera setup for a given display. Both of those topics are discussed in more detail in [IV].

When preparing content for the display, an arbitrary continuous function has to be pre-filtered with a filter aimed at removing all frequency content outside of the selected unit cell (display passband) in order to prevent aliasing errors during sampling. This can be achieved either by using (if possible) a proper continuous-domain filter before sampling the function or first oversampling the continuous function and then performing filtering and downsampling in the discrete domain.

The discrete ray positions and directions, as determined by the geometrical setup are converted to a continuous light-field using a diffuser. Diffusers may have different passband shape. A wider diffuser results in smoother transition between two discrete optical modules, but worse spatial selectivity. For the user the projection setup and the characteristics of the diffuser are typically not known. Therefore it might be beneficial that the capabilities of the display in terms of passband are measured in order to compare them, and to create content for the displays. This can be done by the methods discussed in the following sections.

## 3.2 Objective Measurements

Light field displays attempt to recreate a reference light field representing a scene using a finite number of discrete light sources, each having a finite spatial resolution. All these light rays, which

can be controlled independently spread to the space in front of the display, typically in an uneven distribution across the field of view.

In order to quantify the capabilities of an unknown light field display, we need to measure its bandwidth under different circumstances. On one hand, we can measure the equivalent spatial resolution of the display at the screen plane – this is similar to the spatial resolution as it is known in case of 2D displays. On the other hand, we can characterize the angular resolution, which describes the smallest angle the color of which the display can control individually – this is something 2D displays do not have, as they emit the same color to all directions from a single pixel.

The display metrics commonly used to characterize 2D displays also apply (as 3D displays can typically be used to show 2D content, though in some cases with low fidelity).

### 3.2.1 Previous 3D Display Measurement Methods

Most previous work on 3D displays was focusing on the measurement or characterization of two-view or multi-view autostereoscopic displays. Ref. [131] provides an approach to model multi-view screens in the frequency domain and measure angular visibility using test patterns of increasing frequency. However this work assumes a typical multiview display using a subpixel interleaving topology, which light field displays do not have.

The approach presented in [128] is also targeted for multi-view screens. This method is based on proprietary measurement equipment with Fourier optics, which is costly, and due to the small size of the measurement head, the applicability for large-scale (non-desktop) light field displays is limited. Moreover, it cannot be used for front-projected light field displays as the head would block the light path.

The Information Display Measurement Standard contains measurement methods both for spatial resolution and angular resolution (chapters 17.5.4 and 17.5.1 in [92]). The method described as angular resolution measurement relies on counting local maxima, but also assumes that the display can show two-view test patterns specifically targeting adjacent views, which is not directly applicable for light field displays. Also, it assumes that the number of local maxima can be reliably counted, which is not the case when exceeding the resolution limits of a display. The light field autostereoscopic image resolution measurement is using sinusoidal test patterns, and reports the resolution loss associated with showing objects at different depth. It assumes however that the pixel size of the display is known in advance, it does not provide a method for measuring the pixel size, and is not applicable for a light field display with no discrete pixel structure.

Ref. [129] describes another proprietary measurement instrument, however the 3D display measurement assumes that the display is view-based, which does not apply for typical light field displays. Ref. [130] refers to the equipment in [128] for performing measurements on autostereoscopic 3D screens.

### 3.2.2 Proposed Method

#### 3.2.2.1 Spatial Resolution Measurement

The method proposed by the author and presented in [V] can be used to measure the spatial and angular resolution of light field displays in an automated way, using a commodity camera and a computer running the image processing algorithms. The method has been successfully applied on several projection-based 3D light field displays, screen size ranging from 9” to 140”, with up to 180° field of view.

The spatial resolution measurement method inspects the display in frequency domain, identifying the limits of the display showing sinusoidal test patterns on the light field display’s screen plane, capturing and analyzing the resulting image for distortions.

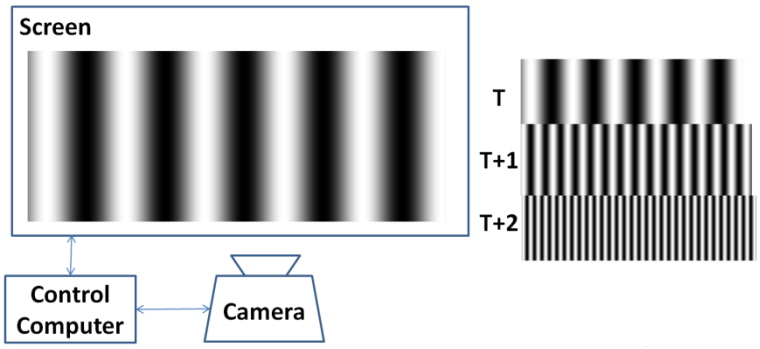



Figure 20. Spatial resolution measurement overview. Left: A sinusoidal test pattern is rendered on the display under test, while a camera attached to the control computer takes a photo. Right: Subsequent measurement iterations show sinusoidals with increasing frequency.

The patterns are generated with special rendering software that allows the operator to determine the color of each light ray based on where it hits the screen plane, regardless of the direction of the light ray. It is assumed that the display under test has a programming interface that allows rendering simple patterns based on the position on the screen and the direction of the generated light ray. Based on the hit position, grayscale sinusoidals are shown on the full screen (see Figure 20), starting from low frequency, in small increments, to very high frequency that well exceeds the resolution of the imaging components by a factor of two. Two sets of test patterns are captured, one with horizontal and one with vertical orientation.

The resulting images are captured with a high-resolution monochrome camera that is positioned to be in line with center of the screen, from a distance that it can capture the whole screen, using manual shooting settings to ensure that the intensity range of the captured pattern is properly represented on the captured images. The resolution of the capture camera is chosen to oversample the theoretical maximum resolution of the display at least two times. The shutter speed is chosen so that the shutter is slower than the time multiplexing frequency of the projection components (as

projectors commonly employ time multiplexing for reproducing color channels). The focus of the camera is on the screen plane.

A single row of samples from the measured pattern is selected from the screen's center from each photo (see Figure 21, left). This row undergoes a fast Fourier transform, resulting in the frequency spectrum showing the amplitude of different frequency components in the image (see Figure 21, right). The frequency spectrum very well shows both the excitation pattern and the distortions introduced by the display.

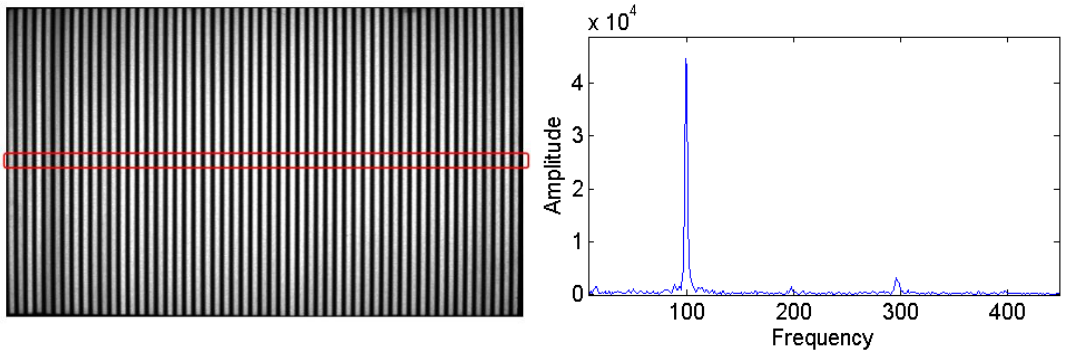



Figure 21. Left: A photo of the screen showing a sinusoidal test pattern. The center row of the photo is used for frequency analysis. Right: Frequency spectrum of a single measurement showing the sinusoidal, with FFT bins on the horizontal axis.

The dominant frequency in each captured image is that of the sinusoidal excitation signal (provided we are within the resolution of the display). As the frequency of the sinusoidal increases, the peak shifts to higher frequencies (see Figure 22). At the same time, the amplitude of the excitation signal decreases as the frequency increases, while some distortions and aliasing also become apparent.

The algorithm measures the amplitude of the sinusoidal excitation signal, and the amplitude of the strongest noise peak. Once the amplitude of noise reaches 20% of the amplitude of the excitation signal, the display is considered to reach its resolution limit in the given direction (see Figure 23). 20% noise threshold is considered as disruptive distortion, the selection of this threshold is based on [131].

The number of peaks of the sinusoidal shown on the screen at this point times two is considered the effective resolution limit of the screen in the given direction (one positive and one negative peak of the sinusoidal representing two pixels). The same measurement is then repeated in the orthogonal direction to obtain the resolution in that direction.

Inspecting the spectrums may reveal some major sources of distortion. The aliasing that occurs after exceeding the resolution limit shows as the mirror image of the excitation signal. The straight lines above the diagonal are the harmonics of the excitation signal, and are caused by the nonlinear intensity profile of the display under test, causing the sinusoidals to appear slightly rectangular.

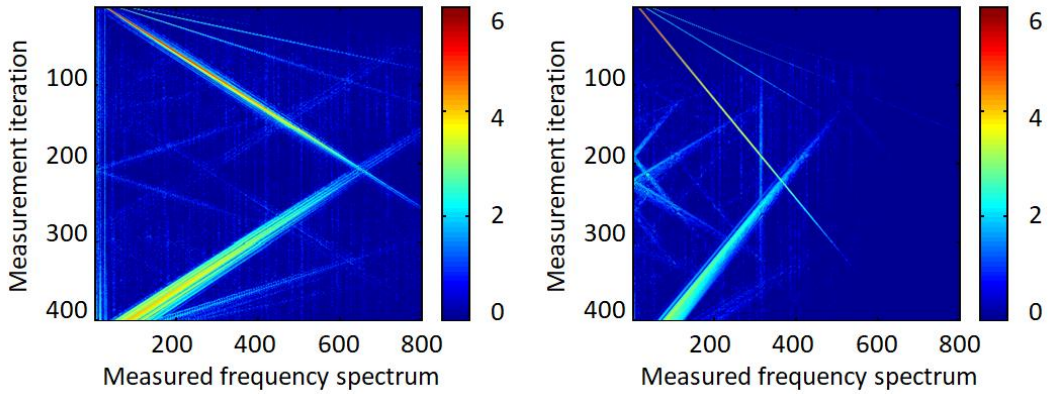



Figure 22. Frequency spectrums of successive measurements stacked in a matrix. Measurement iteration count increases downwards, while the observed frequency increases rightwards. Left: Spectrums of horizontal resolution measurements from a sample display. Major sources of distortion are visible as harmonics, aliasing and constant low-frequency distortion. Right: Spectrums of vertical resolution measurements from the same display.

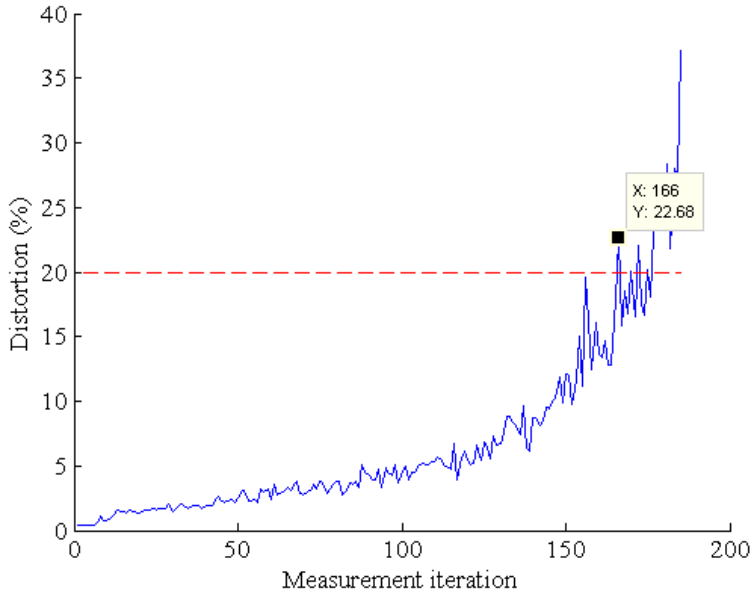



Figure 23. Level of distortion in subsequent measurement iterations. 20% noise threshold is marked with red dashed line.

Constant low frequency components are caused by the slightly non-uniform brightness profile - this is in line with the frequency of the projection modules. These observations made on the spectrum may be useful for improving the performance of light field displays on the long run.

#### 3.2.2.2 Angular Resolution Measurement

The angular resolution measurement method uses a similar approach, but in the angular direction. In order to measure angular resolution, the display has to emit different intensities to different directions, which are then measured from various viewing angles. Using a special rendering algo-

rithm that allows defining the color of each light ray based on its position and direction crossing the screen plane, a small patch is generated on the screen that appears white from some viewing directions, and black from other directions (see Figure 24).

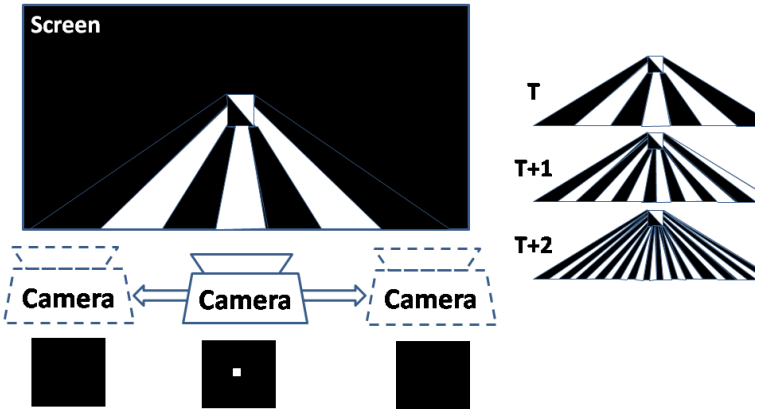



Figure 24. Angular resolution measurement overview. Left: Test setup with moving camera. The rectangle looks black from some locations and white from other locations. Right: Test patterns of increasing angular frequency.

In the initial iterations, changing from white to black and back to white happens slowly as the viewer moves sideways, while in subsequent iterations the frequency is increased. The pattern is captured with a camera moving sideways on a motorized rig from the left extreme to the right extreme of the Field of View. From the captured videos, the center of the patch is selected, on which one can see the intensity of the patch changing following a periodic pattern (see Figure 25, left).

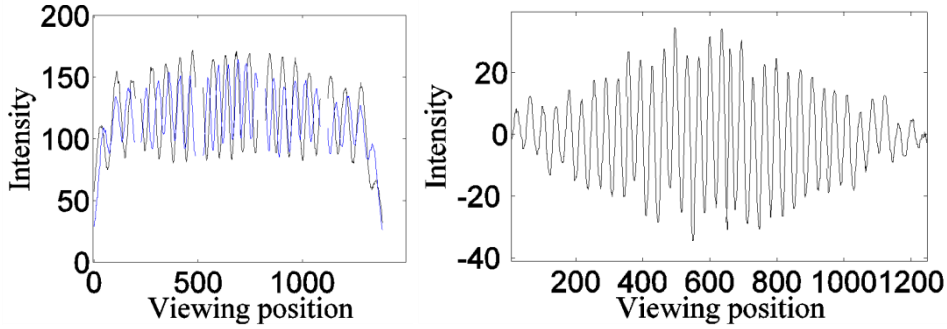



Figure 25. Left: Sample intensity profiles for two different angular frequencies recorded on the same display. Right: Intensity profiles of forced black-white transitions on a light field display.

Running the resulting function through frequency analysis, and stacking measurements made with different excitation signals, the dominant frequency is visibly increasing, representing our excitation frequency in the directional domain (see Figure 26, left), but, similar to the spatial resolution measurement, different sources of distortion are emerging, while the amplitude of the main signal is decreasing.

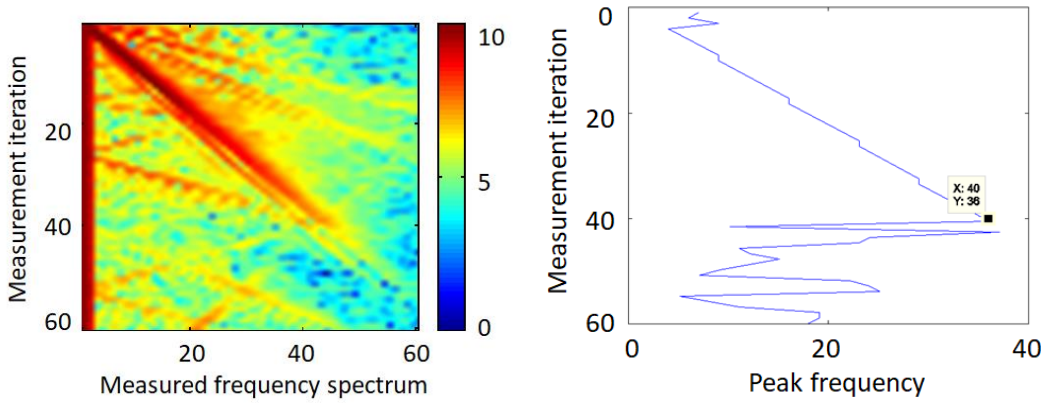



Figure 26. Left: 1D frequency spectrums of angular resolution test patters stacked in a 2D array. Right: Frequency of the peak in subsequent measurement iterations.

The limit of angular resolution is determined as the maximum frequency until the primary peak keeps increasing (see Figure 26 , right) – after this limit, the peak frequency is dominated by noise, as the display is no longer capable of further increasing the number of different rays emitted to different directions from the same screen position. The maximum number of transitions the display was capable of showing is counted, resulting in the number of directions. The total angle of the field of view is divided by the number of directions, resulting in the angular resolution.

The maximum number of directions can also be determined using geometrical considerations if the internals of the display are known. As a light field display can emit light rays from discrete light sources (projection modules), light cannot originate from between two such units. Assuming a display with no internal mirrors to extend the FoV, the maximum number of directions emitted from a single patch of the screen can be determined based on the distribution and emission angle of the projection modules used in the display. To demonstrate this, the display can be forced such that even projection modules emit white, odd engines emit black images. On the captured intensity profile (see Figure 25, right), the peaks can be counted, which matches the result of the angular resolution measurement, as it was verified on multiple light field displays.

## 3.3 Subjective Measurements

The spatial resolution of a 3D display as perceived by users can also be estimated using patters that can determine what viewers can and cannot see. While eye test charts are meant to determine the limits of the visual acuity of a person, turning the thinking around they can also be used as a tool to determine whether the display medium can reproduce these patterns in a way humans can distinguish one pattern from another (assuming the human eye could easily tell them apart with sufficiently detailed visualization in place).

The ’tumbling E’ pattern has been chosen for this subjective test, as it is well known by people performing the test, and because of its rectangular shape that can be represented on 5x5 pixels. That is, the symbol size is 5 times the feature size. Our assumption is that one visual feature corresponds to one pixel of the effective resolution at the limit of visibility.

While in common visual acuity tests 50% recognition threshold is used as the rule-of-thumb limit of visibility, the actual threshold of recognizing visual features might be different. Studies of the human visual system show our capability to recognize symbols whose features are not entirely visible, based on the symbol’s luminance distribution, even when the symbol is heavily distorted [132]. Therefore we determined the limit by testing the correspondence between the objective and subjective methods on multiple displays, including one that had asymmetric pixel aspect ratio (2:1).

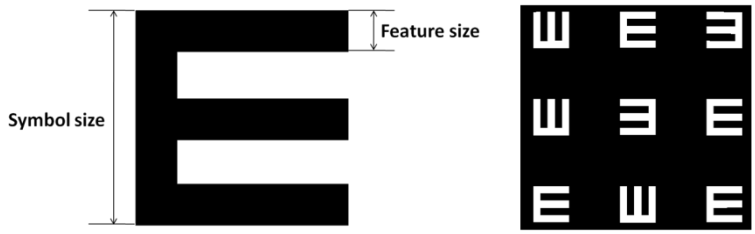



Figure 27. Subjective spatial resolution test overview. Left: One tumbling “E” symbol. Feature size is 1/5 of the total symbol size. Right: A chart of 9 randomized E symbols arranged in a 3x3 matrix.

The experiment is performed as follows: the viewer is presented with sets of tumbling E symbols arranged in a 3x3 matrix (see Figure 27), with random orientation out of four possible orientations (up, down, left, right). The size of all 9 symbols on the screen at a given point in time is the same. The subject is asked to record the orientation of all symbols visible on the screen. After finishing with a set of symbols, a new set of symbols is presented with a different symbol size. Each size appears twice during the experiment, and they appear in random order. The time needed to recognize each set by the subject is also recorded by the operator. It is expected that recognition accuracy of the symbols is around 100% when the symbols appear undistorted (that is, feature size is bigger than or equal to the smallest visual feature the display can show), and drops significantly once the feature size is smaller – that is when the resolution limit of the display is exceeded.

To make sure that the drop in recognition accuracy is caused by the display and not the viewer’s visual acuity, a paper-based reference symbol set is presented before the experiment, which is positioned at the same distance as the display screen. The orientation of the reference set is also recorded to filter those participants who do not have sufficient visual acuity for the experiment.

During our experiments we used 50+ participants of different ages, sex and nationality. Our experiments shown that the recognition rate that corresponds to the pixel size determined by the objective method is approximately 92% to 93%. Confidence intervals of the results, as well as the time to completion grow significantly after reaching this limit (see Figure 28). This means we can use the recognition accuracy to estimate the limit of the visibility of visual features to see at which sym-

bol size the subjects have difficulties to recognize the symbols, but the recoded completion times also give a strong indication.

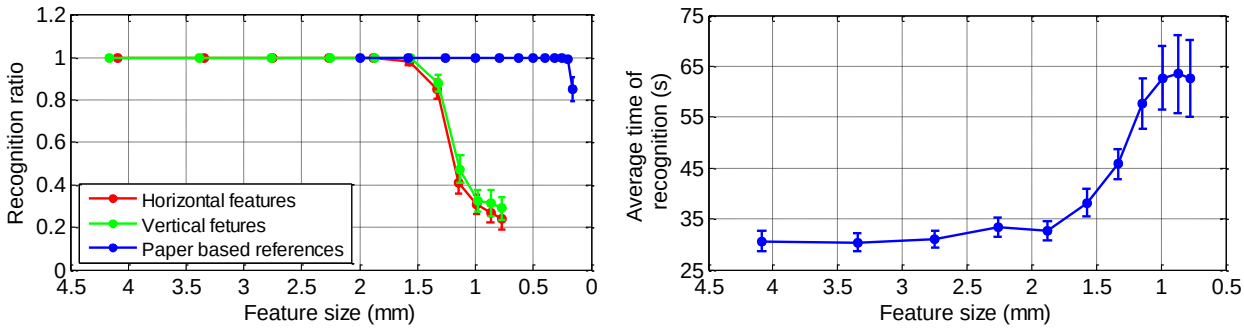



Figure 28. Left: Recognition accuracy of tumbling E symbols on paper and display, for horizontal and vertical features, plotted against feature size, with 95% confidence intervals. Right: Average recognition time for a group of symbols with given feature size.

## 3.4 Discussion

The methods described above, or the earlier versions of them have been used to characterize the capabilities of several light field displays as presented in [II][III][V]. Initially spatial resolution measurement was performed on multiple slanted / diagonal angles [II], which was later found to be redundant – measuring the horizontal and vertical spatial resolutions provides sufficient information about the capabilities of the display, diagonal measurements did not provide significant additional information. The first angular resolution measurement method relied on the observation that the contrast of the angular resolution pattern is diminishing at higher frequency test patterns, and defined the limit as 20% of the contrast ratio of the display when showing 2D patterns. This paper also attempted to characterize perceived depth resolution using a subjective test. The next paper [III] introduced the subjective spatial resolution test using the tumbling E patterns. This paper also made an attempt to characterize the perceived spatial resolution from different viewing angles, as well as with and without motion parallax, and concluded that head parallax slightly improves the perceived spatial resolution. The concluding journal paper [V] introduced the angular resolution measurement based on frequency analysis of the captured intensity, as well as introduced the paper-based reference pattern in subjective spatial resolution tests. This time the subjective tests were performed on a much larger number of subjects, and the time required to record patterns was also recorded. This enabled the identification of the relation between recognition performance and the time spent with recognizing the patterns.

These methods could also be applied for other 3D display technologies (even volumetric ones without a well-defined screen). Also, angular resolution measurement can be performed to profile

horizontal and vertical angular resolution separately when profiling displays that have both horizontal and vertical parallax.

Further details about the measurement methods, subjective tests, and sample measurements from light field displays are described in [V].

# 4 3D Video Representations and Display-Specific Light Field Processing

## 4.1 Choice of Representation

From the formats described in Section 2, the format of choice for the work described is the pure image-based representation from many dense viewpoints. This is the only practical and feasible approach at the time of writing to capture and represent the diversity of real-world scenes with high visual quality. The cost of high visual fidelity is massive information content – transmitting 20 to 180 video streams simultaneously without compression is only possible using very high speed interconnect technologies. To enable future use cases like light field 3DTV or 3D video conferencing using light fields (telepresence), efficient coding techniques are required to reduce the bandwidth between the capture and display side.

## 4.2 3D Video Compression Methods and Their Use for LF Compression

To enable dense light field content to be transmitted efficiently, compression of such imagery was one of the objectives for the FTV group in MPEG [133]. These activities were also supported by test data by Nagoya University [22], Holografika [134] and others. It has been shown using subjective experiments [125] that dense light field data of approx. 720p resolution can be compressed with good quality using MV-HEVC, with a bitrate requirement that is similar to what is required for 4K/8K 2D video streams, which is realistic to achieve in the near future. These experiments however used offline encoding / decoding and view synthesis due to the amount of image data required, and the runtime performance of the reference software implementation used.

The first real-time implementation of light field streaming and rendering was presented in 2010 [VI], which was possible by using a high-performance GPU implementation, and by limiting the number and resolution of the images captured by the cameras. The system described in that paper con-

sisted of 27 USB webcams, which captured VGA resolution images. JPEG encoding was performed by the cameras in hardware, and the JPEG image stream was transmitted to all rendering nodes via Gigabit Ethernet to the display. Light field rendering required the camera system to be precisely calibrated: the calibration data contained the intrinsic and extrinsic camera parameters of each camera, as well as the color differences between the captured images. Camera calibration data was used during the light field rendering process to select the correct light rays during light field interpolation. The light field renderer decoded all 27 images on all rendering nodes, and used light ray interpolation to render the light field. As identified during that work, centralized encoding, decoding and processing of the images becomes prohibitive when using many high resolution cameras in real-time. However, the distributed rendering system typically available in light field displays allows for a distributed rendering approach that requires only partial data to be available in each rendering node. An encoding method that can exploit this distributed processing capability and allows partial decoding was clearly desirable [IX]. The solution is to decouple decoding from rendering and exploiting the typical data access patterns during light field rendering.

A system based on image+depth representation was presented at SIGGRAPH [VII]. This was an attempt to evaluate the image quality of rendering light fields based on image+depth, which did not require decoding many views as the pure image-based approach. That system used 4 cameras, two in the center forming a stereo pair, and two satellite cameras to capture the scene with relatively wide field of view. The system used depth estimation between the four cameras to produce consistent depth maps in real time. The light field rendering system used the four images and depth maps to render a light field. The resulting image quality was good, but suffered from typical artifacts caused by depth estimation. The conclusion was to continue with pure image-based representation, if the objective is to render visually compelling real scenes on light field displays.

Centralized encoding and decoding of the high number of views typically used for light field displays using H.264/MVC, MV-HEVC or 3D-HEVC is not feasible using today's technology. Though encoders and decoders are expected to eventually be available, the resolution of displays is also expected to rapidly increase. As such, both on the short and long term, a distributed approach is desirable and practical. This is especially reasonable considering that in a distributed rendering system, the full image data is not needed, as each rendering process requires different portions of the image data depending on the portion of the light field it is responsible for.

During the light field conversion process, we transform a display-independent light field (that is, many views) to a display-specific light field (that is, the images the projection modules need to project to reproduce the scene). This process involves assembling the images needed by the projection modules from the pixels originating from many camera images (see Figure 29).

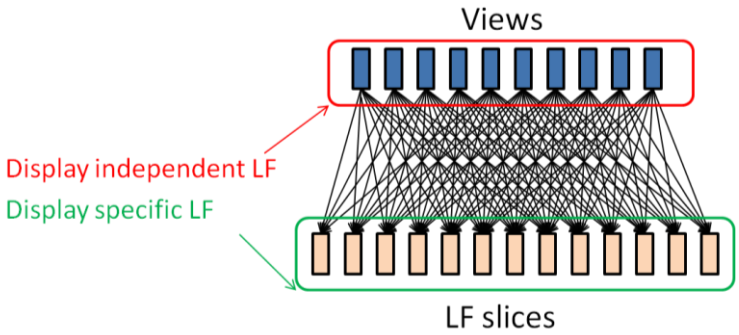



Figure 29. A set of perspective views depicting the same scene are considered as a display independent light field representation, as it can be used to visualize the light field on any suitable 3D display. The images required by a specific display's projection modules together constitute the display specific light field.

Checking the data flow of the light field conversion process, and focusing on the data used by a single rendering node, one can see that the pixels are read from a compact image area from each source image, and pixels outside that area are left unused (see Figure 30). This can be exploited using two different architectures, as described in the next sections.

Figure 30. Left: Adjacent rendering nodes consume adjacent, slightly overlapping parts of a source view. Red, green and blue overlays represent the areas of the image used by three rendering nodes that drive adjacent projection modules. Right: Rendering nodes that drive projection modules positioned further away from each other use a disjoint set of pixels from the same source view.

As an additional benefit, analyzing the light field rendering process revealed that the simple linear camera setups commonly used for rendering super-multiview images for light field rendering is not optimal in terms of how the pixels are used. Therefore an attempt to optimize the camera setup to maximize the number of pixels used during the light field interpolation process has been developed, as described in [VIII].

### 4.2.1 Two-Layer Architecture for Light Field Decoding and Rendering

In this setup, decoding the individually compressed views is distributed to an arbitrary number of processing nodes in the first layer to allow real-time decoding. Once the decoded pixels are available in memory, a high-speed network is used to transmit only those image areas to the nodes in the second layer that are necessary in each node. In this architecture (see Figure 31) an ordinary

2D video codec can be used for the views, but it requires an interconnect between the two layers that has sufficient bandwidth to transmit uncompressed image data.

What image areas are needed for a specific renderer depends on the relation between the capture and the display setup, as well as the Region Of Interest used during light-field rendering. This means that the image areas may change during the sequence, if, for example, we zoom inside the light field, or if there is a change on the capture setup (for example, cameras start to converge).

Considering the typical vertical bar-shape of the image regions in source images, and the typical row-column organization of images in the computer's memory, direct transmission of the bars would require the transmission of many non-continuous memory areas. To enable efficient network transfers, the images can be rotated 90 degrees, and the vertical bars extended to the edges of the image, thus forming a continuous memory region that can be transferred by an efficient interconnect in one batch (e.g. InfiniBand RDMA [135]).

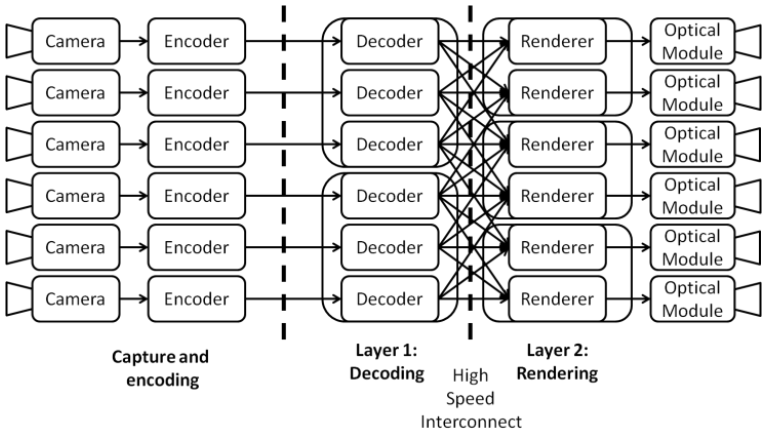



Figure 31. The two-layer decoder-renderer architecture. The first layer decodes video streams in parallel, while the second layer requests portions of uncompressed video data on demand.

### 4.2.2 One-Layer Architecture for Light Field Decoding and Rendering

In this setup, decoding and rendering happens in the same nodes, and decoding time is saved by decoding only those image areas that are actually needed for the specific renderer (see Figure 32).

Support for decoding only parts of the image (referred to as partial decoding), is not common in image and video codecs, as typical use cases do not require accessing only portions of the image, fast. There are some features in the codecs though, that can be used for this purpose.

The H.264 video codec defines slices as regions of the image that can be independently parsed from the bitstream. How the image is partitioned into slices can be defined arbitrarily. While this serves parallel decoding of slice data very well, motion estimation and motion compensation do not respect slice boundaries in any of the available encoders.

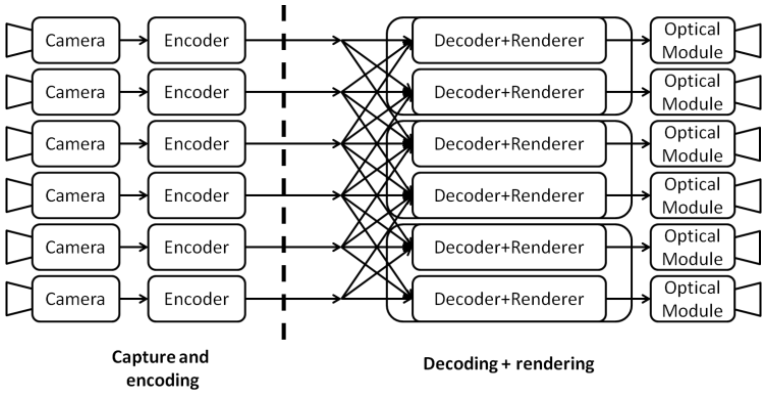



Figure 32. The one layer decoder-renderer architecture. Decoders and renderers are running on the same nodes. Decoders decode those parts of the video that the renderer on the same node will need for rendering. No data exchange except frame synchronization takes place between the nodes.

The result is that the image portions left undecoded may bring in undefined pixels into the image area to be decoded during the motion compensation process (see Figure 33).

Figure 33. When motion vectors point out from the undecoded region (middle slice), they propagate bogus colors from the undecoded region into the slices which we intend to decode

Therefore the H.264 reference encoder implementation has been updated during this research work to respect slice boundaries, i.e. not to perform motion estimation across the boundaries of the slice being encoded. These slices are called self-contained slices, as they can be decoded without decoding other slices in the stream. The effect of restricting motion vectors to go across slice boundaries is very well visible on the motion vectors (see Figure 34).

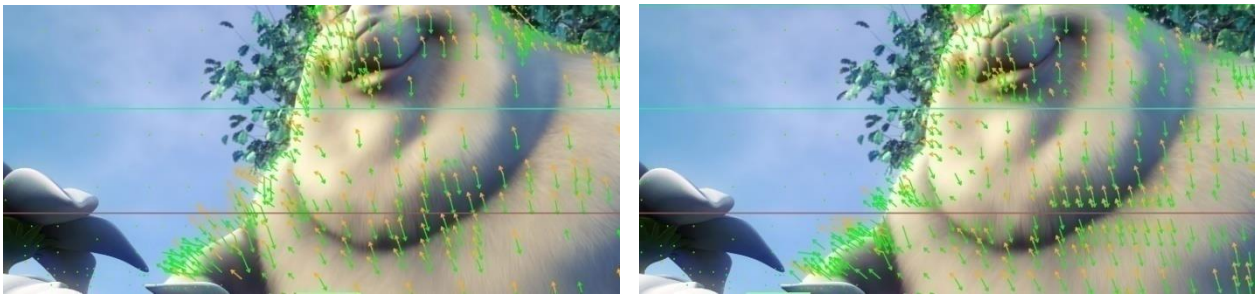


Figure 34. Difference of motion vectors in normal encoding and with self-contained slices. Notice that in the normal case (left) motion vectors cross slice / tile boundaries. In the self-contained case (right) no motion vectors cross the slice / tile boundaries.

Using this restriction on the motion vectors, it is possible to decode only those slices that contain the image areas needed for a specific renderer. Sublinear speedup was observed with the ffmpeg H.264 decoder when decoding partial images [136].

HEVC introduced tiles, which can be used to partition an image into multiple rectangular partitions. In this sense, tiles are similar to slices, however tiles are independent in terms of entropy coding. Similar to H.264 though, inter-frame prediction has to be restricted to remain inside tile boundaries, to enable partial decoding [137].

Though JPEG is a still image codec, it is still a good candidate for video compression due to its simplicity, and availability of high-speed encoders and decoders, despite its relatively low video compression performance (compared to H.264 and HEVC at least). JPEG has an optional feature called Restart intervals, which are portions of the bitstream that can be independently decoded. Restart intervals contain an arbitrary number of 8x8 pixel MCUs (Minimum Coding Unit), and are delimited by Restart Markers. Restart intervals are typically unused, but our use case can take advantage of them to skip decoding image portions that are not needed. While light field regions used by a single renderer typically have the shape of a vertical bar, and MCUs are aligned horizontally, we can still use them if the image is encoded with 90 degree rotation. During this research work, libjpeg-turbo, which is the fastest available CPU-based JPEG decoder has been updated with the capability of skipping Restart intervals. The resulting speedup gained by skipping unnecessary regions is significant.

JPEG2000 is another still image compression standard, and is also used for encoding video frames in digital cinemas. Partial decoding is available in JPEG2000. During this research work, the Comprimato GPU-accelerated JPEG2000 codec [138] has been used, which exposes this functionality. JPEG2000 also showed significant, but sublinear speedup while decoding parts of the image (see Figure 35).

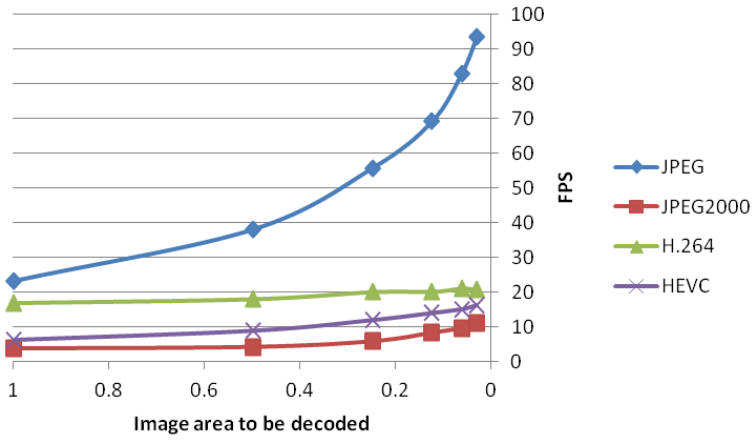



Figure 35. Comparison of overall speed and speedup of different decoders when decoding partial views. In case of JPEG, restart markers are used. In case of H.264 and HEVC, our custom self-contained slices and tiles are used. In case of JPEG2000 no special features are used.

### 4.2.3 Discussion

When comparing all four codecs, unsurprisingly JPEG is the most performant in terms of runtime, but consumes the most bandwidth, too. Introducing restart intervals has almost no effect on the bitrate. JPEG also shows the best relative speedup when decoding only parts of the image. JPEG2000 and HEVC also show reasonable speedup when using partial decoding, and do so with significantly better bitrate (see Figure 36).

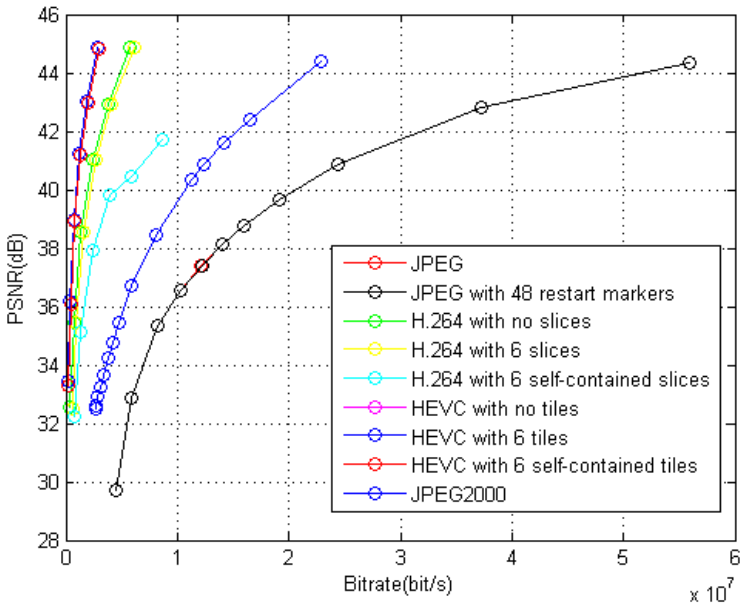



Figure 36. Comparison of overall quality versus bitrate of the different codecs using default settings and configurations that enable partial decoding. Please note reported bitrates are for a single view. JPEG and JPEG with 48 restart markers overlap. The three HEVC curves also overlap.

A very simple alternative solution that works for any image / video codec is to subdivide the images into vertical regions, and encode them separately. This approach has been used for encoding panoramic video with good results [139]. However this approach results in a fixed subdivision, requires even more bandwidth than the solutions described above, and requires the synchronized streaming of a lot of video streams in parallel.

A centralized encoding / decoding solution like H.264/MVC is not suitable for the massive number of views required for proper light field reconstruction, partially because the amount of information cannot be handled by a single processing unit, partially because high pixel-count light field displays can only be served by a distributed rendering system, as discussed in detail in [IX]. The solutions proposed can enable real-time light field transmission using hardware components and codecs available today. The codec can be chosen according to the ratio between processing power and bandwidth in the system. Further details of the performance of the customized codecs can be found in [X].

# 5 Conclusions

The topics discussed in this thesis originate from the desire to enable light field displays to show live 3D content, captured with cameras, reproducing the diversity of real scenes, including realistic, live people; surfaces with real specular reflections, anisotropic effects, transparency, atmospheric effects, subsurface scattering and other phenomena that occur in the real world. The long term goal of this research was to contribute to the success of light field displays to eventually become the future 3DTVs, and to enable high-end use cases like real 3D videoconferencing.

To reach this long term goal, many difficult challenges are to be solved in an end-to-end system. This thesis contributed to two of these challenges: display profiling and light field coding.

## 5.1 Key Findings and Contributions

The first main contribution of this thesis is providing methods to profile the performance of light field displays, using metrics that are generic enough to be used on all sorts of 3D displays, regardless of the underlying technology. The primary motivation to quantify the reproduction capabilities of light field displays was to be able to design optimal capture setups for rendering or live capturing, making sure that the content captured fully exploits the capabilities of the display. The resulting techniques, however, can be used to quantify 3D displays, which enables a fair comparison of different 3D displays using metrics that express what viewers can actually perceive. Display profiling can also be used for end-of-line checking and to verify the precision of display calibration. The methods presented can measure the spatial and angular resolution of light field displays by showing, capturing and analyzing test patterns in the frequency domain.

The spatial resolution measurement method is also supported by a subjective test designed to identify the limits of visibility on the screen.

The second main contribution of this thesis is enabling the coding of the captured light field content in a way that it can be processed and rendered in real-time on hardware available today. Using a pure image-based format was chosen to maximize the visual fidelity of the resulting light field. Decoding and processing multi-megapixel light fields was possible by analyzing how light field rendering algorithms work on a distributed rendering system, and exploiting the characteristics of the data flow by avoiding decoding and processing data that is unnecessary for generating the given slice of the light field. After identifying this possibility, an encoding format was necessary that supports such an unusual use case. Due to the runtime constraints and the concept of centralized encoding / decoding in 3D-HEVC and MV-HEVC, these codecs cannot be used on today's hardware for real-time use cases (though quite efficient when given enough runtime). To facilitate real-time use cases on state-of-the-art hardware, H.264, HEVC, JPEG and JPEG2000 were analyzed in detail, and the codec implementations modified where necessary to enable partial decoding. Experiments were performed to benchmark the behavior of all codecs in terms of bandwidth, runtime, and reduction of runtime when decoding partial data. Based on these findings, live transmission and rendering of high resolution light fields became possible. The results, as well as several additional findings leading to these results have been contributed to the MPEG community, to facilitate the design of future codecs.

In summary, the key findings and contributions of this thesis are as follows:

- The bandwidth of a light field display can be analytically estimated in the frequency domain, considering the display as a bandlimited light field reconstruction device.
- Given a light field display described with its bandwidth, one can create the light field with the same bandwidth based on camera images. In this way, the optimal relation between cameras (light field sampling) and display (light field reconstruction) is uniquely established.
- First measurement methods for light field displays using a single camera: spatial and angular resolution measurement.
- Methods for quantifying the perceived spatial resolution validated by subjective experiments.
- Methods for optimizing the camera setup for capturing light fields optimal for a light field display.
- Analysis of the dataflow of light field rendering on a distributed rendering system, recognition of the locality of data access.

- Analysis of 3D video codecs for suitability in real-time use cases.
- Analysis of commonly used image and video codecs to support real-time use cases involving light field data.
- Proposal for usage and modifications, implementation of modifications in codecs to support real-time use cases involving light field data.

## 5.2 Author's Contributions to the Publications

The book chapter '3D Visual Experience' [I] provided an overview of 3D display technologies, their advantages and disadvantages, and the associated image formats. It included all major 3D display types known at the time of publication: stereoscopic, head mounted, autostereoscopic based on parallax barriers and lenticulars, volumetric displays and light field displays. It supported the understanding and comparison of different 3D display technologies available, as well as the image data required for visualizing content on them. P. Kovacs overviewed and summarized the information about different displays and representations, and completed the scientific writing.

The conference paper 'Quality Measurements of 3D Light-Field Displays' [II] introduced the spatial resolution measurement based on displaying sinusoidal patterns, and used frequency analysis to determine the limits of the display. It also described a method for angular resolution measurement using a spot that changes its apparent color based on the viewing direction, and used relative dynamic range to identify the limit of angular resolution. This approach was replaced with frequency analysis in later works. It also described a small subjective experiment attempting to measure depth resolution. P. Kovacs designed the measurement method, implemented the rendering and analysis software, performed the measurement of several displays, organized the subjective tests, analyzed the results, and completed the scientific writing of the paper except parts of Section 3.1 and 4.1.

The conference paper 'Measurement of Perceived Spatial Resolution in 3D Light field Displays' [III] introduced the subjective experiment for perceived spatial resolution based on the 'tumbling E' pattern. It also checked the perceived resolution from different viewing angles, and the relation between perceived resolution and head parallax. P. Kovacs implemented the rendering and measurement software, organized the subjective tests, analyzed the results, and completed the scientific writing of the paper.

The journal article 'Optimization of light field display-camera configuration based on display properties in spectral domain' [IV] introduced a multidimensional sampling

model for describing light field displays, and proposed a methodology for determining the optimal distribution of light field generators and capture cameras. P. Kovacs contributed to the spatial and frequency domain analysis of light field displays from the point of view of geometrical optics and principles of operations of such displays. In addition, he contributed to the optimization of the camera-display configuration.

The journal article ’Quantifying spatial and angular resolution of light field 3D displays’ [V] summarized those approaches for spatial and angular resolution measurement of light field displays that were found to be the most robust and had the most consistent results. It developed a method based on frequency analysis and signal / noise threshold to determine the limits of a light field display in terms of reconstructing visible image elements (spatial resolution), and a subjective test that uses a pattern similar to ”tumbling E” eye test charts to evaluate the smallest feature humans can distinguish on the screen. It also proposed a method based on frequency analysis to measure the angular resolution of a display using special patterns and a moving camera. P. Kovacs performed all the implementation and measurements related to the spatial and angular resolution measurement, organized the subjective test sessions, and completed the scientific writing of the paper (except Section II. C).

The conference paper ’Real-time 3D light field transmission’ [VI] described the first system capable of capturing, transmitting and visualizing a real scene using a multi-camera system and a light field display in real time. It used 27 USB cameras, a 24-channel HoloVizio display, and highly optimized rendering software to show ~15 frames per second, with approximately 1 second of latency. P. Kovacs designed the architecture of the system, implemented the software to perform synchronized multi-camera capture and transmission, managed the implementation of the other software components, and constructed the capture system. P. Kovacs wrote Sections 3, 4, and 5 of the paper. The importance of this paper is twofold: it not just described the first real-time light field capture and rendering system, but the work raised many interesting challenges that arise when scaling up such a system in terms of number of cameras, resolution, or if used over channels with narrower bandwidth – like in case of a telepresence / 3D videoconferencing system. These research problems were further discussed and addressed during the research work at TUT as described in the publications listed below.

At SIGGRAPH, the exhibition paper at Emerging Technologies titled ‘3D capturing using multi-camera rigs, real-time depth estimation and depth-based content creation for multi-view and light field auto-stereoscopic displays’ [VII] demonstrated the possibility of rendering light field content based on a 4-camera rig (a stereo pair, and two satellite cameras with a wider baseline), using real-time depth estimation and light field synthe-

sis [140]. This experiment in the direction of depth based rendering also very well demonstrated that though depth-based light field rendering from a relatively narrow camera setup is possible, the dense light field capturing approach and rendering without depth estimation presented in the previous paper produces more plausible results, hence research continued in that direction. P. Kovacs designed the architecture of the rendering system and the interface with the external capture system, supported the design of the capture setup, contributed to the development of the light field rendering algorithm and completed the scientific writing of the paper.

The conference paper ’Analysis and Optimization of Pixel Usage of Light field Conversion from Multi-Camera Setups to 3D Light field Displays’ [VIII] inspected the light field conversion process in detail, concluding that many captured pixels are typically left unused, and proposed an optimization method for camera setups that result in better pixel utilization. P. Kovacs performed the analysis of pixel usage of different displays and camera setups, implemented the optimization technique for camera setups, and completed the scientific writing of the paper.

The conference paper ’Overview of the applicability of H.264/MVC for real-time light field applications’ [IX] discussed the issues of using H.264/MVC for encoding, decoding and processing the video streams captured by a multi-camera system and rendered on a light field display. It identified why this encoding method is unfeasible to use on today’s hardware, and that the typical pixel access patterns of the light field image rendering process could be exploited, if the encoding method provided the necessary tools for it. P. Kovacs performed the analysis of the distributed light field rendering algorithm to characterize the typical data access patterns on multiple displays, performed the analysis of H.264/MVC, enumerated and analyzed the available implementations, and completed the scientific writing of the paper.

The journal article ’Architectures and Codecs for Real-Time Light Field Streaming’ [X] then attempted to answer the questions raised in the previous paper, and building on the possible opportunities identified, it introduced two possible processing architectures, and analyzed how JPEG, JPEG2000, H.264 and HEVC can support partial decoding, the primary requirement identified to support efficient, distributed light field decoding and processing. The paper presented results about the impact in terms of bitrate and runtime, and proposed features for next generation video codecs to better support real-time light field rendering. P. Kovacs designed the two architectures, proposed using slicing to enable partial decoding and proposed the modifications to the H.264 and HEVC codecs, implemented the proposed changes to the JPEG codec, performed the measurement and comparison of all four codecs, benchmarked the bandwidths and

processing load typical in a distributed light field rendering system, and completed the scientific writing of the paper.

It is important to mention some MPEG input documents, as a major part of the work on light field video compression techniques directly concerned the work of the MPEG FTV group [141], where P. Kovacs was an active contributor during his PhD.

The MPEG input document ’Requirements of Light field 3D Video Coding’ [142] introduced light field displays to the MPEG community, and described application scenarios where video encoding / decoding is necessary. It referred to draft reports produced by the MPEG FTV group that identify the use cases and requirements to be addressed by the FTV group and amended those by the requirements coming from light field displays and the identified use cases.

The MPEG input document ’Proposal for additional features and future research to support light field video compression’ [143] explained display-specific and display-independent light field processing, and furthermore it introduced the notion of non-linear camera arrays to the MPEG community, which was not considered by the FTV group in previous work.

As the MPEG community was lacking suitable test material to perform light field encoding experiments (except the Nagoya sequences [134]), the author made an effort to enable such experiments by providing test material to the MPEG FTV group in the input document ’Big Buck Bunny light field test sequences’ [134]. The test video sequences were based on the Big Buck Bunny short CGI movie, which is available for free including all source material. It contained three short (3 to 5 seconds) clips from the movie, all rendered from 91 cameras, using both linear and arc camera setups, in both 24-bit YUV and HDR color format, as well as depth maps with 8-bit and floating point format, to enable many different kinds of experiments. This test material contained many novel elements compared to previously available MPEG test material, namely: first test material with nonlinear camera setup, first light field sequence with ground truth depth data, first light field sequence with High Dynamic Range, and first light field sequence that is available both with linear and arc camera setup. P. Kovacs designed and implemented the rendering infrastructure and cluster to enable rendering of the sequences, requested permission from the Blender Foundation to use the material, oversaw the rendering process, provided the material online, and prepared the MPEG input document.

## 5.3 Future Research Directions

While the results presented provide answers to some of the interesting research questions regarding the efficient use of light field displays, there is a large amount of work ahead before such displays will gain widespread use and become the display of choice in every household. Putting display technology, cost and manufacturing challenges aside, the two research angles presented here alone point to further research directions.

On the resolution measurement front, measuring the effective spatial resolution at depth planes different from the screen plane (both inside and outside) would tell a lot about the performance of any 3D display in terms of depth reproduction, and how the depth range can be exploited efficiently during content creation. In case of displays that may not have a uniform distribution of light rays on the screen, it makes sense to quantify the resolution for different areas on the screen / different directions in the field of view. In fact, such a method could be used during light field display calibration to pinpoint any areas where calibration could be improved, and to provide overall quantification of different calibration methods (which at the end have an impact on the effective resolution). While the methods presented here are automatic, reduction of the measurement time using reduced number of test patterns is certainly possible. Another major item of further research may include the use of the presented methods to quantify 3D displays based on other technologies, for example compressive light field displays or volumetric displays.

On the light field video compression and streaming front, the work presented is just the beginning. While we have presented some solutions to implement efficient random access to encoded light field content, this currently requires modification of the involved codecs. Incorporating the necessary changes in the encoding / decoding process of future video codecs means years of work for the standardization groups involved, while the efficient real-time implementation (software or dedicated hardware) of the novel video compression standards takes another several years before it can be used for real use-cases and embedded in 3D display products.

Investigating how Scalable Video Coding (SVC) can support light field displays with different Fields of View is another interesting topic, especially when combined with partial decoding.

# References


[1] P. T. Kovács, N. Murray, G. Rozinaj, Y. Sulema, R. Rybárová, "Application of immersive technologies for education: State of the art," in *Proc. International Conference on Interactive Mobile Communication Technologies and Learning (IMCL)*, Thessaloniki, pp. 283-288, 2015, DOI: 10.1109/IMCTL.2015.7359604

[2] R. Barr, ”Transfer of learning between 2D and 3D sources during infancy: Informing theory and practice,” *Developmental Review: DR*, 30(2), 128–154, 2010, DOI: 10.1016/j.dr.2010.03.001

[3] M. Agus, F. Bettio, A. Giachetti, E. Gobbetti, J. A. Iglesias Guitián, F. Marton, J. Nilsson, G. Pintore, ”An interactive 3D medical visualization system based on a light field display,” *Vis Comput* (2009) 25: 883, 2009, DOI:10.1007/s00371-009-0311-y

[4] F. Steinicke, T. Ropinski, G. Bruder, K. Hinrichs, ”Towards Applicable 3D User Interfaces for Everyday Working Environments,” in *Human-Computer Interaction – INTERACT 2007* by C. Baranauskas, P. Palanque, J. Abascal, S. D. J. Barbosa (eds), Lecture Notes in Computer Science, vol 4662. Springer, Berlin, Heidelberg, DOI:10.1007/978-3-540-74796-3_55

[5] P. Lincoln, A. Nashel, A. Ilie, H. Towles, G. Welch, H. Fuchs, ”Multi-view lenticular display for group teleconferencing,” in *Proc. of the 2nd International Conference on Immersive Telecommunications (IMMERSCOM '09)*, 2009

[6] T. Balogh, “The HoloVizio system,” in *Proc. SPIE 6055, Stereoscopic Displays and Virtual Reality Systems XIII*, 60550U, 2006, DOI:10.1117/12.650907

[7] M. S. Banks, J. R. Read, R. S. Allison and S. J. Watt, "Stereoscopy and the Human Visual System," *SMPTE 2nd Annual International Conference on Stereoscopic 3D for Media and Entertainment*, New York, NY, USA, 2011, pp. 2-31, DOI:10.5594/M001418

[8] M. Schuck, G. Sharp, ”3D digital cinema technologies,” *SID Tech. Digest*, 43,629, 2012

[9] C. van Berkel, D. W. Parker, A. R. Franklin, "Multiview 3D-LCD," in *Stereoscopic Displays and Virtual Reality Systems III, Proc. SPIE 2653*, pp. 32-39, 1996

[10] G. E. Favalora, J. Napoli, D. M. Hall, R. K. Dorval, M. Giovinco, H. J. Richmond, W. S. Chun, "100 Million-voxel volumetric display," in *Cockpit Displays IX: Displays for Defense Applications, Proc. SPIE 4712*, pp. 300-312, 2002

[11] S. A. Benton, "The Second Generation of the MIT Holographic Video System," in *Proc. TAO First International Symposium on Three Dimensional Image Communication Technologies*, Tokyo, Japan, 1993

[12] M. Lucente, "Interactive three-dimensional holographic displays: seeing the future in depth," in *ACM SIGGRAPH Computer Graphics*, 31(2), pp. 63-67, 1997

[13] F. Yaras, H. Kang, L. Onural, "Real-time color holographic video display system," in *Proc. 3DTV-Conference: The True Vision Capture, Transmission and Display of 3D Video*, Potsdam, Germany, 2009

[14] M. S. Banks, D. M. Hoffman, J. Kim, G. Wetzstein, ”3D Displays,” in *Annual Review of Vision Science* 2016 2:1, pp. 397-435, 2016, DOI:10.1146/annurev-vision-082114-035800

[15] Holografika 3D light field displays. Online: http://holografika.com/ Visited:2018-09-15

[16] A. Maimone, H. Fuchs, “Encumbrance-free telepresence system with real-time 3d capture and display using commodity depth cameras,” in *10th IEEE ISMAR*, October 2011, pp. 137–146

[17] B. Petit, J-D. Lesage, C. Menier, J. Allard, J-S. Franco, B. Raffin, E. Boyer, F. Faure, “Multicamera real-time 3d modeling for telepresence and remote collaboration,” *International Journal of Digital Multimedia Broadcasting*, vol. 2010, pp. 247108–12, 2009

[18] B. S. Wilburn, M. Smulski, H-H. Kelin Lee, M. A. Horowitz, "Light field video camera," in *Proc. SPIE vol. 4674, Media Processors 2002*, DOI:10.1117/12.451074, 2002

[19] R. Bregović, P.T. Kovács, T. Balogh, and A. Gotchev, “Display-specific light-field analysis,” in *Proc. SPIE 9117*, 15 pages, 2014

[20] T. Fujii, K. Mori, K. Takeda, K. Mase, M. Tanimoto, Y. Suenaga, "Multipoint Measuring System for Video and Sound - 100-camera and microphone system," *2006 IEEE International Conference on Multimedia and Expo, Toronto*, Ont., 2006, pp. 437-440, DOI:10.1109/ICME.2006.262566

[21] E. Ekmekcioglu, V. De Silva, G. Nur, A. M. Kondoz, "Immersive 3D Media," in *Media Networks Architectures, Applications, and Standards* by Hassnaa Moustafa, Sherali Zeadally (eds.), Taylor&Francis, DOI: 10.1201/b12049-20, 2012

[22] MPEG-FTV test sequences from Tanimoto Lab at Nagoya University. Online: http://www.fujii.nuee.nagoya-u.ac.jp/multiview-data/mpeg/mpeg_ftv.html Visited:2018-09-09

[23] A. Vetro, T. Wiegand, G. J. Sullivan, "Overview of the Stereo and Multiview Video Coding Extensions of the H.264/MPEG-4 AVC Standard," in *Proc. of the IEEE*, vol. 99, no. 4, pp. 626-642, DOI:10.1109/JPROC.2010.2098830, 2011

[24] G. Tech, Y. Chen, K. Müller, J. Ohm, A. Vetro, Y. Wang, "Overview of the Multiview and 3D Extensions of High Efficiency Video Coding," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 35-49, DOI:10.1109/TCSVT.2015.2477935, 2016

[25] E. H. Adelson, J. R. Bergen, ”The plenoptic function and the elements of early vision,” in *Computational models of visual processing* by M. S. Landy, J. A. Movshon (Eds.), pp. 3-20, Cambridge, MA, US: The MIT Press, 1991

[26] C.K. Liang, Y.C. Shih, and H.H. Chen, “Light field analysis for modeling image formation,” *IEEE Trans. Image Processing* 20(2), pp. 446-460, 2011

[27] P. Sturm, "Pinhole Camera Model," in *Computer Vision* by Ikeuchi K. (ed), Springer, Boston, MA, DOI:10.1007/978-0-387-31439-6_472, 2014

[28] P. Green, L. MacDonald, *Colour Engineering: Achieving Device Independent Colour*, Wiley, ISBN:978-0-471-48688-6, 2002

[29] B. E. Bayer, "Color imaging array," US patent US3971065A, 1975

[30] A. Hornberg, *Handbook of Machine and Computer Vision: The Guide for Developers and Users, 2nd Edition*, Wiley, ISBN:978-3-527-41339-3, 2017

[31] Z. Zhang, “A Flexible New Technique for Camera Calibration,” Technical Report MSR-TR-98-71, Microsoft Research, 1998

[32] R. Tsai, "A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses," *IEEE J. Robotics and Automation*, vol. 3, no. 4, pp. 323-344, 1987

[33] D. C. Brown, "Decentering distortion of lenses," in *Photogrammetric Engineering*, 32 (3), pp. 444–462, 1966

[34] D. Scaramuzza, A. Martinelli, R. Siegwart, "A Toolbox for Easily Calibrating Omnidirectional Cameras," *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Beijing, pp. 5695-5701, DOI:10.1109/IROS.2006.282372, 2006

[35] N. Joshi, B. Wilburn, V. Vaish, M. Levoy, M. Howoritz, "Automatic Color Calibration for Large Camera Arrays," 2005

[36] P. Debevec, T. Hawkins, C. Tchou, H-P. Duiker, W. Sarokin, M. Sagar, "Acquiring the reflectance field of a human face," in *Proc. 27th annual conference on Computer graphics and interactive techniques (SIGGRAPH '00)*. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, pp.145-156, DOI:10.1145/344779.344855, 2000

[37] P. Debevec, "The Light Stages and Their Applications to Photoreal Digital Actors," in *Proc. SIGGRAPH Asia*, 2012

[38] V. Vaish, M. Levoy, R. Szeliski, C. L. Zitnick, and S. B. Kang, "Reconstructing Occluded Surfaces Using Synthetic Apertures: Stereo, Focus and Robust Measures," in *Proc. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR '06)*, Vol. 2., Washington, DC, USA, 2331-2338, DOI:10.1109/CVPR.2006.244, 2006

[39] M. Levoy, P. Hanrahan, "Light field rendering," in *Proc. 23rd annual conference on Computer graphics and interactive techniques (SIGGRAPH '96)*. ACM, New York, NY, USA, 31-42. DOI:10.1145/237170.237199, 1996

[40] R. Hartley, A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, ISBN:978-0521540513, 2004

[41] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, P. Hanrahan, "Light Field Photography with a Hand-held Plenoptic Camera," Stanford Tech Report CTSR 2005-02, 2005

[42] D. R. Proffitt, C. Caudek, “Depth perception and the perception of events,” in *Handbook of Psychology*, Wiley, New York, DOI: 10.1002/0471264385.wei0408, 2002

[43] A. Stern, Y. Yitzhaky, B. Javidi, “Perceivable Light Fields: Matching the Requirements Between the Human Visual System and Autostereoscopic 3-D Displays,” *Proc. IEEE*, vol. 102, no. 10, pp. 1571-1587, DOI:10.1109/JPROC.2014.2348938, 2014

[44] L. Goldmann, T. Ebrahimi, “Towards reliable and reproducible 3-D video quality assessment,” in *Proc. SPIE Int. Soc. Opt. Eng.*, vol. 8043, DOI:10.1117/12.887037, 2011

[45] A. Boev, R. Bregović, A. Gotchev, “Signal processing for stereoscopic and multi-view 3D displays,” in *Handbook of Signal Processing Systems, 2nd edition*, S. Bhattacharyya, E. Deprettere, R. Leupers, and J. Takala, eds., Springer, 2013

[46] H. Hirschmuller, "Stereo Processing by Semiglobal Matching and Mutual Information," in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328-341, DOI:10.1109/TPAMI.2007.1166, 2008

[47] L. Jimenez-Ortega, N. F. Troje, ”Differential motion parallax as a monocular depth cue?,” *Journal of Vision*, 3(9): 855, 855a, DOI:10.1167/3.9.855, 2003

[48] W. Vanduffel, D. Fize, H. Peuskens, K. Denys, S. Sunaert, J. T. Todd, G. A. Orban, “Extracting 3D from motion: differences in human and monkey intraparietal cortex,” in *Science*, 298(5592), pp. 413-415, 2002

[49] G. Yi, L. Jianxin, Q. Hangping, W. Bo, "Survey of structure from motion," *Proc. International Conference on Cloud Computing and Internet of Things*, Changchun, pp. 72-76, DOI:10.1109/CCIOT.2014.7062508, 2014

[50] D. M. Hoffman, A. R. Girshick, K. Akeley, and M. S. Banks, “Vergence–accommodation conflicts hinder visual performance and cause visual fatigue,” *Journal of Vision*, vol. 8, no. 33, 2008

[51] M. Mehrabi, E. M. Peek, B. C. Wuensche, C. Lutteroth, “Making 3D work: a classification of visual depth cues, 3D display technologies and their applications,” in *Proc. Australasian User Interface Conference (AUIC '13)*, vol. 139, pp. 91-100, Australian Computer Society, Darlinghurst, Australia, 2013

[52] B. Perroud, R. Regnier, A. Kemeny, F. Merienne, “Model of realism score for immersive VR systems,” in *Transportation Research Part F: Traffic Psychology and Behaviour*, in press, DOI:10.1016/j.trf.2017.08.015, 2017

[53] C. Wheatstone, F. R. S. XVIII, ”Contributions to the physiology of vision. —Part the first. On some remarkable, and hitherto unobserved, phenomena of binocular vision,” in *Phil. Trans. R. Soc. Lond*. 1838 128 371-394, DOI:10.1098/rstl.1838.0019 2053-9223

[54] Niagara Falls (1915).
Online: http://www.silentera.com/PSFL//data/N/NiagaraFalls1915-1.html
Visited 2018-09-09

[55] The Power of Love (1922). Online: https://www.imdb.com/title/tt0013506/
Visited 2018-09-09

[56] L. Hammond, ”Stereoscopic Motion Picture,” US patent 1,435,520
L. Hammond, ”Stereoscopic-picture-viewing apparatus,” US patent 1,658,439

[57] F. E. Ives, ”Parallax stereogram and process of making same,” US patent 725567A

[58] G. Lippmann, "La Photographie Integrale," *Comptes-Rendus Academie des Sciences* 146, pp. 446-451, 1908

[59] C. van Berkel, D. W. Parker, A. R. Franklin, ”Multiview 3D-LCD,” in *Proc. Stereoscopic Displays and Virtual Reality Systems III, Proc. SPIE 2653*, pp. 32-39, 1996

[60] D. Gabor, ”A new microscopic principle,” *Nature*, 1948 May 15;161(4098):777

[61] E. Leith, J. Upatnieks, ”Holographic Imagery Through Diffusing Media,” *Journal of the Optical Society of America*, vol. 56, Issue 4, pp. 523-523, 1966, DOI:10.1364/JOSA.56.000523

[62] P. St-Hilaire, S. A. Benton, M. E. Lucente, M-L Jepsen, J. Kollin, H. Yoshikawa, J. S. Underkoffler, ”Electronic display system for computational holography," in *Proc. SPIE 1212, Practical Holography IV*, 1990, DOI:10.1117/12.17980

[63] H. Sato, T. Kakue, Y. Ichihashi, Y. Endo, K. Wakunami, R. Oi, K. Yamamoto, H. Nakayama, T. Shimobaba, T. Ito, ”Real-time colour hologram generation based on ray-sampling plane with multi-GPU acceleration,” *Scientific Reports*, 8., 2018, DOI:10.1038/s41598-018-19361-7

[64] V. Bianco, P. Memmolo, M. Leo, S. Montresor, C. Distante, M. Paturzo, P. Picart, B. Javidi, P. Ferraro, ”Strategies for reducing speckle noise in digital holography,” *Light: Science & Applications*, 7, Article number: 48, 2018

[65] H. Urey, K. V. Chellappan, E. Erden and P. Surman, "State of the Art in Stereoscopic and Autostereoscopic Displays," in *Proc. of the IEEE*, vol. 99, no. 4, pp. 540-555, April 2011, DOI:10.1109/JPROC.2010.2098351

[66] D. Minoli, *3D Television (3DTV) Technology, Systems, and Deployment: Rolling Out the Infrastructure for Next-Generation Entertainment*, CRC Press, ISBN:9781439840665, 2010

[67] J. L. Fergason, S. D. Robinson, C. W. McLaughlin, B. Brown, A. Abileah, T. E. Baker, P. J. Green, "An innovative beamsplitter-based stereoscopic/3D display design," in *Stereoscopic Displays and Virtual Reality Systems XII, Proc. SPIE 5664*, 488-494, 2005

[68] S. D. Robinson, A. Abileah, P. J. Green, "The StereoMirror™: A High Performance Stereoscopic 3D Display Design," in *Proc. SID Americas Display Engineering and Applications Conference (ADEAC '05)*, Portland, Oregon, USA, 2005

[69] H. L. Morton, "Stereoscopic-television apparatus for individual use," United States Patent 2955156, 1960

[70] I. E. Sutherland, "A head-mounted three dimensional display," in *Proc. of the December 9-11, 1968, fall joint computer conference, part I (AFIPS '68 (Fall, part I))*. ACM, New York, NY, USA, 757-764. DOI: https://doi.org/10.1145/1476589.1476686, 1968

[71] T. P. Caudell, D. W. Mizell, "Augmented reality: an application of heads-up display technology to manual manufacturing processes," *Proc. of the Twenty-Fifth Hawaii International Conference on System Sciences*, Kauai, HI, USA, pp. 659-669 vol.2., DOI:10.1109/HICSS.1992.183317, 1992

[72] D. Cheng, Y. Wang, H. Hua, M. M. Talha, "Design of an optical see-through head-mounted display with a low f-number and large field of view using a freeform prism," *Appl. Opt.*, 48, 2655-2668, 2009

[73] P. Milgram, F. Kishino, "A taxonomy of mixed reality visual displays," *in IEICE Trans. on Information and Systems* 77.12, pp. 1321-1329, 1994

[74] J. L. Olson, D. M. Krum, E. A. Suma and M. Bolas, "A design for a smartphone-based head mounted display," *2011 IEEE Virtual Reality Conference*, Singapore, pp. 233-234, DOI:10.1109/VR.2011.5759484, 2011

[75] A. Jones, M. Lang, G. Fyffe, X. Yu, J. Busch, I. McDowall, M. Bolas, P. Debevec, "Achieving Eye Contact in a One-to-Many 3D Video Teleconferencing System," in *ACM Transactions on Graphics, Proc. SIGGRAPH*, 28(3), 64, 2009

[76] H. Gotoda, ”A multilayer liquid crystal display for autostereoscopic 3D viewing,” in *Proc. SPIE 7524, Stereoscopic Displays and Applications XXI*; 75240P, DOI:10.1117/12.840286, 2010

[77] M. Sayinta, S. O. Isikman and H. Urey, "Scanning Led Array Based Volumetric Display," *2008 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, Istanbul, 2008, pp. 21-24, DOI:10.1109/3DTV.2008.4547798

[78] D. Wyatt, ”A Volumetric 3D LED display,” Project final report, MIT, 2005

[79] D. Ezra, G. J. Woodgate, B. A. Omar, N. S. Holliman, J. Harrold, L. S. Shapiro, "New autostereoscopic display system," in *Stereoscopic Displays and Virtual Reality Systems II, Proc. SPIE 2409*, 31-40 (1995)

[80] D. Sandin, T. Margolis, J. Ge, J. Girado, T. Peterka, T. DeFanti, "The Varrier autostereoscopic virtual reality display," in *ACM Transactions on Graphics, Proc. ACM SIGGRAPH*, 24(3), pp. 894-903, 2005

[81] H. Isono, M. Yasuda, H. Sasazawa, "Autostereoscopic 3‐D display using LCD‐generated parallax barrier," in *Electronics and Communications in Japan* 76, 7, pp. 77-84, 1993

[82] N. Inoue, M. Kawakita, K. Yamamoto, "200-Inch glasses-free 3D display and electronic holography being developed at NICT," in *Lasers and Electro-Optics Pacific Rim (CLEO-PR)*, Kyoto, pp.1-2, 2013, DOI:10.1109/CLEOPR.2013.6600199

[83] K. Nagano, A. Jones, J. Liu, J. Busch, X. Yu, M. Bolas, P. Debevec, “An autostereoscopic projector array optimized for 3D facial display,” in *Proc. SIGGRAPH 2013 Emerging Technologies*, Anaheim, 2013, DOI:10.1145/2503368.2503371

[84] M. Agus, E. Gobbetti, A. Jaspe, G. Pintore, R. Pintus, ”Automatic Geometric Calibration of Projector-based Light Field Displays,” in *Proc. EuroVis 2013*, DOI:10.2312/PE.EuroVisShort.EuroVisShort2013.001-005

[85] H. Huang, H. Hua, "Modeling of Eye's response in Viewing 3D Light Field Display," in *Imaging and Applied Optics*, OSA Technical Digest, Optical Society of America, paper DTu3F.2., 2017

[86] Voxon VX1. Online: https://voxon.co/voxon-vx1-available-for-purchase/ Visited: 2018-09-09

[87] Could we be watching the 2022 World Cup on a 3D Volumetric Display? – Voxon blog. Online: https://voxon.co/volumetric-display-world-cup-2022/ Visited: 2018-09-09

[88] H. S. El-Ghoroury, Z. Alpaslan, "Quantum Photonic Imager (QPI): A New Display Technology and Its Applications," in *Proc. The International Display Workshops*, 21, 2014

[89] D. Lanman, D. Luebke, "Near-eye light field displays," in *Proc. ACM SIGGRAPH 2013 Emerging Technologies (SIGGRAPH '13)*. ACM, New York, NY, USA, Article 11, 2013, DOI:10.1145/2503368.2503379

[90] G. Wetzstein, D. Lanman, M. Hirsch, R. Raskar, "Tensor displays: compressive light field synthesis using multilayer displays with directional backlighting," in *Proc. ACM Trans. Graph*. 31, 4, Article 80, 2012, DOI:10.1145/2185520.2185576

[91] Y. Han, H. Y. Lin, C. Chen, "Visual fatigue for laser-projection light-field 3D display in contrast with 2D display," *24th International Workshop on Active-Matrix Flatpanel Displays and Devices (AM-FPD)*, Kyoto, 2017, pp. 9-12.

[92] The International Display Measurement Standard v1.03, Society for Information Display, 2012

[93] O. Doronin, A. Barsi, P. A. Kara, M. G. Martini, "Ray tracing for HoloVizio light field displays," *International Conference on 3D Immersion (IC3D)*, Brussels, 2017, pp. 1-8, DOI:10.1109/IC3D.2017.8251894

[94] A. Ouazan, P. T. Kovacs, T. Balogh, A. Barsi, "Rendering multi-view plus depth data on light field displays," *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, Antalya, 2011, pp. 1-4, 2011, DOI:10.1109/3DTV.2011.5877220

[95] Ł. Dąbała, M. Ziegler, P. Didyk, F. Zilly, J. Keinert, K. Myszkowski, H.-P. Seidel, P. Rokita, T. Ritschel, "Efficient Multi-image Correspondences for On-line Light Field

Video Processing," in *Computer Graphics Forum* 35(7), 2016, DOI:10.1111/cgf.13037

[96] A. Smolic, K. Müller, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "Intermediate View Interpolation Based on Multiview Video Plus Depth for Advanced 3D Video Systems," *Proc. ICIP 2008, IEEE International Conference on Image Processing*, San Diego, CA, USA, October 2008.

[97] Multimedia Scalable 3D for Europe. Online: https://cordis.europa.eu/project/rcn/93783_en.html Visited:2018-09-09

[98] S. B. Kang, Y. Li, X. Tong, et al., "Image-based rendering," *Found. Trends. Comput. Graph. Vis.* 2, 3, 2006, pp. 173-258, DOI:10.1561/0600000012

[99] H-Y. Shum, S-C. Chan, S-B. Kang, *Image-Based Rendering*, Springer, ISBN:978-0-387-32668-9, 2007

[100] F. Marton, E. Gobbetti, F. Bettio, J. A. Iglesias Guitían, R. Pintus, "A real-time coarse-to-fine multiview capture system for all-in-focus rendering on a light field display," *2011 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, Antalya, pp. 1-4, 2011, DOI:10.1109/3DTV.2011.5877176

[101] Q. H. Nguyen, N. D. Minh, J. P. Sanjay, "Depth image-based rendering from multiple cameras with 3D propagation algorithm," in *Proc. 2nd International Conference on Immersive Telecommunications (IMMERSCOM '09)*, Brussels, Belgium, 2009

[102] F. Klose, K. Ruhl, C. Lipski, et al., "Stereoscopic 3D view synthesis from unsynchronized multi-view video," in *Proc. European Signal Processing Conference (EUSIPCO)*, Barcelona, Spain, pp. 1904–1909, 2011

[103] C. Zitnick, S.B. Kang, M. Uyttendaele, et al., "High-quality video view interpolation using a layered representation," *ACM Trans. Graph*. 23, 3, pp. 600-608, 2004, DOI:10.1145/1015706.1015766

[104] D. Scharstein, R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," *Int. J. Comput. Vision* 47, 1-3, pp.7-42, 2002, DOI:10.1023/A:1014573219977

[105] M. Hansard, S. Lee, O. Choi, et al., "Time-of-flight cameras: Principles, Methods and Applications," *Springer Briefs in Computer Science*., 2012, DOI:10.1007/978-1-4471-4658-2, ISBN 978-1-4471-4657-5

[106] P. Fechteler, P. Eisert, J. Rurainsky, "Fast and High Resolution 3D Face Scanning," in *Proc. IEEE International Conference on Image Processing*, vol.3, no., pp.III - 81,III - 84, 2007, DOI:10.1109/ICIP.2007.4379251

[107] M. Georgiev, A. Gotchev, M. Hannuksela, "Joint de-noising and fusion of 2D video and depth map sequences sensed by low-powered tof range sensor," in *Proc. IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2013, pp.1,4, 2013, DOI:10.1109/ICMEW.2013.6618264

[108] K. Müller, A. Smolic, K. Dix, et al., “View Synthesis for Advanced 3D Video Systems,” *EURASIP Journal on Image and Video Processing, Special Issue on 3D Image and Video Processing*, vol. 2008, Article ID 438148, 11 pages, 2008, DOI:10.1155/2008/438148

[109] L. Jiangbo, S. Rogmans, G. Lafruit, et al., "Stream-Centric Stereo Matching and View Synthesis: A High-Speed Approach on GPUs," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol.19, no.11, pp.1598,1611, 2009, DOI: 10.1109/TCSVT.2009.2026948

[110] D. Tian, P. Lai, P. Lopez et al., "View synthesis techniques for 3D video," in *Proc. SPIE 7443, Applications of Digital Image Processing XXXII*, 74430T, 2009, DOI:10.1117/12.829372

[111] N. Stefanoski, O. Wang, M. Lang, et al., "Automatic View Synthesis by Image-Domain-Warping," in *IEEE Transactions on Image Processing*, vol.22, no.9, pp.3329,3341, 2013, DOI: 10.1109/TIP.2013.2264817

[112] M. Bertalmio, G. Sapiro, V. Caselles, C. Ballester, "Image inpainting,” in K. Akeley (Ed.), *Proceedings of the 27th annual conference on Computer graphics and interactive techniques SIGGRAPH 00* (Vol. 2, pp. 417-424), ACM Press, 2000

[113] M. Bertalmio, L. Vese, G. Sapiro, S. Osher, "Simultaneous structure and texture image inpainting," *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol.2, no., pp. II- 707-12 vol.2, 18-20 June 2003 DOI:10.1109/CVPR.2003.1211536

[114] S. Jianbing, J. Xiaogang, C. Zhou, C. CL Wang, "Gradient Based Image Completion by Solving the Poisson Equation," *Computers & Graphics*, Elsevier Science Press, 2007, 31(1):119-126

[115] T. Akenine-Moller, E. Haines, N. Hoffman, A. Pesce, M. Iwanicki, S. Hillaire, *Real-Time Rendering, Fourth Edition*, New York: A K Peters/CRC Press, 2018

[116] A. Smolic et al., "Representation, coding, and rendering of 3D video objects with MPEG-4 and H.264/AVC," *IEEE 6th Workshop on Multimedia Signal Processing*, 2004., Siena, Italy, pp. 379-382, DOI:10.1109/MMSP.2004.1436572, 2004

[117] K. Mamou, C. Dehais, F. Chaieb, F. Ghorbel, "Multi-resolution 3D Mesh Coding in MPEG," *2011 Visual Communications and Image Processing (VCIP)*, Tainan, 2011, pp. 1-4, DOI:10.1109/VCIP.2011.6116054

[118] K. Kolev, T. Brox, D. Cremers, "Fast Joint Estimation of Silhouettes and Dense 3D Geometry from Multiple Images," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 34, 2012

[119] A. Mustafa, H. Kim, J-Y. Guillemaut, A. Hilton, "General Dynamic Scene Reconstruction from Multiple View Video," in *Proc. International Conference on Computer Vision (ICCV)*, 2015

[120] ISO/IEC 10918-1:1994 Information technology -- Digital compression and coding of continuous-tone still images: Requirements and guidelines

[121] S. Winkler, M. Kunt, L. C. J. van den Branden, "Vision and Video: Models and Applications," in *Vision Models and Applications to Image and Video Processing* by L. C. J. van den Branden (ed), Springer, Boston, MA, DOI:10.1007/978-1-4757-3411-9_10, 2001

[122] ISO/IEC 15444, Information technology -- JPEG 2000 image coding system: Core coding system

[123] ISO/IEC 14496–10:2012, Information technology -- Coding of audio-visual objects -- Part 10: Advanced Video Coding

[124] ISO/IEC 23008–2:2013, Information technology -- High efficiency coding and media delivery in heterogeneous environments -- Part 2: High efficiency video coding

[125] A. Dricot, J. Jung, M. Cagnazzo, B. Pesquet, F. Dufaux, P. Kovacs, V. Kiran, “Subjective evaluation of Super Multi-View compressed contents on high-end 3D displays,” in *Signal Processing: Image Communication*, vol. 39, Part B, pp. 369-385, 2015

[126] E. Dubois, “The sampling and reconstruction of time-varying imagery with application in video systems,” in *Proc. IEEE* 73, pp. 502–522, 1985

[127] E. Dubois, “Video sampling and interpolation,” in *The Essential Guide to Video Processing* by J. Bovik, ed., Academic Press, 2009

[128] P. Boher, T. Leroux, T. Bignon, V. Colomb-Patton, “A new way to characterize autostereoscopic 3D displays using Fourier optics instrument,” *Proc. SPIE 7237, Stereoscopic Displays and Applications XX*, 72370Z, 2009

[129] R. Rykowski, J. Lee, “Novel Technology for View Angle Performance Measurement,” *IMID 2008*, Ilsan, Korea, 2008

[130] A. Abileah, “3D Displays –Technologies & Testing Methods,” *SCIEN Workshop on 3D Imaging*, Stanford University, 2011

[131] A. Boev, R. Bregović, A. Gotchev, “Visual-quality evaluation methodology for multiview displays,” *Displays*, vol. 33, no. 2, pp. 103-112, 2012

[132] S. P. Heinrich, M. Back, “Resolution acuity versus recognition acuity with Landolt-style optotypes,” *Graefes Arch Clin Exp Ophthalmol*, vol. 251, no. 9. pp. 2235–2241, 2013

[133] G. Lafruit, M. Domański, K. Wegner, T. Grajek, T. Senoh, J. Jung, P. T. Kovács, P. Goorts, L. Jorissen, A. Munteanu, B. Ceulemans, P. Carballeira, S. García, M. Tanimoto, ”New visual coding exploration in MPEG: Super-MultiView and Free Navigation in Free viewpoint TV,” in *Proc. Stereoscopic Displays and Applications XXVII*, pp. 1-9(9), 2016, DOI:10.2352/ISSN.2470-1173.2016.5.SDA-426

[134] P. T. Kovács, A. Fekete, K. Lackner, V.K. Adhikarla, A. Zare, T. Balogh, “Big Buck Bunny light field test sequences,” ISO/IEC JTC1/SC29/WG11 M36500, Warsaw, Poland, Jun. 2015

[135] RDMA and RoCE for Ethernet Network Efficiency Performance. Online: http://www.mellanox.com/page/products_dyn?product_family=79. Visited:2018-09-16

[136] A. Zare, P. T. Kovács, A. Gotchev, "Self-contained slices in H.264 for partial video decoding targeting 3D light field displays," *2015 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, Lisbon, pp. 1-4, 2015, DOI:10.1109/3DTV.2015.7169379

[137] A. Zare, P. T. Kovacs, A. Aminlou, M. M. Hannuksela, A. Gotchev, "Decoding complexity reduction in projection-based light field 3D displays using self-contained HEVC tiles," *2016 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, Hamburg, pp. 1-4, 2016, DOI:10.1109/3DTV.2016.7548965

[138] Comprimato JPEG2000 encoder and decoder. Online: http://www.comprimato.com/ Visited: 2018-09-09

[139] P. Quax, P. Issaris, W. Vanmontfort, W. Lamotte, “Evaluation of distribution of panoramic video sequences in the eXplorative television project,” in *Proc. 22nd International Workshop on Network and Operating System Support for Digital Audio and Video*, Toronto, pp. 45-50, 2012

[140] F. Zilly, C. Riechert, M. Müller, P. Eisert, T. Sikora, P. Kauff, ”Real-time generation of multi-view video plus depth content using mixed narrow and wide baseline,” in. *J. Vis. Comun. Image Represent*. 25, 4 (May 2014), 632-648, 2014, DOI:10.1016/j.jvcir.2013.07.002

[141] M. P. Tehrani, S. Shimizu, G. Lafruit, T. Senoh, T. Fujii, A. Vetro, M. Tanimoto, “Use Cases and Requirements on Free-viewpoint Television (FTV),” ISO/IEC JTC1/SC29/ WG11 MPEG2013/N14104, Geneva, Switzerland, November 2013.

[142] P. T. Kovács, T. Balogh, J. Konieczny, G. Cordara, “Requirements of Light field 3D Video Coding,” ISO/IEC JTC1/SC29/WG11 M31954, San Jose, US, Jan. 2014

[143] P. T. Kovács, Zs. Nagy, A. Barsi, V.K. Adhikarla, R. Bregovic, “Proposal for additional features and future research to support light field video compression,” ISO/IEC JTC1/SC29/WG11 MPEG2013/m37434, Geneva, Switzerland, Oct. 2015

# ORIGINAL PAPERS

# I

## 3D VISUAL EXPERIENCE

by

P. T. Kovács, T. Balogh, 2010

High-Quality Visual Experience: Creation, Processing and Interactivity of High-Resolution and High-Dimensional Video Signals (eds M. Mrak, M. Grgic, M. Kunt), Springer, Signals and Communication Technology,
DOI: 10.1007/978-3-642-12802-8

# 3D Visual Experience

**Péter Tamás Kovács, Tibor Balogh**

Holografika Kft

Baross u. 3. H-1192 Budapest

Hungary

p.kovacs@holografika.com, t.balogh@holografika.com

## *Abstract*

The large variety of different 3D displaying techniques available today can be confusing, especially since the term “3D” is highly overloaded. This chapter introduces 3D display technologies and proposes a categorization that can help to easily grasp the essence of specific 3D displays that one may face, regardless of the often confusing and ambiguous descriptions provided by manufacturers. Different methods for creating the illusion of spatial vision, along with the advantages and disadvantages will be analyzed. Specific examples of stereoscopic, autostereoscopic, volumetric and light-field displays emerging or already available in the market are referenced. Common uncompressed 3D image formats preferred by each display technology are also discussed.

## *1. Introduction*

The chapter will go through the main technologies used for implementing 3D displays using the four top level categories of the "3D display family tree" created by the 3D@Home Consortium, Steering Team 4 [1]. It will take a different approach from that of the family tree detailing the main categories based on selected driving technologies that the authors think the most important. Other categorizations of 3D displays might exist, hopefully this one helps to understand the main trends and easily grasp the technology underlying different 3D displays.

The chapter strictly focuses on technologies that generate spatial vision, so it does not cover for example displays that project a floating 2D image using a fresnel lens, or displays that project 2D images on the sides on a cube.

## *2. Stereoscopic displays*

Stereoscopic displays [2][3] simulate 3D vision by showing different images to the eyes. The two images are either shown on a traditional 2D display, projected onto a special surface, or projected separately to the eyes. Stereoscopic displays by definition all require some kind of eyewear to perceive 3D (otherwise they are called autostereoscopic, as seen later). Separation of the two images, corresponding to the left and right eye happens either time-sequentially, or by means of differentiating wavelength or polarization.

### 2.1. Time sequential separation

In the time sequential case, left and right images are displayed on LCD or PDP or projected one after the other, and then separated by shutter glasses that block incoming light to one eye at a time, alternating the blocked eye with the same frequency as the display changes the images. Such shutter glasses are usually implemented with LCDs, which become transparent and opaque synchronized with the display. Several companies provide shutter glasses based 3D solutions including LG [4], Panasonic [5], Toshiba [4], eDimensional [6] and NVIDIA[9], projectors with high refresh rate for stereoscopic operation [7][8], and NVIDIA also provides a stereo driver to use the glasses with PC games [9]. A stylish NVIDIA shutter glass can be seen in Fig. 1, with the IR sensor used for synchronization in the frame of the glasses.

Fig. 1 NVIDIA 3D Vision Glasses. Image courtesy of NVIDIA Corporation

### 2.2. Wavelength based separation

Wavelength based separation is achieved by tinting the left and right images using different colours, overlaying the two and displaying the resulting 2D image. Separation is done by glasses with corresponding colour filters in front of the eyes, as done in the well known red-blue or red-green glasses. This method of creating stereoscopic vision is often referred to as the anaglyph method. The main advantage of anaglyph is that all signals and displaying requirements match 2D displaying requirements, thus existing storage, transmission and display systems can readily be used to show 3D imagery, only coloured glasses are needed (which is inexpensive, and often packaged together with an anaglyph “3D” DVD). This is possible because the left and right images are overlapped and separated by means of colour differences. A sample anaglyph image is shown in Fig. 2 where the two differently tinted overlapped images are clearly visible. This causes the main disadvantage of this technology, that is, colours are not preserved correctly, and ghosting artefacts are also present. Because of its simplicity, anaglyph stereoscopic videos are also appearing on YouTube, and also hundreds of games support anaglyph mode using NVIDIA 3D Vision™ Discover.

Fig 2 Anaglyph image. Image courtesy of Kim Scarborough.

A similar method better preserving colours apply narrow-band colour filters, separating the left and right images with wavelength triplets biased in a few 10 nm range, less visible to human perception [10].

### 2.3. Polarization based separation

Polarization based separation exploits the possibility of polarizing light and filtering them with polar filters. The two images are projected through different polarization filters onto a surface that reflects light toward viewers, keeping the polarization of the incoming light (mostly) unmodified. Viewers wearing glasses with the respective filters in front of the eyes can then perceive a stereoscopic view. A popular example of this technology can be experienced in most 3D cinemas. [11][12]

Light can be polarized either linearly or circularly. In the first case, the left and right images pass through two perpendicular linear polarizers and then projected onto a surface. The reflected images then pass through the respective polarizing filters that are embedded into glasses, separating the left and right images. The downside of linear polarization is that the image degrades when a user tilts her head, as separation does not work as intended with this orientation. Circular polarization overcomes this problem being invariant to head tilt. In this case one image is polarized with clockwise, the other with counter-clockwise direction.

The advantage of the polarization based stereoscopic technique is that it keeps image colours intact (unlike anaglyph), with glasses that are relatively cheap, however the overall brightness is challenged and some cross-talk is always present.

One way of generating a pair of polarized images is by using two projectors, one projecting the left eye image with a polarizing filter in front of it, the other projecting the right eye image with orthogonal polarization [13][14]. There is also a single-projector technique, in which a rotating polarizator wheel or an LCD polarization modulator is used in the projector to change the direction of polarization of every second frame [15]. One needs a special projection screen to reflect polarized images, as surfaces used for 2D projection do not maintain the polarization of the reflected light. Previously silver screens have been used, now specialized materials are available for this purpose [16]. Polarized stereo images can also be created using two LCD monitors with perpendicular polarization arranged with a passive beamsplitter (half-mirror) at a bisecting angle between the displays. The resulting stereo image pair can be seen directly with polarizing glasses [17,18], as shown in. Fig. 3.

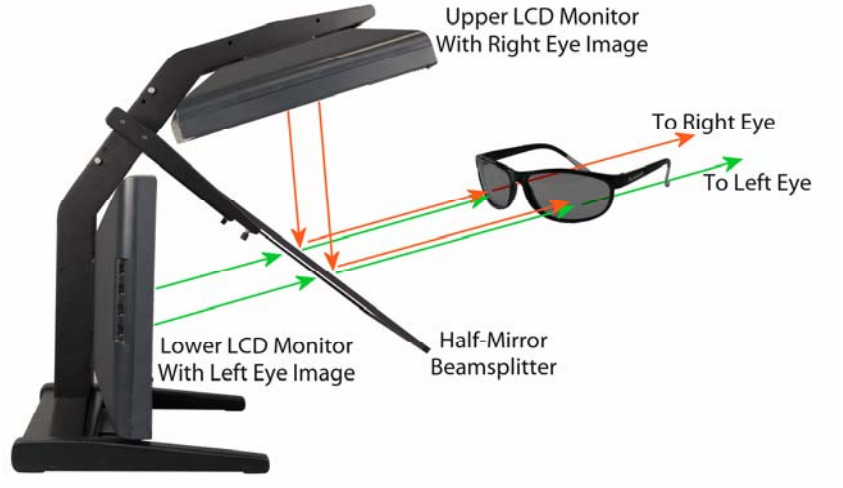



Fig. 3 Polarization based stereoscopy using two flat screens. Image courtesy of Planar Systems, Inc.

Another approach to create polarized images is using a patterned micro-polarizer sheet (also called x-pol or micro-pol), which is placed on the surface of a 2D LCD panel. The sheet is aligned with the rows on the LCD panel so that pixels in the even row will be polarized clockwise, pixels in the odd row will polarized in reverse, as shown in Fig. 4. Providing corresponding line interleaved stereoscopic images for the display will result in a 3D effect when using circularly polarized glasses (although with resolution reduced by half). Some manufacturers providing such displays are LG [4] and Zalman [19], but 3D laptops using this technology also appeared from Acer [20].

### 2.4. Discussion of stereoscopic systems

Stereoscopic techniques are definitely the simplest and cheapest, thus the most widespread methods to generate 3D vision. On the other hand, they come with several drawbacks. A stereo image with glasses provides correct 3D images only from a single point of view. Observing the same image from other locations results in distorted views, which is most visible while moving in front of the screen, when the image ”follows” the viewer. Although this limitation can be overcome by tracking the position / orientation / gaze of the user and updating images in response to movements [21], some latency will inherently be introduced [22], significantly compromising immersiveness and limiting the correct view to a single (tracked) user. This and other missing 3D cues result in effects like discomfort, sea sickness, nausea and headache which make them inconvenient for long-term use according to some users [23].

One possible explanation comes from neuroscientists’ research in the field of human perception of 3D. They found that showing each eye its relevant image is not enough for the brain to understand the 3D space [24]. For getting the 3D picture of the environment, humans rely on two main visual cues: the slightly different image seen by each eye and the way the shape of an object changes as it moves. A brain area, the anterior intraparietal cortex (AIP), integrates this information [25]. With a stereoscopic display the image becomes 3D, but as soon the brain thinks that it does see a 3D image, it starts working like in a normal 3D world, employing micro head movements to repeatedly and unconsciously check the 3D model built in our brain. When an image on a stereo display is checked and the real 3D world mismatches the 3D image, the trick is revealed. Presumably the AIP cortex never got used to experience such 3D cue mismatch during its evolution and this produces glitches which result in unwanted effects.

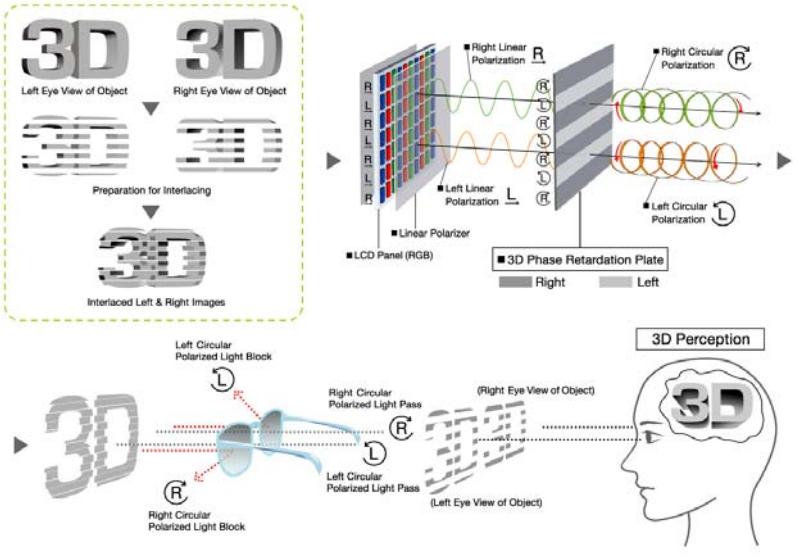



Fig. 4 Principle of micro-polarization. Image courtesy of Zalman Tech Co., Ltd.

## 2.5. Stereoscopic 3D uncompressed image formats

Stereoscopic displays need two images as input (left eye and right eye image), which seems to be simple, yet various formats exist. The most straightforward solution is having two different images making up a 3D frame (see Fig. 5.), but this requires double bandwidth compared to the 2D case.


Fig. 5 Left-right image pair. Image courtesy of NXP Semiconductors.

Another common approach uses an image and a corresponding depth image often called 2D + Depth (see Fig. 6.), which may consume less bandwidth depending on the bit depth of the depth map, but needs metadata to map depth information to the 3D context, and still consumes more than a 2D image.


Fig. 6 Image plus depth map. Image courtesy of NXP Semiconductors.

The 2D + Delta format stores the left (or right) video stream intact, and adds the stereo disparity or delta image that is used to reconstruct the other view. The advantage is that compressed Delta information can be embedded into an MPEG stream in a way that does not affect 2D players, but provides stereoscopic information to compatible 3D decoders [26].

To make the transition from 2D to 3D easier, broadcasters and manufacturers preferred stereoscopic image formats that can be fit into a 2D frame compatible format, in order to defer the upgrade of the transmission infrastructure. Some examples of such formats include frame doubling, side-by-side, interleaved and checkerboard, which can be seen in Fig. 7.

The frame doubling approach uses a single 2D stream to transmit alternating left and right images, halving the effective frame rate. This is the most suitable format for shutter-glass based systems and 3D projectors using rotating polarizers.

Side-by-side places the left and right images next to each other. This either requires doubled horizontal resolution, or halves the horizontal resolution of left and right images, fitting them in the original 2D image size. A very similar image configuration is over/under.

Interleaving places rows of the left view into even lines, and rows of the right view into odd lines (or the same reversed). As with side-by-side, two possibilities are doubling image size and keeping the resolution of the images or halving the resolution of the component images to fit them into a 2D frame with the same size. Interleaving can also work in a vertical configuration. This representation is the best choice for a 3D display based on micro-polarizers.

The checkerboard format mixes pixels of the left and right images so that they alternate in one row, and alternate the reverse way in the next row. This makes

better interpolation of the missing pixels possible when reconstructing the left and right images. This representation is used by Texas Instruments DLPs.

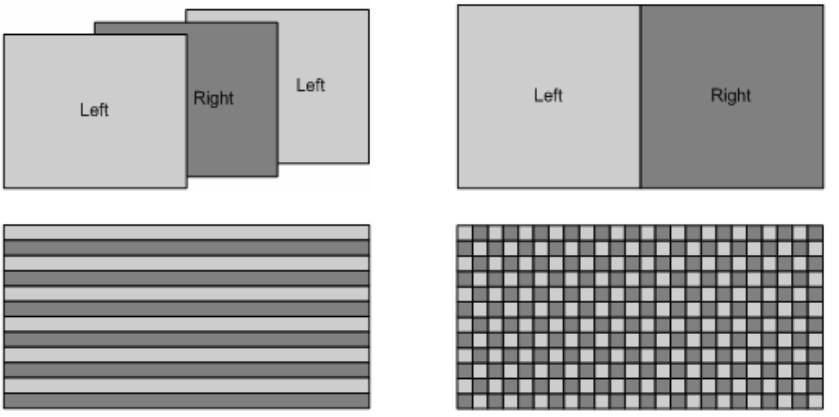



Fig. 7 Stereoscopic image formats (from left to right, top to bottom): Frame doubling, Side-by-side, Interleaved and Checkerboard. Image courtesy of NXP Semiconductors.

The High-Definition Multimedia Interface (HDMI) supports stereoscopic 3D transmission starting from version 1.4 of the HDMI specification. It defines common 3D formats and resolutions for supporting 3D up to 1080p resolution and supports many 3D formats including frame doubling, side-by-side, interleaving and 2D+depth. There are two mandatory 3D formats defined, which must be supported by all 3D display devices: 1080p@24Hz and 720p@50/60Hz [27].

### 2.6. Multi-user stereo and CAVE systems

A common extension of stereoscopic projection systems is using them in CAVEs [28] that use three to six walls (possibly including the floor and ceiling) as stereoscopic 3D projection screens. The users entering the CAVE wear glasses for stereoscopic viewing, one of them (commonly referred to as “leader” or “driver”) wearing extra equipment for tracking. Since the stereo pairs are generated for a single point of view that of the driver, using stereoscopic 3D for multiple users is problematic, as only the driver will perceive a correct 3D image, all others will see a distorted scene. Whenever the driver moves, the images are updated, thus all other users will see the scene moving (according to the movement of the driver), even is they stay at the same place not doing any movements, resulting in disturbing effects. Stereoscopic CAVEs are widely used for providing immersive 3D experience, but unfortunately carry all the drawbacks of stereoscopic systems.

### 2.7. Head Mounted Displays

A head mounted display [29] is a display device worn on the head or as part of a helmet that has a small display optic in front of both eyes in case of a binocular HMD (monocular HMDs also exist but unable to produce 3D images). A typical HMD has two small displays with lenses embedded in a helmet or eye-glasses. The display units are miniaturized and may include CRT, LCDs, LCOS, or OLED. Some HMDs also allow partial see-through thus super-imposing the virtual scene on the real world. Most HMDs also have head tracking functionality integrated. From the 3D vision point of view, they are equivalent to glasses based systems. HMD manufacturers include Cybermind [30], I-O [31], Rockwell Collins [32], Trivisio [33], Lumus [34].

## *3. Autostereoscopic displays*

Autostereoscopic displays provide 3D perception without the need for wearing special glasses or other head-gear, as separation of left / right image is implemented using optical or lens raster techniques directly above the screen surface. In case of two views, one of the two visible images consists of even columns of pixels; the second image is made up of odd columns (other layouts also exist). The two displayed images are visible in multiple zones in space. If the viewer stands at the ideal distance and in the correct position he or she will perceive a stereoscopic image (sweet spot). Such passive autostereoscopic displays require the viewer to be carefully positioned at a specific viewing angle, and with her head in a position within a certain range. The downside is that there is a chance of the viewer being in the wrong position (invalid zone) and seeing an incorrect image. This means that the viewer is forced to a fixed position, reducing the ability to navigate freely and be immersed.

To overcome the problem of invalid zones head and/or eye tracking systems can be used to refresh the images whenever the viewer is about to enter such a zone and experience an incorrect 3D image [35]. Even though there could be latency effects, such a system provides the viewer with parallax information and it is, therefore, a good solution for single user applications. Multi-user extensions of this technique are also developed [36].

Some autostereoscopic displays show stereoscopic 3D (consisting of two images), others go beyond that and display multiview 3D (consisting of more than two views). Multiview displays [37] project different images to multiple zones in space. In each zone only one image (view) of the scene is visible. The viewer's two eyes are located in different zones, seeing different images thus 3D perception is enabled. When the user moves, entering different zones will result in different views, thus a somewhat limited horizontal motion parallax effect is achieved. As the number of views ranges from 4 to 9 in current multiview displays, the transi-

tion to adjacent zones is discrete, causing „jumps” as the viewer moves. Multiview displays allow multiple simultaneous viewers, restricting them, however, to be within a limited viewing angle. The image sequences are periodically repeated in most multi-view displays, thus enabling more diamond shaped viewing positions at the expense of invalid zones in between.

Autostereoscopic displays typically use parallax barrier, lenticular sheet or wavelength selective filter which divide the pixels of the underlying, typically LCD display into two or more sets corresponding to the multiple directions.

### 3.1. Parallax barrier

Parallax barrier [38] is an array of slits spaced at a defined distance from a high resolution display panel. The parallax effect is created by this lattice of very thin vertical lines, causing each eye to view only light passing through alternate image columns, allowing the well-positioned viewer to perceive stereoscopic 3D, as shown In Fig. 8. Parallax barrier-based displays typically show stereoscopic 3D made up of two images, but with the proper choice of distance and width of the slit multi-view effect can be provided. Parallax barrier systems are less efficient in terms of light output, thus the image gets darker than in 2D, especially in case of multiple views.

Parallax barrier displays are making their way to mobile devices, as they can be easily implemented in small size. One example is a 3.07” size WVGA 3D LCD from Masterimage with an integrated, configurable parallax barrier layer on top of the LCD (2D, portrait or landscape 3D). Such displays make the manufacturing of 3D-enabled handheld devices like the Hitachi Wooo H001 possible [39].

Parallax barrier display manufacturers include Spatial View [40], Tridelity [41] and NewSight [42].

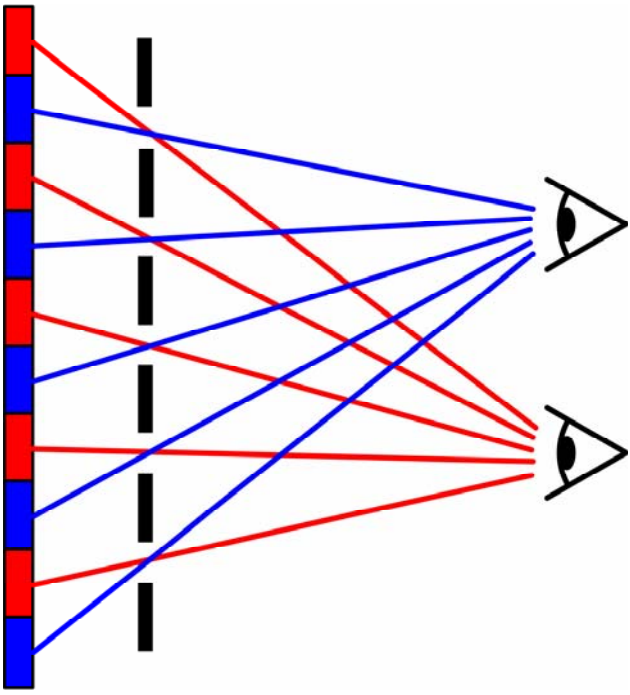


Fig. 8 Principle of parallax barrier based stereoscopic vision

### 3.2. Lenticular lens

Lenticular lens [37] based displays, which are the most common for implementing multiview 3D, use a sheet of cylindrical lens array placed on top of a high resolution LCD in such a way that the LCD image plane is located at the focal plane of the lenses. The effect of this arrangement is that different LCD pixels located at different positions underneath the lenticular fill the lenses when viewed from different directions. Provided these pixels are loaded with suitable 3D image information, 3D effect is obtained in which left and right eyes see different but matching information, as shown in Fig. 9. Both parallax barrier and lenticular lens based 3D displays require the user to be located at a specific position and distance to correctly perceive the stereoscopic image, as incorrect positioning results in incorrect images reaching the eye. A major disadvantage of lenticular lens based systems is their inability to use the displays in 2D with full resolution.

Lenticular 3D display manufacturers include Alioscopy [43], Philips (now retired from 3D display business) [44], NEC [4] and Tridelity [41].

Since both parallax barrier and lenticular lens based displays require a flat panel display underneath, the size of such 3D displays is always limited by the maximum size of such panels manufactured. As of November 2009, the maximum size is slightly more than 100 inches diagonal. Since tiling such displays is not seamless, these technologies are not scalable to arbitrary large sizes.

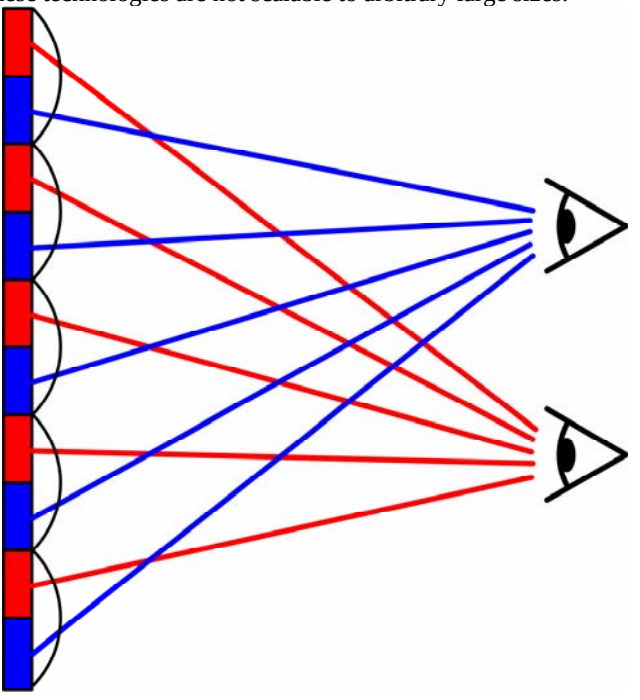


Fig. 9 Principle of lenticular lens based stereoscopic vision

### 3.3. Wavelength selective filters

Another possible implementation is using wavelength selective filters for the multi-view separation. The wavelength-selective filter array is placed on a flat LCD panel oriented diagonally so that each of the three colour channels correspond to a different direction, creating the divided viewing space necessary for 3D vision. A combination of several perspective views (also combining colour channels) is displayed. The filter array itself is positioned in front of the display and transmits the light of the pixels from the combined image into different directions, depending on their wavelengths. As seen from the viewer position different spectral components are blocked, filtered or transmitted, separating the viewing space into several zones where different images can be seen. [45]

### 3.4. Multiview 3D uncompressed image formats

Common image formats used by multi-view displays include multiple images on multiple links, 2D+Depth (described earlier), 2D+Depth with two layers, and the extension of frame-doubling, side-by-side and interleaving to the multi-view case.

Using multiple links, the same number of display interfaces are provided as many views the display have (possibly combined with side-by-side). When used for multi-view, the 2D + Depth approach is often criticized for missing parts of the scene behind occluded objects. This effect is somewhat reduced by using two layers, that is 2D + Depth + Occluded 2D + Occluded depth, what Philips calls Declipse format. An example 3D image in Declipse format can be seen in Fig. 10.

Frame doubling, side-by-side and interleaving (either horizontal or vertical), as described at stereoscopic displays can be naturally extended for using with multiple views. However, if the resolution of the image is to be kept, even more significant reduction in the resolution of the component images is required. We have to note that in case of multi-view displays, the resolution of the individual views is divided anyway as it cannot have more pixels than the underlying LCD panel.

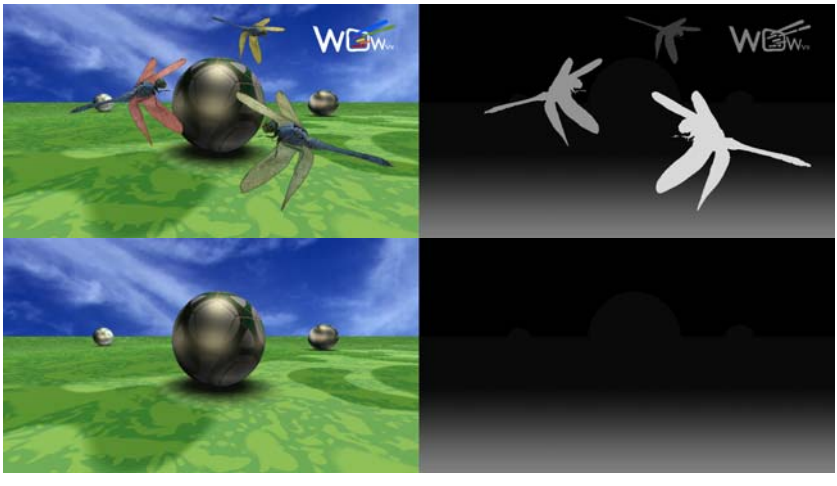


Fig. 10 2D image + depth + occluded image + occluded depth. Image courtesy of Philips Electronics N.V.

As a general rule for multi-view systems, the resolution seen in a direction is equal to the native resolution of the underlying display panel divided by the number of views.

## *4. Volumetric displays*

Volumetric displays use a media positioned or moved in space on which they project/reflect light beams so they are scattered/reflected from that point of space. The media used is generally a semi-transparent or diffuse surface. Among volumetric displays there are exotic solutions like the laser induced plasma explosions [46]. In general they less conform to displaying conventions and in most cases follow the “looking into” instead of “looking out” philosophy.

One possible solution is a moving screen on which different perspectives of the 3D object are projected. A well known solution [47] is a lightweight screen sheet that is rotated at very high speed in a protecting globe and the light beams from a DLP microdisplay are projected onto it. Such a display is shown in Fig. 11a. Employing proper synchronization it is possible to see 3D objects in the globe [48]. Such systems can be considered time-multiplexing solutions, where number of the displayable layers or voxels is determined by the speed of the projection component. A similar solution is the usage of rotated LED arrays as the emissive counterpart of the reflective moving media, as done in the display shown in Fig. 11b.

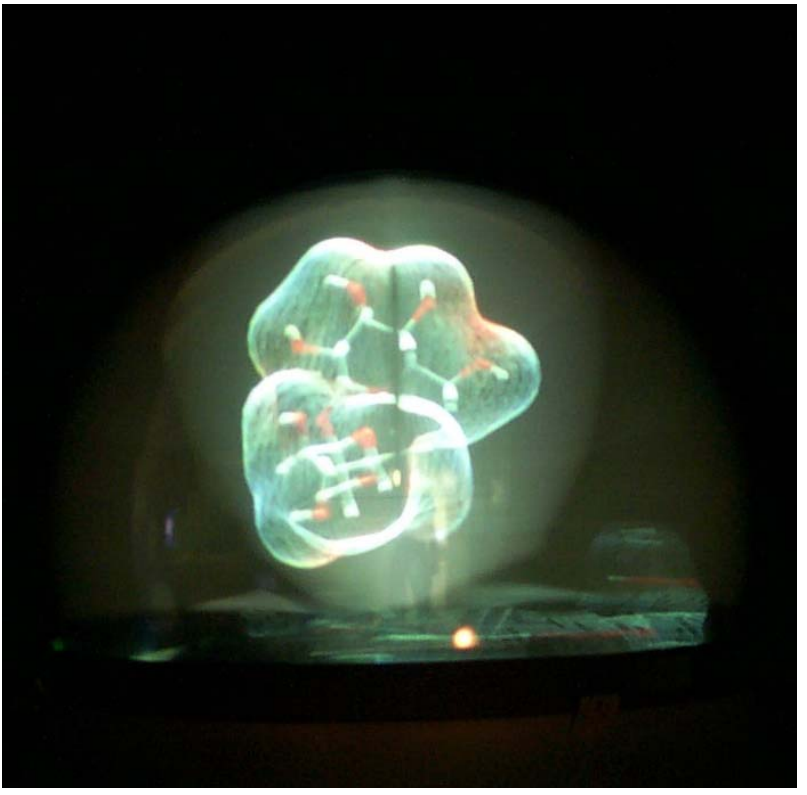


Fig. 11 Perspecta volumetric display from Actuality. Image courtesy of Actuality Systems, Inc.

Another technique in volumetric display technology is using two or more LCD layers as a projection screen, creating the vision of depth. Deep Video Imaging and PureDepth [49] produced a display consisting two LCDs. The depth resolution equals 2, enabling special foreground-background style content only, which is hard to qualify as 3D. The DepthCube display [50] from LightSpace Technologies shown in Fig. 12 has 20 layers inside. The layers are LCD sheets that are transparent / opaque (diffuse) when switched on/off, and are acting as a projection screen positioned in 20 positions. Switching the 20 layers is synchronized to the projection engine, inside which an adapting optics is keeping the focus.

Fig. 12 DepthCube volumetric display. Image courtesy of LightSpace Technologies Inc. [provisional permission to be finalized]

Disadvantages of volumetric displays are scalability and the ability to display occlusion, since the light energy addressed to points in space cannot be absorbed by foreground pixels. The problem of occlusion has been recently solved by using an anisotropic diffuser covering a rapidly spinning mirror [51]. As of advantages, both vertical and horizontal parallax is provided by principle.

The natural data format for volumetric displays is layered images (in the layered case) or image sequence showing the scene from all around (in the rotating case).

## *5. Light field displays*

### 5.1. Integral imaging

Integral imaging [52] 3D displays use a lens array and a planar display panel. Each elemental lens constituting the lens array forms each corresponding elemental image based on its position, and these elemental images displayed on the display panel are integrated forming a 3D image. Integral imaging can be though of as a 2D extension of lenticular lens based multiview techniques, providing both

horizontal and vertical parallax. Real-time generation of integral images from live images has been demonstrated [53].

Its disadvantages are narrow viewing angle and reduced resolution. The viewing angle within which observers can see the complete image is limited due to the restriction of the area where each elemental image can be displayed. Each elemental lens has its corresponding area on the display panel. To prevent image flipping the elemental image that exceeds the corresponding area is discarded optically in direct pick up method or electrically in computer-generated integral imaging method. Therefore the number of the elemental images is limited and observers outside the viewing zone cannot see the integrated image.

### 5.2. Holographic displays

Pure holographic systems [54] have the ability to store and reproduce the properties of light waves. Techniques for creating such holographic displays include the use of acusto-optic material and optically addressed spatial light modulators [55]. Pure hologram technology utilises 3D information to calculate a holographic pattern [56], generating true 3D images by computer control of laser beams and a system of mirrors. Compared to stereoscopic and multi-view technologies the main advantage of a hologram is the good quality of the generated 3D image. Practical application of this technology today is hampered by the huge amount of information contained in the hologram which limits its use to mostly static 3D models, in limited size and narrow viewing angle.

### 5.3. HoloVizio type light-field displays

Such displays follow hologram geometry rules, however direction selective light emission is obtained by directly generating the light beams instead of interference. In this way the huge amount of redundant information present in a hologram (phase, speckle) is removed and only those light beams are kept which are needed to build up the 3D view. Each point of the holographic screen emits light beams of different colour and intensity to the various directions in a controlled manner. The light beams are generated through a specially arranged light modulation system and the holographic screen makes the necessary optical transformation to compose these beams into a 3D view. The light beams cross each other in front of the screen or they propagate as if they were emitted from a common point behind the screen, as shown in Fig. 13. With proper control of the light beams viewers see objects behind the screen or floating in the air in front of the screen just like with a hologram.

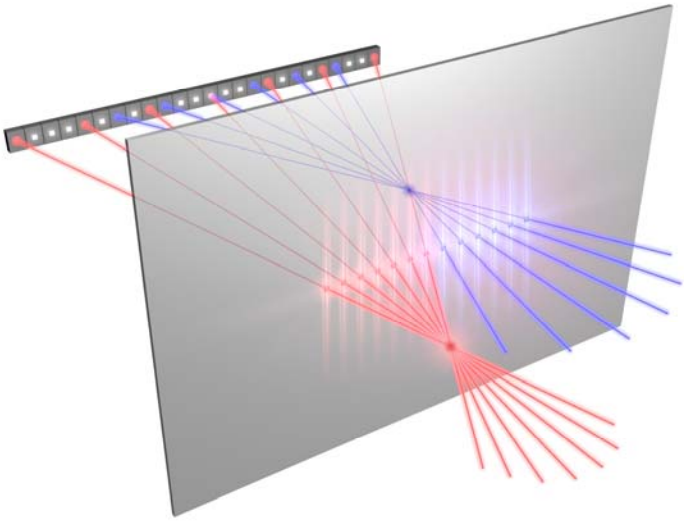


Fig. 13 Principle of HoloVizio light-field displays.

The main advantage of this approach is that, similarly to the pure holographic displays, it is able to provide all the depth cues for multiple freely moving users within a reasonably large field of view. Being projection based and using arbitrary number of projection modules, this technique is well scalable to very high pixel count and display size, not being limited to the resolution of a specific display technology (like the ones using a single LCD panel). 2D compatibility is implicitly solved here, as light rays making up a 2D image are also easy to emit without any reconfiguration. These systems are fairly complex because of the large number of optical modules and the required driving/image generation electronics.

The natural data format for this kind of display is the light field [57] that is present in a natural 3D view. HoloVizio 3D displays are an implementation of this technology [58, 59].

## *6. Conclusion*

Very different ideas have been used so far to achieve the goal of displaying realistic 3D scenes, the ultimate goal being a virtual 3D window that is indistinguishable from a real window. Most implementations of the approaches mentioned have found their specific application areas where they perform best, and they are gaining an ever growing share in visualization. 3D data is already there in

a surprisingly large number of industrial applications, still visualized in 2D in most cases.

As for home use, the natural progression of technology will bring the simplest technologies to the mainstream first, and with advances in technology, cost effectiveness and increased expectations regarding 3D will eventually bring more advanced 3D displays currently only used by professionals to the homes.

## Reference list


1. 3D@Home Consortium Display Technology Steering Team 4: 3D Display Technology Family Tree. Motion Imaging Journal, 118(7), insert (2009)
2. Holmes, O.W.: The Stereoscope and Stereoscopic Photographs. Underwood & Underwood, New York (1906)
3. Ezra, D., Woodgate, G.J., Omar, B.A., Holliman, N.S., Harrold, J., Shapiro, L.S.: New autostereoscopic display system. In: Stereoscopic Displays and Virtual Reality Systems II, Proc. SPIE 2409, 31-40 (1995)
4. Insight Media: 3D Displays & Applications at SID'09. http://www.insightmedia.info. Accessed 10 Nov. 2009
5. Panasonic: Full HD 3D Technology. http://www.panasonic.com/3d/default.aspx. Accessed 10 Nov. 2009
6. eDimensional 3D Glasses. http://www.edimensional.com/images/demo1.swf. Accessed 10 Nov. 2009
7. Barco: Stereoscopic projectors. http://www.barco.com/en/productcategory/15. Accessed 10 Nov. 2009
8. Viewsonic: PJD6241 projector. http://ap.viewsonic.com/in/products/productspecs.php?id=382. Accessed 10 Nov. 2009
9. NVIDIA 3D Vision Experience. http://www.nvidia.com/object/3D_Vision_Overview.html. Accessed 10 Nov. 2009
10. Jorke, H., Fritz, M.: INFITEC a new stereoscopic visualisation tool by wavelength multiplex imaging. In: Proc. Electronic Displays. Wiesbaden (2003)
11. RealD. http://www.reald.com/Content/about-reald.aspx. Accessed 10 Nov. 2009
12. IMAX. http://www.imax.com/corporate/theatreSystems/. Accessed 10 Nov. 2009
13. Valley View Tech: VisDuo. http://www.valleyviewtech.com/immersive.htm#visduo. Accessed 10 Nov. 2009
14. Barco: Passive 3D display systems with two projectors. http://www1.barco.com/entertainment/en/stereoscopic/passive.asp. Accessed 10 Nov. 2009
15. DepthQ Polarization Modulator. http://www.depthq.com/modulator.html. Accessed 10 Nov. 2009

16. Brubaker, B.: 3D and 3D Screen Technology (Da-Lite 3D Screen Whitepaper). http://www.3dathome.org/files/products/product.aspx?product=1840. Accessed 10 Nov. 2009
17. Fergason, J.L., Robinson, S.D., McLaughlin, C.W., Brown, B., Abileah, A., Baker, T.E., Green, P.J.: An innovative beamsplitter-based stereoscopic/3D display design. In: Stereoscopic Displays and Virtual Reality Systems XII, Proc. SPIE 5664, 488-494 (2005)
18. Robinson, S.D., Abileah, A., Green, P.J.: The StereoMirror™: A High Performance Stereoscopic 3D Display Design. In: Proc. SID Americas Display Engineering and Applications Conference (ADEAC '05). Portland, Oregon, USA (2005)
19. Zalman Trimon 2D/3D Convertible LCD Monitor. http://www.zalman.co.kr/ENG/product/Product_Read.asp?idx=219. Accessed 10 Nov. 2009
20. Insight Media: 3D Displays & Applications, August 2009. http://www.insightmedia.info/. Accessed 10 Nov. 2009
21. Woodgate, G.J., Ezra, D., Harrold, J., Holliman, N.S., Jones, G.R., Moseley, R.R.: Observer-tracking autostereoscopic 3D display systems. Stereoscopic Displays and Virtual Reality Systems IV, Proc SPIE 3012, 187-198 (2004)
22. Wul, J.R., Ouhyoung, M.: On latency compensation and its effects on head-motion trajectories in virtual environments. The Visual Computer, 16(2), 79-90 (2000)
23. Takada, H., Fujikake, K., Miyao, M.: On a Qualitative Method to Evaluate Motion Sickness Induced by Stereoscopic Images on Liquid Crystal Displays. In: Proc. 3rd International Conference on Virtual and Mixed Reality. San Diego, CA, USA (2009)
24. Baecke, S., Lützkendorf, R., Hollmann, M., Macholl, S., Mönch, T., Mulla-Osman S., Bernarding, J.: Neuronal Activation of 3D Perception Monitored with Functional Magnetic Resonance Imaging. In: Proc. Annual Meeting of Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (gmds). Leipzig, Germany (2006)
25. Vanduffel, W., Fize, D., Peuskens, H., Denys, K., Sunaert, S., Todd, J.T., Orban, G.A.: Extracting 3D from motion: differences in human and monkey intraparietal cortex. Science, 298(5592), 413-415 (2002)
26. TDVision Knowledgebase contribution: 3D Ecosystem. http://www.tdvision.com/WhitePapers/TDVision_Knowledgbase_Public_Release_Rev_2.pdf. Accessed 19 Nov. 2009
27. Park, J.: 3D over HDMI – New feature of the HDMI 1.4 Specification. In: Proc. DisplaySearch TV Ecosystem Conference. San Jose, CA, USA (2009)
28. Cruz-Neira, C., Sandin, D.J., DeFanti, T.A., Kenyon, R.V., Hart, J.C.: The CAVE: Audio Visual Experience Automatic Virtual Environment. Communications of the ACM, 35(6), 65-72 (1992)
29. Heilig, Morton L.: Stereoscopic-television apparatus for individual use. United States Patent 2955156 (1960)

30. Cybermind Interactive Nederland. http://www.cybermindnl.com/index.php?option=com_content&task=view&id=19&Itemid=49. Accessed 11 Nov. 2009
31. i-O Display Systems. http://www.i-glassesstore.com/hmds.html. Accessed 11 Nov. 2009
32. Rockwell Collins: Soldier Displays. http://www.rockwellcollins.com/products/gov/surface/soldier/soldier-systems/index.html. Accessed 11 Nov. 2009
33. Trivisio Prototyping. http://trivisio.com/index.php/products/hmdnte/options/hmd-options. Accessed 11 Nov. 2009
34. Lumus. http://www.lumus-optical.com/index.php?option=com_content&task=view&id=5&Itemid=8. Accessed 11 Nov. 2009
35. Boev, A., Raunio, K., Georgiev, M., Gotchev, A., Egiazarian, K.: Opengl-Based Control of Semi-Active 3D Display. In proc. 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video. Istanbul, Turkey (2008)
36. HELIUM 3D: High Efficiency Laser-Based Multi-User Multi-Modal 3D Display. http://www.helium3d.eu. Accessed 10 Nov. 2009
37. van Berkel, C., Parker, D.W., Franklin A.R.: Multiview 3D-LCD. In: Stereoscopic Displays and Virtual Reality Systems III, Proc. SPIE 2653, 32-39 (1996)
38. Sandin, D., Margolis, T., Ge, J., Girado, J., Peterka, T., DeFanti, T.:The Varrier autostereoscopic virtual reality display. In: ACM Transactions on Graphics, Proc. ACM SIGGRAPH, 24(3), 894-903 (2005)
39. Insight Media: 3D Displays & Applications, April 2009. http://www.insightmedia.info/. Accessed 10 Nov. 2009
40. SpatialView. http://www.spatialview.com. Accessed 10 Nov. 2009
41. Tridelity Display Solutions. http://www.tridelity.com. Accessed 11 Nov. 2009
42. Newsight Advanced Display Solutions. http://www.newsight.com/support/faqs/autostereoscopic-displays.html. Accessed 11 Nov. 2009
43. Alioscopy – glasses-free 3D displays. http://www.alioscopyusa.com/content/technology-overview. Accessed 19 Nov. 2009
44. Philips 3D Display Products. http://www.business-sites.philips.com/3dsolutions/3ddisplayproducts/index.page. Accessed 19 Nov. 2009
45. Schmidt, A., Grasnick, A.: Multiviewpoint autostereoscopic dispays from 4D-Vision GmbH. In: Stereoscopic Displays and Virtual Reality Systems IX, Proc. SPIE 4660, 212-221 (2002)
46. Advanced Industrial Science and Technology: Three Dimensional Images in the Air, Visualization of “real 3D images” using laser plasma.

http://www.aist.go.jp/aist_e/latest_research/2006/20060210/20060210.html. Accessed 19 Nov. 2009
47. Favalora, G.E., Napoli, J., Hall, D.M., Dorval, R.K., Giovinco, M., Richmond, M.J., Chun, W.S.: 100 Million-voxel volumetric display. In: Cockpit Displays IX: Displays for Defense Applications, Proc. SPIE 4712, 300-312 (2002)
48. Jones, A., Lang, M., Fyffe, G., Yu, X., Busch, J., McDowall, I., Bolas, M., Debevec, P.: Achieving Eye Contact in a One-to-Many 3D Video Teleconferencing System. In: ACM Transactions on Graphics, Proc. SIGGRAPH, 28(3), 64, (2009)
49. PureDepth multi layer display. http://www.puredepth.com/technologyPlatform_ip.php?l=en. Accessed 03 Dec. 2009
50. Sullivan, A: 3 Deep. http://www.spectrum.ieee.org/computing/hardware/3-deep. Accessed 05 Oct 2009
51. Jones, A., McDowall, I., Yamada, H., Bolas, M., Debevec, P.: Rendering for an interactive 360° light field display. ACM Transactions on Graphics 26 (3), 40 (2007)
52. Davies, N., McCormick, M.: Holoscopic Imaging with True 3D-Content in Full Natural Colour. In: Journal of Photonic Science, 40, 46-49 (1992)
53. Taguchi, Y., Koike, T., Takahashi, K., Naemura, T.: TransCAIP: Live Transmission of Light Field from a Camera Array to an Integral Photography Display. IEEE Transactions on Visualization and Computer Graphics, 15(5), 841-852 (2009)
54. Lucente M.: Interactive three-dimensional holographic displays: seeing the future in depth. ACM SIGGRAPH Computer Graphics, 31(2), 63-67 (1997)
55. Benton, S.A.: The Second Generation of the MIT Holographic Video System. In: Proc. TAO First International Symposium on Three Dimensional Image Communication Technologies. Tokyo, Japan (1993)
56. Yaras, F., Kang, H., Onural, L.: Real-time color holographic video display system. In: Proc. 3DTV-Conference: The True Vision Capture, Transmission and Display of 3D Video. Potsdam, Germany (2009)
57. Levoy, M., Hanrahan, P.: Light Field Rendering. In: Proc. ACM SIGGRAPH. New Orleans, LA, USA. (1996)
58. Balogh, T.: Method & apparatus for displaying 3D images. U.S. Patent 6,201,565, EP0900501 (1997)
59. Balogh, T.: The HoloVizio System. In: Stereoscopic displays and virtual reality systems XIII, Proc. SPIE 6055, 60550U (2006)

**II**

# QUALITY MEASUREMENTS OF 3D LIGHT-FIELD DISPLAYS

by

P. T. Kovács, A. Boev, R. Bregović, A. Gotchev, 2014

in Proc. Eighth International Workshop on Video Processing and Quality Metrics
for Consumer Electronics (VPQM 2014)

# QUALITY MEASUREMENTS OF 3D LIGHT-FIELD DISPLAYS

*Péter Tamás Kovács* [1, 2], *Atanas Boev*[2], *Robert Bregović*[2], *Atanas Gotchev*[2]

[1]Holografika, Budapest, Hungary
[2]Department of Signal Processing, Tampere University of Technology, Tampere, Finland

**ABSTRACT**

We present methods to measure the spatial, angular and depth resolution of LF displays using off-the-shelf equipment and performing a subjective experiment. The spatial resolution is measured in circles per degree and the challenge is to display and quantify sinusoidal patterns with varying spatial frequencies on displays, which in general do not exhibit regular or pixel-like structure. Being specific for 3D displays, the angular resolution represents the number of unique directions that can be emitted from a point, and measured in circles per degree. The paper presents the experimental setup and discusses the results. The depth resolution shows the minimum distinguishable depth that can be reproduced by the display used, and is estimated by a subjective experiment.

## 1. INTRODUCTION

3D displays are expected to reconstruct 3D visual scenes with certain level of realism relying on various 3D visual cues. Auto-stereoscopic (AS) and multi-view (MV) displays generate a discrete set of *views* (two or more) forming stereo-pairs thus providing binocular cues without the need of wearing 3D glasses [1]. Observers are supposed to find the proper viewpoint (sweet spot) where the views from a stereo pair are best visible by the corresponding eyes. AS displays do not provide head parallax and MV displays provide a very limited one resulting from the limited set of discrete views available. This is usually accompanied by the effect of transition (jump) between those views and known as *image flipping*.

The next generation of 3D displays, denoted as light-field (LF) displays aim at providing a continuous head parallax experience over a wide viewing zone with no use of glasses. In general, this requires a departure from the 'discrete view' formalism. Instead, a certain continuous reconstruction of the light field emanated or reflected from the 3D visual scene is targeted. Correspondingly, the scene is represented by a discrete set of light rays, which serve as generators for the subsequent process of continuous light field reconstruction.

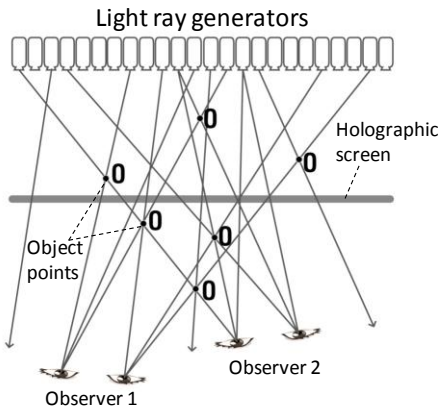



**Figure 1: Light-field display architecture**

Technologically, this two-stage process (scene representation and scene reconstruction) can be implemented by using an array of projection modules emitting light rays toward a custom-made LF reconstruction surface (screen). The latter makes the optical transformation that composes rays into a continuous LF [9].

With proper design of the LF display, light rays leaving the screen spread in multiple directions, as if they were emitted from points of 3D objects at fixed spatial locations. This gives the illusion of points appearing either behind the screen, on the screen, or floating in front of it, achieving an effect similar to holograms, as illustrated in Figure 1. Therefore, the LF reconstruction screen is sometimes denoted as a *holographic* screen.

With the emergence of LF displays, the issue of their quality and its characterization becomes of primary importance. For 2D displays, display quality is directly characterized by their spatial resolution and observation angle, which quantification and measurement are standardized [4] and therefore, easily available to and interpretable by the end users. Manufacturers of 3D displays do not follow a single metric. Instead, they either describe display quality using 2D-related parameters or provide no such information at all. For example, the visualization capabilities of a 3D display are given in terms of the underlying TFT matrix resolution and the number of unique views [16].

Most previous work related to characterization of 3D displays is focused on stereoscopic and multi-view (MV) displays. The work presented in [2] provides an approach to model multi-view displays in the frequency domain using test patterns with various density and orientation. However, the work assumes a MV display with a sub-pixel interleaving topology – something that LF displays do not have. The approach presented in [3] targets MV displays as well. This method is based on proprietary measurement equipment with Fourier optics, and due to the small size of the instrument, the applicability for large-scale (non-desktop) LF displays is limited. Moreover, it cannot be used for front-projected LF displays as the head would block the light path. The Information Display Measurement Standard [4] contains measurement methods for spatial and angular resolution (chapters 17.5.4 and 17.5.1 in [4]). The method described as angular resolution measurement relies on counting local maxima of a test pattern, but also assumes that the display can show two-view test patterns specifically targeting adjacent views, which is not directly applicable for LF displays. The method also assumes that the pixel size of the display is known in advance, which is not applicable for an LF display. The authors of [5] describe another proprietary measurement instrument, but the measurement assumes that the display is view-based, which does not apply for typical LF displays. We are not aware of a method that can measure the spatial and angular resolution of LF display with no discrete pixel structure.

In this work, we identity three parameters with direct influence on the perceptual quality of a LF display. *Spatial resolution* and *angular resolution*, both quantified in circles per degree (CPD) can be measured by off-the-shelf equipment. The minimum perceivable depth at the screen level, referred to as *perceptual depth resolution* is estimated by a subjective experiment. We specifically experiment with displays produced using HoloVizio technology [8] however the methodology can be easily adapted to other types of LF displays.

## 2. SPATIAL RESOLUTION

### 2.1. Background

The 2D spatial resolution of a LF display cannot be directly measured in terms of horizontal and vertical pixel count. This is due to the specific interaction between the discrete set of rays coming from projectors and the continuous LF reconstruction screen. For the proper generation of 3D effect, rays coming from different projectors might not form a regular structure. Correspondingly, the group of rays visible from a given direction do not appear as "pixels" on a rectangular grid [2].

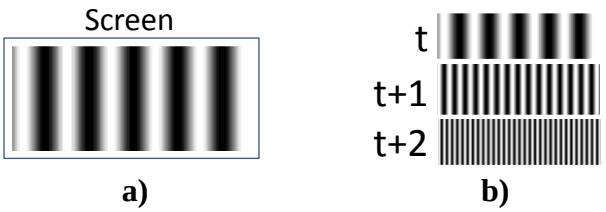



**Figure 2: a) Sinusoidal pattern for spatial resolution measurement; b) Sinusoidals of increasing frequency used for the spatial resolution measurement**

Therefore, in our approach, the spatial resolution is quantified through the display capability to produce fine details with a given spatial orientation. In practice, this means measuring the display's capability to reliably reproduce sinusoidal patterns in various directions.

One way to measure the 2D resolution would be to have an objective metric that analyses the contrast ratio of a sinusoidal pattern visualized on the screen. By measuring the contrast ratios of patterns with various density and orientations, one can find the threshold in each case, thereby determining the maximal resolution for the direction in question. However, rendering a sinusoidal pattern onto a non-rectangular grid produces imaging and aliasing distortions, which manifest themselves as Moiré patterns [7]. Therefore, measuring the contrast alone is not enough to assess how these distortions affect the image.

A more perceptually correct way to measure the 2D resolution would be to measure the so-called "pass-band" of the display [7]. In essence, pass-band measurement consists of a series of pass/fail tests, where each test analyses the distortions introduced by the display on a given test pattern. One starts with a test pattern with a given frequency and orientation, visualizes it on the screen, photographs it and analyses the output in the frequency domain. If the input (desired) frequency is still the dominant frequency on the output (distorted) image, the frequency of the pattern under test is considered to belong to the pass-band of the display. By repeating the pass/fail test for multiple test patterns, one can discover a large set of "passing" input frequencies. Union of all those frequencies is the pass-band of the display [7].

### 2.2. Experimental setup

Spatial resolution measurement involves showing sinusoidal black and white patterns of increasing frequency on the screen, taking photos of the resulting images and analyzing these photos in the frequency domain to determine the limits of visibility. The full-screen patterns projected on the screen show a sinusoidal change in intensity in one direction, but constant intensity in the orthogonal direction. The pattern used to measure horizontal resolution is shown in Figure 2a.

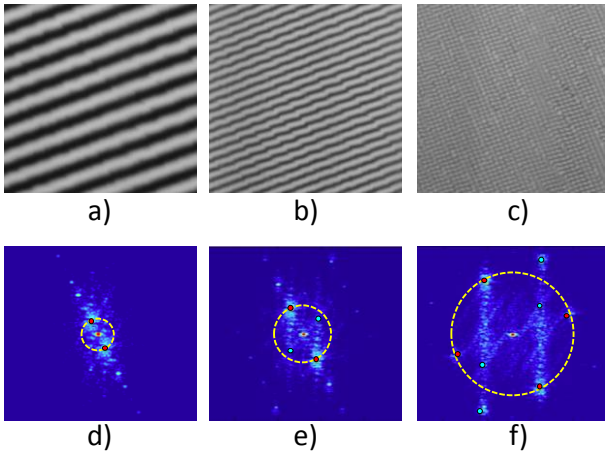



**Figure 3: Gratings with slant 22.5 degrees, tested for frequency dominance: a) grating with negligible distortions, b) grating with visible distortions, c) grating with ambiguous dominant frequency, d)-e) frequency analysis of the three gratings**

The frequency of the sinusoidal pattern is then increased and with every increment a new photo is taken (see Figure 2b). This procedure is then repeated with other measured directions (8 in our experiments), and a sequence of photos is taken starting from low frequency, with the same increments.

The measured directions of the test patterns are 0º, 22.5º, -22.5º, 45º, -45º, 67.5º, -67.5º and 90º. These provide a good estimation of the display's pass-band.

As the LF image is comprised of a set of light rays originating from various sources, sampling and visualizing this pattern is not trivial. For this purpose, we have developed a pattern generator software that enumerates all the light rays emitted by the display. By knowing where the light rays originate from and where they cross the screen's surface, the intensity of the specific light ray is determined, much like a procedural texture. The direction and frequency of the pattern can be changed interactively or automatically. In automatic operation mode the software is also able to control a DSLR camera attached to the computer controlling the display, which takes photos corresponding to each pattern. The camera in this experiment is on a static stand, and is positioned so that it is pointing to the center of the screen and is in line with the screen center both horizontally and vertically. Camera settings shutter speed is set up so that the shutter speed is longer than the double of the refresh rate used by the display. The resolution of the camera is order of magnitude higher than what is needed for sampling of the grating with the highest frequency. The ISO and aperture are set so that is set so that the white and black intensities do not over saturate the camera, but still exploit most of its dynamic range.

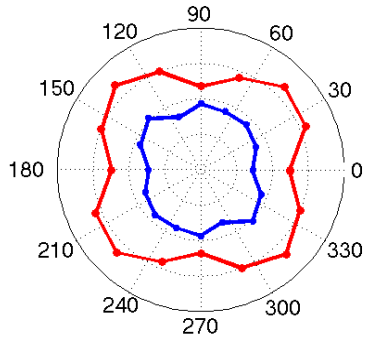



**Figure 4: Sample polar plot of resolution limits in different directions on the screen**

The resolution of the camera is set so that it oversamples the test pattern with the highest frequency. The camera is linearized as described in [15].

### 2.3. Analysis

Signals with measured frequencies lower than the frequency of the generated signal (i.e. aliased frequencies) are classified as distortions [7]. We consider distortion with amplitude 5% of the original (input) signal to be barely visible, and distortion with amplitude of 20% as unacceptable.

In other words, distortion with amplitude between 5% and 20% will produce visible distortions, but the original signal is still recognizable; distortion with amplitude greater than 20% results in perceptual loss (masking) of the original signal [7].

We start the analysis with cropping the acquired photos - we keep only the part depicting the visualized test pattern. Each cropped image is then windowed and a 2D FFT is executed on it. In the spectrum we can identify the peak that corresponds to the original frequency, and other peaks, which are created by display distortions. Next, we create the so-called unit circle around the point of origin, with a radius equal to the distance between the original peak and the point of origin. We search for peaks inside the unit circle. If the amplitude of such peaks is between 5% and 20% of the amplitude of the original one, we deem the patch to exhibit visible distortions. If the amplitude of the extra peaks is higher than 20% of the amplitude of the original one, then the dominant frequency is lost, and we deem the patch to have intolerable distortions.

An example of sinusoidal gratings under test is shown in Figure 3 a)-c) and frequency domain representations of these gratings can be seen in Figure 3 d)-f). The unit circle is plotted with yellow dashed line. The first grating exhibits minor distortions, and all peaks in the unit circle have negligible amplitude. The second grating has visible distortions, which appear as minor (5-20%) peaks inside the unit circle. In the third grating the dominant frequency is lost, and this can be detected by finding large peaks

inside of the unit circle. Such analysis is repeated for gratings with increasing density slanted in all preselected directions. The gathered data allows one to estimate the resolution of the display for details with different orientation in terms of cycles-per-degree (CPD). The resolution in a certain direction is estimated from the threshold for lines in the orthogonal direction – e.g. the horizontal resolution is estimated using a grating with vertical lines. Two sets of resolutions can be derived. One is what we call *distortion-free resolution*, i.e. the amount of cycles per degree the display can reproduce in a given direction, without introducing visible distortions. The other is *peak resolution*, which characterizes the maximum amount of CPD for which the introduced distortions do not mask the original signal. An example of these two resolutions derived for a LF display for different directions is given in Figure 4 – the blue line marks distortion-free resolution in CPD, while the red one marks the peak resolution. The data points in the figure indicate display resolution in a given direction.

## 3. ANGULAR RESOLUTION

### 3.1. Background

The angular resolution of a multi-view display can be directly derived from the geometrical properties of the display, that is, it is related to the number of views the display can generate throughout its Field Of View (FOV). It can be calculated by dividing the FOV of the display with the number of views, or alternatively, it can be measured, for example, by the approach proposed in [2].

In the case of LF displays, the concept of views is abandoned. Instead, the angular resolution is determined by the minimal angle of change that rays can reproduce with respect to a single point on the screen. In a simplified form, the minimal angle that an ideal display should reproduce can be estimated based on the properties of the human visual system, as [13]

$$\theta_{min}^{(ideal)} = \tan^{-1}\frac{d_p}{D} \qquad (1)$$

where $d_p$ is the pupil diameter, $D$ is distance to the screen (more precisely to the visualized point under consideration) and $\theta_{min}$ is the minimum angle between two rays originating from a single point that the eye can discriminate. Having smaller angular resolution limits the capabilities of the display, i.e. proper continuous-like parallax is limited to objects closer to the screen level. By knowing the angular resolution of an LF display, it is possible to prepare the content for the display (e.g. keeping the depth of the scene inside depth budget) that will result in higher visual quality.

As discussed earlier, due to different specifications provided by display manufacturers as well as the fact that rays might not originate from the same point on the surface of the screen, it is not straightforward to evaluate the angular resolution based on that data. We are not aware of a systematical approach to measure the angular resolution of an LF display. Therefore, in the following section we describe an experimental setup for measuring the angular resolution. Similar principle as in the case of the spatial resolution measurement described in Section 2 is applied. First, a set of signals is generated that change the pixel intensity with observation angle, and then, second, the highest amount of change is evaluated for which these changes can be reliably detected by an observer. These changes are then related to the angular resolution.

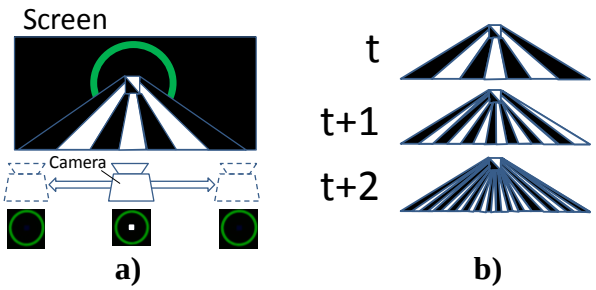



**Figure 5: a) Angular resolution test pattern, as seen by the camera from various angles. b) Angular resolution test patterns with increased frequency**

### 3.2. Experimental setup

Angular resolution measurement consists of showing a pattern that has a different appearance when viewed under different angles, that is, in this measurement, the direction selective light emission property of the 3D display is evaluated. We project a rectangular area that appears white from some viewing angles, and appears black from other angles. This black / white transition is repeated over the FOV of the display. A camera moving in front of the screen, pointing at the rectangle will see one or more white-to-black and black-to-white transitions during its travel (see Figure 5a). The duty cycle of black and white areas is always 50%, thus they appear to have the same size. Using a camera that moves parallel to the screen from one end of the display's FOV to the other end, a video is recorded. This video shows the intensity transitions of the measured spot as the screen is observed from various angles. The measurement spot is surrounded by a circle of solid green color that can be tracked in the recorded video. The angle between the observation direction where the measured spot is seen white and the direction it is seen black, is initially chosen to be large (e.g. two transitions over the whole FOV). Then the distance is gradually decreased, thus increasing the angular frequency of the pattern in subsequent iterations (see Figure 5b). Camera settings are as in the spatial resolution measurement.

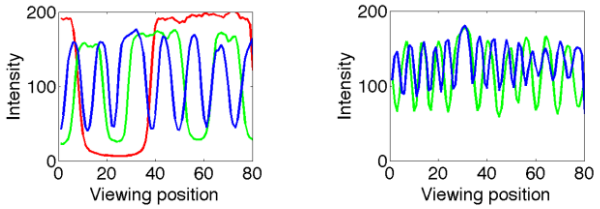



**Figure 6: a) Intensity of the measured spot with test patterns of increasing frequency ($f_{red} < f_{green} < f_{blue}$) b) Decreased dynamic range for higher frequencies**

### 3.3. Analysis

From the video recordings, frames are extracted. On each frame, the center of the green marker circle is located giving the position of the measured spot. The intensity of the measured spot changes from low to high and high to low in successive frames (change in observation angle is related to the camera movement). This is shown in Figure 6a. In this figure we can also see the increased rate of transitions with patters of increased frequency. As we proceed to higher frequency test patterns, the dynamic range between blacks and whites is decreasing (see Figure 6b and Figure7a). We define the 2D dynamic range of the display to be difference between the black and white levels when a 2D image is shown. In that case, the dynamic range of different angular frequencies can be expressed proportionally to the 2D dynamic range, as shown in Figure 7b. Finally, we find the threshold where the angular dynamic range drops below 20% of the 2D dynamic range. We define the angular resolution of the display $\theta_{min}$ to be equal to that threshold.

## 4. PERCEIVED DEPTH RESOLUTION

### 4.1. Background

As a result from the continuous head parallax an LF display can provide, the recreated scene appears visually as a 3D volume seen from various angles. Apart from planar (2D, or x-y) resolution, one might be interested also in the resolution in z-direction, i.e. what is the minimum depth difference that can be reliably reproduced by the display. The available parallax is characterized directly by the spatial resolution and the display FoV. The angular resolution naturally specifies how much one can move objects of size of one spatial element in front of the display before starting loosing spatial resolution. Thus, the minimum perceivable depth is a function of the spatial and the angular resolution, but cannot be fully characterized by those since there are subjective factors such as motion speed, memory and temporal masking along with other degradation factors, such as inter-perspective crosstalk and

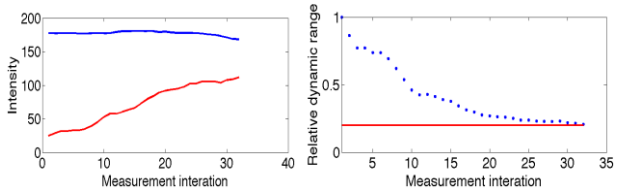



**Figure 7: a) Highest and lowest intensities as observed with successive angular resolution test patters (blue: highest intensity, green: lowest intensity) b) Relative dynamic range**

texture blur, which influence the ability of the observers to discriminate depth layers [6][9]. Besides, human vision is less sensitive to depth variations than to spatial variations of the same magnitude [12].

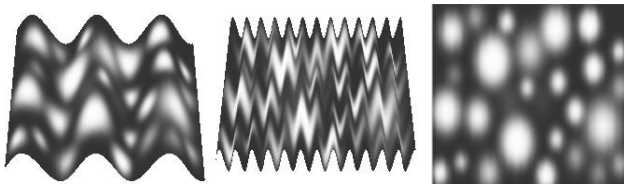
Usually, studies on stereoscopic perception use random dot stereograms to assess the thresholds in perceiving depth [10][12]. Experiments involving sinusoidal depth plane discrimination have been used to study disparity [10] or motion parallax [11] thresholds. In this work, we aim at finding the minimal step in z-direction that can be observed on a particular LF display, providing certain level of continuous parallax, by means of a direct subjective experiment.

### 4.2. Experimental setup

In this measurement we show a 3D object with a sinusoidal depth grating having a random dot texture. The grating is a surface with sinusoidal depth profile, as shown in Figure 8a and Figure 8b. The texture shows smooth circles of various sizes. It is projected orthogonally onto the surface, so that for an observer staying in front of the display the texture has no projective distortions and bears no pictorial depth cues (see Figure 8c). The grating is visualized as being parallel to the screen, and is scaled so that its borders appear outside of the screen area. A grating with zero amplitude appears as a flat surface parallel to the screen, while a grating with depth variations appears as sinusoidal surface changing alternatively towards and away from the observer. The only depth cues of the scene are the interocular and head parallax created by the LF display.

The experiment uses custom software that allows the density and the amplitude of the grating to be changed interactively. The experiment starts by showing a low-density sinusoidal grating with zero amplitude. The test subject is encouraged to move around and observe the grating from different perspectives, and is asked to distinguish whether the visualized grating is flat or grooved.

The amplitude of the grating is increased by 10% at a time, till the observer notices the depth changes of the grating. In should be noted, that the threshold sensitivity of



**a)** **b)** **c)**

**Figure 8: Sinusoidal depth gratings: a) low density grating, b) high density grating, c) orthogonal observation of the high density grating**

the human vision to sinusoidal gratings varies with grating density [12]. We have selected planar frequencies in the range between 0.5 and 4 CPD, where the human vision has constant sensitivity to depth variations.

### 4.3 Analysis

The experiment has been performed with three volunteers, and using three grating frequencies. The perceptual thresholds for depth variations were recorded in terms of absolute distance values, as provided to the rendering engine. After the experiments, these were converted to relative values, relative to the total depth range provided by the display. The results are shown in Figure 9. Apparently, even though high angular frequency is needed for the display to provide continuous head parallax, much lower depth resolution would be sufficient for acceptable 3D representation of 3D data.

## 5. CONCLUSIONS

We have presented methods for measuring three characteristic parameters of 3D LF displays – namely spatial, angular and perceived depth resolution. The methods are suitable for LF displays, which do not have discrete pixels. Spatial resolution measurement is fully automatic, while the angular resolution measurement requires moving the camera. The proposed methods provide valuable information about the visualization capabilities of a given LF display, which can be utilized in capturing, compressing and visualizing of LF data.

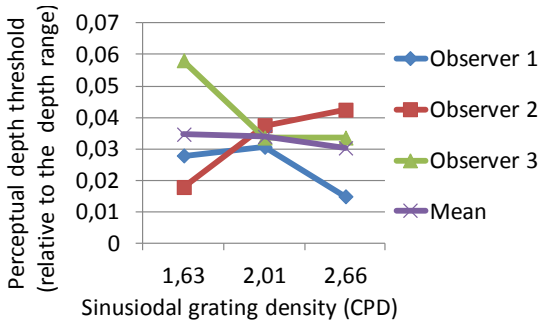



**Figure 9: Perceived depth resolution of the display**

## 6. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the PROLIGHT-IAPP Marie Curie Action of the People programme of the European Union's Seventh Framework Programme, REA grant agreement 32449.

## 7. REFERENCES

[1] 3D@Home Consortium and International 3D Society, "3D Display Technology Matrix- April 2012", http://www.3dathome.org/images/Tech_Matrix_120329.html, retrieved 23.10.2013

[2] A. Boev, R. Bregovic, and A. Gotchev, "Measuring and modeling per-element angular visibility in multi-view displays," *J. of the Society for Information Display*, 18: 686–697. 2010, doi: 10.1889/JSID18.9.686

[3] P. Boher, T. Leroux, T. Bignon, et al., "A new way to characterize autostereoscopic 3D displays using Fourier optics instrument," Proc. SPIE 7237, Stereoscopic Displays and Applications XX, 72370Z, February 17, 2009

[4] The International Display Measurement Standard v1.03, The Society for Information Display, 2012

[5] R. Rykowski, J. Lee, "Novel Technology for View Angle Performance Measurement," IMID 2008, Ilsan, Korea, 2008

[6] F. Kooi, A. Toet, "Visual comfort of binocular and 3D displays", *Displays* 25 (2-3) (2004) 99–108. ISSN:0141-9382, doi:10.1016/j.displa.2004.07.004

[7] A. Boev, R. Bregović, and A. Gotchev, "Visual-quality evaluation methodology for multiview displays," *Displays*, vol. 33, 2012, pp. 103-112, doi:10.1016/j.displa.2012.01.002

[8] T. Balogh, "The HoloVizio system," Proc. SPIE 6055, Stereoscopic Displays and Virtual Reality Systems XIII, 60550U (January 27, 2006); doi:10.1117/12.650907

[9] P.J.H. Seuntiëns, L.M.J. Meesters, W.A. IJsselsteijn, "Perceptual attributes of crosstalk in 3D images", Displays, Volume 26, Issues 4-5, October 2005, Pages 177-183, ISSN 0141-9382, 10.1016/j.displa.2005.06.005

[10] M. Nawrot, "Depth from motion parallax scales with eye movement gain" Journal of Vision December 18, 2003

[11] H. Ujike, H. Ono, "Depth thresholds of motion parallax as a function of head movement velocity", J. of Vision Res. 2001 Oct;41(22):2835-43.

[12] D. Kane, P. Guan, M. Banks (in press). "The limits of human stereopsis in space and time". Journal of Neuroscience. Manuscript submitted for publication.

[13] S. A. Benton and V. M. Bove Jr. *Holographic Imaging*, John Wiley and Sons, New Jersey, 2008.

[14] S. Pastoor, "3D displays", in (Schreer, Kauff, Sikora, edts.) *3D Video Communication*, Wiley, 2005.

[15] P. Debevec, J. Malik, "Recovering High Dynamic Range Radiance Maps from Photographs," in Proc. ACM SIGGRAPH, 1997

[16] A. Schmidt and A. Grasnick, "Multi-viewpoint autostereoscopic displays from 4D-vision", in Proc. SPIE Photonics West 2002: Electronic Imaging, vol. 4660, pp. 212-221, 20023D

# III

# MEASUREMENT OF PERCEIVED SPATIAL RESOLUTION IN 3D LIGHT FIELD DISPLAYS

by

P. T. Kovács, K. Lackner, A. Barsi, Á. Balázs, A. Boev, R. Bregović, A. Gotchev, 2014

in Proc. IEEE International Conference on Image Processing (ICIP) 2014

# MEASUREMENT OF PERCEIVED SPATIAL RESOLUTION IN 3D LIGHT-FIELD DISPLAYS

*Péter Tamás Kovács* [1, 2], *Kristóf Lackner* [1, 2], *Attila Barsi* [1], *Ákos Balázs* [1],
*Atanas Boev* [2], *Robert Bregović*[2], *Atanas Gotchev*[2]

[1]Holografika, Budapest, Hungary
[2]Department of Signal Processing, Tampere University of Technology, Tampere, Finland

## ABSTRACT

Effective spatial resolution of projection-based 3D light-field (LF) displays is an important quantity, which is informative about the capabilities of the display to recreate views in space and is important for content creation. We propose a subjective experiment to measure the spatial resolution of LF displays and compare it to our objective measurement technique. The subjective experiment determines the limit of visibility on the screen as perceived by viewers. The test involves subjects determining the direction of patterns that resemble tumbling E eye test charts. These results are checked against the LF display resolution determined by objective means. The objective measurement models the display as a signal-processing channel. It characterizes the display throughput in terms of passband, quantified by spatial resolution measurements in multiple directions. We also explore the effect of viewing angle and motion parallax on the spatial resolution.

***Index Terms***— 3D displays, light-field displays, spatial resolution, resolution measurement, subjective experiment

## 1. INTRODUCTION

3D light-field (LF) displays [1] are capable of providing 3D images with a continuous motion parallax over a wide viewing zone, and viewers can experience spatial vision inside this zone without wearing 3D glasses. Instead of showing separate 2D views of a 3D scene, they reconstruct the 3D light field describing a scene as a set of light rays. One way to implement such a display is using an array of projection modules emitting light rays and a custom holographic screen [2]. The light rays generated in the projection modules hit the holographic screen at different points and the holographic (reconstruction) screen as seen Figure 1, which makes the optical transformation that composes rays into a continuous 3D view. Each point of the holographic screen emits light rays of different color to the various directions. However, it is important to note that such screens do not have discrete pixels since the light rays can pass through the screen at arbitrary positions.

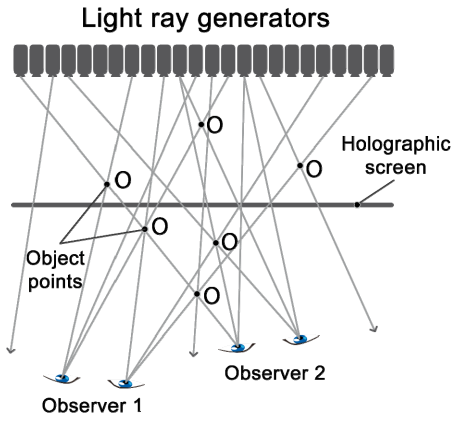



**Figure 1: Light-field display architecture**

When using properly designed LF displays, light rays leaving the screen spread in multiple directions, as if they were emitted from points of 3D objects at fixed spatial locations. This gives the illusion of points appearing either behind the screen, on the screen, or floating in front of it, achieving an effect similar to holograms.

For 2D displays, essential information such as spatial resolution and observation angle is standardized [3] and easily available to end users. For 3D displays in general, a common standard does not exist yet, and manufacturers rarely provide information about the display capabilities or provide 2D-related parameters. For example, the capabilities of a 3D display are given in terms of the underlying TFT matrix resolution and the number of unique views, which does not explicitly define what viewers can see [4].

## 2. RELATED WORK

Most previous work on 3D display characterization is only applicable for stereoscopic and multi-view (MV) displays. An approach to model multi-view displays in the frequency domain is presented in [5], where test patterns with various density and orientation are used. However, the inherent assumption about 3D displays having a sub-pixel interleaving topology does not apply in the case of LF displays. The method presented in [6] is based on proprietary measurement equipment with Fourier optics,

and is targeting MV displays. The Information Display Measurement Standard [3] provides a number of methods to measure spatial and angular resolution (in chapters 17.5.4 and 17.5.1 of [3]). However, the angular resolution measurement method relies on counting local maxima of a test pattern, and assumes that the display can show two-view (stereoscopic) test patterns, which is a very unnatural input in case of LF displays, not having discrete views. This method also assumes that the pixel size of the display is known, which, not having discrete pixels, is also not applicable for LF displays.

In this work, we present a subjective experiment to evaluate the spatial resolution of LF displays, which has a direct influence on the perceptual quality of a LF display. We analyze the effect of different viewing angles on the measured and perceived resolution, as well as the effect of motion on the perceived resolution.

We specifically experiment with displays produced based on the HoloVizio technology [2], however the methodology can be directly adapted to other LF displays [11][12][13].

## 3. OBJECTIVE MEASUREMENT OF SPATIAL RESOLUTION

The method we have used for measuring the objective resolution has been presented in our previous work [7], therefore it is not detailed here. To be able to show the correspondence between the objective and subjective approach, a short summary of the algorithm follows. What is new, is that we have executed the resolution measurement from different viewing angles relative to the screen.

The 2D spatial resolution of a LF display cannot be directly measured in terms of horizontal and vertical pixel count. This is due to the specific interaction between the discrete set of rays coming from the projectors and the continuous LF reconstruction screen. In order to recreate a valid 3D effect, rays coming from different projectors might not form a regular structure. Correspondingly, the group of rays visible from a given direction does not appear as pixels on a rectangular grid [5].

Therefore we chose to quantify the display's capability to produce fine details in a given spatial direction. This is achieved by measuring the so-called "pass-band" of the display [10]. Pass-band measurement consists of a series of pass/fail tests, where each test analyses the distortions introduced by the display on a given test pattern. It starts with a test pattern showing sinusoidal black and white patterns (see Figure 2a) on the screen with a given frequency and orientation. The image on the screen is photographed and analyzed in the frequency domain. If the input (desired) frequency is still the dominant frequency on the output (distorted) image, the frequency of the current test pattern is considered to belong to the pass-band of the display.

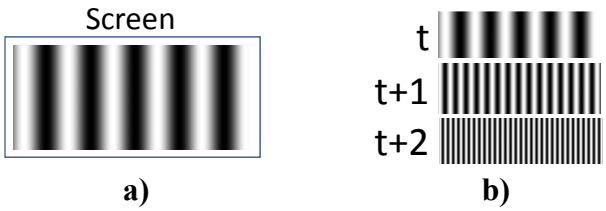



**Figure 2: a) Sinusoidal pattern for spatial resolution measurement; b) Sinusoidals of increasing frequency used for the spatial resolution measurement**

By repeating the pass/fail test for multiple test patterns with different frequency (see Figure 2b) and directions (0º, 22.5º, -22.5º, 45º, -45º, 67.5º, -67.5º and 90º in our experiments), one can find the range of input frequencies that can pass through the system. Signals with measured frequencies lower than the frequency of the generated signal (i.e. aliased frequencies) are classified as distortions [10]. We consider distortion with amplitude of 5% of the original (input) signal to be barely visible, and distortion with amplitude of 20% as unacceptable. Based on these thresholds, two sets of resolution values can be derived. One is what we call *distortion-free resolution*, i.e. the amount of cycles per degree the display can reproduce in a given direction, without introducing visible distortions (5%). The other is *peak resolution*, which characterizes the maximum resolution for which the introduced distortions do not mask the original signal (20%).

As the LF image is comprised of a set of light rays originating from various sources, sampling and visualizing this pattern is not trivial. For this purpose, and for visualizing the test patterns of the subjective experiment, we have developed pattern generator software that enumerates all the light rays emitted by the display. By knowing where the light rays originate from and where they cross the screen's surface, the intensity of the specific light ray is determined, much like a procedural texture.

This measurement has been performed on the same LF display from a central viewpoint, as well as from the edge of the Field of View (FOV).

## 4. SUBJECTIVE TEST OF PERCEIVED SPATIAL RESOLUTION

To quantify what resolution human viewers are able to see, we have developed a test that involves subjects distinguishing small details on the screen, and iteratively finding the detail size they cannot properly see anymore. We aimed at constructing a test that subjects are familiar and feel comfortable with. The perceived spatial resolution has been measured using a test similar to the tumbling E eye test charts [8]. In our case, symbols of different size have been shown on the screen of a LF display, and viewers have been asked to record the orientation of the symbols.

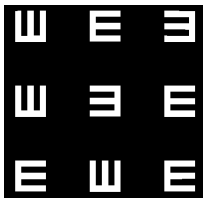


**Figure 3: 3x3 E signs with randomized direction**

In each iteration a viewer can see nine equally sized symbols on the screen, arranged in a 3x3 layout, the orientation of each symbol randomly chosen from the four possible orientations (left, right, up, down), as shown on Figure 3. The test software records the rendered symbols along with the symbol size, timestamp and the ID of the test subject. The subject is asked to record the nine orientations on paper, that is, draw the symbols into a 3x3 grid with the same orientation he/she can see.

In the next iteration, the size of the symbols is decreased 1.3 times, and the next set of randomized symbols is presented. The viewer records the orientation of the new set of symbols. The rendered symbol orientations and the perceived orientations recorded by test subjects are then compared to see where the orientations cannot be seen by the test subjects. When ophthalmologists perform eye tests, a line of a specific symbol size is considered read when more than half of the characters are read correctly, thus we use the same criteria [9].

The application used to render the symbols on the LF display is based on the same technique we used to render the spatial resolution test patters in the sense that the color of the emitted light rays is determined procedurally, that is, a GPU shader is executed for each light ray, which, based on the ray parameters and the pattern to be rendered, calculates if the specific light ray should be white or black. This is in contrast with rendering approaches that start with a flat texture depicting the intended test pattern, and generate the light rays by applying a set of transformation and filtering steps. This rendering method ensures that the test patterns are rendered with the highest possible fidelity with no degradations caused by the rendering process.

The subjective tests were conducted with nine subjects, sitting 5m away from a 140" diagonal LF screen. The room has been darkened so that external light reflected from the screen does not affect the perception of patterns. The analysis of the results show that on average, subjects started to introduce recognition errors in iteration 6, and have fallen below the 50% threshold in iteration 8, close to the level of random guess (25%), as shown on Figure 4.

## 5. COMPARISON OF OBJECTIVE AND SUBJECTIVE RESULTS

The results of the subjective tests show that the level where subjects started to introduce recognition errors roughly corresponds to 150% of the peak resolution as determined

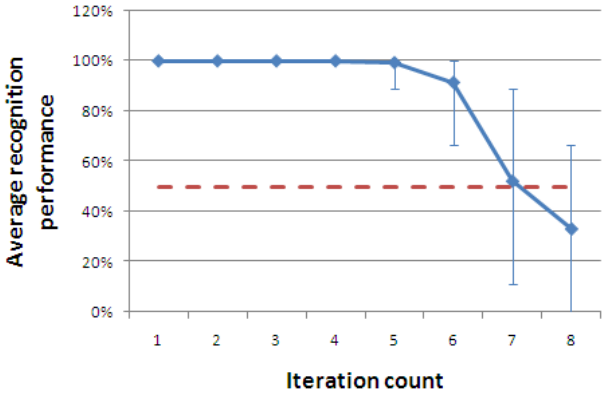



**Figure 4: Average recognition performance. Dashed line marks 50% threshold.**

by the objective measurement, even though the average performance at this level is still 92%.
Interestingly, even in the next iteration (1.3x smaller size) subjects performed slightly higher than the 50% recognition threshold. The reason for the higher perceived resolution might be that the human vision system is very good at determining shapes even when they suffer from distortions.

We have found that the objective measurement is strict in the sense that when the 20% distortion level is reached, the original patterns are still visible at some areas and some viewing directions on the screen, although heavily distorted or even invisible in other areas.

The correspondence between the resolution represented by the sinusoidal pattern and the E symbols is determined in the following way: the feature size in case of the tumbling E test pattern is the thickness of one line segment in the E, while the feature size of the sinusoidal is a half period.

## 6. VIEWING ANGLE DEPENDENCE

In all displays, the perceived resolution when watched from the center or from other angles is different. Due to the way the emitted rays are typically distributed in a LF display we expected that the resolution will be slightly lower at the sides of the FOV.

To check the viewing angle dependence of the measured and perceived resolution of the display, we have performed both the objective and subjective resolution measurements from the center of the viewing area, and the side of the viewing area.

The results of the objective resolution test show that the horizontal resolution is slightly lower when perceived from the edge of the FOV, see Figure 5. The results of the subjective tests also show that the performance of subjects in recognizing the correct orientation of symbols is slightly lower when they were positioned on the side of the viewing zone. The decrease of accuracy starting at iteration 6 is steeper in this case. Moreover, some subjects made a mistake with relatively large symbol sizes, which did not occur when they were positioned in the center.

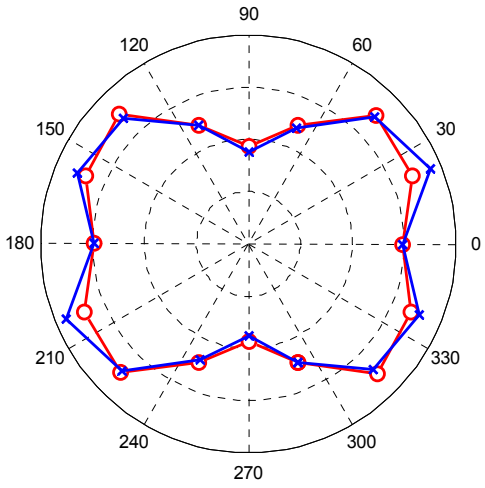



**Figure 5: Difference between measured resolution from the center (blue / x) and the side (red / o) of the FOV. The plot shows the measured resolution in test patterns of different directions.**

With the smallest symbol size we can see subjects performed slightly better from the side view, but we consider this irrelevant, as both results (31% and 33%) are close to random guess (25%), and are only shown here for completeness.

## 7. MOTION PARALLAX DEPENDENCE

During the subjective tests we have realized that subjects, although sitting, have been moving their heads excessively, especially when observing very small symbols. When asked, they confirmed that head parallax helps them to see the correct orientation of small symbols on the screen. In order to check the importance of motion parallax on the perceived resolution, we repeated the experiment with no head movements allowed. Comparing the results of the eye chart test with and without head movements shows that looking at the same display from multiple directions let subjects see finer details, thus increasing the perceived spatial resolution. While subjects performed slightly above the 50% threshold in iteration 7 in the normal test, their performance dropped below 50% in the last two iterations when no head movements were allowed.

The reason for this effect is rooted in the non-uniform pixel structure of LF displays, that is, the light rays visible from one viewing angle may sample the 3D scene at slightly different locations than from different angles. That is, when viewers are moving their heads, they are looking for the positions where the direction of the symbol can be seen.

We should note that the measured resolution as seen by a still camera can be just as high as the perceived resolution of a viewer with no head movements allowed.

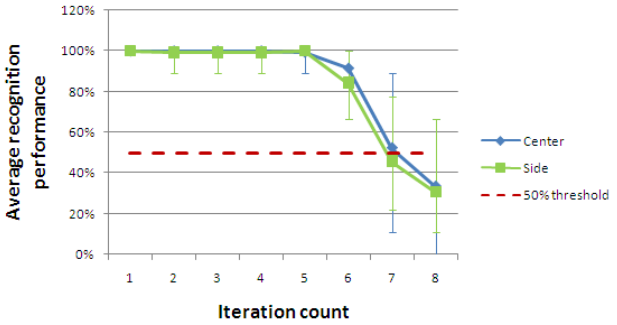



**Figure 6: Average recognition performance from the center, and from the side of the LF display's FOV**

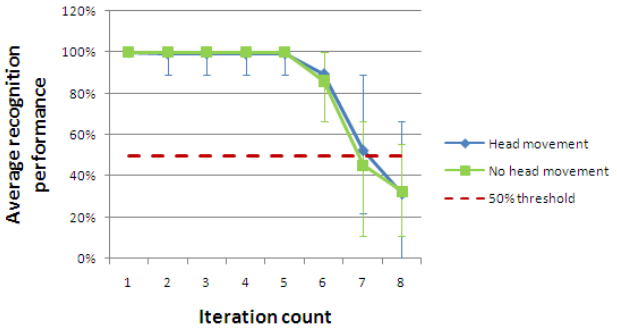



**Figure 7: Average recognition performance with and without head movements**

## 8. CONCLUSIONS AND FUTURE WORK

Both an objective measurement method and a subjective test have been presented for measuring the spatial resolution of LF displays, which can be applied to any LF 3D display. The results of the measurement and subjective tests have been compared, and a difference has been found. The dependence of resolution on viewing angle has been checked and confirmed. It has also been shown that an observer could see finer details when head movements were allowed compared to the case when the head position was fixed on a LF 3D display, and commented on the possible causes. These results highlight some of the many differences between 2D displays and 3D LF displays. These differences have consequences on content creation, processing, compression and rendering for LF displays.

Future work will aim at the development of an objective measurement or estimation method for determining the perceived resolution of moving viewers.

## 9. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the PROLIGHT-IAPP Marie Curie Action of the People programme of the European Union's Seventh Framework Programme, REA grant agreement 32449.
The authors would like to thank the active participation of test subjects.

## 10. REFERENCES


[1] 3D@Home Consortium and International 3D Society, “3D Display Technology Matrix- April 2012”, http://www.3dathome.org/images/Tech_Matrix_120329.html, retrieved 28.01.2014

[2] T. Balogh, "The HoloVizio system," Proc. SPIE 6055, *Stereoscopic Displays and Virtual Reality Systems XIII*, 60550U (January 27, 2006). doi:10.1117/12.650907

[3] The International Display Measurement Standard v1.03, The Society for Information Display, 2012

[4] A. Schmidt and A. Grasnick, "Multi-viewpoint autostereoscopic displays from 4D-vision", in *Proc. SPIE Photonics West 2002: Electronic Imaging*, vol. 4660, pp. 212-221, 20023D

[5] A. Boev, R. Bregovic, and A. Gotchev, “Measuring and modeling per-element angular visibility in multi-view displays,” *J. of the Society for Information Display*, 18: 686–697. 2010, doi: 10.1889/JSID18.9.686

[6] P. Boher, T. Leroux, T. Bignon, et al., “A new way to characterize autostereoscopic 3D displays using Fourier optics instrument,” *Proc. SPIE 7237, Stereoscopic Displays and Applications XX*, 72370Z, February 17, 2009

[7] P. T. Kovács, A. Boev, R. Bregović, A. Gotchev, “Quality Measurements of 3D Light-Field Displays”, in *Proc. Eighth International Workshop on Video Processing and Quality Metrics for Consumer Electronics* (VPQM 2014), January 31, 2014, Chandler, Arizona, USA

[8] B. K. Samar, *Ophthalmology Oral and Practical (3rd edition)*, E;sevier, 2009, ISBN 81-86793-66-6

[9] International Society for Low vision Research and Rehabilitation, “Guide for the Evaluation of Visual Impairment”, International Low Vision Conference (VISION-99)

[10] A. Boev, R. Bregović, and A. Gotchev, “Visual-quality evaluation methodology for multiview displays,” *Displays*, vol. 33, 2012, pp. 103-112, doi:10.1016/j.displa.2012.01.002

[11] J.-H. Lee, J. Park, D. Nam, S. Y. Choi, D.-S. Park, C. Y Kim, “Optimal Projector Configuration Design for 300-Mpixel Light-Field 3D Display”, *SID Symposium Digest of Technical Papers*, 44: 400–403. doi: 10.1002/j.2168-0159.2013.tb06231.x

[12] G. Wetzstein, D. Lanman, M. Hirsch, R. Raskar, “Tensor Displays: Compressive Light Field Synthesis using Multilayer Displays with Directional Backlighting”, In *Proc. SIGGRAPH* 2012

[13] D. Lanman, D. Luebke, "Near-Eye Light Field Displays", In *ACM Transactions on Graphics (TOG)*, Volume 32 Issue 6, November 2013 (Proceedings of SIGGRAPH Asia), November 2013

# IV

# OPTIMIZATION OF LIGHT FIELD DISPLAY-CAMERA CONFIGURATION BASED ON DISPLAY PROPERTIES IN SPECTRAL DOMAIN

by

R. Bregović, P. T. Kovács, A. Gotchev, 2016

Optics Express, vol. 24, no. 3, pp. 3067-3088

R. Bregović, P. T. Kovács, and A. Gotchev, "Optimization of light field display-camera configuration based on display properties in spectral domain," *Opt. Express*, vol. 24, no. 3, pp. 3067-3088, Feb. 2016.

# Optimization of light field display-camera configuration based on display properties in spectral domain

**Robert Bregović,[1,*] Péter Tamás Kovács,[1,2] and Atanas Gotchev[1]**

[1]*Department of Signal Processing, Tampere University of Technology, P.O. Box 553, FIN-33101 Tampere, Finland*
[2]*Holografika, P.O. Box 100, H-1704 Budapest, Hungary*
[*]*robert.bregovic@tut.fi*

**Abstract:** The visualization capability of a light field display is uniquely determined by its angular and spatial resolution referred to as display passband. In this paper we use a multidimensional sampling model for describing the display-camera channel. Based on the model, for a given display passband, we propose a methodology for determining the optimal distribution of ray generators in a projection-based light field display. We also discuss the required camera setup that can provide data with the necessary amount of details for such display that maximizes the visual quality and minimizes the amount of data.



**OCIS codes:** (100.6890) Three-dimensional image processing; (120.2040) Displays; (110.6880) Three-dimensional image acquisition; (100.2000) Digital image processing; (100.0100) Image processing.

## References and links

1. A. Boev, R. Bregović, and A. Gotchev, "Signal processing for stereoscopic and multi-view 3D displays," in *Handbook of Signal Processing Systems, 2nd edition*, S. Bhattacharyya, E. Deprettere, R. Leupers, and J. Takala, eds. (Springer, 2013).
2. W.A. IJsselsteijn, P.J.H. Seuntiëns, and L.M.J. Meesters, "Human factors of 3D displays," in *3D Video Communication*, O. Schreer, P. Kauff, and T. Sikora, eds. (Wiley, 2005).
3. T. Balogh, "The HoloVizio system," Proc. SPIE 6055, 12 pages (2006).
4. J.H. Lee, J. Park, D. Nam, S.Y. Choi, D.S. Park, and C.Y. Kim, "Optimal projector configuration design for 300-Mpixels multi-projection 3D display," Opt. Express **21**(22), 26820–26835 (2013).
5. E. Adelson and J. Bergen, "The plenoptic function and the elements of early vision," in *Computational Models of Visual Processing*, M. Landy and J.A. Movshon, eds. (MIT, 1991).
6. A. Stern, Y. Yitzhaky, and B. Javidi, "Perceivable light fields: Matching the requirements between the human visual system and autostereoscopic 3-D displays," Proceedings of the IEEE **102**(10), 1571-1587 (2014).
7. S.A. Benton and V.M. Bove, *Holographic Imaging*, (Willey, 2008).
8. R. Bregović, P.T. Kovács, T. Balogh, and A. Gotchev, "Display-specific light-field analysis," Proc. SPIE 9117, 15 pages (2014).
9. S.J. Gortler, R. Grzeszczuk, R. Szeliski, and M.F. Cohen, "The lumigraph," SIGGRAPH (Computer Graphics), 43-54 (1996).
10. M. Levoy and P. Hanrahan, "Light field rendering." SIGGRAPH (Computer Graphics), 31-42 (1996).
11. C.K. Liang, Y.C. Shih, and H.H. Chen, "Light field analysis for modeling image formation," IEEE Trans. Image Processing **20**(2), 446-460 (2011).
12. C. Zhang and T. Chen, "Spectral analysis for sampling image-based rendering data," IEEE Trans Circuits and Systems for Video Technology **13**(11), 1038-1050 (2003).
13. E. Dubois, "The sampling and reconstruction of time-varying imagery with application in video systems," Proc. IEEE 73, 502–522 (1985).
14. E. Dubois, "Video sampling and interpolation," in *The Essential Guide to Video Processing*, J. Bovik, ed. (Academic Press, 2009).
15. P.Q. Nguyen and D. Stehlé, "Low-dimensional lattice basis reduction revisited," ACM Trans. Algorithms **5**(4), 46 pages (2009).
16. F. Aurenhammer, "Voronoi diagrams – A survey of a fundamental geometric data structure," ACM Computing Surveys **23**, 245-405 (1991).

17. E.B. Tadmor and R.E. Miller, *Modeling Materials: Continuum, Atomistic and Multiscale Techniques*, (Cambridge University, 2011).
18. X. Cao, Z. Geng, and T. Li, "Dictionary-based light field acquisition using sparse camera array," Opt. Express **22**(20), 24081–24095 (2014).
19. M. Zwicker, W. Matusik, F. Durand, and H. Pfister, "Antialiasing for automultiscopic 3D displays", Proc. of Eurographics Symposium on Rendering, 10 pages (2006).
20. A. Boev, R. Bregović, and A. Gotchev, "Methodology for design of antialiasing filters for autostereoscopic displays," IET Signal Processing **5**, 333–343 (2011).
21. N. Holliman, N. Dodgson, G. Favalora, and L. Pockett, "Three-dimensional displays: A review and application analysis," IEEE Trans. Broadcasting, **57**(2), 362-371 (2011).
22. Y. Takaki and N. Nago,"Multi-projection of lenticular displays to construct a 256-view super multi-view display," Opt. Express **18**(9), 8824-8835 (2010).
23. Y. Takaki, Y. Urano, S. Kashiwada, H. Ando, and K. Nakamura "Super multi-view windshield display for long-distance image information presentation," Opt. Express **19**(2), 704-716 (2011).
24. G. Wetzstein, D. Lanman, M. Hirsch, and R. Raskar, "Tensor displays: Compressive light field synthesis using multilayer displays with directional backlighting," ACM Trans. Graph. **31**(4), 11 pages (2012).
25. D. Kane, P. Guan, and M. Banks, "The limits of human stereopsis in space and time", Journal of Neuroscience **34**(4), 1397-1408 (2014).

## 1. Introduction

Most of the commercially available, stereoscopic as well as autostereoscopic, 3D displays concentrate on reproducing the binocular visual cue for single or multiple observers thereby giving the illusion of 3D [1]. However, a typical consumer display is not capable or has difficulty in reproducing other cues important for 3D vision, most notable one being the continuous head parallax [2]. There are two practical ways for achieving the illusion of continuous head parallax. First way is by performing user's eye tracking and rendering parallax-correct views depending on user's position. This can be achieved by either using a head mounted display (e.g. Oculus Rift, Samsung Gear VR, Zeiss VR) or a custom built display with eye-tracking capabilities (e.g. zSpace). Disadvantage of those lies in the fact that, typically, only one user is supported. Second way of achieving a reasonably convincing continuous parallax is by using so called light field (LF) displays [1,3,4].

A LF display strives to reproduce the underlying plenoptic function describing the scene that is visualized [5]. It can be observed by multiple users simultaneously without a need of user tracking or glasses. In order to support continuous parallax, a large and dense set of light rays have to be generated to reconstruct the underlying LF function. In today's LF displays this is achieved by using projection-based systems [3,4]. There are two major drawbacks of such LF displays. First, only a finite number of light rays can be generated in practice. Based on the properties of the human visual system (HVS) it is possible to estimate the optimal (required) number of rays needed to achieve a level of detail that is sufficient for a human observer [6,7]. Unfortunately, achieving that level of detail is impractical with today's technology. Second, due to the multiple sources of rays, it is very difficult to achieve the desired uniform density of rays (position wise as well as intensity wise) on the screen surface [4]. Both drawbacks reduce the perceived resolution of the display. Therefore it is important to optimize the display setup and properly preprocess data sent to the display in order to mitigate the aforementioned two drawbacks as much as possible.

We have shown earlier [8] that by performing a frequency domain analysis of a typical LF display it is possible to determine the throughput of the display in terms of its spatial and angular resolution. This enables one to calculate the optimal amount of data that has to be captured and sent to the display to maximally utilize its visual capability. Moreover, it gives a user a good idea what to expect from the display in terms of visual quality. In this paper, we build on some of the ideas presented in [8] in order to achieve a deeper understanding of the relations of various hardware and software parts building a LF display. We present an analysis assuming a desired ray-sampling pattern at the screen plane that will define display specifications. For such display, we estimate the throughput of the display in terms of its

angular-spatial bandwidth. Having the display specifications, we develop a methodology for determining the optimal distribution of ray generators that will result in the desired display properties as well as a camera setup that can provide data with required amount of details. This is achieved by developing an optimization / estimation method for determining the required display / camera parameters.

Outline of this paper is as follows: Section 2 introduces the LF concepts and notations followed by the description of the principle of operation and properties in spatial and frequency domain of projection based LF displays. The proposed display-camera system optimization is introduced in Section 3, with several examples given in Section 4. Finally, concluding remarks are given in Section 5.

## 2. Light field displays

### *2.1. Light field basics*

In the most general case, by using ray-optics assumptions, the propagation of light in space can be described by a 7D continuous plenoptic function $R(\theta,\varphi,\lambda,\tau,A_x,A_y,A_z)$, where $(A_x,A_y,A_z)$ is a location in the 3D space, $(\theta,\varphi,)$ are directions (angles) of observation, $\lambda$ is wavelength, and $\tau$ is time [5]. For practical reasons, the continuous plenoptic function is typically simplified to its 4D version, which describes the static and monochromatic light ray propagation in half space. This 4D approximation of the plenoptic function is referred to as LF [10]. In this approximation, the LF ray positions are indexed either by their Cartesian coordinates on two parallel planes, the so-called two-plane parameterization $L(x,y,s,t)$ or by their one plane and direction coordinates $L(x,y,\varphi,\theta)$ [9,10].

In this paper, without loss of generality and in line with today's display technology, we concentrate on the so-called horizontal parallax only (HPO) case, ignoring the vertical parallax and subsequently dropping variables $t$ or $\theta$ in the aforementioned parameterization. Furthermore, we assume that the relation between planes parameterized by $(x,s)$ and $(x,\varphi)$ , is given by $s = \tan\varphi$ with $x$ being the same in both representations. In this parameterization, the origin of the $s$ axis is relative to the given $x$ coordinate.

The position of two parallel planes $x$ and $s$ can be chosen depending on the application. Two such positions, where the distances between parameterizing planes are taken as unit, are given in Fig 1. According to the figure, the propagation of light rays through space can be mathematically expressed as [8,11]

$$L_2\left(\begin{bmatrix} x_2 \\ s_2 \end{bmatrix}\right) = L_1\left(\begin{bmatrix} x_1 \\ s_1 \end{bmatrix}\right) = L_1\left(\begin{bmatrix} 1 & -d \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_2 \\ s_2 \end{bmatrix}\right) \tag{1}$$

$$L_2\left(\begin{bmatrix} x_2 \\ \varphi_2 \end{bmatrix}\right) = L_1\left(\begin{bmatrix} x_1 \\ \varphi_1 \end{bmatrix}\right) = L_1\left(\begin{bmatrix} x_2 - d\tan\varphi_2 \\ \varphi_2 \end{bmatrix}\right) \tag{2}$$

with $L_1$ and $L_2$ referring to LFs on plane position 1 and plane position 2, respectively, and $d$ being the distance between the plane positions along the $z$ axis. As can be seen from Eq. (2), when considering propagation of light rays in plane and direction representation, the relation between parameters on both planes is not strictly linear. However, for small angles, this nonlinearity can be ignored. More detailed evaluation on light ray propagation can be found in [8,11].

The continuous LF function has to be sampled in a way, which allows its reconstruction from samples. The plenoptic sampling theory, that considers the LF as a multidimensional bandlimited function has been developed in [12]. In general, it states that LF frequency support depends on the min and max depth of the visual scene, and sampling along $x$ and $s$ creates the usual replication of the baseband, which should be taken into account when designing the end-to-end LF camera to display system. While the sampling physically occurs

at the LF acquisition (sensing) stage, it is the LF display, which recreates the LF originating from a visual 3D scene. In the sampling theory formalism, an LF display can be considered as a discrete-to-continuous (D/C) converter that converts a sampled LF into its continuous version, thereby achieving a continuous visualization of a 3D scene, with continuous parallax being part of it. Consequently, we can consider a LF display as a multidimensional sampling system and as such apply multidimensional sampling theory when analyzing LF displays.

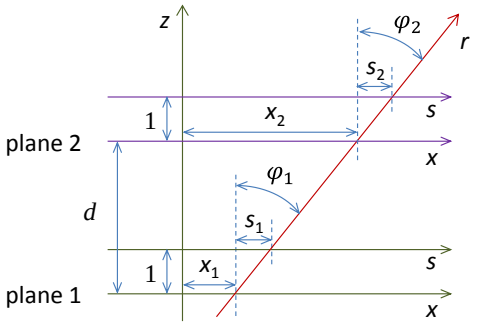



Fig. 1. Light ray ($r$) propagation through space (representation on two planes).

### *2.2. Light field display as sampling-reconstruction system*

In our general model, we consider the LF display being composed by a set of ray generators and a continuous LF reconstruction optical module. The ray generators act as discrete sources of light rays and the module is the D/C converter that converts the set of samples (rays) into its continuous representation that is observed by a viewer. While different display settings can fall into this general model, we specifically concentrate on a LF display consisting of a set of projection engines and a special screen, dubbed as holographic screen as illustrated in Fig. 2(a). Each light ray generated by a ray generator, hits this screen from a different angle at a different position, and the screen converts (diffuses) each light ray into an angular beam around the main direction of the ray. The span of the beam after diffusion is anisotropic with narrow horizontal angle $\delta_x$ and wide vertical angle $\delta_y$, as illustrated in Fig. 2(b) [3]. The screen does not have an explicit pixel structure. A finite area on it emits different light rays to different directions. The properties of such screen are described in more detail in [3]. In this paper we will assume that the screen is a perfect D/C converter. In practice it introduces some low-frequency selectivity that additionally smooths the reconstructed LF. However, this can be ignored for the purposes of our work.

From the observer viewpoint, a point (object) in space is reconstructed by the interaction of rays originating from different sources (i.e. coming from different directions). This is illustrated in Fig. 2(a) for two observers and several points in space. Each ray can be traced from its origin (ray generator) to the screen surface and it is uniquely described by its starting position and angle or its starting position and the place it hits the screen surface. This is reminiscent to the two-plane LF parameterization discussed in the previous section.

The overall throughput of the display is directly related to the number of light rays the display can generate. Denser set of rays produces finer spatial and angular details. Technology limitations prevent us from achieving the resolution power of the HVS [6, 7]. Therefore, it is important to take these limits into account when building the display and/or processing the visual data to be represented on it. Frequency domain analysis of the sampled and reconstructed light field is the proper tool for doing this.

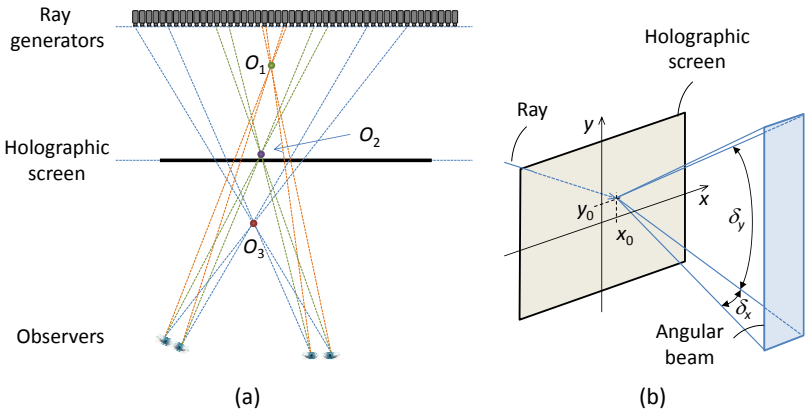



Fig. 2. Projection based light field display – principle of operation. (a) Different ray-generators participate in forming a point in space (O) depending on the place of the point and the position of the observer. (b) The holographic screen acts as diffusor with a narrow spread in horizontal and a wide spread in vertical direction.

### *2.3. Spatial and frequency domain analysis of light field displays*

A typical LF display under consideration is illustrated in Fig. 3. It consists of $N_p$ projection engines uniformly distributed on the ray generators (RG) plane ($p$ - plane) over distance $d_p$ thereby making the distance between engines $x_p = d_p / (N_p - 1)$. Each projection engine generates $N_x$ rays over its field of view $FOV_p$. We assume that the rays hit a certain plane (screen plane, $s$ - plane) parallel to the RG plane at equidistant points. As a consequence, the angular distribution of rays inside the FOV is not uniform. Nevertheless, for small angles we can assume that this is uniform and approximate the angular resolution at the RG plane as $\alpha_p = FOV_p / N_x$.

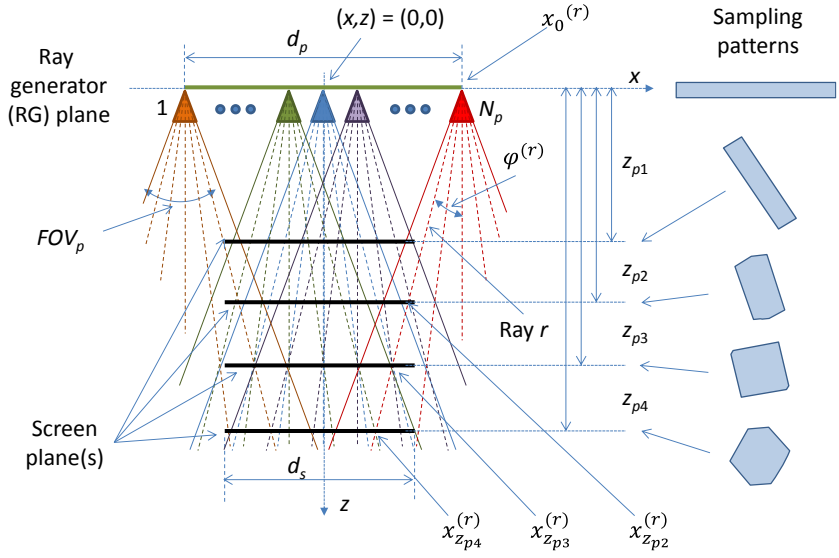



Fig. 3. Ray propagation in a light field display – different sampling patterns are illustrated for different positions of the screen plane.

The 'trajectory' of a ray can be uniquely defined by its origin $x_0^{(r)}$ at the RG plane and its direction determined by angle $\varphi^{(r)}$. The position of the ray at a distance $z$ from the display is given as

$$\begin{bmatrix} x_z^{(r)} \\ \varphi_z^{(r)} \end{bmatrix} = \begin{bmatrix} x_0^{(r)} + z\tan\left(\varphi^{(r)}\right) \\ \varphi^{(r)} \end{bmatrix} \tag{3}$$

which is according to the propagation of rays in $(x,\varphi)$ LF parameterization – see Eq. (2).

The screen of the display is where rays recombine to reconstruct the desired continuous LF function to be observed by a viewer. In Fig. 3, several positions for the screen are illustrated with thick black lines. As seen in the figure, the ray ($r$) crosses those 'screens' at different horizontal positions ( $x_{z_{p2}}^{(r)}, x_{z_{p3}}^{(r)}, x_{z_{p4}}^{(r)}$ ) and, due to a finite width of the screen $d_s$, it even does not contribute to the screen at distance $z_{p1}$. In practice this means that the ray would contribute to a different part of the screen depending on the screen position. Moreover, at different screen positions, it intersects with different rays originating from different ray generators, that is, depending on the screen position, a different combination of rays will be responsible for forming a multiview pixel at that position. As a consequence, the uniform distribution of rays we had at the RG plane is lost.

Rays are indexed (parameterized) by their spatial position and direction $(x,\varphi)$ and thus represented as samples in the corresponding ray space. This parameterization has been selected among several possibilities (e.g. $\varphi$ vs. $x$, $\tan\varphi$ vs. $x$, $z\tan\varphi$ vs. $x$) since both ray-space axes can be allocated with measurable (quantifiable) units (position can be expressed in mm and angle in degrees) that are easy to understand by a user. Consequently, at the screen plane, the display can be quantified by its spatial resolution (e.g. number of pixels per mm or per screen size) and its angular resolution (e.g. number of rays per degree or FOV of the display $FOV_{disp}$).

For the need of frequency analysis, each ray is considered as a sample, positioned in the 2D ray-space plane for fixed $z$ (in the case of full parallax, this turns into a 4D plane). Since the position of the ray is changing along $z$, as given by Eq. (3), for a set of ray generators, different sampling patterns are obtained at different distances from the screen. This is illustrated by means of an example in Fig. 4 (see also Fig. 3). The figures on the top row for $z = 0, z_{p2}, z_{p4}$ show how the whole LF that the display is capable of generating is sheared along the $x$-axis as the screen plane moves away from the RG plane. For better visualization, one set of rays is marked in blue. The figures in bottom row show zoomed in versions of the LF at different distances from the RG plane. One can observe that for every distance, the sampling pattern is regular although not rectangular. The fact that the sampling patterns are regular, enables us to utilize the multi-dimensional sampling theory [13,14].

Any regular 2D pattern can be uniquely described through a notion of sampling lattice $\Lambda$. The elements of the lattice are calculated as a linear combination of two linearly independent vectors

$$\Lambda(\boldsymbol{V}) = \left\{ n_1\boldsymbol{v}_1 + n_2\boldsymbol{v}_2 \mid n_1, n_2 \in \mathrm{Z} \right\} \tag{4}$$

with $\boldsymbol{v}_k = \begin{bmatrix} v_k^{(x)} & v_k^{(\varphi)} \end{bmatrix}^{\mathrm{T}}$ for $k = 1,2$ referred to as basis vectors and $^{\mathrm{T}}$ being the transpose operator. The vectors building the lattice can be expressed in matrix form as

$$\boldsymbol{V}(\boldsymbol{v}_1, \boldsymbol{v}_2) = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{bmatrix} = \begin{bmatrix} v_1^{(x)} & v_2^{(x)} \\ v_1^{(\varphi)} & v_2^{(\varphi)} \end{bmatrix} \tag{5}$$

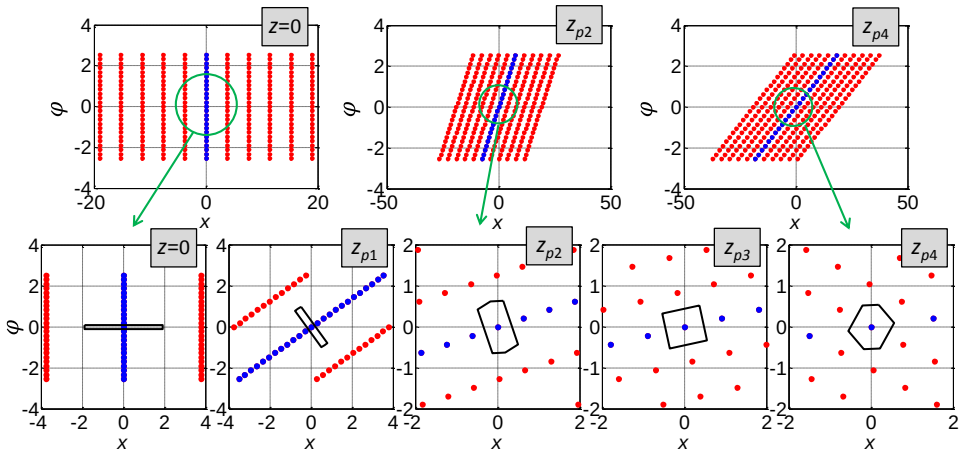



Fig. 4. Light field display – ray space spatial sampling patterns at different distances from the RG plane (c.f. Fig. 3)

with $\boldsymbol{V}$ being referred to as the sampling matrix. It is important to point out that the sampling matrix is not unique for a given sampling pattern since $\Lambda(\boldsymbol{V})=\Lambda(\boldsymbol{EV})$ where $\boldsymbol{E}$ is any integer matrix with $|\det \boldsymbol{E}|=1$. Consequently, there are multiple basis vectors describing the same lattice. In practice the set of basic vectors with minimum length (norm) is preferred. Therefore, given a set of basis vectors ($\boldsymbol{v}_1,\boldsymbol{v}_2$), one should find a pair of vectors ($\tilde{\boldsymbol{v}}_1,\tilde{\boldsymbol{v}}_2$) such that $\Lambda(\boldsymbol{V})=\Lambda(\tilde{\boldsymbol{V}})$. Here, tilde denotes the sampling matrix with minimized basis vectors (length $\|\boldsymbol{v}_1\|+\|\boldsymbol{v}_2\|$ is minimized) – see Fig. 5(a). The problem of finding such vectors is known in literature as the lattice basis reduction problem [15]. The solution applicable to our 2D case can be obtained using the following Lagrange's algorithm applied to a pair of basis vectors ($\boldsymbol{v}_1$,$\boldsymbol{v}_2$):

$$\begin{aligned}&\text{do}\\&\quad \text{if } \|\boldsymbol{v}_1\|>\|\boldsymbol{v}_2\| \text{ then swap } \boldsymbol{v}_1 \text{ and } \boldsymbol{v}_2\\&\quad \boldsymbol{v}_2=\boldsymbol{v}_2-\left\lceil\frac{\langle\boldsymbol{v}_2,\boldsymbol{v}_1\rangle}{\|\boldsymbol{v}_1\|^2}\right\rfloor\boldsymbol{v}_1\\&\text{until } \|\boldsymbol{v}_1\|\le\|\boldsymbol{v}_2\|\end{aligned} \tag{6}$$

The pair of vectors $(\boldsymbol{v}_1,\boldsymbol{v}_2)$ resulting from the algorithm are with the smallest norm for the given sampling pattern, that is, $(\boldsymbol{v}_1,\boldsymbol{v}_2)\equiv(\tilde{\boldsymbol{v}}_1,\tilde{\boldsymbol{v}}_2)$.

For a regular grid described with a lattice $\Lambda$, one can also define a unit cell *P* that is a set in $\mathrm{R}^2$ such that the union of all cells centered on each lattice point covers the whole sampling space without overlapping or leaving empty space. Similar to the basis vectors, the unit cell is not unique, as illustrated in Fig. 6. The figure illustrates three possibilities out of an infinite set of valid unit cells describing the same lattice. The shapes become even stranger if the underlying sampling pattern in not rectangular.

In this paper we use the Voronoi cell as the unit cell representing a given sampling pattern [16]. As illustrated in Fig. 5(b), the Voronoi cell, denoted by *P* (green shaded area in the

figure), is a set in $\mathrm{R}^2$ such that all elements of the set are closer, based on Euclidean distance, to the one lattice point that is inside the cell than to any other lattice point – this makes it the most compact unit cell. In the literature, Voronoi cells are also known as Wigner-Seitz cell – e.g. in solid-state physics [17 p.122]. By using the minimum length basis vectors, the construction of the Voronoi cell is straightforward and is illustrated in Fig. 5(b) (in Fig. 6 the Voronoi cell is the one shown by the leftmost example).

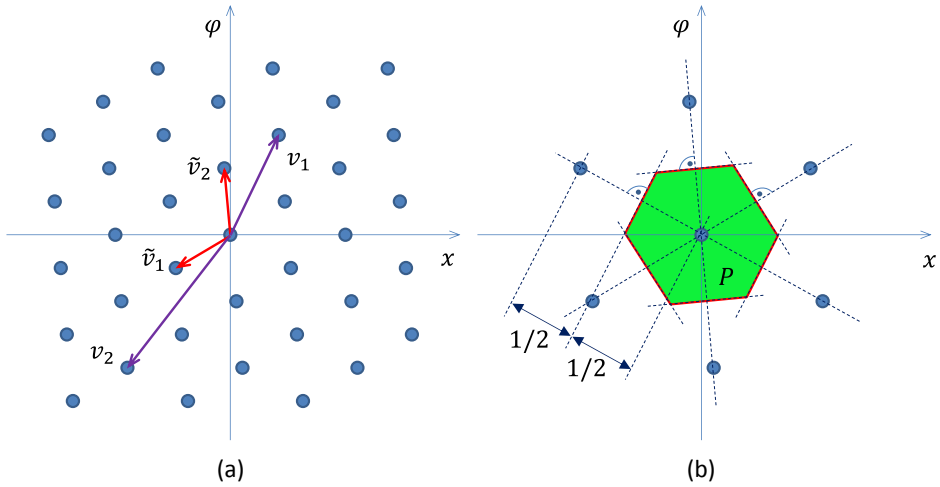



Fig. 5. (a) Example of two sets of basis vectors, $(\boldsymbol{v}_1,\boldsymbol{v}_2)$ and $(\tilde{\boldsymbol{v}}_1,\tilde{\boldsymbol{v}}_2)$, describing the same lattice. (b) Illustration of Voronoi cell construction.

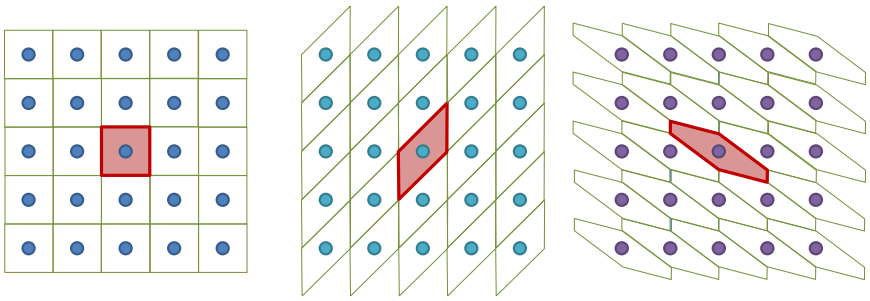


Fig. 6. Some possible unit cells *P* (shaded area) for a given lattice Λ (points).

The samples in ray space forming regular sampling patterns at different depths *z* represent a bandlimited function. In frequency domain, it has periodic structure with multiple replicas of the baseband. The periodicity and at the same time the baseband frequency support is defined through the reciprocal lattice $\Lambda^*$, that can be evaluated as [8,14]

$$\Lambda^*(\boldsymbol{V}) = \Lambda\left(\left(\boldsymbol{V}^T\right)^{-1}\right). \tag{7}$$

There are many possible unit cells for a given lattice $\Lambda^*$. Each possible unit cell describes a set of bandlimited functions that can be represented by the sampling pattern and can be reconstructed from a given discrete representation assuming that the reconstruction filter has the shape of the selected unit cell. Furthermore, this also means that an arbitrary continuous

function has to be pre-filtered with a filter aimed at removing all frequency content outside of the selected unit cell in order to prevent aliasing errors during sampling. This can be achieved either by using (if possible) a proper continuous-domain filter before sampling the function or first oversampling the continuous function and then performing filtering and down sampling in the discrete domain. It should be pointed out that in the case under consideration, it might not be possible to perform pre-filtering in the continuous domain since this would require an optical filter in spatial and angular direction. Therefore, in this paper we assume that we oversample the continuous function at the sampling stage and perform all filtering in the discrete domain. If the scene is captured by sparse cameras, the dense (oversampled) LF can be reconstructed by compressive sensing approaches, e.g. [18].

The most compact (isotropic) unit cell for a given sampling pattern is, as in the spatial domain, a Voronoi cell, denoted in this paper as $P^*$. The importance of this unit cell is twofold. First, it will represent frequency support that treats equally both directions (spatial and angular direction in ray space representation) – this is beneficial from the HVS viewpoint. Second, the screen in the display that will perform the D/C conversion has for practical reasons a 'low-pass' type characteristics (typically it is rectangular with Gaussian type weights [3]) that has to be matched to available ray distribution or vice versa. As such, the Voronoi cell will be the most convenient unit cell to match the screen reconstruction filter.

The Voronoi cell of a sampling pattern can be considered equivalently in spatial or frequency domain. Given its isotropic behavior, it is precisely the quantity, which characterizes the properties of the reconstructed bandlimited function. Therefore, the estimation of the optimal display and camera setup can be done by comparing Voronoi cells formed on the screen plane. From one side, there is the sampling pattern of the rays generated by the display; from another side there is the sampling pattern of the rays as captured by cameras. Both sampling patterns and the respective bandlimited LF are compared for similarity through their Voronoi cells in ray-space domain at the screen plane. This makes the overall optimization procedure computationally less demanding and thus faster. The frequency bandwidth of the system can be easily estimated once the optimal configuration is determined.

## 3. Light field display–camera configuration optimization

In an ideal case, one would require that a display perfectly reconstructs the underlying plenoptic function or at least up to the level of detail supported by the HVS. With limited resources, one can target the best possible continuous LF approximation out of a given discrete set of rays. In such a case, it is important to determine the optimal display and camera setup that maximizes the visual capabilities of the display.

We tackle the problem in two steps. First, we evaluate the optimal setup of ray generators for a given or desired density of rays at the screen plane. Second, we estimate the bandwidth for such system from the perspective of the scene, that is, what kind of capture setup and pre-processing is required to sense enough data for the given display setup. It should be pointed out that step two can be applied to an arbitrary display setup, as long as the basic setup parameters (ray generators, distances, screen properties, etc.) are available. The complete display-camera setup considered in this paper, with all adjustable parameters, is illustrated in Fig. 7 and is discussed in more detail in the following two sections. To streamline the text in the rest of this paper, we use the notations for various sampling patterns emerging from the display setup as in Fig. 7. Subscripts *p*, *s*, and *c* are used to denote the parameters related to the RG, screen, and camera/viewer plane, respectively, with *z* increasing in the direction of the observer and $z = 0$ being relative to the parameter's origin, e.g. for parameters originating on the screen plane, $z = 0$ is on the screen plane. Practical angles are denoted by $\alpha$ in contrast to 'theoretical' angles $\varphi$ used in the LF parameterization. The estimated and optimized parameters are denoted by hat and bar, respectively, e.g. $\hat{\alpha}_p$ and $\bar{x}_p$. Finally, tilde is used to denote parameters after the lattice basis reduction operation.

The proposed optimization technique can be extended for other display-camera configurations, than the one shown in Fig. 7, as long as those configurations result in regular sampling patterns in the angular-spatial domain and as such can be described by sampling matrices as illustrated for cases under consideration next.

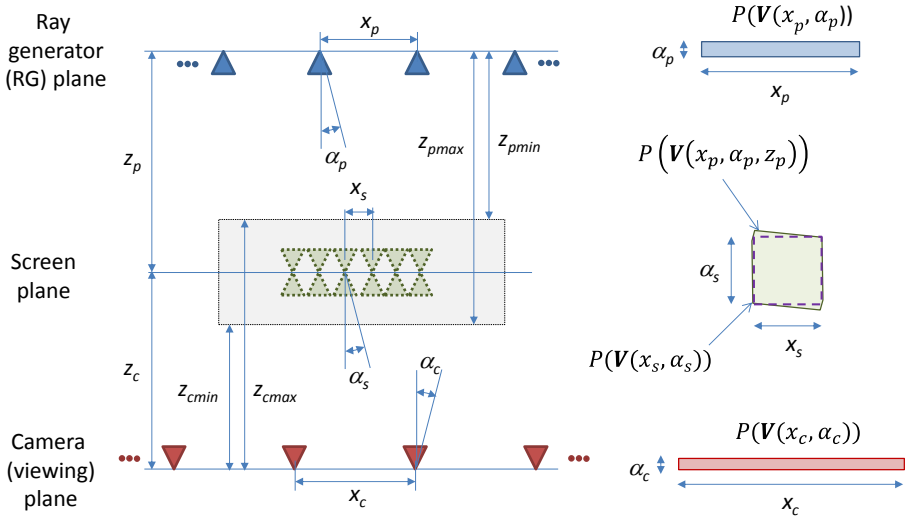



Fig. 7. Light field display–camera setup together with notations for expected sampling patterns.

### *3.1. Light field display configuration optimization*

For the purpose of stating the problem under consideration, we start from the center of Fig. 7, namely, the screen plane. We require that the display should be able to reproduce a LF with a desired bandwidth or, equivalently, a LF with a given density at the screen plane – the density being defined by the spatial and angular resolution. This determines the values $(x_s,\alpha_s)$ in the ray space representation, and in turn, it determines the desired sampling pattern at the screen plane. We assume that the pattern is rectangular – this is realistic assumption due to the properties of the screen and the requirements that both directions (spatial and angular) should be treated in a similar manner. The sampling pattern is uniquely defined through the following sampling matrix:

$$\boldsymbol{V}\left(x_s,\alpha_s\right)=\begin{bmatrix} x_s & 0 \\ 0 & \alpha_s \end{bmatrix}. \tag{8}$$

Having the sampling pattern at the screen plane, the problem is to determine optimal parameters of the ray generators $(x_p,\alpha_p)$ and distance between the RG plane and the screen plane $z_p$ for which the sampling pattern mapped to the screen plane will match the desired one, that is, grids described by sampling lattices $\Lambda\left(\boldsymbol{V}\left(x_p,\alpha_p,z_p\right)\right)$ and $\Lambda\left(\boldsymbol{V}\left(x_s,\alpha_s\right)\right)$ should match. With reference to Eq. (7), this will ensure the same Fourier domain bandwidth of the desired LF. Mismatches between the lattices $\Lambda\left(\boldsymbol{V}\left(x_p,\alpha_p,z_p\right)\right)$ and $\Lambda\left(\boldsymbol{V}\left(x_s,\alpha_s\right)\right)$ will manifest themselves either as aliasing effects in the reconstructed continuous LF or as inefficient utilization of the display bandwidth. The targeted lattice matching is done by an optimization technique presented below aimed at mitigating the aforementioned two problems.

The ray generators' sampling matrix mapped to screen plane is defined as

$$\boldsymbol{V}\left(x_p,\alpha_p,z_p\right)=\begin{bmatrix} x_p & z_p\tan\alpha_p \\ 0 & \alpha_p \end{bmatrix}. \tag{9}$$

Two sampling patterns will match in the case they have identical unit cells or will be similar if the difference between the unit cells that can be expressed as $\left\|P\left(\boldsymbol{V}\left(x_p,\alpha_p,z_p\right)\right)-P\left(\boldsymbol{V}\left(x_s,\alpha_s\right)\right)\right\|$ is small. Here we assume that the similarity criterion is defined via the difference in the area (size) of the unit cells and the difference in the shape of the unit cell. Furthermore, based on the sampling theory [14] a unit cell is uniquely described by its (lattice basis reduced) sampling matrix. Consequently, the similarity measure of two sampling grids (and correspondingly the underlying unit cells) can be expressed through the similarity between the corresponding sampling matrices. In summary, the problem under consideration is, for given $(x_s,\alpha_s)$, to find $(x_p,\alpha_p,z_p)$ that minimizes $\delta_p$

$$\delta_p=\left\|\tilde{\boldsymbol{V}}\left(x_p,\alpha_p,z_p\right)-\boldsymbol{V}\left(x_s,\alpha_s\right)\right\| \tag{10}$$

with $\boldsymbol{V}(x_s,\alpha_s)$ being the desired sampling matrix at the screen plane and $\tilde{\boldsymbol{V}}\left(x_p,\alpha_p,z_p\right)$ being the lattice basis reduced sampling matrix of the ray generators $\boldsymbol{V}(x_p,\alpha_p)$ mapped to the screen plane. It should be pointed out that $\boldsymbol{V}\left(x_s,\alpha_s\right)=\tilde{\boldsymbol{V}}\left(x_s,\alpha_s\right)$. Furthermore, when implementing Eq. (10), it should be kept in mind that the reduced matrix is unique up to the sign and sequence of basis vectors, that is, $\Lambda\left(\boldsymbol{V}\left(\boldsymbol{v}_1,\boldsymbol{v}_2\right)\right)\equiv\Lambda\left(\boldsymbol{V}\left(\pm\boldsymbol{v}_1,\pm\boldsymbol{v}_2\right)\right)\equiv\Lambda\left(\boldsymbol{V}\left(\pm\boldsymbol{v}_2,\pm\boldsymbol{v}_1\right)\right)$.

The lattice basis reduced sampling matrix of the ray generators mapped to the screen plane can be expressed as

$$\tilde{\boldsymbol{V}}\left(x_p,\alpha_p,z_p\right)=\begin{bmatrix} x_1 & x_2 \\ \alpha_1 & \alpha_2 \end{bmatrix}=\begin{bmatrix} x_s+\Delta x_1 & \Delta x_2 \\ \Delta\alpha_1 & \alpha_s+\Delta\alpha_2 \end{bmatrix}=\begin{bmatrix} x_s & 0 \\ 0 & \alpha_s \end{bmatrix}+\begin{bmatrix} \Delta x_1 & \Delta x_2 \\ \Delta\alpha_1 & \Delta\alpha_2 \end{bmatrix}. \tag{11}$$

In this notation, minimizing the difference between $\boldsymbol{V}(x_s,\alpha_s)$ and $\tilde{\boldsymbol{V}}\left(x_p,\alpha_p,z_p\right)$ corresponds to minimizing

$$\delta_p=\left\|\tilde{\boldsymbol{V}}\left(x_p,\alpha_p,z_p\right)-\boldsymbol{V}\left(x_s,\alpha_s\right)\right\|=\left\|\begin{matrix} \Delta x_1 & \Delta x_2 \\ \Delta\alpha_1 & \Delta\alpha_2 \end{matrix}\right\| \tag{12}$$

with $\Delta x_k$ and $\Delta\alpha_k$ (for $k=1,2$) depending on the unknowns $(x_p,\alpha_p,z_p)$, and in the ideal case should be zero (in practice they can never be zero but one can attempt making them small enough). The minimization of the measure $\delta_p$ in Eq. (12) is illustrated in Fig. 8(a). The lattice basis reduction procedure $\boldsymbol{V}\left(x_p,\alpha_p,z_p\right)\rightarrow\tilde{\boldsymbol{V}}\left(x_p,\alpha_p,z_p\right)$ is an iterative procedure with no analytical solution and there is no analytical relation between $\Delta x_k$ and $\Delta\alpha_k$ (for $k=1,2$) and unknowns $(x_p,\alpha_p,z_p)$.

The above problem can be tackled by fixing one of the parameters $(x_p,\alpha_p,z_p)$ and finding a solution for the other two that achieves the smallest $\delta_p$. Unfortunately, the optimization problem is not convex and has multiple local minima, which complicates finding the global minimum. However, since there are only three unknowns, a good practical approach is to do a grid search over a reasonable range of the unknown variables. This is a time consuming yet a reliable way to obtain the global minima. We will illustrate this by an example in Section 4.

Practical limitations of a projection-based light field display are related with the physical size and resolution of the ray generators, the number of generated rays, and other screen

properties – see [3] for more details. These limitations translate to a finite number of ray sources with high angular and lower spatial density at the RG plane, i.e. small $\alpha_p$, and larger $x_p$ Desired spatial resolution at the screen plane is higher, which can be achieved by reducing the angular resolution. This leads to practical limitations expressed as

$$\alpha_p \gg \alpha_s \text{ and } x_s \gg x_p, \tag{13}$$

where the difference between parameters at the RG and screen planes is at least one order of magnitude. This is illustrated in Fig. 9(a) and it is the starting point for considering the effect of shearing when moving from RG to screen plane.

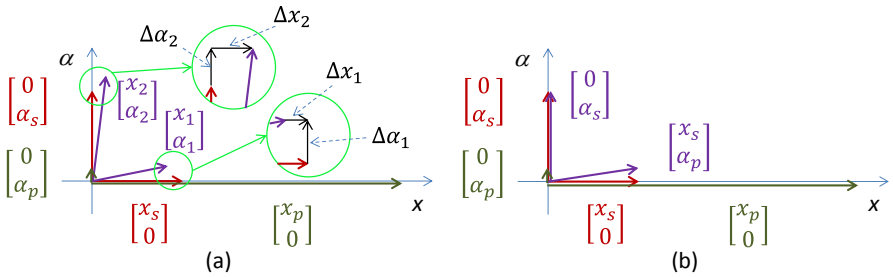



Fig. 8. Display optimization – matching $\boldsymbol{V}(x_s,\alpha_s)$ and $\tilde{\boldsymbol{V}}(x_p,\alpha_p,z_p)$. (a) General optimization solution obtained by minimizing Eq. (12). (b) Expected approximation, according to Eq. (14), for the setup under consideration.

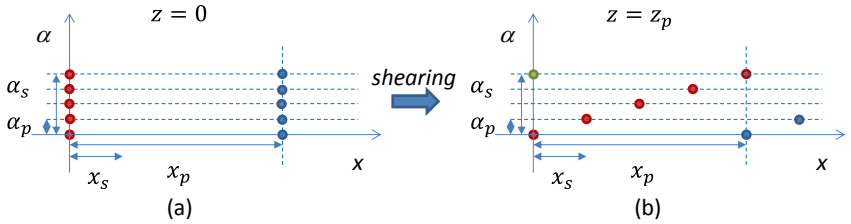



Fig. 9. Display optimization – shearing of $\Lambda(\boldsymbol{V}(x_p,\alpha_p))$ to $\Lambda(\boldsymbol{V}(x_p,\alpha_p,z_p))$ to match $\Lambda(\boldsymbol{V}(x_s,\alpha_s))$. (a) Sampling grid on RG plane. (b) Sheared sampling grid on the screen plane.

The sampling grid at the RG plane is described by $\boldsymbol{V}(x_p,\alpha_p)$. After shearing that grid by distance $z_p$ it turns into the sampling grid at the screen plane described by sampling matrix $\boldsymbol{V}(x_p,\alpha_p,z_p)$ as given in Eq. (9). The question is: Which sampling points in the original grid contribute to the basis vectors after shearing and lattice basis reduction? The approach for finding a good candidate can be graphically visualized as shown in Fig. 9. The original pattern corresponding to $\boldsymbol{V}(x_p,\alpha_p)$ in Fig. 9(a) is sheared to position $z=z_p$ in Fig. 9(b). The best approximation of the pattern $\boldsymbol{V}(x_s,\alpha_s)$ is achieved when (see also Fig. 8(b) for illustration)

$$\tilde{\boldsymbol{V}}(x_p,\alpha_p,z_p)=\begin{bmatrix} x_s & 0 \\ \alpha_p & \alpha_s \end{bmatrix}=\begin{bmatrix} z_p\tan\alpha_p & 0 \\ \alpha_p & \dfrac{x_p}{x_s}\alpha_p \end{bmatrix}. \tag{14}$$

This leads to the following estimates of two out of three unknown parameters $(x_p,\alpha_p,z_p)$:

$$\hat{z}_p = \frac{x_s}{\tan\alpha_p} \quad \Leftrightarrow \quad \hat{\alpha}_p = \tan^{-1}\frac{x_s}{z_p} \tag{15}$$

$$\hat{x}_p = x_s\frac{\alpha_s}{\alpha_p} \quad \Leftrightarrow \quad \hat{\alpha}_p = \alpha_s\frac{x_s}{x_p}. \tag{16}$$

Since we still have two equations and three unknowns, we need to select one of them by some other means. In the case under consideration, a good selection for $\alpha_p$ is

$$\hat{\alpha}_p \approx \alpha_s / L \quad \text{for} \quad L \in \mathrm{N}. \tag{17}$$

The reason for this selection lies in the fact that the sampling grid on the screen plane is a sheared version of the sampling grid at the RG plane with shearing performed only in the horizontal direction according to Eq. (3). Under these circumstances, the selection of $\alpha_p$ according to Eq. (17) will ensure that there exist a point in the sheared grid that approximately matches the desired sampling vector $\begin{bmatrix}0 & \alpha_s\end{bmatrix}^T$ thereby minimizing $\Delta\alpha_2$ and $\Delta x_2$. This is illustrated in Fig. 9(b).

The estimated parameters, as illustrated in Section 4 by means of examples, will be very close to the optimal ones, e.g. the optimal value of $\tilde{x}_p$ will be in the range $\hat{x}_p \pm x_s/2$. Based on this, we can formulate the optimization technique as follows:

1. Select a value $\hat{\alpha}_p$ according to available hardware resources and Eq. (17).
2. Use estimation formulas given by Eqs. (15) and (16) to get $\hat{x}_p$ and $\hat{z}_p$.
3. Refine the result by applying iterative search / general purpose optimization in range $\hat{x}_p \pm x_s/2$ thereby obtaining an optimal set of parameters $(\bar{x}_p, \bar{\alpha}_p, \bar{z}_p)$.

The evaluated sampling density in ray space $(\bar{x}_p, \bar{\alpha}_p)$ determines the spatial and angular resolution of the display. This technique will be illustrated by means of examples in Section 4.

*3.2. Camera setup optimization*

The camera setup should provide the rays required by the display for proper recreation of the LF of the scene. While the display is a band-limited device, the 3D visual scene is not (except for simple scenes with low-frequency spatial content, limited depth, and no occlusions) This means that when discussing the optimization of the camera setup, we have two problems to consider. First, how to estimate the optimal camera setup in terms of minimal amount of data that will provide the information needed for rendering all rays generated by the display. Second, how to ensure an alias-free capture of the scene to be recreated by the display.

Both problems are directly related to the display parameters and the corresponding display bandwidth they determine. The ultimate goal is to match that bandwidth with an optimal camera setup which allows rendering all rays needed by the display in an anti-aliassed pass band manner. The solution goes through matching the sampling patterns of the display and cameras at the screen plane. With reference to Fig. 7, the optimization problem is formulated as follows: for a given display sampling pattern described by ($x_p$,$\alpha_p$), find ($x_c$,$\alpha_c$,$z_c$) that minimizes

$$\delta_c = \left\| \tilde{\boldsymbol{V}}\left(x_p, \alpha_p, z_p\right) - \tilde{\boldsymbol{V}}\left(x_c, \alpha_c, -z_c\right) \right\| \tag{18}$$

with $\tilde{\boldsymbol{V}}\left(x_p, \alpha_p, z_p\right)$ and $\tilde{\boldsymbol{V}}\left(x_c, \alpha_c, -z_c\right)$ being the lattice basis reduced sampling matrices of the ray generators and cameras mapped to the screen plane, respectively – the argument ($-z_c$) indicates that the camera sampling matrix $\boldsymbol{V}(x_c,\alpha_c)$, is mapped to the screen plane with the

minus being there due to the orientation of the $z$ axis. Two comments regarding the above optimization criteria. First, we are doing the matching on the screen plane since this is the place where the D/C conversion takes place – sampling criteria must be satisfied at that plane. Second, in order to speed up the optimization, instead of $\tilde{\boldsymbol{V}}(x_p, \alpha_p, z_p)$ we could also use $\boldsymbol{V}(x_s, \alpha_s)$, assuming that the display sampling grid at the screen plane approximates well enough the desired one. This is perfectly fine in practice since it is expected that practical limitations (e.g. anti-aliasing filter, screen's D/C conversion) will affect the overall visual performance much more than mismatch between the desired and obtained display properties.

By applying an iterative optimization as described in the previous section, we can determine the optimal camera setup in terms of camera parameters $(\bar{x}_c, \bar{\alpha}_c, \bar{z}_c)$. In comparison to display optimization, there are additional restrictions that have to be taken into account, e.g. reasonable distance of a viewer from the screen, practical camera resolutions, $FOV_c \geq FOV_p$, and camera-to-camera distance that cannot be too small.

After determining the minimal camera sampling pattern $\Lambda\left(\boldsymbol{V}\left(\bar{x}_c, \bar{\alpha}_c\right)\right)$ for a given $\bar{z}_c$, we map the optimized display unit cell in the frequency domain at the screen plane $P^*\left(\boldsymbol{V}\left(\bar{x}_p, \bar{\alpha}_p, \bar{z}_p\right)\right)$ to the camera plane where it turns into $S_{\bar{z}c}\left(P^*\left(\boldsymbol{V}\left(\bar{x}_p, \bar{\alpha}_p, \bar{z}_p\right)\right)\right)$ where the $S_z\left(P^*\right)$ is the mapping (shearing) operator. Since $P^*$ is a convex set with points being the vertices of the unit cell in frequency domain that can be defined as

$$P^* = \left\{\left(\omega_k, \psi_k\right) \in \mathrm{R}^2\right\}_{k=1,2,\ldots K} \tag{19}$$

the mapping operator $S_z\left(P^*\right)$ maps each of those points such that

$$\begin{aligned} S_z\left(P\right) &= \left\{F_z\left(\omega, \psi\right) \,|\, \left(\omega, \psi\right) \in P\right\} \\ F_z\left(\omega, \psi\right) &= \left(\omega_k,\ \psi_k - z\tan\omega_k\right). \end{aligned} \tag{20}$$

The obtained shape of the cell determines the bandwidth of the display as observed from the camera plane. This is illustrated in Fig. 10. As can be seen, the shape is very different from the unit cell on the screen plane. Comparing this with the plenoptic sampling theory, it is obvious that, with respect to the overall plenoptic function, the display can only reproduce a finite amount of data concentrated around a particular distance from the viewer / camera plane. The limitation applies to angular and spatial coordinates. It should be pointed out that area (bandwidth) wise, $S_{\bar{z}c}\left(P^*\left(\boldsymbol{V}\left(\bar{x}_p, \bar{\alpha}_p, \bar{z}_p\right)\right)\right)$ and $P^*\left(\boldsymbol{V}\left(\bar{x}_c, \bar{\alpha}_c\right)\right)$ are of approximately same size (depending on how good minima have been found), but they cover different set of frequencies. In practice this means that for proper preparation (sensing) of content to be shown on the display, one has to do the following steps:

1. Capture the scene with sampling rate (large number of cameras) that will ensure proper anti-alias capture. This depends on the scene. However, the smallest bandwidth that has to be captured is marked by $P^*\left(V\left(x_c^{BIG}, \alpha_c^{BIG}\right)\right)$.
2. Filter the captured content with filter having the passband determined by $S_{\bar{z}c}\left(P^*\left(\boldsymbol{V}\left(\bar{x}_p, \bar{\alpha}_p, \bar{z}_p\right)\right)\right)$.
3. Down-sample the filtered signal to $P^*\left(\boldsymbol{V}\left(\bar{x}_c, \bar{\alpha}_c\right)\right)$.

This sensing procedure will result in properly pre-processed minimal amount of data that at the same time maximally utilizes the visualization capabilities of the display.

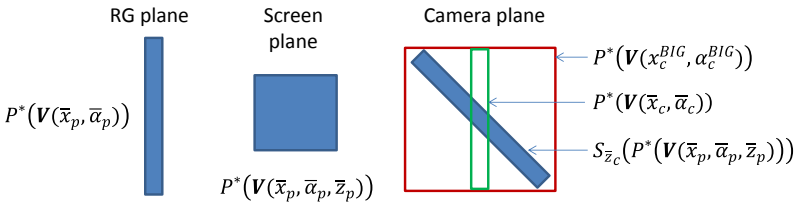



Fig. 10. Illustrations of display/camera sampling unit cell in frequency domain at different planes.

## 4. Examples

We illustrate the proposed optimization procedure on a 'realistic' display with reasonable quality as it can be built today, illustrate the optimization approach for optimal capture setups, and finally show / discuss what would be the setup of a display matching the requirements of the HVS.

### *4.1. Display optimization examples*

First we illustrate the proposed display configuration optimization on a display having desired spatial and angular resolution at the screen plane such that $x_s = 1\,\text{mm}$ and $\alpha_s = 1°$. We fix the angular resolution of the ray generators at the RG plane to $\alpha_p = 0.0391°$ (this resolution corresponds to a spatial resolution of 1024px over FOV of 40 degrees). For fixed $\alpha_p$, we evaluate the matching error $\delta_p$ on the screen plane for various values of $x_p$ $(10\,\text{mm} \le x_p \le 40\,\text{mm})$ and $z_p$ $(600\,\text{mm} \le z_p \le 1800\,\text{mm})$. The results of the optimization are shown in Fig. 11. In the figure, the left column shows the overall optimization range and the right column shows a zoomed-in range around the minimal value. Top row shows the result of overall optimization whereas middle and bottom row show the best solutions for a given $z_p$ and $x_p$. As it can be seen, there is a dominant minimum at $(\bar{x}_p, \bar{z}_p) = (26.01\,\text{mm}, 1465.90\,\text{mm})$ with an error value of $\delta_p = 0.04249$. The Voronoi unit cell $P\left(\boldsymbol{V}\left(\bar{x}_p, \bar{\alpha}_p, \bar{z}_p\right)\right)$ of such optimized display is shown in Fig. 12. As it can be seen, the match with the desired $P\left(V\left(x_s, \alpha_s\right)\right)$ one is almost perfect. The downside of such grid-based search is in the need to evaluate many combinations of $(x_p, z_p)$ not knowing which one will result in optimal solution. This can be considerably speeded-up by using the estimation approach described in Section 3.1. Following Eqs. (15) and (16), the estimated values for the problem under consideration are $(\hat{x}_p, \hat{z}_p) = (25.58\,\text{mm}, 1465.36\,\text{mm})$. As can be seen they are very close to the ones above obtained by the grid search. By performing single gradient-based optimization from the estimate, we end up with $(\bar{x}_p, \bar{z}_p) = (26.00\,\text{mm}, 1465.34\,\text{mm})$ with an error value of $\delta_p = 0.04249$. This is almost identical to the one obtained by the grid-based search and is obtained with a small amount of computational resources – fraction of a second instead of 10-15 min needed by the grid-based approach. Since almost identical result is obtained with both approaches, we can conclude that our proposed estimation method is correct and useful.

By using the fast estimation method, we can easily calculate optimal display setups for various screen parameters. First, Fig. 13 shows display optimization results for $x_s = 1\,\text{mm}$ and $\alpha_s = 1°$ for various values of $\alpha_p$. It is seen that for a good approximation we need small values of $\alpha_p$. However, very small values of $\alpha_p$ require impractically large values of $z_p$ and $x_p$. Therefore, in practice, a compromise between those has to be made. For illustration, the unit cells for optimal solutions for several values of $\alpha_p$ are shown in Fig. 14.

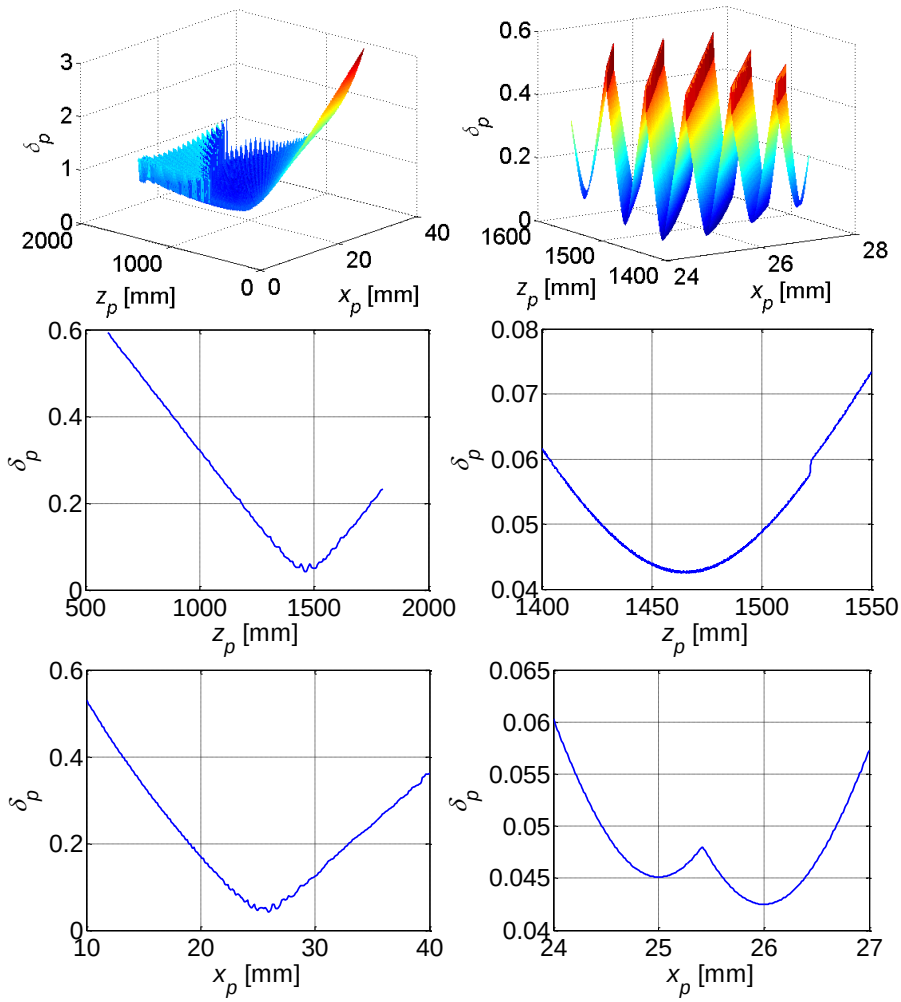



Fig. 11. Display optimization example based on grid search for $x_s = 1$ mm, $\alpha_s = 1°$, and $\alpha_p = 0.0391°$ – figures in the right column are zoomed in version of figures in the left column.

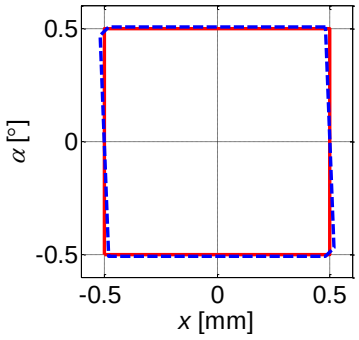



Fig. 12. Unit cells at screen plane for the optimized display setup solution for $x_s = 1$ mm, $\alpha_s = 1°$, and $\alpha_p = 0.0391°$ – $P\left(\boldsymbol{V}\left(\bar{x}_p, \bar{\alpha}_p, \bar{z}_p\right)\right)$ dashed/blue and $P\left(\boldsymbol{V}\left(x_s, \alpha_s\right)\right)$ solid/red.

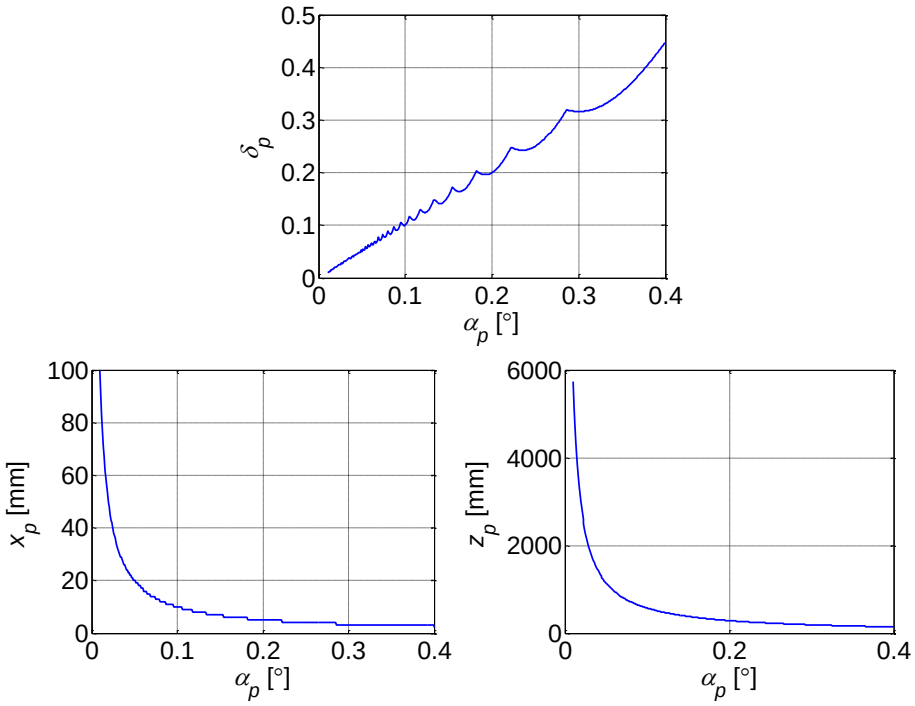



Fig. 13. Display optimization example for $x_s = 1$ mm and $\alpha_s = 1°$ for various values of $\alpha_p$.

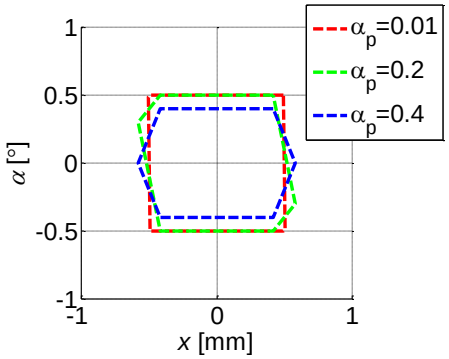



Fig. 14. Normalized ray-generator unit cells at the screen plane, $P\left(\boldsymbol{V}\left(\bar{x}_p, \bar{\alpha}_p, \bar{z}_p\right)\right)$ for various values of $\alpha_p$.

Next, we investigate the influence of different values of $(x_s, \alpha_s)$ on $(z_p, x_p)$ for fixed $\alpha_p = 0.0391°$. As seen in Fig. 15 a similar reconstruction error $\delta_p$ can be achieved independently of the choice for $x_s$ and $\alpha_s$. Furthermore, distance $z_p$ is influenced only on the desired $x_s$ and finally, $x_p$ has to be increased if either $x_s$ or $\alpha_s$ is increased. These figures give us a good understanding about the relation between involved parameters and can help us in making proper selection decisions.

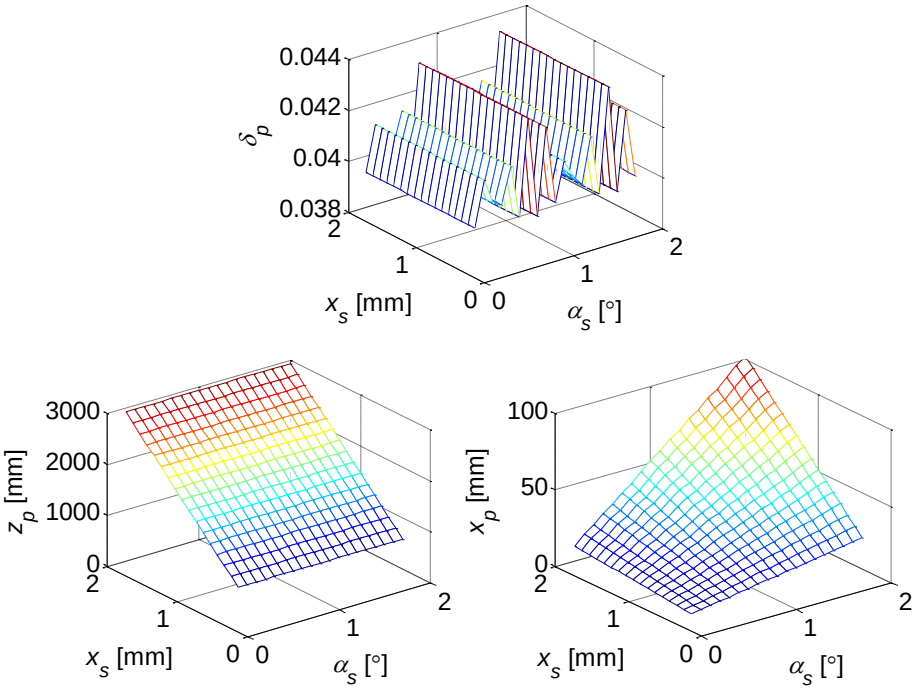



Fig. 15. Display optimization example for $\alpha_p = 0.0391°$ and various values of $x_s$ and $\alpha_s$.

*4.2. Camera optimization examples*

For an optimized display as described in the previous section, the display bandwidth is uniquely defined by $P\left(\boldsymbol{V}\left(\bar{x}_p, \bar{\alpha}_p, \bar{z}_p\right)\right)$ with a good approximation being described by $P\left(\boldsymbol{V}\left(x_s, \alpha_s\right)\right)$. Content captured by any means has to be pre-filtered to this bandwidth. The question here is what is the optimal camera/viewer setup, that is, what are the optimal parameters ($x_c$,$\alpha_c$,$z_c$) that would support the display bandwidth in the best possible way. In comparison to display optimization where it was logical to fix parameter $\alpha_p$, here it is more convenient to fix the screen to viewer distance $z_c$ since the viewer distance is typically 'fixed' / 'selected' by the user preferences / general recommendation for 'TV' watching. For a fixed distance $z_c = 2000$ mm, the result of optimization are shown in Fig. 16. Matching is performed again at the screen plane. There is dominant minimum at $(\bar{x}_c, \bar{\alpha}_c) = (35.72\text{ mm}, 0.0284°)$ with an error value of $\delta_c = 0.06832$. For comparison purpose, optimized ray generators and camera unit cells are shown in Fig. 17.

The grid search can be made faster by a better initial estimation. This can be done by assuming that the unit cell at the screen distance is ideal, that is, it is defined by ($x_s$,$\alpha_s$). By following the approach presented in Section 3.2, we obtain $(\hat{x}_c, \hat{\alpha}_c) = (34.91\text{mm}, 0.0286°)$. This is very close to the aforementioned optimal solution. Due to a high nonlinearity (see Fig. 16, middle row, right), one cannot use gradient based optimization but can perform a grid search only in the vicinity of the estimated values. Since this drastically limits the search space, it can be performed much faster than the full grid search.

The sampling pattern in the spatial domain can be converted to the frequency domain by using Eq. (7). By converting the frequency domain unit cell belonging to optimized display pattern from the screen plane to the camera plane, we obtain the bandwidth of the display –

shown in blue in Fig. 18. As discussed before, one should sample the scene with wide enough bandwidth to avoid aliasing, then pre-filter and then downsample. After downsampling, one obtains the maximum amount of data required by the display – display cannot show more and as such there is no point to provide more. It should be pointed out that this is in line with similar analysis performed for autostereoscopic displays [19,20].

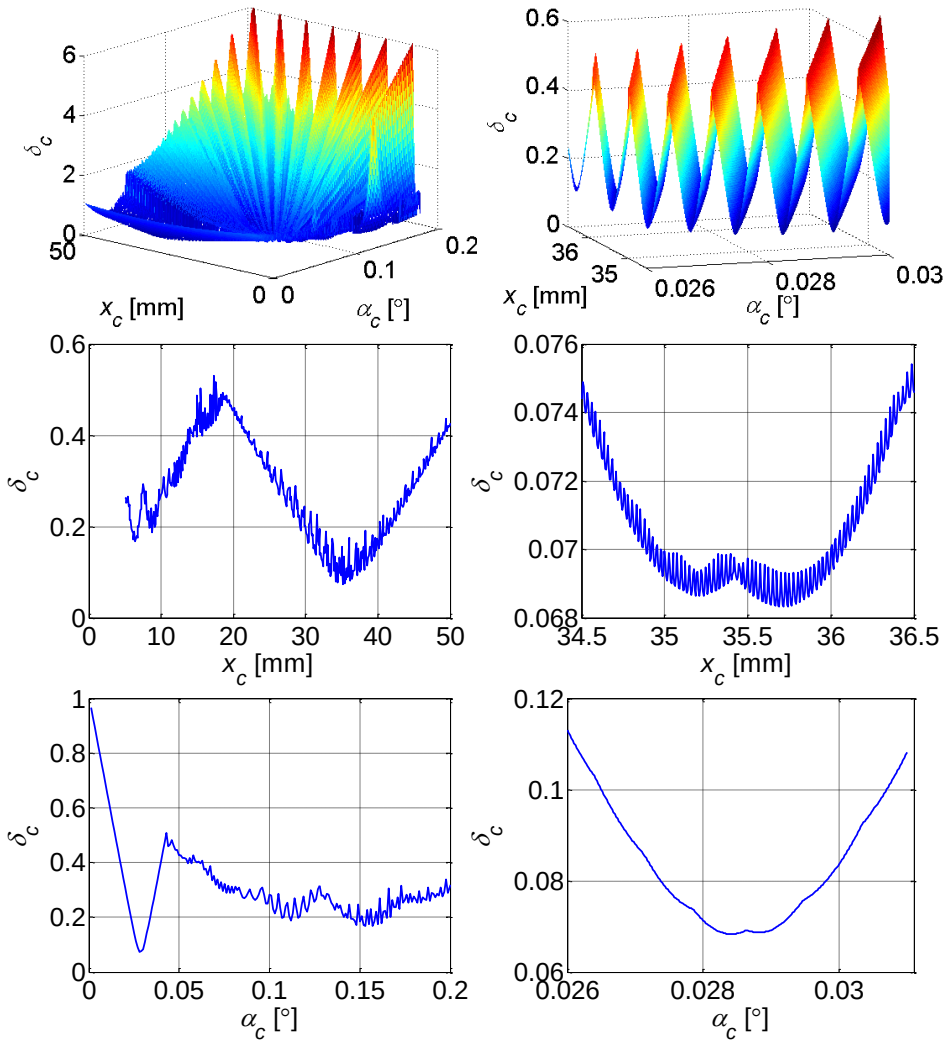



Fig. 16. Camera optimization example based on grid search for optimal $P\left(\boldsymbol{V}\left(\bar{x}_p,\bar{\alpha}_p\right)\right)$ optimized for $x_s = 1$ mm, $\alpha_s = 1°$, and $\alpha_p = 0.0391°$ with $z_c = 2000$ mm – figures in the right column are zoomed in version of figures in the left column.

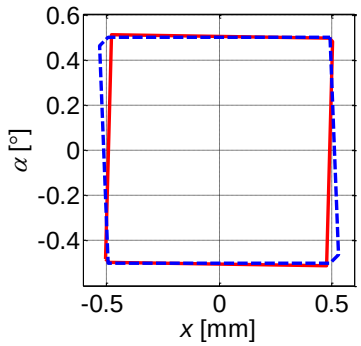



Fig. 17. Unit cells at screen plane for the optimized display and camera setup solution for $x_s = 1\,\text{mm}$, $\alpha_s = 1°$, $\alpha_p = 0.0391°$ and $z_c = 2000\,\text{mm}$ – $P\left(\boldsymbol{V}\left(\bar{x}_p, \bar{\alpha}_p, \bar{z}_p\right)\right)$ dashed/blue and $P\left(\boldsymbol{V}\left(\bar{x}_c, \bar{\alpha}_c, -\bar{z}_c\right)\right)$ solid/red.

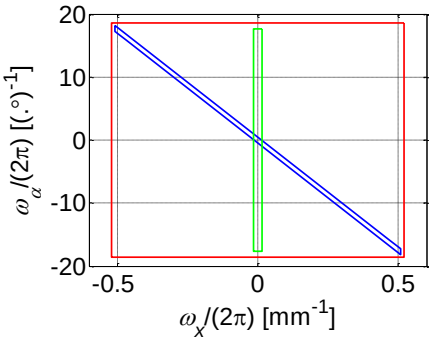



Fig. 18. Unit cells (bandwidths) at the camera (viewer) plane – $S_{\bar{z}_c}\left(P^*\left(V\left(\bar{x}_p, \bar{\alpha}_p, \bar{z}_p\right)\right)\right)$ blue, $P^*\left(V\left(\bar{x}_c, \bar{\alpha}_c\right)\right)$ green, and $P^*\left(V\left(x_c^{BIG}, \alpha_c^{BIG}\right)\right)$ red.

### *4.3. Consideration related to an 'ideal' HPO 3D display*

An ideal display should deliver the resolution required by the HVS. For estimating the required display angular-spatial resolution, in this section, we follow the discussion presented in [7, p 219-220]. It is assumed that an eye at distance $z_c$ from the display can differentiate spatial changes equal to 1/60° – this is equivalent to resolution of 30cpd (cycles per degree). This maps to

$$x_s = z_c \tan 1/60°. \tag{21}$$

Furthermore, the angular deviation that the eye can distinguish depends on the pupil size $d_p$ and can be estimated as [7]

$$\alpha_s = \tan^{-1}\left(d_p / z_c\right). \tag{22}$$

Assuming that average pupil size, as reported in the literature, is $d_p = 3\,\text{mm}$ and the viewing distance is fixed at $z_c = 2000\,\text{mm}$, we end up with required display resolution of $(x_s, \alpha_s) = (0.58\,\text{mm}, 0.086°)$. This means that an 'ideal' HPO display with 60 degree FOV, for

the assumed fixed distance, is required to reproduce at least $2 \cdot 10^9$ rays per square meter of the screen surface.

Following the proposed display optimization, one can determine that for fixed $\alpha_p = 0.0313°$, the optimal parameters of the ray generators should be $(\bar{x}_p, \bar{z}_p) = (1.74 \text{ mm}, 1062.86 \text{ mm})$ with the matching error being $\delta_p = 0.0322$. By mapping this values to the camera plane (c.f. Fig. 19), we can determine the necessary sampling rates as discussed in the previous section.

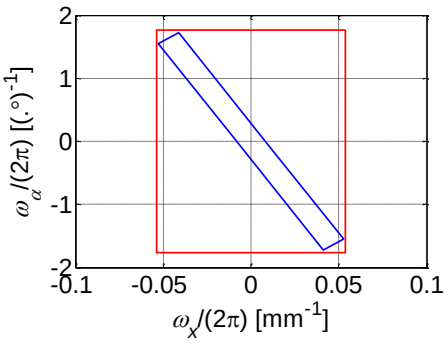



Fig. 19. Unit cells (bandwidths) at the camera (viewer) plane – $S_{\bar{z}_c}\left(P^*\left(\boldsymbol{V}\left(\bar{x}_p, \bar{\alpha}_p, \bar{z}_p\right)\right)\right)$ blue and $P^*\left(\boldsymbol{V}\left(x_c^{BIG}, \alpha_c^{BIG}\right)\right)$ red.

## 5. Concluding remarks

In this paper we presented a sampling model of the LF display-camera channel. We have shown that, from the sampling theory viewpoint, we can start with the required properties of the display specified in the ray space at the screen plane and then calculate the display setup fulfilling those requirements. Having the display setup, we can estimate the minimal set of data that the display needs to maximally utilize its visualization capabilities together with filter bandwidth for data pre-filtering aimed at alias-free reproduction.

Several points should be emphasized beyond the scope of this paper. First, we did not discuss all practical (hardware) aspects of implementing such displays, e.g. additional limitations to the design might be enforced by the available components and space like overall size, available ray sources, etc. Nevertheless, the same methodology presented in the paper still applies. Second, we assumed ideal D/C properties of the (holographic) screen. In practice, the screen will introduce additional smoothing that will further band limit the content the display can reproduce. Third, it should be always kept in mind that while the display is a bandlimiting device, a typical visual scene is not bandlimited. This means that special care has to be taken when sensing a scene and preparing the content for its optimal anti-aliasing filtering prior of its visualization on the LF display.

The discussion in the paper concentrated on projection-based displays employing diffusion-based holographic screen for reconstructing the continuous light field. This specific setting allows to clearly demonstrate the importance of ray sampling patterns for characterizing the display bandwidth and to directly relate it with the ray acquisition setting. However, the proposed approach can be used with any type of display system that attempts recreating a continuous light field and has an underlying (not necessary uniform) sampling of the input light field in the angular-spatial domain. Examples include autostereoscopic [21] and super multi-view displays [22, 23]. Further work and a more comprehensive analysis is required for displays having a non-uniform density of input light rays or /and ones that are

capable of changing the density of light rays based on content (e.g. tensor displays [24]) and it will be a topic of further research.

The estimates of 'ideal' projection-based LF display parameters as obtained in Section 4.3 were based on geometrical assumptions about the resolution power of the human eye. They show that a projection-based LF display matching the sampling density of the HVS is possible given that the individual optical modules are spaced 1,74 mm apart, while each module delivers rays with 0.03 rad angular step. Such a display is still difficult to produce and can be attempted in the future. Any other display designs with lower resolutions shall greatly benefit from the solution presented in this paper for delivering alias-free imagery. While the resolution power of the HVS estimated by geometrical assumptions is quite high, further studies are needed to characterize the perceptual threshold of continuous parallax, in the fashion how the window of visibility in disparity domain has been estimated [25]. Such perceptual characterization of continuous parallax would be more instructive when specifying the desired LF display bandwidth. A similar problem does exist with LF content creation. To cope with the anti-aliasing requirements, a very dense set of cameras is required for capturing content to be further processed for the specific display. Future development of intermediate view generation out of a set of sparsely captured views and employing signal processing sparsification approaches is of great interest.

## Acknowledgments

The research leading to these results has received funding from the PROLIGHT-IAPP Marie Curie Action of the People programme of the European Union's Seventh Framework Programme, REA grant agreement 32449 and from the Academy of Finland, grant No. 137012: High-Resolution Digital Holography: A Modern Signal Processing Approach.

# V

# QUANTIFYING SPATIAL AND ANGULAR RESOLUTION OF LIGHT-FIELD 3-D DISPLAYS

by

P. T. Kovács, R. Bregović, A. Boev, A. Barsi, A. Gotchev, 2017

IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 7, pp. 1213-1222,
DOI: 10.1109/JSTSP.2017.2738606

# Quantifying spatial and angular resolution of light field 3D displays

Péter Tamás Kovács, *Student member, IEEE*, Robert Bregović, *Member, IEEE*, Atanas Boev, Attila Barsi, and Atanas Gotchev, *Member, IEEE*

***Abstract*— Light field 3D displays are expected to play an important role as terminal devices, visualizing 3D objects apparently floating in the air, or letting viewers see through a window with a scene behind it. Currently, there are neither methods nor practical tools to quantify light field display's effective resolution or the perceived quality of the presented imagery. Most 3D displays are simply characterized by the total number of pixels or light rays; however this number does not properly characterize the distribution of the emitted light rays, nor the level of detail that the display can visualize properly. This paper presents methods to measure the spatial (i.e. 2D equivalent) and angular (i.e. directional) resolution of a given light-field display. The frequency domain analysis of recorded test patterns gives the spatial resolution limit of the display under test, while angular resolution is determined by the display's ability to pass through patterns of uniform, angle dependant patterns. It also presents a subjective experiment to corroborate the objectively measured spatial resolution.**

***Index Terms*—Displays, Three-dimensional television, Image resolution, Spatial resolution, Distortion, Spectral analysis**

## I. Introduction

VISUAL information has always been the primary and richest kind of information perceived by humans. Consequently, displays are primary output devices of computers and other electronic devices, including handheld, desktop and large-scale surfaces for representing visual information, some of which have 3D capabilities in recent products. Examples include 3D enabled phones, TVs and computer monitors. With recent advances in display technology, users expect immersiveness and realism when perceiving and interacting with visual information. Thus, displays are aimed at reproducing an increasing subset of visual cues present in reality: vivid colors, high resolution, binocular effect (when two eyes see different images, resulting in stereopsis), and the parallax effect (when an observer can see different perspectives while moving in front of the screen) all contribute to a higher realism of the displayed information [1]. Reproducing binocular effects and motion parallax are unique to autostereoscopic (glasses-free) 3D displays, which are implemented based on various technologies, such as lenticular-lens based multiview [2] or projection-based light-field displays [3][4][5][6].

### A. Motivation

Given different display technologies and models, one needs quantitative measures in order to know what to expect from a given display, what they can deliver in terms of visual realism. The characterization of the throughput of light-field 3D displays is an important yet challenging issue [7].

In case of 2D displays, spatial resolution is one of the major factors affecting visual realism, and is equally important for 3D displays. However, most 3D display manufacturers describe their products using metrics originating from 2D displays, such as the resolution of the underlying display panel, which does not quantify the distribution of these pixels in the spatial / directional domain. Light-field displays do not have a regular pixel structure, as will be described in Section II.A. Therefore measuring the number of features that can be faithfully represented by the display on the screen plane from a single viewing position (i.e. equivalent spatial resolution) is an important metric to judge the visual quality and level of detail presented by the 3D display. The effective resolution away from the screen plane (that is, for content that appears inside the display, or floating in front of the screen) can be calculated from the resolution at the surface using geometrical considerations [3].

Angular resolution is a new metric specific for 3D displays. It does not exist for 2D displays, as those do not exhibit viewing angle dependent behavior. For 3D displays, angular resolution quantifies the quality of the parallax effect. Visible discrete transitions in the motion parallax caused by insufficient angular resolution can hinder viewing experience and may hide important details of the presented information. The smoothness of motion parallax largely depends on the number of rays one can see over unit length when moving in front of the display's screen. Even more importantly, angular resolution has a direct effect on the depth range, that is, the range of depth shown with reasonable image quality. By

Manuscript received XXX; accepted XXX. Date of publication XXX.

The research leading to these results has received funding from the PROLIGHT-IAPP Marie Curie Action of the People programme of the European Union's Seventh Framework Programme, REA grant agreement 32449. Participants of the subjective experiment are gratefully acknowledged.

P. T. Kovács is with Tampere University of Technology, Tampere, Finland and Holografika, Budapest, Hungary (e-mail: peter.t.kovacs@tut.fi).

R. Bregović is with Tampere University of Technology, Tampere, Finland (e-mail: robert.bregovic@tut.fi).

A. Boev is with Tampere University of Technology, Tampere, Finland and Huawei ERC, Munich, Germany (e-mail: atanas.boev@huawei.com).

A. Barsi is with Holografika, Budapest, Hungary (e-mail: a.barsi@holografika.com).

A. Gotchev is with Tampere University of Technology, Tampere, Finland (e-mail: atanas.gotchev@tut.fi).

measuring the angular resolution, we are indirectly measuring the maximum depth that can be shown on the display [3].

Spatial and angular resolutions together imply the properties and the necessary processing [8] of the light field content the display should be supplied with.

### B. Organization of this paper

In this paper we propose methodologies for objective measurements to quantify spatial and angular resolutions. Furthermore, we propose and conduct subjective tests to corroborate the results of the spatial resolution measurements. Angular resolution measurement results are validated by technology insights from the display used in the experiments.

The rest of this paper is organized as follows. Section II describes related work in display measurement and briefly introduces light field displays (the 3D display technology this paper is focused on), and the concept for estimating display passband based on the display's geometry. Section III describes the objective measurement methods and how the results are analyzed to calculate spatial and angular resolutions, while Section IV describes the subjective test aimed at corroborating the objective measurement results. Section V presents sample results for a specific light-field display and compares the measurement results with the results of subjective experiment. Section VI concludes the paper.

## II. Related work

### A. Light-field 3D displays

Projection-based light-field 3D displays as described by Balogh [3] aim to reproduce the light-field of a real or synthetic scene by creating a surface with direction selective light emission. This is achieved by stacking many projection modules in a regular arrangement, so that these modules project light rays onto the screen, typically from behind. All projection engines beam light rays onto the whole screen area and these light rays hit the screen surface in slightly different directions. The special holographic screen onto which the rays are projected lets these rays pass through without changing their direction or mixing the color of the different light rays, as shown in Fig. 1.

The holographic screen applies a limited amount of horizontal diffusion, effectively applying discrete-to-continuous conversion on the discretized sources of light [9].

Using such optical arrangement, it is possible to show different images of the same scene to slightly different directions, so that the two eyes of the viewer can see different perspectives, resulting in stereopsis. Also, yet other images are shown to other directions that are seen when the viewer is moving in front of the screen, resulting in motion parallax. The number of different directions emitted from a single screen position can vary and reproducing hundreds of directions is not uncommon. The color of each light ray emitted by the projection modules is determined based on the light field to be represented, and the geometry of the optical arrangement by sampling the desired light field.

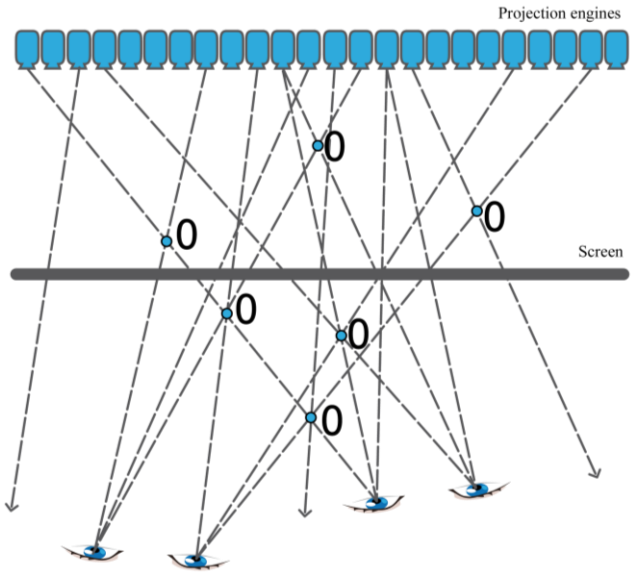



Fig. 1. Principle of operation of projection-based light-field 3D displays. Projection engines project light rays from behind, which may cross each other behind or in front of the holographic screen.

The light field is reproduced up to a certain spatial and angular resolution, upper bounded by the total number of light rays emitted by the display.

Projection engines are computer controlled, and thus generating and outputting arbitrary light-fields is possible by software means. The methods described in the next sections are supported by software for generating test light-field patterns, as well as for capturing and analyzing the displayed patterns.

### B. Limiting factors of light-field displays

There are factors that impose an upper limit on the capabilities of projection-based light-field displays, as well as other factors that can affect the final perceived image quality. First, the projection engines generating the light rays, typically containing an imaging component like DLP (Digital Light Processing) or LCoS (Liquid Crystal on Silicon), have finite resolution, such as VGA, XGA, WVGA, HD, or 4k, posing an upper limit on the number of light rays emitted from a single source.

Mounting a large number of projection engines requires a mechanical system, which cannot always ensure pixel-precise matching of light ray hit points on the screen. As a result, positional and rotational misalignments may occur. Also, the optical system of projection engines might have lens distortions, resulting in a non-perfectly rectangular image with equally spaced pixels. While these distortions can be compensated by measuring and pre-distorting the light field, sub-pixel differences still occur, resulting in light rays emitted at fractional positions. Because of this, an observer cannot count pixels on a light-field display's screen in the way it is possible on a 2D display having a discrete pixel structure.

The finite number of projection engines implies that the number of different directions reproduced is finite, and thus angular resolution has a limit.

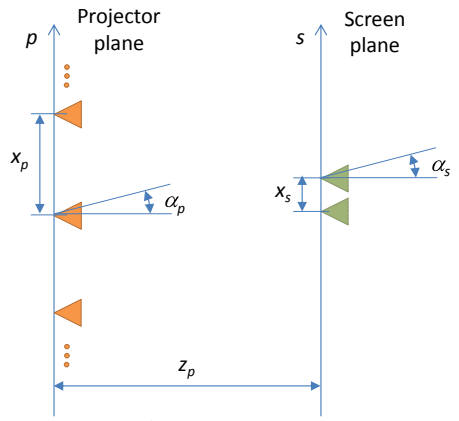



Fig. 2. Projection based light field display setup with notations.

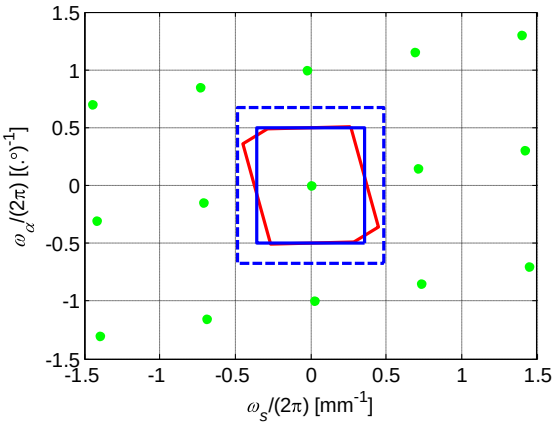



Fig. 3. Frequency domain sampling pattern (green) with estimated display passband (red) and shapes of different diffusors (blue).

While one projection engine can contribute to multiple emitted directions simultaneously using side mirrors which reflect a subset of light rays back to the screen from angles not covered by projection engines, the maximum number of directions cannot be higher than three times the number of projection engines in the system in case of a Horizontal Parallax Only (HPO) system (center, left reflection, right reflection).

### C. *Display passband*

Spatial and angular resolution are parameters of a more general concept, namely the bandwidth or passband of the display [10]. The passband characterises the display as a light field generator in Fourier domain and as such contains those frequencies, both in the spatial and angular domain (or, equivalently in the ray phase-space domain), which the display is able to reproduce without excessive spatial or inter-perspective aliasing.

As demonstrated in our previous work [9], the passband of a projection-based light field display can be estimated based on the geometrical configuration of the display and applying light field formalism on the display-generated light field. In practice, this requires knowledge about the properties and position of the projection engines and the screen (diffusor). Such approach is particularly useful when designing a display. Based on the desired light field, one can determine the optimal configuration of the display optical elements (which in turn can then be followed by determining optimal camera setup and the necessary pre-filtering) for approximating that desired light field.

In either case (building a display or evaluating the properties of an existing display), the assumption is that the reconstruction support of the diffusor in Fourier domain follows the optimal passband shape. However, in practice, the diffusor cannot have arbitrary shape. It is more-or-less restricted to rectangular shape similar to the ones discussed in [9]. The geometrical analysis considers the ray (spatial-angular) positions and interpret them as sample positions in ray space. As such, it enables estimating the preferable (optimal) reconstruction filter at the diffusor plane.

For illustrative purpose we present here an example. Consider a projection-based display with 22 cpd (cycles per degree) resolution at viewing distance of 3.5 m having 70 views over a field-of-view of 70 °.[1]

Following the notations in Fig. 2, this corresponds to spatial-angular sampling at the display plane such that $(x_s, \alpha_s) = (1.40\ \text{mm}, 1\ °)$. Rewriting the relation between the plane with projection units and the screen plane (see Eq. 14 in [9]), the parameters describing the sampling grid on the projection plane are evaluated as

$$\alpha_p = \text{atan}\left(\frac{x_s}{z_p}\right) \text{ and } x_p = \frac{\alpha_s}{\alpha_p} x_s.$$

Selecting the distance between screen and projection units as $z_p = 2500\ \text{mm}$, the display parameters are estimated as $(x_p, \alpha_p) = (43.633\ \text{mm}, 0.0321\ °)$. By performing the analysis as proposed in [9] one obtains sampling pattern in the Fourier domain as shown in Fig. 3. On the same figure, the display passband is shown in red, that in turn would be the optimal shape for the reconstruction filter of the diffusor. In practice the diffusor's Fourier-domain bandwidth is more like the ones of the blue squares with the one shown with solid line being the best candidate for the given sampling pattern. It offers the best overlap with the display bandwidth as determined by the geometry of the ray generators. A diffusor with wider bandwidth would imply worsened spatial selectivity of rays and diffusor with a narrower Fourier-domain support would be unnecessarily restrictive thereby removing finer details.

In summary, there are two issues with the theoretical analysis based on the configuration of optical elements: first, the need to know the correct physical configuration and, second, to know the properties of the diffusor. From the point of view of a general display user, these are not provided by the manufacturer. Then, the practical solution is to measure the parameters of the display passband in a way that would include the real contribution of all elements building the

[1] The display has been selected such to be close to the real display that will be measured / analyzed later on and for which the real geometrical data is not available.

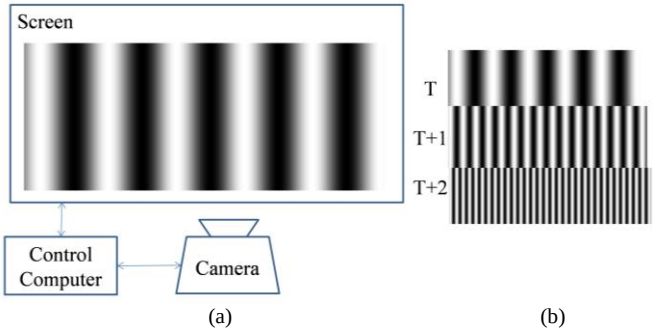



Fig. 4. Spatial resolution measurement overview. (a) A sinusoidal test pattern is rendered on the display under test, while a camera attached to the control computer takes a photo. (b) Subsequent measurement iterations show sinusoidals with increasing frequency.

display. Two measurement approaches aimed at this are introduced in this paper.

### *D. 3D Display measurement and quantification*

Most previous work on 3D display measurement have targeted the measurement and quantification of stereoscopic and multiview displays. Boev et al. developed a method [11] for modeling multiview displays in the frequency domain and identifying their limits using test patterns of different orientation and frequency. Their work, however only applies to multiview displays and only concerns spatial resolution.

Boher et al. developed measurement equipment using Fourier optics to measure the views emitted by desktop multiview displays [12]. While this gives precise results for the targeted class of displays, its applicability for large-scale 3D displays, as well as light-field displays using front projection is rather limited due to the size of the equipment and the way it is used (it would block the light path).

The International Display Measurement Standard (IDMS) published by the Society for Information Display [13] provides guidelines for measuring spatial and angular resolution of 3D displays in general. The recommended angular resolution measurement method relies on showing two-view patterns, which is not applicable for light-field displays, as these do not have discrete views to be controlled. It is also assumed that pixel size is known, which cannot be ensured when measuring an unknown display, which limits the applicability of this method. It further implicitly assumes that the pixels are rectangular, which is not true for several projection technologies (for example diamond shaped pixels in some DLPs).

From the methods presented in the IDMS for spatial resolution measurement for 2D displays, "Resolution from contrast modulation" is probably the closest to our method. This method uses grille lines of discrete sizes and measures the contrast ratio for these patterns. The effective resolution is estimated based on  where the contrast ratio is expected to fall below 50%, which is typically located between two discrete measurements. Our method is more precise, because on LF displays patterns with arbitrary size can be shown, thus interpolation is not necessary. The other advantage of our method is that it does not rely on contrast ratio only, but takes into account other sources of noise than decrease in image contrast (regardless of its source).

The authors are not aware of any comprehensive measurement method capable of quantifying spatial and angular resolution of light-field displays or any other 3D display with irregular pixel structure. Our previous work [14] presented an earlier version of spatial resolution measurement of light-field displays and an early method for angular resolution measurement, which used intensity loss for characterizing angular resolution. In this work we expand it by applying frequency analysis for characterizing angular resolution.

## III. MEASUREMENT METHODOLOGY

### *A. Methodology for spatial resolution measurement*

Our aim is to devise a methodology for measuring the equivalent spatial resolution of a light-field display using commodity camera and processing tools. The approach is inspired by the methodology presented in [11] for the case of multi-view displays. The measurement is accomplished by (1) visualizing sinusoidal test patterns with different frequency; (2) capturing the visible output by a camera; (3) analyzing the captured images in the frequency domain in order to check if the display reproduces the intended pattern without excessive distortions, and (4) converting frequencies to equivalent spatial resolution.

The test patterns are rendered with custom software that colors each light ray based on the hit point with the display's screen, regardless of its angle.

The rendered test pattern is a black-white sinusoidal that is shown over the whole screen surface (see Fig. 4. (a)), initially showing only a few periods on the screen. The attached camera takes a photo of the pattern as shown on the display, and the frequency of the sinusoidal is increased (see Fig. 4. (b)). The process is repeated until the frequency of the sinusoid well exceeds the theoretical limit of resolution of the light-field display, which is determined by the resolution of the imaging components. We use frequencies which are up to twice higher than the frequency corresponding to the maximum resolution of the imaging components. The same measurement is repeated for vertical sinusoidals.

The camera recording the test patterns as they are shown is set up on a tripod facing the display; the height of the sensor matches the center of the screen, and the use of manual shooting settings to ensure the captured intensity range matches the brightness of the display, so that the blacks and whites in the displayed patterns are visible on the photos (not saturating the dynamic range of the camera). The camera must be set to a resolution at least 2x times higher than the theoretical maximum resolution of the display (in practice, our algorithm uses four times as many samples to oversample the picture, considering any possible super-resolution effects caused by the multi-projection system). The shutter speed must be slower than the time-multiplexing frequency of the projection components (as projection engines typically employ

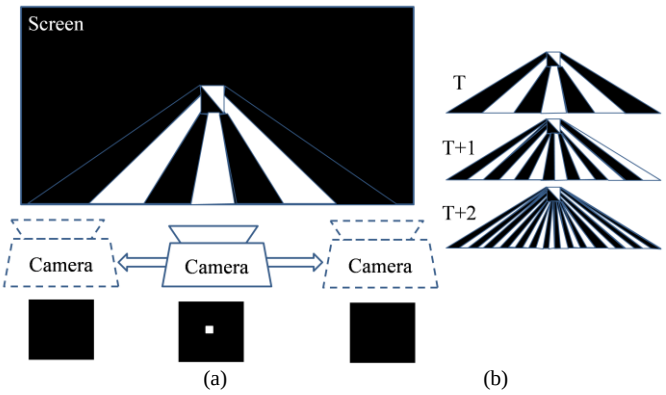



Fig. 5. Angular resolution measurement overview. (a) Test setup with moving camera. The rectangle looks black from some locations and white from other locations. (b) Test patterns of increasing angular frequency.

time multiplexing to emit R, G and B color components using a single imaging component; this frequency is available in the projection engine's data sheet), and the camera focus is set on the screen plane.

### *B. Analysis of spatial resolution*

The analysis of the photos of spatial resolution measurement is based on the observation that the frequency spectrum of the cropped photos, showing only the sinusoidal patterns and the distortions introduced by the display, clearly indicates the limit where the display is incapable of reproducing the test pattern. The test pattern is displayed on the whole screen, photographed, and 1D frequency analysis is performed on a single line of samples in the center of the screen. By increasing the frequency, the amplitude of the dominant frequency decreases, while at the same time distortion and after a certain frequency aliasing are introduced.

The algorithm determines the limit of resolution where the amplitude of distortion exceeds 20% of the amplitude of the dominant frequency. The horizontal spatial resolution then matches with the number of transitions of the test pattern that has been shown before the threshold was exceeded. The 20% threshold is associated with the level of disruptive distortion - this is the level of distortion (originating from aliasing or other nonlinear effects) which makes it impossible for the viewer to precisely identify all features of the presented visual data. The arguments for selecting 20% threshold came from [15], which in turn took insights from previous works, that had reported 10% [16], 15% [17], and 25% [18], respectively.

The same analysis is repeated for the orthogonal direction.

### *C. Methodology for angular resolution measurement*

For performing the desired measurements, one needs to emit different colors or intensities to different directions, and measure the screen from various angles. For a relatively small field of view produced by multiview displays, this may be accomplished by placing a relatively large sensor close to the screen as done in [12], but for the wider field of view typically produced by light-field displays, a moving camera is necessary.

On the display side, one needs to present a light field that has a well recognizable part that looks differently when observed from different angles (see Fig. 5. (a)).

In practice, the renderer driving the display uses a GPU shader that receives the parameters of the ray to be colored as input, based on which it calculates the position of the hit point on the screen and forms a rectangle-shaped measured spot. It also calculates the emitted direction, according to which rays are colored either black or white. The screen area that is colored in this way appears black from some viewing angles, and appears white from others, alternating as the observer or camera moves sideways. A camera moving in parallel with the screen plane records the transitions on video, the frames of which are subsequently analyzed. The setup of the camera is similar to that of the spatial resolution test, but the camera is mounted on a motorized rig that allows precise positioning of the camera.

In subsequent measurement iterations the angle of black / white features is decreased, thus more transitions are emitted per unit angle, and a new sliding video is recorded (see Fig. 6. (b)).

There is a possibility to check the angular resolution relying on information about the engine's order and topology in case of projection-based light-field displays, determining the maximum number of distinct light ray directions that can be emitted. Knowing the order of optical engines, one can force every even projection unit to project white, and odd ones to project black. This presents an intensity profile, while reversing the pattern presents the complement of this profile. Overlaying the two profiles shows peaks at every distinct direction that can be reproduced, and these peaks can be counted (see Fig. 10. (b) for an example, where one profile is shown for clarity). As no light can originate from between two discrete light sources inside the display, this limit cannot be exceeded by any test pattern. As our aim is to devise a general methodology for measuring the angular resolution, we use this alternative only to validate the first approach, which in turn requires no direct control of individual projection engines.

### *D. Analysis of angular resolution*

The display is considered to be capable of showing the pattern with the given angular resolution if the black / white transitions can be still recognized. In our analysis this is formulated by maximum achievable frequency in the directional domain. As we increase the angular frequency of the test pattern, we can observe that the transitions are still present; however, after a given frequency, the analysis shows that the dominant frequency is disappearing, or one may even observe decreasing apparent frequency. Frequency analysis is performed on the sequence of photos taken during the measurement. We define the limit of angular resolution where the dominant frequency reaches its upper limit (as seen later in Section V.B).

In case of HPO light-field displays, this measurement is performed in horizontal direction only. However, for 3D displays which have both a horizontal and vertical parallax, the measurement shall be repeated in a vertical direction to characterize the angular resolution in both directions. This is

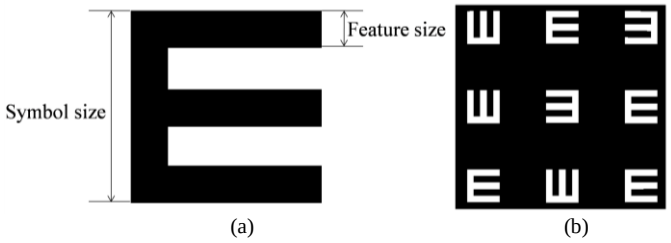



Fig. 6. Subjective spatial resolution test overview. (a) One tumbling "E" symbol. Feature size is 1/5 of the total symbol size. (b) A chart of 9 randomized E symbols arranged in a 3x3 matrix.

because horizontal and vertical angular resolution might be substantially different.

## IV. Subjective Test of Perceived Spatial Resolution

For the case of angular resolution measurement, there is a direct way to verify the correctness of the results of our proposed methods, as discussed in Section III.C. However, there is no such direct method for the case of equivalent spatial resolution measurement. Therefore, we resort to designing and conducting a subjective test to corroborate the proposed objective measurements for that case.

### *A. Methodology of subjective test of perceived spatial resolution*

In this test, the perceived spatial resolution of a light-field display is determined by showing visual features with various feature size to participants (see Fig. 6), and asking them to record what they can see on the screen. The visual feature chosen is the tumbling "E" symbol, which is often used by ophthalmologists to measure people's visual acuity [19], thus participants might already be familiar with them.

The methodology, as proposed, consists of two subsequent tests. The first test is done by using symbols printed on paper. It acts as pre-screening aimed at filtering participants with insufficient visual acuity. As such, it ensures that when in later tests people cannot recognize small symbols, this is caused by the peculiarities of the display, and not by the participant's limited visual acuity. In this first test, participants are asked to record the orientation of E's with randomized orientation, printed on paper, arranged in rows of 5, sized from 4 mm down to 0.85 mm, symbol size shrinking with 4/5 in 12 steps (halving every 3 steps), like in real tumbling E charts. From 80 cm viewing distance the 4 mm symbol size corresponds to 0.29 degree symbol size and 0.058 degree (~3.52 arcmins) feature size (also see Fig. 6 (a)). The smallest symbol size (0.8 mm) from the same distance corresponds to 3.69 minutes of arc, and a feature size of 0.73 minutes of arc. Only participants that pass the first test successfully (100% recognition of bigger symbols on paper) proceed to second test.

In the second test, people are presented with the 3D display at the same distance where the paper was shown. The display shows 9 randomized symbols, arranged in a 3 x 3 matrix (as shown in Fig. 6. (b)), and participants copy the orientation of the E's onto a 3x3 matrix on paper. 26 sets are presented in total and the first 4 are considered training sets, while the remaining 22 tests show 11 symbol sizes, both shown twice. Since we are not assessing the visual quality but the ability of the display to discriminate resolution variations, we use similar content for training and the actual experiment in order to allow people to familiarize with the test pattern and the task. The answers given for the training sets are not used during analysis. Symbol sizes are randomized to avoid habituation or anticipation that may affect results obtained by monotonically increasing or decreasing symbol sizes. The time elapsed between the presentation of a new set of symbols and reporting readiness (when all symbols have been recognized and recorded on paper) is measured and logged.

### *B. Analysis of subjective test of perceived spatial resolution*

First the results of the paper-based visual acuity tests are calculated, based on which participants with insufficient visual acuity to perform the test are eliminated. After removing the training sets, the randomized pattern sizes are reordered based on the recorded log files, resulting in an ordered list of accuracy and completion time for all participants (except those who failed on the paper based visual acuity test) for 11 symbol sizes of decreasing feature size. The accuracy values for all participants are averaged and the standard deviations and the 95% confidence intervals over recognition data are calculated.

Our hypothesis is that the participants will correctly recognize the symbols up to a given symbol size that corresponds to perceived display resolution and start making significant mistakes (wrong guesses) once the symbols are below the resolution of the display. We aim to identify the subjectively perceived spatial resolution corresponding to the smallest feature size participants are still able to recognize properly. We also check whether the time spent to recognize the symbols increases significantly, which would indicate that the resolution limit has been reached (when participants have difficulties in recognizing symbol orientations).

Although common visual acuity tests determine the limit of visibility where recognition accuracy falls below 50%, as recommended by [19], recent results such as [20] suggest that there may be significant differences between the recognition accuracy of different symbols (optotypes) due to the fact that the unbalanced symbols can be recognized based on their luminance distribution, even if they are heavily blurred or distorted. This suggests that such a subjective test cannot practically determine an absolute limit of visibility. We aim, however, to relate the uncertainty of visibility with the objective measurements.

## V. Results and Discussion

We demonstrate the applicability of the proposed measurement approaches on a light-field display. The display under test was a large-scale prototype light-field display with 180 cm screen diagonal. The display is controlled by a rendering cluster with high-end GPUs connected via HDMI connections. The setup for subjective experiments is described in Section V.C.

### *A. Spatial resolution measurement*

During the measurement, 411 photos have been taken for both horizontal and vertical directions and analyzed as described in Section III.B. The analysis procedure has been implemented in

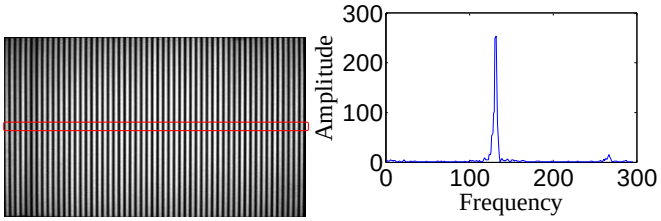



Fig. 7.(a) A photo of the screen showing a sinusoidal test pattern. The center row of the photo is used for frequency analysis. (b) Frequency spectrum of a single measurement showing the sinusoidal, with FFT bins on the horizontal axis.

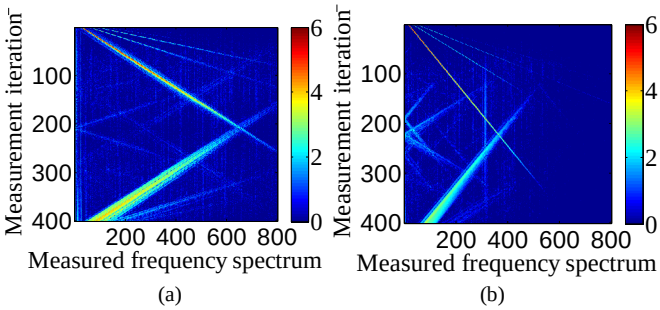



Fig. 8. Frequency spectrums of successive measurements stacked in a matrix. Measurement iteration count increases downwards, while the observed frequency increases rightwards. (a) Spectrums of horizontal resolution measurements. Major sources of distortion are visible as harmonics and constant low-frequency distortion. (b) Spectrums of vertical resolution measurements.

a Matlab script.

A photo of the light-field display showing a sinusoidal test pattern is shown in Fig. 7. (a). One line of samples is extracted from the center of the screen, on which 1D frequency analysis is performed. The frequency spectrum of a single measurement is shown in Fig. 7. (b), with FFT bins shown on the horizontal axis. How these relate to real resolution will be explained later in this section.

Using a series of photos and stacking the resulting frequency spectrums to form a 2D matrix, the frequency response of the display under test with increasing input frequencies is obtained, as shown in Fig. 8. Measurement iterations with increasing input frequency are shown on the vertical axis from top to bottom, while the resulting frequency spectrum is on the horizontal axis. The strong diagonal in the spectrum shows the input frequency being output by the display as the dominant frequency.

There are however other frequency components in the spectrum, the causes of which will be discussed. The decrease of the amplitude of the dominant signal is also visible on the diagonal. After finding the primary and secondary peak for all iterations, the ratio of the amplitude of the primary peak and the amplitude of the secondary peak can be plotted, which shows the level of distortion for each iteration. Such a plot is shown in Fig. 9, where distortion level is plotted in blue against the 20% threshold. From the plot one can see that the first iteration where distortion is above 20% is number 166.

The iteration number can be easily mapped to effective display resolution by plotting the frequency of the primary

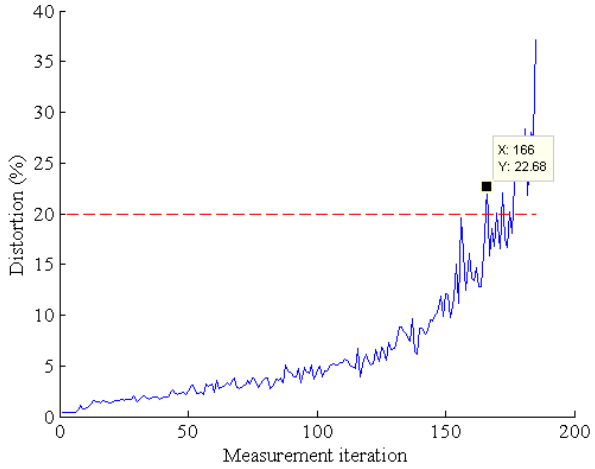



Fig. 9. Level of distortion in subsequent measurement iterations. 20% noise threshold is marked with red dashed line.

peak against the measurement iteration.

The resolution is the double of the measured frequency (529 in our case), as one period is considered to be created by the equivalent of two pixels per period in the classical discrete to analog signal conversion (one black and one white). Based on this correspondence between resolution and the frequency the horizontal resolution for the display under test is 2*(529-1) = 1056 pixels. One is subtracted from the frequency value, as the first bin in the spectrum corresponds to the DC component of the signal.

While determining the effective resolution is a useful result by itself, closely checking the stacked spectrums reveals some sources of distortion. Due to the slightly nonlinear intensity transfer function of the display, sinusoidals appear slightly rectangular, causing harmonics at odd multiples of the dominant frequency, appearing as weaker diagonals, as shown in Fig. 8. (a). Also, constant low frequency components on the left-hand side of the spectrum are visible in Fig. 8. (a). These are caused by the characteristic of the holographic screen, which results in slightly nonuniform brightness over the screen surface. This manifests itself in a fixed low frequency across all measurements.

Frequency aliasing can also be observed when exceeding the frequency throughput of the display with a large margin. Fig. 8. (b) shows the stacked spectrum of the vertical resolution of the measured display, which happens to be lower in the vertical direction. Starting around iteration 170, the mirrored image of the peak appears. Using the same calculation as with the horizontal resolution, the vertical resolution of the display is determined to be 636 pixels.

### *B. Angular resolution measurement*

Results for the same light-field display as used for the spatial resolution measurement are presented below. During this measurement, 60 videos have been taken with increasing angular frequencies, and analyzed as described in Section III.D. Sample intensity profiles of two different frequencies are shown in Fig. 10. (a). Each curve corresponds to the intensity observed by the camera at the center of the captured

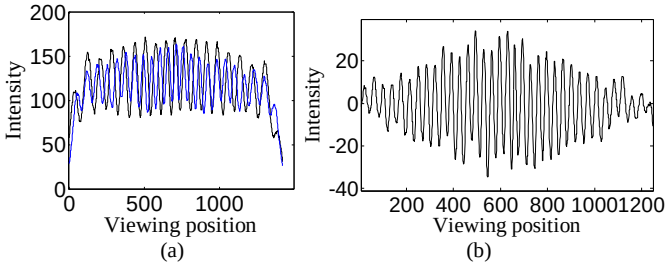



Fig. 10. (a) Sample intensity profiles for two different angular frequencies recorded on the same display. (b) Intensity profiles of forced black-white transitions on a light-field display.

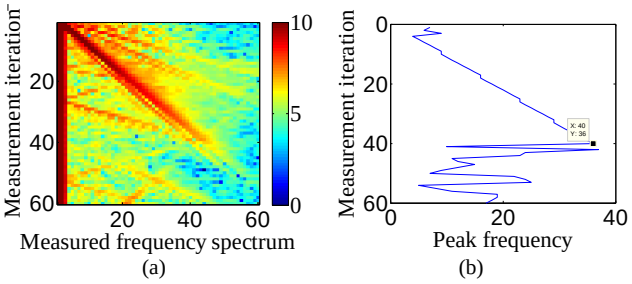



Fig. 11. (a) 1D frequency spectrums of angular resolution test patters stacked in a 2D array. (b) Frequency of the peak in subsequent measurement iterations.

image as the camera was moving sideways in front of the screen, showing a constant angular resolution test pattern.

By performing frequency analysis on each of the captured intensity profiles and stacking the spectrums below each other (as seen in Fig. 11. (a)), the increasing frequency of the test pattern can be clearly observed as a primary peak, with increasing frequency until approx. iteration 45. In subsequent iterations we can see the peak disappear and noise appear instead. Finding the peak in subsequent iterations (as shown in Fig. 11. (b)) gives the maximum number of periods one can observe over the field of view of the display. The peak in our case is at 36, which corresponds to 35 full periods (as the first bin in the spectrum corresponds to the DC component of the signal). The number of distinct directions is twice the number of full periods, in our case 70.

Dividing the total viewing angle of the display with this number, the angular resolution can be obtained. This display's total viewing angle is 65.4 degrees according to the measured intensity profile; therefore, the average angular resolution is 0.93 degrees.

Using the direct method for identifying the number of distinct directions that can be reproduced (as described in Section III.C), one can count 35 peaks on Fig. 10. (b), showing the intensity profile of the display under test. This alternative approach demonstrates that our method succeeded in finding the maximum number of directions emitted by the display.

### *C. Subjective test of perceived spatial resolution*

Subjective tests have been performed on 53 participants, 43 male and 10 female. The age range of participants was between 22 and 52 years. 21 of them used glasses. All participants claimed to have good eye sight and this was

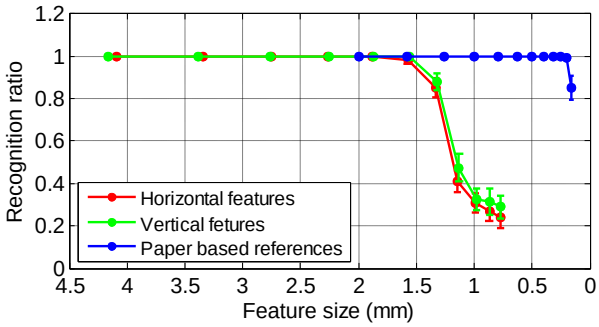



Fig.12. Recognition accuracy of tumbling E symbols on paper and display, for horizontal and vertical features, plotted against feature size, with 95% confidence intervals.

verified using the paper based charts as the first step of the tests. None of the participants were native English speakers (the study being performed in Finland), but represented several nationalities. We selected 50+ participants to participate in the subjective experiment in order to meet the requirements in the ITU Recommendation [21], that specifies that 'at least fifteen participants, non-experts, should be employed'. The distance between the display's screen and the participant's eye was 80 cm during the spatial resolution tests.

In our experiment, 3 out of 53 participants could not reliably record the orientation of the paper based patterns - those were removed from all analysis related to estimating the spatial resolution.

The weighted average of recognition accuracies and corresponding 95% confidence intervals have been calculated for all used symbol sizes for symbols with horizontal and vertical detail orientation (U/D and L/R). Based on symbol size, the feature sizes, and the corresponding resolution values for each symbol size have been calculated. Completion times have been averaged and 95% confidence intervals calculated. The averaged recognition accuracies are shown in Fig. 12. The figures clearly show several effects: the most obvious is that participants recognized the paper based reference almost perfectly (please note that the paper based symbols started smaller, though there is some overlap in symbol sizes). This means the limits visible in recognition accuracy in the case of the display are caused by the display's resolution limit, and not because of the participant's visual acuity limits.

The second noticeable result is that there is a statistically significant drop in recognition around feature size 1.4 mm: there, the confidence intervals do not overlap between measurement steps, and the p-values are very small ($p_{horizontal}$ = 3.91e-07, $p_{vertical}$ = 3.54e-07). This proves our hypothesis that there is a limit on the resolution the display is capable to visualize and it is close to symbol size where the recognition ratios have a major drop.

From charts in Fig. 13 and Fig. 14 it can be read that the measured resolution by the objective method (1056x636) corresponds to 93% and 92% recognition ratios in the horizontal and vertical direction, respectively. This is much higher than the 50% threshold commonly used in visual acuity tests, which can be attributed to the previously mentioned capability of humans recognizing symbols whose features are

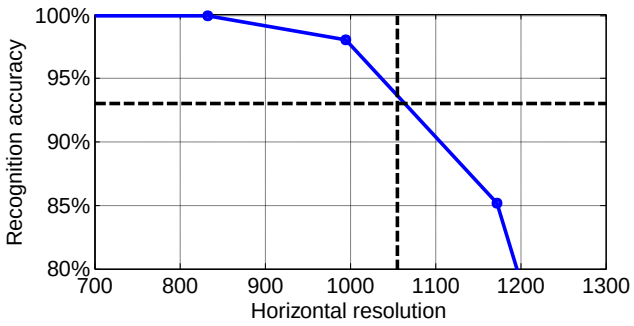



Fig.13. Zoomed in version of recognition accuracy versus resolution plot for horizontal resolution

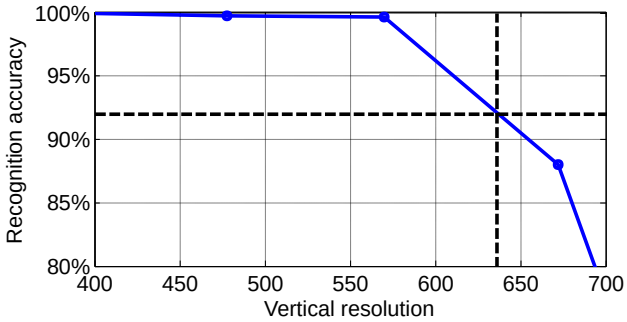



Fig.14. Zoomed in version of recognition accuracy versus resolution plot for vertical resolution

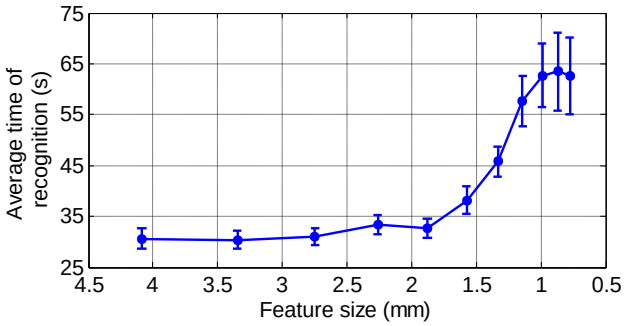



Fig.15. Average recognition time for a group of symbols with given feature size

not entirely visible [20][22][23]. Still, the thresholds being so close in the horizontal and vertical directions indicate that the measured resolution values and subjective perception of the visible features are proportional. From a practical perspective, this means that if a content producer provides content complying with the measured spatial resolution, this will ensure good suppression of visually annoying 3D artifacts for the greatest majority of viewers.

Note that the way of carrying out the subjective experiments imposed the use of certain finite set of symbol sizes. Therefore, we used linear interpolation between measured values to estimate the expected recognition accuracy for intermediate resolutions.

A summary of times participants needed to perform each iteration of the test are shown in Fig. 15. It is clearly visible that the average time to recognize and record a group of symbols stays almost constant until feature size ~1.4 mm. Symbols smaller than this limit seem to be increasingly difficult to recognize, indicated by the increased average recognition time as well as the larger confidence intervals. Incidentally, the resolution limit obtained by the objective method is reached at this feature size respectively. This shows that a steep increase in recognition time also indicates that the limit of visibility has been reached.

## VI. Conclusions

This paper addressed the proper characterization of light-field 3D displays. Two important limiting factors, namely spatial and angular resolution have been discussed, and camera-based objective measurement methodologies have been proposed. A subjective test for corroborating the results of the spatial resolution measurement has also been proposed. These measurement methods and subjective tests have been performed on a light-field display. We have identified limiting factors in light-field displays and how these manifest themselves in the measurements, and how human participants react when these limits are reached. Using the proposed measurement methods, light-field displays can be objectively quantified.

Further work will address any of the advancements over the methods presented here, such as: measuring the perceived resolution on depth planes different from the screen plane; the detection of potential non-uniformities in spatial and angular resolution; or the reduction of measurement time by using smaller number of photos to find the limits.

## References

[1] M. S. Banks, D. M. Hoffman, J. Kim, G. Wetzstein, “3D Displays,” *Annual Review of Vision Science*, vol. 2, pp. 397-435, Oct. 2016.

[2] C. van Berkel, D. W. Parker, A. R. Franklin, “Multiview 3D-LCD,” *Proc. SPIE 2653, Stereoscopic Displays and Virtual Reality Systems III*, Apr. 1996

[3] T. Balogh, “The HoloVizio system,” in *Proc. SPIE 6055, Stereoscopic Displays and Applications XIII*, 60550U, Jan. 2006.

[4] G. Wetzstein, D. Lanman, M. Hirsch, R. Raskar, “Tensor Displays: Compressive Light Field Synthesis using Multilayer Displays with Directional Backlighting,” *ACM Trans. Graphics, Proc. SIGGRAPH 2012*, vol. 31, no. 4, Jul. 2012.

[5] J.-H. Lee, J. Park, D. Nam, S.-Y. Choi, D.-S. Park, C.-Y. Kim, “Optimal Projector Configuration Design for 300-Mpixel Light-Field 3D Display,” *Optics Express*, vol. 21, no. 22, pp. 26820-26835, 2013.

[6] M. Kawakita, S. Iwasawa, R. Lopez-Gulliver, M. Makino, M. Chikama, M. P. Tehrani, “Glasses-free 200-view 3D Video System for Highly Realistic Communication,” *Proc. Digital Holography and Three-Dimensional Imaging 2013*, Apr. 2013, pp. DM2A.1.

[7] A. Stern, Y. Yitzhaky, B. Javidi, "Perceivable Light Fields: Matching the Requirements Between the Human Visual System and Autostereoscopic 3-D Displays," *Proc. of the IEEE*, vol. 102, no. 10, pp. 1571-1587, Oct. 2014.

[8] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wong, Q. Dai, T. Chai, Y. Liu, “Light field image processing: An Overview,” *IEEE Journal of Selected Topics in Signal Processing*, 2017

[9] R. Bregović, P. T. Kovács, A. Gotchev, “Optimization of light field display-camera configuration based on display properties in spectral domain,” *Optics Express*, vol. 24, no. 3, pp. 3067-3088, Feb. 2016.

[10] M. Zwicker, W. Matusik, F. Durand, H. Pfister, C. Forlines, “Antialiasing for automultiscopic 3D displays,” *ACM Trans. Graphics, Proc. SIGGRAPH 2006*, Aug. 2006.

[11] A. Boev, R. Bregovic, A. Gotchev, “Measuring and modeling per-element angular visibility in multi-view displays,” *Journal of the Society for Information Display*, vol. 18, no. 9, pp. 686–697, Sep. 2010.

[12] P. Boher, T. Leroux, T. Bignon, V. Colomb-Patton, “A new way to characterize autostereoscopic 3D displays using Fourier optics

instrument," *Proc. SPIE 7237, Stereoscopic Displays and Applications XX*, 72370Z, Feb. 2009
[13] The International Display Measurement Standard v1.03, Society for Information Display, 2012.
[14] P. T. Kovács, A. Boev, R. Bregović, A. Gotchev, "Quality measurements of 3D light-field displays," *Proc. 8th Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics 2014*, Jan. 2014.
[15] A. Boev, R. Bregović, A. Gotchev, "Visual-quality evaluation methodology for multiview displays," *Displays*, vol. 33, no. 2, pp. 103-112, Apr. 2012.
[16] L. Wang, K. Teunissen, T. Yan, C. Li, Z. Panpan, Z. Tingting, I. Heynderickx, "Crosstalk Evaluation in Stereoscopic Displays," *Display Technology*, vol. 7, no. 4, pp. 208-214, Mar. 2011.
[17] P. J. H. Seuntiëns, L. M. J. Meesters, W. A. IJsselsteijn, "Perceptual attributes of crosstalk in 3D images," *Displays*, vol. 26, no. 4-5, pp. 177-183, Oct. 2005.
[18] F. Kooi, A. Toet. "Visual comfort of binocular and 3D displays," *Displays*, vol. 25, no. 2-3, pp. 99–108. Aug. 2004.
[19] International Society for Low vision Research and Rehabilitation, "Guide for the Evaluation of Visual Impairment," *Proc. International Low Vision Conference (VISION-99)*, Jul. 1999.
[20] S. P. Heinrich, M. Back, "Resolution acuity versus recognition acuity with Landolt-style optotypes," *Graefes Arch Clin Exp Ophthalmol*, vol. 251, no. 9. pp. 2235–2241, Sep. 2013.
[21] Methodology for the subjective assessment of the quality of television pictures, ITU-R BT.500-13, Jan. 2012.
[22] J. S. Pointer, "Recognition versus Resolution: a Comparison of Visual Acuity Results Using Two Alternative Test Chart Optotype," *J Optom* 2008, vol 1, pp. 65-70. 2008
[23] L. N. Reich, M Ekabutr, "The Effects of Optical Defocus on the Legibility of the Tumbling-E and Landolt-C," *Optometry and Vision Science*, vol. 79, no. 6, pp. 389–393

**Péter Tamás Kovács** (Student member, IEEE) received the M.Sc. degree in computer science from the Budapest University of Technology, Budapest, Hungary in 2004. He is currently pursuing the Ph.D. degree in signal processing at Tampere University of technology, Tampere, Finland.

From 2004 to 2006, he was a Software Engineer with Archi-Data. Since 2006, he has been Software Engineer, Lead Software Engineer, then CTO of Holografika. He has been a visiting researcher at Tampere University of Technology from 2013 to 2014. He is the author or co-author of four book chapters, four journal papers and more than 30 conference papers. His research interests include 3D displays, more specifically light-field displays, and the capture / compression / rendering of light fields.

He has served as PC member for numerous IEEE conferences, Local Organizing Chair of 3DTV-Con 2014, and is a contributing member of the International 3D Society and the International Committee for Display Metrology (ICDM), where he contributed to the first IDMS standard. He was the Head of Delegation to MPEG for Hungary.

**Robert Bregović** (Member, IEEE) received the Dipl. Ing. and MSc degrees in electrical engineering from University of Zagreb, Zagreb, Croatia, in 1994 and 1998, respectively, and the Dr. Sc. (Tech.) degree (with honors) in information technology from Tampere University of Technology, Tampere, Finland, in 2003.

From 1994 to 1998, he was with the Department of Electronic Systems and Information Processing of the Faculty of Electrical Engineering and Computing, University of Zagreb. Since 1998, he is with the Laboratory of Signal Processing, Tampere University of Technology. His research interests include the design and implementation of digital filters and filterbanks, multirate signal processing, and topics related to acquisition, processing, modeling, and visualization of 3D content.

**Atanas Boev** is a researcher with expertise in light field displays and human stereopsis. He got the best demo award at Tampere Innovation Days 2010, best student paper award in Electronic Imaging 2011, and Nokia Scolarship award in 2010. He defended his PhD in Tampere University of Technology, Finland in 2012. In 2013 he was visiting researcher at Holografika KFT, Hungary, doing development and implementation of light field rendering algorithms. In 2014 he was post-doctoral researcher in Tampere University of Technology working on modern signal processing methods for lightfield displays. Currently, he is a research engineer in the Huawei European Research Centre, working on light field display technology.

**Attila Barsi** received the M.Sc. degree in computer science from the Budapest University of Technology, Budapest, Hungary in 2004.

From 2005 to 2006, he was a Software Engineer with DSS Hungary. Since 2006, he has been Software Engineer, then Lead Software Engineer of Holografika. He is the author or co-author of several conference and journal papers. His research interests include light fields, real-time rendering, ray tracing, global illumination and GPU computing.

**Atanas Gotchev** (Member, IEEE) received the M.Sc. degrees in radio and television engineering (1990) and applied mathematics (1992) and the Ph.D. degree in telecommunications (1996) from the Technical University of Sofia, and the D.Sc.(Tech.) degree in information technologies from the Tampere University of Technology (2003).

He is a Professor at the Laboratory of Signal Processing and Director of the Centre for Immersive Visual Technologies at Tampere University of Technology. His research interests have been in sampling and interpolation theory, and spline and spectral methods with applications to multidimensional signal analysis. His recent work concentrates on algorithms for multisensor 3-D scene capture, transform-domain light-field reconstruction, and Fourier analysis of 3-D displays.

# VI

## REAL-TIME 3D LIGHT FIELD TRANSMISSION

by

T. Balogh, P. T. Kovács, 2010

in Proc. SPIE 7724, Real-Time Image and Video Processing 2010, 772406,
DOI:10.1117/12.854571

# Real-time 3D light field transmission

Tibor Balogh, Péter Tamás Kovács
Holografika Kft, Baross u 3., Budapest, Hungary

## ABSTRACT

Although capturing and displaying stereo 3D content is now commonplace, information-rich light-field video content capture, transmission and display are much more challenging, resulting in at least one order of magnitude increase in complexity even in the simplest cases. We present an end-to-end system capable of capturing and real-time displaying of high-quality light-field video content on various HoloVizio light-field displays, providing very high 3D image quality and continuous motion parallax. The system is compact in terms of number of computers, and provides superior image quality, resolution and frame rate compared to other published systems. To generate light-field content, we have built a camera system with a large number of cameras and connected them to PC computers. The cameras were in an evenly spaced linear arrangement. The capture PC was directly connected through a single gigabit Ethernet connection to the demonstration 3D display, supported by a PC computation cluster. For the task of dense light field displaying massively parallel reordering and filtering of the original camera images is required. We were utilizing both CPU and GPU threads for this task. On the GPU we do the light-field conversion and reordering, filtering and the YUV-RGB conversion. We use OpenGL 3.0 shaders and 2D texture arrays to have an easy access to individual camera images. A network-based synchronization scheme is used to present the final rendered images.

**Keywords:** Lightfield, Light-field, 3D video, HoloVizio, 3D capture, Rendering, Real-time, Camera array, 3D display

## 1. INTRODUCTION

Displaying live 3D imagery is a major step towards realistic visualization. Today most common solutions for 3D displaying are active or passive stereoscopic glasses, which are cheap and easily available. However, like all stereoscopic systems they can only provide 3D view for a single, fixed position (this is also true when multiple people are watching, as they all see the same image). Autostereoscopic displays can show different 3D images to multiple directions, but the most widespread displays (lenticular or parallax barrier systems) have limited light ray count (3D resolution). These can provide continuous view in a narrow FOV, however viewers moving may experience jumps when leaving or entering valid zones. HoloVizio light-field display technology is capable of providing 3D images featuring continuous motion parallax for a wide viewing zone for multiple viewers.

HoloVizio displays are capable of displaying high quality horizontal parallax light fields. The HoloVizio principle can also be extended to have both horizontal and vertical parallax, but this has not been demonstrated yet. As the information content of a light-field is very high, displaying large FOV light field videos is a challenging task. Our camera array consisting of 27 cameras (capturing 18 Million light rays) was connected a 10 MPixel HoloVizio system (HoloVizio 240P) and later a 30 MPixel HoloVizio system (HoloVizio 720RC). This shows the flexibility of our software system, as it can handle an arbitrary number of incoming and outgoing light rays in practically any camera and display configuration. The captured natural content was processed and displayed in real time. Storage and playback of captured 3D content is also possible with the system.

## 2. HOLOVIZIO TECHNOLOGY

The approach used by HoloVizio technology is quite different from that of stereoscopic, multiview, volumetric and holographic systems. It uses a specially arranged array of optical modules and a holographic screen. Each point of the holographic screen emits light beams of different color and intensity to the various directions. The light beams generated in the optical modules hit the screen points in various angles and the holographic screen makes the necessary optical transformation to compose these beams into a perfectly continuous 3D view, as shown in Figure 1. With proper software

control, light beams leaving the pixels propagate in multiple directions, as if they were emitted from the points of 3D objects at fixed spatial locations[1,2].

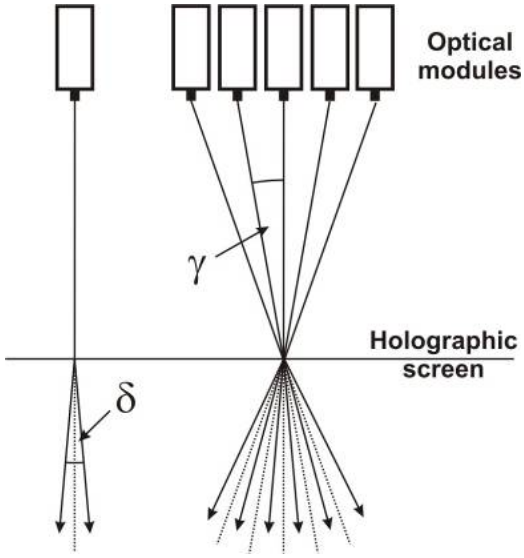
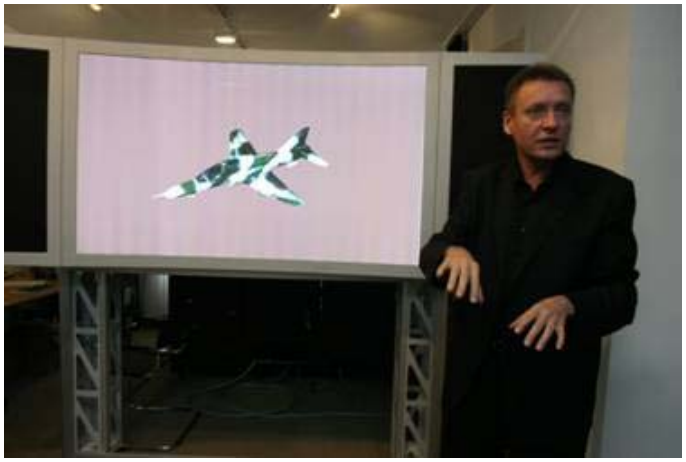



Figure 1. HoloVizio light-field generation principle. The light beams generated in the optical modules hit the holographic screen in various angles.

### 2.1 HoloVizio Displays

HoloVizio can be implemented in both small-scale and large-scale display systems. A 50 Mpixel large-scale system[3] has been developed with a screen diagonal above 1.8m. The display's optical system consists of compact projection modules arranged in horizontal rows. The system has a high angular resolution; approximately 50 independent light rays originate from each pixel. A PC-based render cluster provides 50Mpixels the display in real-time (using GPUs for most image generation tasks), a control system controls the projectors, PCs, the network, power supplies and monitors all system parameters.



Video 1. Large-scale HoloVizio light-field 3D display projecting 50Mpixels. http://dx.doi.org/doi.number.goes.here

Using HoloVizio technology is possible to build displays that have excellent image resolution of 1920x1080 or beyond, large FOV above 100 degrees, large Field-of-Depth, and at the same time the number of pixels being in the range of hundreds of millions, which demonstrate the scalability of the system very well. Being projection based, and using a high number of optical engines pointing towards the same screen, this technology will always dominate 3D display

technologies based on flat screens in terms of pixel count by at least one order of magnitude. Our desktop 3D display is available in 32" size and features 10 Mpixels[4]. This model is in the dimensions of normal TV sets. The size in between is implemented in the HoloVizio 240P display, the first HoloVizio featuring a slim optical design, which, despite being a projection based display, allows it to be only 70 cm deep, and is controlled by 3 built-in PCs.

### 2.2 HoloVizio Software

There are several possibilities for displaying 3D data on the HoloVizio, the most important is interfacing interactive graphics applications to the holographic displays through the HoloVizio OpenGL wrapper. This library is able to display existing applications without any modification, recompiling, or relinking, thus users can continue using the applications they are used to, but now in 3D. A 3D converter and video player application is also provided, which can be used to create computer generated 3D videos for the display. An application development framework to render directly to the display is also under development, which allows users willing to develop HoloVizio-enabled applications to avoid using an intermediate library, and to use the rendering node's resources arbitrarily.

As clearly visible from the above, our existing software system focused on displaying synthetic content, which was the dominant use case of HoloVizio displays. This is sufficient for professional applications that are working with 3D data, like scientific visualization, engineering, prototyping, oil&gas exploration (see Figure 2) or digital signage. However, to target the public with a 3D display, displaying live 3D content is a necessity, constituting a major step towards 3D Cinema and later on 3DTV.

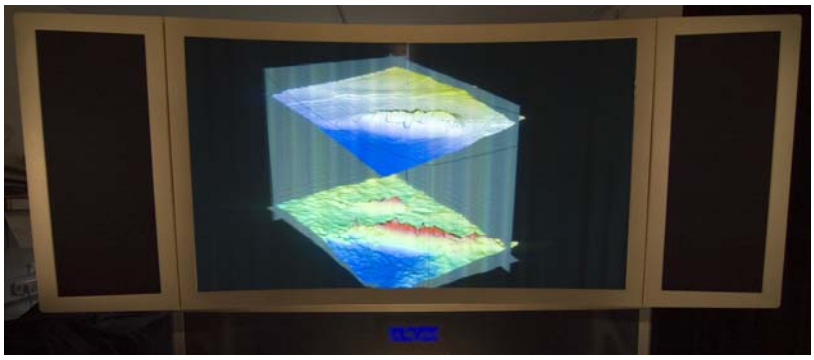


Figure 2. HoloVizio display running an oil&gas exploration application in real-time.

## 3. REAL-TIME LIGHT-FIELD CAPTURING

### 3.1 Light-field content

Not surprisingly, publicly available light-field content is not a common resource (some groups providing such content to the public are[6,7]). Using multi-view content results in suboptimal viewing experience on HoloVizios. As these are targeted for multi-view displays, usually very narrow FOV is used for capturing. Angular resolution is also lower than desirable, at least when compared to the capabilities of light-field displays, which provide both wide FOV, and fine angular resolution. The issue of angular resolution can be somewhat compensated by using Depth Based Rendering, Image Based Rendering, or a hybrid approach[8], but on the other hand, increasing the FOV is very challenging after a certain extent, and provides incorrect results. There are companies offering Time Freeze shooting (names may vary)[9], but these are headed towards the movie industry, where real-time acquisition, transmission and playback is not an objective.

### 3.2 State Of The Art

A number of papers describing real-time 3D video or light-field capture and display have been published in recent years, achieving significant advances. The random access light-field camera provides very good results with 2D or stereo displays due to selective transmission[10], but as the authors pointed out, this approach is less applicable with 3D displays which typically need access to all captured light-field data. On the other hand, as shown later, their observation still holds for light-field displays too, if the rendering process is distributed between a number of processing units. The TransCAIP system[11] uses a single PC and GPU algorithms to achieve interactive speeds with an impressive number of cameras,

however, our system provides better results in terms of resolution, frame rate, and angular resolution. Moreover, the strength of that system is that everything is handled in a single GPU (avoiding frequent bus transfers), in contrast our system is designed to be highly scalable to be able to serve a number of cameras and 3D displays with very high light-ray count. The impressive MERL 3D TV System[12] uses a symmetrical system with a high number of PCs and a lenticular-lens based display to create live 3D visuals, however the HoloVizios we used have far better 3D image quality compared to their 3D display, moreover they use excessive number of PCs for capturing, processing and rendering: 16 cameras and 16 projectors are served by 8+8 PCs. In contrast, our system could easily serve such a configuration with using only 3 PCs.

### 3.3 3D acquisition system and calibration

Our camera system is made up from 27 USB CCD cameras. In the first stages, these were connected to 9 computers, but later on this was reduced to a single capture PC. The cameras are evenly spaced in a linear arrangement on a camera rig, each one capable of capturing at 640x480@15 FPS or 960x720@10 FPS resolutions (see Figure 3).

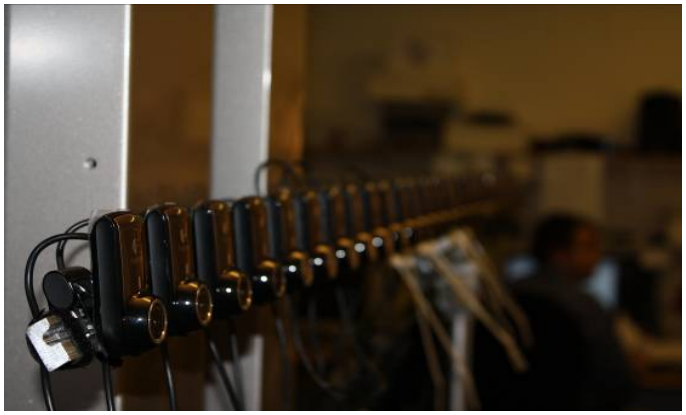


Figure 3. Camera array consisting of 27 CCD USB cameras. Together they capture up to 18MPixels.

The capture computer is directly connected to the demonstration 3D display (HV240RC) through a single gigabit Ethernet connection, supported by a 3 PC computation cluster, which is an integral part of the HoloVizio display. The camera system was calibrated off line using a semi-automatic calibration method, using images of a previously known reference object[14,15]. During the calibration, the parameters of the light field captured by the cameras are estimated based resulting in intrinsic and extrinsic camera calibration parameters. These estimates are further refined by a third refinement step to minimize the error of the estimated model, resulting in very good 3D image quality and field of depth.

The cameras can stream uncompressed YUV or MJPEG output out of which MJPEG has been chosen, as it can also be used for transmission over the network (although not very efficient).

## 4. REAL-TIME LIGHT-FIELD RENDERING

### 4.1 Incoming images

The original MJPEG images are arriving to the cluster's nodes on a single gigabit Ethernet channel with approx 10% link utilization. Each individual channel has its own CPU thread that decodes the Huffman encoding and does the inverse DCT algorithm for the incoming JPEG image. This yields a YUV image on the CPU. IDCT has also been implemented on the GPU, but that is not the real bottleneck. These images are then uploaded to GPU memory, where the light-field reordering takes place.

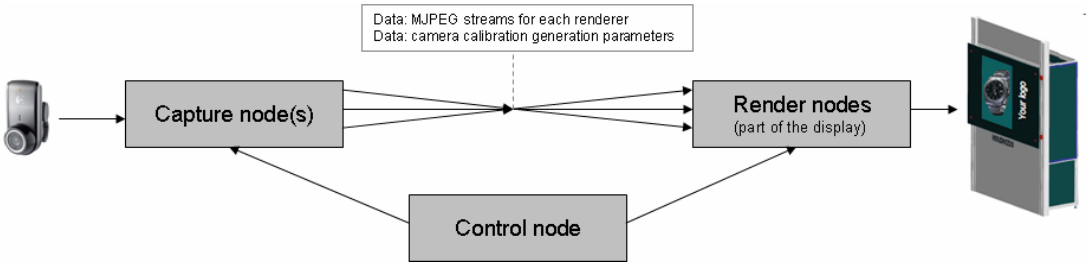



Figure 4. Architecture of the first generation light-field transmission system.

### 4.2 Light-field reordering and calibration

For the task of dense light field displaying a massively parallel reordering and filtering of the original camera images is required. On one hand, based on camera calibration information, we know which light rays of the scene are captured exactly. On the other hand, we know which light rays are emitted from the display, based on the arrangement of the optical engines and the display calibration information. Thus, we can derive which incoming light rays need to be used for the generation of each outgoing light ray. Once we have that mapping, the reordering of pixels can happen on the GPU very rapidly. On the GPU we do the light-field conversion and reordering, filtering and the YUV-RGB conversion. We use OpenGL 3.0 shaders and 2D texture arrays to have an easy access to individual camera images. This yields the correct bilinear filtering. The additional advantage is that display-specific calibration can also be handled in the same step. There is also a network-based synchronization scheme for displaying the final rendered images.


Video 2. HoloVizio 240P 3D display showing live 3D light-field content. http://dx.doi.org/doi.number.goes.here

## 5. RESULTS

Our 27 cameras light-field capture system streamed 960x720 resolution video streams with 10 FPS. The image data was converted on-the-fly to light field format and the continuous parallax 3D image stream was displayed on various HoloVizio displays. To our knowledge, this is the first system capable of displaying live 3D video with such quality.

A single PC was used for capturing 27 cameras, and 3 PCs for rendering (which are integral part of the HoloVizio display) were used. With this setup, we reached 15 frames per seconds playback speed for the 640x480 resolution stream and 10 frames per second for the 960x720 resolution stream. The bottleneck here was the camera acquisition speed.

## 6. FUTURE WORK

Although the bottleneck is camera acquisition speed now, using more and better cameras is desirable in the future. Once we do that, the two performance-critical points will be the Huffman decoding of the JPEG images and the upload speed to the GPU.

There are several ways to improve the solution to overcome these bottlenecks. We have observed that when rendering is distributed to multiple computers, none of them need all parts of all the images (not even when they serve multiple optical engines). Thus, the capture node could transmit only parts of the images needed, based on information received from the renderers. To do that, either uncompressed images, or a compressed image format that can be partially decoded should be used. The capture nodes could then transcode the MJPEG stream to this intermediate format, and transmit only parts requested from the renderers.

To transmit 3D light-field to longer distances, a more efficient compression approach is desirable. Even H.264[16] simulcast coding would help reducing the bandwidth requirement, but applying MultiView Coding (MVC[17]) could decrease the amount of transmitted image information even further.

Combining this two will result in a highly bandwidth efficient and future-proof system. A layered approach for using different encoding during transmission and light-field rendering is being developed, providing good compression on one part, and fast rendering on the other hand, allowing arbitrary number of cameras and very high resolution.

Such a system can be used to implement the most realistic 3D telepresence system ever made, which – being three dimensional and providing very fine angular resolution – also overcomes the problem of missing eye contact between participants [13].

Video 3. HoloVizio 720RC 3D display showing a telepresence situation. http://dx.doi.org/doi.number.goes.here


## ACKNOWLEDGEMENTS

The development of the 10Mpixel test display system and camera system has been supported by EU IST-FP6 Integrated Project OSIRIS (IST-33799 IP)[5]. The FP6 project OSIRIS aims to create novel display systems including a high resolution LED based compact display that is capable of real-time playback of live captured natural content. In the same project HoloVizio technology is applied to create a 3D Cinema application.

## REFERENCES


[1] Balogh, T., “Method and apparatus for displaying three-dimensional images,” U.S. Patent 6,201,565, EP 0900501, Feb 04, 1997.

[2] Balogh, T., “The HoloVizio system, ” Proc. SPIE 6055, 60550U (2006).

[3] Balogh, T., Forgacs, T., Agocs, T., Bouvier, E., Bettio, F., Gobbetti, E., Zanetti, G., ”A Large Scale Interactive Holographic Display,” IEEE VR2006, Virginia, USA.

[4] Balogh, T., et al, “A Scalable Hardware and Software System for the Holographic Display of Interactive Graphic Applications,” EuroGraphics 2005, Dublin.

[5] OSIRIS Project – Original System for Image Rendition via Innovative Screens, EU IST-FP6 Integrated Project OSIRIS (IST-33799 IP), http://www.osiris-project.eu/

[6] UCSD/MERL Light Field Repository, http://vision.ucsd.edu/datasets/lfarchive/

[7] The (New) Stanford Light Field Archive, http://lightfield.stanford.edu/

[8] Megyesi, Z., Barsi, A., Balogh, T., “3D Video Visualization On The Holovizio™ System,” 3DTV-CON 2008, Istanbul, Turkey

[9] Upstage Productions: Time Freeze, http://www.upstage.com.au/content/standard.asp?name=Time_freeze

[10] Yang, J. C., Everett, M., Buehler, C., McMillan, L., “A Real-Time Distributed Light Field Camera,” 13th Eurographics Workshop on Rendering, 2002.

[11] Taguchi, Y., Koike, T., Takahashi, K., Naemura, T., “TransCAIP: Live Transmission of Light Field from a Camera Array to an Integral Photography Display,” ACM SIGGRAPH ASIA 2008.

[12] Matusik, W., Pfister, H., “3D TV: A Scalable System for Real-Time Acquisition, Transmission, and Autostereoscopic Display of Dynamic Scenes,” ACM SIGGRAPH 2004.

[13] Lincoln, P., Nashel, A., Ilie, A., Towles, H., Welch, G., Fuchs, H., ”Multi-view lenticular display for group teleconferencing,” IMMERSCOM 2009.

[14] Zhang, Z., “A Flexible New Technique for Camera Calibration,” Technical Report MSR-TR-98-71, Microsoft Research, 1998.

[15] Tsai, R., "A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses," IEEE J. Robotics and Automation, vol. 3, no. 4, pp. 323-344, Aug. 1987.

[16] Wiegand, T., Sullivan, G. J., Bjontegaard, G., Luthra, A., "Overview of the H.264/AVC Video Coding Standard," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 13, No. 7, pp. 560-576, July 2003.

[17] Merkle, P., Muller, K., Smolic, A., Wiegand, T., "Efficient Compression of Multi-View Video Exploiting Inter-View Dependencies Based on H.264/MPEG4-AVC," IEEE International Conference on Multimedia and Expo, pp.1717-1720, 9-12 July 2006.

# VII

# 3D CAPTURING USING MULTI-CAMERA RIGS, REAL-TIME DEPTH ESTIMATION AND DEPTH-BASED CONTENT CREATION FOR MULTI-VIEW AND LIGHT FIELD AUTOSTEREOSCOPIC DISPLAYS

by

P. T. Kovacs, F. Zilly, 2012

in ACM SIGGRAPH 2012 Emerging Technologies (SIGGRAPH '12),
DOI: 10.1145/2343456.2343457

# 3D Capturing using Multi-Camera Rigs, Real-time Depth Estimation and Depth-based Content Creation for Multi-view and Light-field Auto-Stereoscopic Displays

Peter Tamas Kovacs[1], Frederik Zilly[2]

## Overview

The wide variety of commercially available and emerging 3D displays - such as stereoscopic, multi-view and light-field displays - makes content creation challenging as each displays technology requires a different number of views available of the scene. As consequence, the content creation pipelines differ considerably and involve different camera setups such as beam-splitter rigs with small baselines and high quality cameras used for stereo 3D productions or camera arrays for auto-stereoscopic displays which usually use small lower quality cameras in a side-by-side arrangement. Converting content shot for a specific display technology into a different format usually impairs the image quality and is very labor-intensive.

Against this background a generic method for capturing and rendering live 3D footage for stereoscopic, multi-view and light-field displays is presented. The system consists of a wide-baseline multi-camera rig, a camera assistance system, a real-time depth estimator, a real-time view generation and rendering engine, and multiple displays, one multi-view auto-stereoscopic and a light-field display. The system features several innovative components: the professional-grade multi-camera assistance and calibration system; a real-time depth estimator producing convincing depth maps; a real-time and generic depth-image based rendering (DIBR) engine that is suitable for generating imagery for a range of 3D displays; and the largest auto-stereoscopic light-field display to date.

## The Experience for SIGGRAPH Attendees

The exhibited system will allow attendants (watching one of the 3D displays) to see other attendants (standing in front of the multi-camera rig) in 3D, without glasses, with the possibility to walk around the perceived 3D image, experience smooth motion parallax and large depth-range. Visitors are interested in the technical details will also have the opportunity to see the camera calibration and assistance system in operation, as well as the output of the real-time depth estimation system (as gray-scale depth maps), and also other 3D content – including interactive 3D applications – on the light-field display.

## Core Technical Innovations

The first 4-camera rig built from professional HD cameras (Sony HDC-P1) is presented. In order to support a wide range of 3D displays while being backwards compatible with already established stereo displays, two of the captured views are shot using a beam-splitter which allows showing them directly on a stereoscopic 3D display without any further processing. The additional satellite cameras placed outside the mirror box provide the information that is needed to create a generic depth-based 3D representation format and content for other wide baseline applications.

The first professional-grade multi-camera calibration and assistance system is demonstrated as part of the system. The precise calibration of a multi-camera system is a demanding task as the system has many degrees of freedom. However, to keep the system easy to use and robust, a dedicated assistance system has been developed as an extension of the stereoscopic analyzer (STAN) towards a quadrifocal setup. For the multi-camera setup we bring all four cameras in a position such that each pair of two cameras is rectified. The system has shown its maturity and suitability for productions during two field trials.

We present the first real-time multi-view depth estimation system based on line-recursive matching that generates depth maps for the visualization components. We have extended the existing depth estimator Hybrid Recursive Matcher (HRM) towards parallelization. Although the HRM is able to generate depth maps in real-time for smaller resolutions, its recursive structure prevents it to take advantage of multiple CPU cores. We broke the recursive structure of the HRM and limited the recursion on a line-wise level. Thus, each line can be processed in a different thread, resulting in a significant speed-up when executed on multi-core CPUs. The estimation is applied independently for the images with subsequent filtering of left/right consistent disparities. Temporal stability avoiding flickering artifacts is achieved by incorporating temporal disparity candidates in the estimation process.

The first real-time Multi-View plus Depth (MVD) based view generation & rendering system targeted for wide-baseline light-field displays is presented. The view generator renders interpolated views (between original cameras) as well as heavily extrapolated (outside the original cameras) novel views. The interpolation process detects and keeps gap area information from the content using depth layers. Extrapolation is hierarchical, using each image from the closest to the furthest. Holes are filled using information coming from the other images, where available. Inpainting techniques are used where no information is available, during which texture and structure is rendered, propagating contour gradients with prioritized matching costs.

The largest glasses-free light-field 3D display to date (140” screen diagonal) will be shown. The display presents natural 3D light-field to a larger audience on a cinema-sized screen size previously not possible with auto-stereoscopic displays. The display itself consists of a complex hardware and software system, being the first front-projected light-field 3D display, controlling 63 Mega-Pixels in total. It consists of an array of optical engines, projecting light rays onto a reflective holographic screen, in front of which viewers can see 3D content with an exceptionally wide Field Of View and depth range.

## The Future of this Work

Today’s glasses-based stereoscopic 3D display systems can be seen as stepping stones towards more advanced 3D display technologies. The generic Multi-View plus Depth (MVD) representation used inside the system can serve as the future 3DTV format, which is generic enough to drive a multitude of 3D displays, independent of the underlying technology. The presented approach is also in line with MPEG’s efforts towards future 3DTV formats.

## Acknowledgments

The demonstrated system is based on work partially supported by the MUSCADE European FP7 project (EU-FP7-247010).

[1] Holografika
[2] Fraunhofer HHI

# VIII

## ANALYSIS AND OPTIMIZATION OF PIXEL USAGE OF LIGHT FIELD CONVERSION FROM MULTI-CAMERA SETUPS TO 3D LIGHT FIELD DISPLAYS

by

P. T. Kovács, K. Lackner, A. Barsi, V. K. Adhikarla, R. Bregović, A. Gotchev, 2014

in Proc. IEEE International Conference on Image Processing (ICIP) 2014

# ANALYSIS AND OPTIMIZATION OF PIXEL USAGE OF LIGHT-FIELD CONVERSION FROM MULTI-CAMERA SETUPS TO 3D LIGHT-FIELD DISPLAYS

*Péter Tamás Kovács* [1, 2], *Kristóf Lackner* [1, 2], *Attila Barsi* [1], *Vamsi Kiran Adhikarla* [1,3], *Robert Bregović* [2], *Atanas Gotchev* [2]

[1]Holografika, Budapest, Hungary
[2]Department of Signal Processing, Tampere University of Technology, Tampere, Finland
[3]Pazmany Peter Catholic University, Faculty of information Technology, Budapest, Hungary

## ABSTRACT

Light-field (LF) 3D displays require vast amount of views representing the original scene when using pure light-ray interpolation to convert multi-camera content to display-specific LF representation. Synthetic and real multi-camera setups are both used to feed these displays with image-based data, however the layout, number, frustum, and resolution of these cameras are mostly suboptimal. Storage and transmission of LF data is an issue, especially considering that some of the captured / rendered pixels are left unused while generating the final image. LF displays can have significantly different requirements for camera setups due to differences in Field of View (FOV), angular resolution and spatial resolution. An analysis of typical camera setups and LF display setups, and the typical patterns in pixel usage resulting from the combination of these setups are presented. Based on this analysis, an optimization method for virtual camera setups is proposed. As virtual cameras have wide range of adjustment possibilities, highly optimized setups for specific displays can be achieved.

***Index Terms***— 3D display, light-field display, multi-camera capture, camera rig optimization

## 1. INTRODUCTION

The current generation of Light-field (LF) 3D displays [1][12][13][14] typically reconstruct the equivalent of 100+ viewing directions, and up to 100 million light rays today. One of the possible input formats for such 3D displays is a multitude of images, captured by means of real or virtual cameras. By using a multi-camera setup one can capture or render the necessary number of images as well as estimate or calculate camera calibration information that allows transforming the camera pixels into a common 3D space, and consider the pixels as light-rays captured by the cameras. As there is no direct correspondence between cameras and imaging components in the LF display (even if the number of cameras matches the number of light ray emitters [2]), correspondences between light rays emitted by the display and light rays captured by cameras have to be found, and based on these correspondences, ray interpolation is used to generate the light rays generated by the display. In this case a large number of input views is assumed, and thus additional information like depth maps are not used to perform view synthesis.

In this study, Horizontal Parallax Only (HPO) LF displays are used. Such displays show different views of the scene when the viewer is moving horizontally in front of the display, however the image does not change with vertical movements. This implies that the considered capture setups do not have vertical displacements either.

In Section 2, analysis of previous work on camera setups is presented, mostly in relation with stereoscopic and multiview (MV) displays. Section 3 shows why LF displays require different, most notably wider capture setups. It also describes typical camera setups in use today for content creation. Section 4 provides an analysis of how many pixels are actually utilized during a typical LF conversion process. These lead us to the discussion of ideal camera setups in Section 5. Such ideal camera setups are only applicable in the synthetic case, and even then, only practical in special cases. Therefore, Section 6 discusses our requirements for practical camera setups, thus defining the constraints and search space that are utilized in Section 7 to generate optimized camera setups. In Section 8, the effectiveness of the optimized camera setups are analyzed, and Section 9 concludes the paper.

## 2. RELATION TO PRIOR WORK

Content creation for stereoscopic 3D displays is mostly concerned with providing the human visual system with the two images directly presented to the eyes [3] without causing discomfort. Stereoscopic 3D shooting is nowadays assisted with automatic tools to ensure that the cameras are properly aligned, that disparity range and convergence are

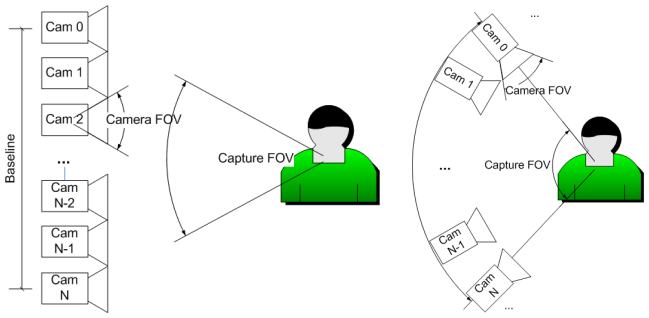



**Figure 1: Left: Linear camera setup Right: Arc camera setup**

within the desired limits [4], and tools that assist the adjustment of stereoscopic content after it has been shot [5].

MV 3D displays present multiple views directly (two of which are visible at the same time with the two eyes), thus, generating the necessary views for such displays is similar to stereoscopic content creation in the sense that two selected views, which are supposed to be seen at the same time, should obey the same rules as stereoscopic content. However the views are created over a bigger baseline (the distance between leftmost and rightmost camera) [6][7].

Current LF content creation and conversion tools [8][9] use multiple-camera setups (linear, arc), and perform ray interpolation based on the multitude of images and their supplementary information of the cameras used (position, orientation, FOV, distortion), generating a display-specific LF. However, the authors are not aware of any literature that discusses camera setups used for content generation for LF displays.

## 3. LF DISPLAY AND CAMERA SETUPS

LF displays require substantially different content compared to those required by stereoscopic and MV displays. The reasons for this are twofold.

First, there is a large difference in viewing angles: due to the displacement between human eyes, the viewing angle reproduced by stereoscopic displays is typically fairly small (few degrees). MV displays, as they provide some amount of motion parallax, reproduce a slightly wider viewing angle, thus requiring a bigger capture baseline. However, this baseline is far from the baseline required by LF displays that have viewing angles between 45º and 180º, 70º being a typical value. To avoid view extrapolation, capture setups shall reflect this wide angle, resulting in wide capture setups. In practice, camera setups based on rule of thumb have been used. This typically means a linear or arc camera setup (see Figure 1), the capture FOV which roughly corresponds to the display's FOV, and the number of cameras matching or closely matching the number of directions reproduced by the LF display. In this case, capture FOV means the angle between the leftmost and rightmost cameras as visible from the scene center, as

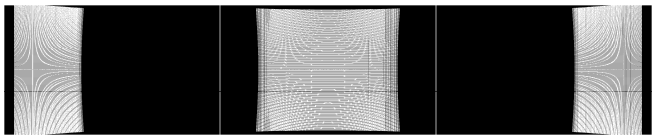


**Figure 2: Pixels used during LF conversion from camera number 15, 45 and 75, respectively. Camera setup is 45º arc with 0.5º angular resolution (91 cameras), LF display has 45º FOV. Pixels marked with white are used.**

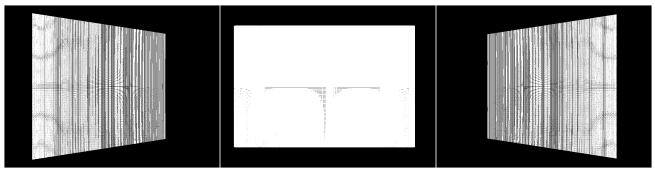


**Figure 3: Pixels used during LF conversion from camera number 45, 90 and 135, respectively. Camera setup is 180º arc with 1º angular resolution (181 cameras), LF display has 180º FOV.**

opposed to the opening angle of the individual cameras. Typical examples include a 112-camera linear rig, a 180-degree, 180-camera arc rig, and a 45 degree 90-camera rig, each targeting specific LF display layouts.

Second, there is no direct correspondence between cameras and viewing directions, thus the captured views are not used "as is", but have to go through ray interpolation.

## 4. ANALYSIS OF PIXEL USAGE

As it has been noticed [2] and exploited [11] earlier, such simple camera setups result in suboptimal usage of pixels. Typically only portions of the rendered images are actually used for generating the displayed light field, and the rest of pixels are left unused. Figure 2 and Figure 3 show some typical patterns of pixel usage. Pixels which are used are shown in white, while black areas are unused. From these figures it is clear that the ratio of used pixels is relatively low especially for the side cameras, and thus it is expected that the camera layouts could be improved to enhance the efficiency of both real and synthetic content generation in terms of having the best possible ratio of used pixels.

## 5. IDEAL CAMERA SETUPS

Based on the above, an ideal camera setup would be one which has many single-pixel sensors exactly matching each emitted ray, and capturing light in the required direction.

Such a set of pixels / light rays can be calculated in a virtual environment provided that rendering single pixels does not have much overhead (for example, via ray tracing [15]). However it is unusual to use such a complex camera setup in a rendering tool due to the scene set-up time typically associated with rendering each image (even when that image is a single pixel). When using real cameras to capture live scenes, such setups consisting of millions of

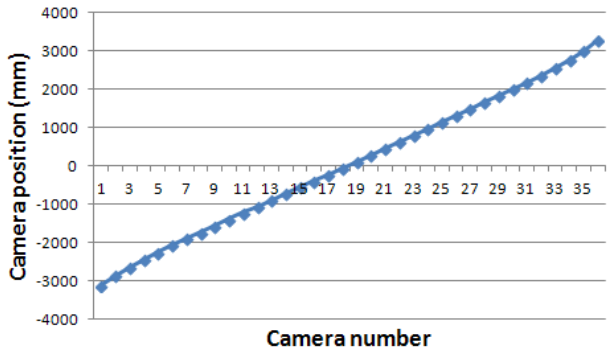



**Figure 4: Horizontal position of 36 cameras after optimization for a sample 36-channel LF display**

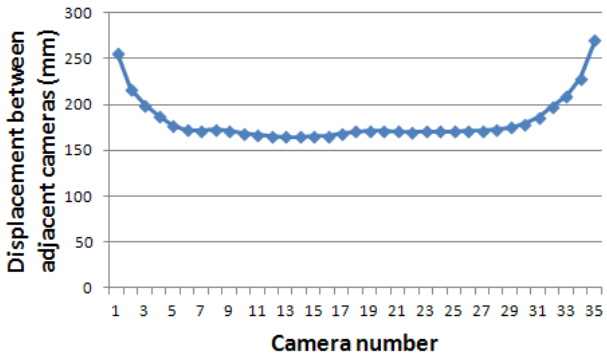



**Figure 5: Horizontal displacement between 36 cameras after optimization for a sample 36-channel LF display**

## 6. PRACTICAL CAMERA SETUPS

distinct sensors are clearly out of scope. The desired camera setups rather consist of some tens to hundred cameras (on the order of the number of the directions reconstructed by the display), but are arranged so that the captured pixels are better utilized, and also better match the individual displayed rays, resulting in less interpolated values.

Autodesk 3ds max has been used as a use case for synthetic content generation, as it is commonly used for creating content for 3D displays. While it is not possible to render arbitrary rays by overriding the ray generation step in the rendering engines bundled with 3ds max, it is possible to generate a camera setup consisting of an arbitrary number of cameras, each having custom (even highly asymmetric) FOV and custom resolution. These obviate the need to move the cameras out of the linear setup, as the same effect can be achieved by adjusting the FOV. Such arrangements can be utilized to create highly optimized synthetic camera rigs that render the 3D scene from multiple viewpoints.

## 7. OPTIMIZING CAMERA SETUPS

As a closed form solution for optimal camera positions could not be derived for real LF displays, optimization has been performed for synthetic camera setups. The set of rays emitted by the display in question has been generated. For simplicity, 1 to 1 matching between the display's physical space and the captured real or synthetic scene is assumed (that is, the scene is depicted in its real scale). A linear camera system model is used with a fixed number of cameras, optimize camera positions and calculate all other parameters. The rays to be captured are those that start from the display and cross the line where the viewer is assumed to be moving [10].

The objective function is defined so that for each displayed ray, the closest matching camera is found, and the distance determined. The sum of squared differences for the distance between all rays and the closest camera is minimized. The ParadisEO metaheuristics framework [16] has been utilized to implement a genetic algorithm that, starting from a large population of randomized linear camera setups, finds the optimal solution that best satisfies the objective function. As the vector describing camera positions is real valued, the genetic algorithm uses mutations to shift the camera positions, and crossovers to combine camera setups. To make convergence faster, the amount with which positions are updated is gradually decreased during the optimization. That is, cameras are randomly displaced with a bigger offset, and when the optimization cannot generate a better population over many iterations, the extent of shifts is decreased, similar to simulated annealing.

The optimized camera positions reflect the slightly higher ray density in the center of the FOV, thus cameras are placed more densely in this area. Figure 4 shows the horizontal positions of the cameras, while Figure 5 shows the displacement between adjacent cameras over the linear rig. This suggests that camera setups better than the equidistant linear setup can be found, and that cameras can be slightly sparser at the sides compared to the center.

After the position of cameras is optimized, the FOV of cameras is determined so that their leftmost and rightmost rays horizontally enclose the display's screen, as rays outside that area are clearly not used. An example of such a setup with calculated FOVs is shown on Figure 6. The vertices on the top of the figure represent the 36 cameras placed, the blue lines represent the edges of the cone captured by each camera, and the red box is the volume of the display. As visible from the figure, the sides of the captured frustum correspond to the sides of the screen of the LF display.

Once the geometry of the capture cameras is found, the desired resolution of the cameras is determined to avoid oversampling the scene by rendering many pixels, which are then left unused. If the resolution of the rendered image is higher than necessary, the unused pixels appear as holes in the pixel usage patterns shown on Figure 2 and Figure 3. The same tool that generates pixel usage maps is capable of determining how many times each pixel has been used during the LF conversion process. If pixels are read multiple times, that is caused by the resolution of the rendered image

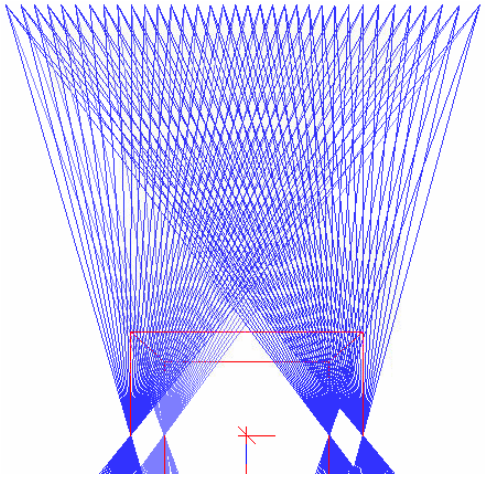


**Figure 6: Camera FOVs enclosing the LF display's screen. The top-left, top-right, bottom-left and bottom-right captured ray of each camera is visualized**

being lower than desired. Therefore, by summing the number of used pixels over a horizontal line of the camera, a good approximation of the needed horizontal camera resolution can be found,and the virtual camera can be configured to render just as many pixels as necessary.

## 8. RESULTS

The resulting camera setups can improve the utilization of rendered and captured pixels during LF conversion, therefore improving the efficiency of the content rendering process. As shown on Figure 7, the utilization of pixels is high even in side cameras after the optimization, which is visible from the lack of solid black areas.

The generated camera setup has been created with the assumption that the physical scene is represented in its full size. However, LF displays can be used to visualize scenes of different physical sizes, which are controlled during the LF conversion process with the Region of Interest (ROI) box: the ROI box represents the volume that is reconstructed by the 3D display. Assuming that the aspect ratio of the ROI box does not change, the optimized camera setup determined for the display's real screen size can be used for capturing scenes with a different scale after rescaling the camera system, therefore the camera setup need to be calculated only once for a specific LF display, and can be applied to any scene.

## 9. CONCLUSIONS AND FUTURE WORK

The presented results will be included in our LF content generation tool chain [8], complementing the 3ds max tools with optimized non-equidistant linear camera setups with sheared frustums, enabling shorter rendering times for synthetic content. The natural next step is to extend our analysis to real camera setups. In capture setups involving

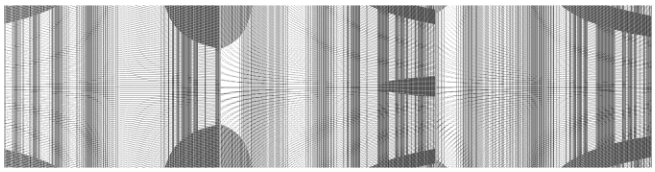


**Figure 7: Pixels used during LF conversion from the optimized 36-camera setup. Camera number 9, 20 and 30 are shown.**

real cameras, optimization possibilities are more restricted. The number of cameras is scalable, but within budgetary limits, and resolution of the cameras has an upper limit imposed by the resolution of the sensor.

With cameras having a lens mount, FOV can be selected from a set of available lenses, however the FOV is typically symmetric and equal among all cameras. While cropping the captured images with Area Of Interest setting is possible in some cameras (thus creating smaller capture frustums), this does not bring practical benefits in our case. In the past a 27-camera rig has been used [2] for live LF capture, which has been assembled to form a 1.5m wide, equidistant, parallel linear rig. The number, resolution and FOV of these cameras is fixed, but optimizing the placement of these cameras (position and orientation) is planned to provide a better coverage for specific LF displays. In case of real cameras, using an arc (or other nonlinear) camera setup can be advantageous to increase the coverage of the cameras to compensate for the lack of custom FOV settings. In case of modeling and optimizing rigs of real cameras, more complex modeling and optimization is necessary, also considering the vertical position and direction of rays.

The approach used for optimizing synthetic camera setups does improve the utilization of pixels, however future work should prove that this also results in improved perceived image quality for the same amount of cameras. As an extreme example, if the number of cameras is small, and the camera spacing becomes bigger than a stereo baseline, viewers may prefer having the small number of cameras in the center, and lose the sides of the FOV, instead of not experiencing a stereoscopic effect at all. Subjective tests will be performed to check that the perceived quality of rendered LFs does improve with the optimized camera setups, keeping the number of rendered pixels constant.

## 10. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the PROLIGHT-IAPP Marie Curie Action of the People programme of the European Union's Seventh Framework Programme, REA grant agreement 32449.

The research leading to these results has also received funding from the DIVA Marie Curie Action of the People programme of the European Unions Seventh Framework Programme under REA grant agreement 290227.

## 11. REFERENCES


[1] T. Balogh, "The HoloVizio system," Proc. *SPIE 6055, Stereoscopic Displays and Virtual Reality Systems XIII*, 60550U (January 27, 2006). doi:10.1117/12.650907

[2] T. Balogh, P. T. Kovács, "Real-time 3D light field transmission". *Proc. SPIE 7724, Real-Time Image and Video Processing* 2010, 772406 (May 04, 2010); doi:10.1117/12.854571.

[3] F. Zilly, J. Kluger, P. Kauff, "Production Rules for Stereo Acquisition," *Proceedings of the IEEE*, vol.99, no.4, pp.590,606, April 2011. doi: 10.1109/JPROC.2010.2095810

[4] F. Zilly, K. Muller, P. Eisert, P. Kauff, "The Stereoscopic Analyzer — An image-based assistance tool for stereo shooting and 3D production," Image Processing (ICIP), 2010 17th IEEE International Conference on , vol., no., pp.4029,4032, 26-29 Sept. 2010. doi: 10.1109/ICIP.2010.5649828

[5] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, M. Gross, "Nonlinear disparity mapping for stereoscopic 3D". In *ACM SIGGRAPH 2010* papers (SIGGRAPH '10), Hugues Hoppe (Ed.). ACM, New York, NY, USA, , Article 75 , 10 pages. doi: 10.1145/1833349.1778812

[6] A. Boev, K. Raunio, M. Georgiev, A. Gotchev, K. Egiazarian, "Opengl-Based Control of Semi-Active 3D Display," *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, 2008 , vol., no., pp.125,128, 28-30 May 2008. doi: 10.1109/3DTV.2008.4547824

[7] C. González, J. Martínez Sotoca, F. Pla, M. Chover, "Synthetic content generation for auto-stereoscopic displays", In *J. Multimedia Tools and Applications*, Feb. 2013. doi: 10.1007/s11042-012-1348-x

[8] Holografika Software System. http://www.holografika.com/Software-and-system-compatibility/Software-system-and-compatibility.html Visited 06 June 2014

[9] J. Park, D. Nam, S. Y. Choi, J.-H. Lee, D. S. Park, C. Y. Kim, "Light field rendering of multi-view contents for high density light field 3D display". *SID Symposium Digest of Technical Papers*, 44: 667–670, 2013. doi: 10.1002/j.2168-0159.2013.tb06300.x

[10] A. Said, B. Culbertson, "Virtual object distortions in 3D displays with only horizontal parallax," *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, vol., no., pp.1,6, 11-15 July 2011. doi: 10.1109/ICME.2011.6012200

[11] V. K. Adhikarla, ABM T. Islam, P. T. Kovács, O. Staadt, "Fast and Efficient Data Reduction Approach for Multi-Camera Light-Field Display Telepresence Systems". In *Proceedings 3DTV Conference*. October 2013

[12] J.-H. Lee, J. Park, D. Nam, S. Y. Choi, D.-S. Park, C. Y Kim, "Optimal Projector Configuration Design for 300-Mpixel Light-Field 3D Display", *SID Symposium Digest of Technical Papers*, 44: 400–403. doi: 10.1002/j.2168-0159.2013.tb06231.x

[13] G. Wetzstein, D. Lanman, M. Hirsch, R. Raskar, "Tensor Displays: Compressive Light Field Synthesis using Multilayer Displays with Directional Backlighting", In *Proc. SIGGRAPH 2012*

[14] D. Lanman, D. Luebke, "Near-Eye Light Field Displays", In *ACM Transactions on Graphics (TOG)*, Volume 32 Issue 6, November 2013 (Proceedings of SIGGRAPH Asia), November 2013

[15] J. A. Iglesias Guitián, E. Gobbetti, F. Marton, "View-dependent Exploration of Massive Volumetric Models on Large Scale Light Field Displays". *The Visual Computer*, 26(6--8): 1037-1047, 2010

[16] S. Cahon, N. Melab and E-G. Talbi, "ParadisEO: A Framework for the Reusable Design of Parallel and Distributed Metaheuristics", *Journal of Heuristics*, vol. 10(3), pp.357-380, May 2004.

# IX

# OVERVIEW OF THE APPLICABILITY OF H.264/MVC FOR REAL-TIME LIGHT FIELD APPLICATIONS

by

P. T. Kovács, Z. Nagy, A. Barsi, V. K. Adhikarla and R. Bregović, 2014

in Proc. 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON 2014), Budapest
DOI: 10.1109/3DTV.2014.6874744

# OVERVIEW OF THE APPLICABILITY OF H.264/MVC FOR REAL-TIME LIGHT-FIELD APPLICATIONS

*Péter Tamás Kovács [1, 2], Zsolt Nagy [1], Attila Barsi [1], Vamsi Kiran Adhikarla [1,3], Robert Bregović [2]*

[1]Holografika, Budapest, Hungary
[2]Department of Signal Processing, Tampere University of Technology, Tampere, Finland
[3]Pazmany Peter Catholic University, Faculty of information Technology, Budapest, Hungary

**ABSTRACT**

Several methods for compressing light-fields (LF) and multiview 3D video content have been proposed in the literature. The most widely accepted and standardized method is the Multi View Coding (MVC) extension of H.264, which is considered appropriate for use with stereoscopic and multiview 3D displays. In this paper we will focus on light-field 3D displays, outline typical use cases for such displays, analyze processing requirements for display-specific and display-independent light-fields, and see how these map to MVC as the underlying 3D video compression method. We also provide an overview of available MVC implementations, and the support these provide for multiview 3D video. Directions for future research and additional features supporting LF video compression are presented.

***Index Terms*** **—** light-field, 3D video, compression, multi-view coding, MVC, H.264

## 1. INTRODUCTION

Future 3D displays will go far beyond stereoscopic and multi-view, as demonstrated in currently existing prototype and commercial 3D displays [1][2][3]. Some of the existing displays aim to reproduce light-fields having both horizontal and vertical parallax, while others omit vertical parallax in order to provide better resolution and higher number of viewing directions horizontally, typically resulting in wider horizontal Field Of View (FOV) for the same number of light rays.

Wide-angle LF displays may have hundreds of viewing directions, but typically only in the horizontal direction (Horizontal Parallax Only, HPO). To achieve wide field-of-view and still maintain a reasonable resolution, these displays operate with large pixel counts (nowadays, up to 100 megapixels). The storage, compression, transmission and rendering of light-fields of this size is a major challenge, which needs to be solved to pave the way towards the wide adoption of such advanced 3D display technologies.

There have been a lot of effort directed towards supporting 3D displays with effective 3D video compression standards [4][5]. In this paper we give an insight into the computational background of LF displays, and analyze how the results of standardized 3D video coding technology can be exploited. Based on this analysis, we identify areas that need attention in future research in 3D LF video coding. In this paper we focus on H.264/MVC, since that is the current accepted standard for coding 3D video data, and is more likely to have mature implementations than work-in-progress 3D HEVC.

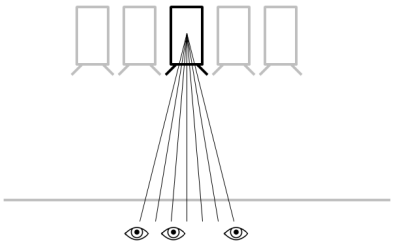


Figure 1. Light rays emitted by a single projection module are spread over screen positions and viewing directions, thus cannot be seen from a single viewing position

## 2. LF DISPLAY ARCHITECTURE

We focus our discussion on HoloVizio light-field displays [1], but the results presented in this paper are directly applicable to any LF display that is driven by a distributed projection and rendering system. Considering the gap between pixel / light ray counts and the rendering capacity available in a single computer / GPU, using a distributed rendering system for these systems is a necessity today and in the foreseeable future. Therefore LF displays are typically driven by multiple processing nodes.

LF displays are capable of providing 3D images with a continuous motion parallax on a wide viewing zone, without wearing glasses. Instead of showing separate 2D views of a 3D scene, they reconstruct the 3D light field as a set of light rays. In most LF displays this is achieved by using an array of projection modules emitting light rays and a custom made holographic screen. The light rays generated in the projection modules hit the holographic screen at different points and the holographic screen makes the optical transformation to compose these light rays into a continuous 3D view. Each point of the holographic screen emits light rays of different color to various directions.

Light rays leaving the screen spread in multiple directions, as if they were emitted from points of 3D objects at fixed spatial locations. However, the most important characteristic of this distributed projection architecture is that the individual projection modules do not correspond to discrete perspective views, in the way *views* are defined in a typical multi-view setting. What the projection modules require on their input depends on the exact layout of the LF display, but in general, a single projection module is responsible for light rays emitted at different screen positions, and in different directions at all those positions. The whole image projected by a single projection module cannot be seen from a single viewing position, as shown on Figure 1. As such, one projection module represents a *LF slice*, which is composed of many image fragments that will be perceived from different viewing positions.

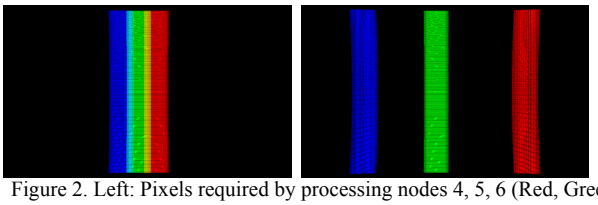


Figure 2. Left: Pixels required by processing nodes 4, 5, 6 (Red, Green and Blue channels). Right: Pixels required by processing nodes 0, 5, 9 (Red, Green and Blue channels)

Although these LF slices can be composed based on the known geometry of a multi-camera setup and the geometry of the LF display, this mapping is nonlinear and typically requires accessing light rays from a large number of views, even when generating the image for a single projection module.

The layout of the typical rendering cluster, made up of processing nodes (nodes for short), is such that a single computer is attached to multiple projection modules (2, 4, 8 or more), and as such, a single computer is responsible for generating adjacent LF slices. During LF conversion, individual nodes do not require all the views, nor all the pixels from these views. Although there is some overlap between the camera pixels required by nodes, those that are responsible for distant parts of the overall light-field require a disjoint set of pixels from the camera images.

To demonstrate this arrangement visually, Figure 2 shows which parts of the input perspective views are actually required for generating specific LF slices. A simulation has been run on a 45º large-scale light-field display with 80 projection modules, which has 10 processing nodes for generating the light-field. The display has been fed with 91-view input. What we can see is that adjacent processing nodes use adjacent, somewhat overlapping parts of the views, while processing nodes that are further away in the sense of LF slices will require completely different parts of the same view to synthesize the light field. These results are shown for the central camera, the pattern for other views is similar.

## 3. USE CASES

Two general use cases are defined to evaluate the applicability of specific 3D video coding tools, as the requirements imposed by these use cases are substantially different. The use cases identified by MPEG [6][7] can be classified into one of these, depending on whether the content is stored / transmitted in a display-specific or display-independent format. In both use cases, the requirement for real-time playback (as seen by the viewers) is above all other requirements.

The first and least demanding use case is *playback of pre-processed LF* content. In this case content has been prepared for a specific LF display model in advance, and must be played back in real time. In this setting the content is stored in *display specific LF* format. Display specific LF means the light rays are stored in a way that the individual slices of the full LF already correspond to the physical layout (projection modules) of the display on which the content should be played back. In other words, the LF in this case has already gone through the ray interpolation step that transforms it from camera space to display space. The implication is that the LF slices correspond to the layout of the distributed system driving the LF display, and as such, no ray interpolation is needed during playback, and no image data needs to be exchanged between nodes. As an example, in case of an 80-channel LF display, we may consider this data to be 80 separate images or videos making up a 3D image or video, for example 80 times WXGA (~78 MPixels).

The second use case we consider is *broadcast LF video* transmission, with the possibility to target different LF displays. 3D LF displays can differ in multiple properties, but spatial resolution and FOV have the most substantial effect on the content. The goal is to support different LF displays with the same video stream in a scalable way. In order to support different displays, we need to use *display independent LF*, which is not parametrized by display terms, but using some other terms (for example capture cameras), which is subsequently processed on the display side during playback. In this paper we consider this display independent LF to be a set of perspective images representing a scene from a number of viewpoints. Please note there are many other device-independent LF representations which lay between these two, however these two are the closest to practical hardware setups (camera rigs and LF displays).

The analysis that follows focuses on the decoder / display side, and does not consider encoder complexity.

## 4. PROCESSING DISPLAY-SPECIFIC LIGHT-FIELDS

In this case, as LF preprocessing is performed offline, the encoding process is not time critical, i.e. there is no real-time requirement for the encoder. Visual quality should be maximized wrt. bitrate, to be able to store the largest amount of LF video. On the decoding side, the goal is to be able to decompress separately the LF slices that correspond to the individual projection engines contained in the display, in real-time. The simplest solution to this problem is simulcoding all the LF slices independently using a 2D video codec (ie. H.264), and distribute the decoding task to the processing nodes corresponding to the mapping between processing nodes and projection engines. Take 80 optical engines and 10 nodes as an example: if all nodes are able to decompress 8 videos in real-time, simultaneously, we have a working solution (provided we can maintain synchronized playback). The complexity of H.264 decoding typically allows running several decoders on a high-end PC, and 25 FPS can be achieved. This solution is currently used in production LF displays.

However, in this case we do not exploit similarities between the LF slice images which have similar features, like multiview imagery. On the other extreme, compressing all 80 LF streams with MVC would require that a single processing node can decompress all of them simultaneously in real-time, which is typically prohibitive. The complexity of MVC decoding is expected to increase linearly with the number of views in terms of computing power. Furthermore it also requires a larger Decoded Picture Buffer (DPB) depending on the number of views. Assuming that having enough RAM for the DPB is not an issue, decoding a 80-view MV stream on a single node in real-time is still an issue, especially as there is no real-time implementation available that can perform this task (see Section 7). Even considering parallelization techniques [8], decoding all views in real-time on a single node is out of reach.

A reasonable tradeoff is to compress as many LF module images that are mapped to a single processing element, and do this as many times as necessary to contain all the views. As an example, we may use 10 separate MVC streams, each having 8 LF slices inside. We can increase the number of views contained in one MVC stream as long as a single processing node can maintain real-time decoding speed.

## 5. PROCESSING DISPLAY-INDEPENDENT LIGHT-FIELDS

As discussed in Section 2, and in [9], not all views are required for interpolating a specific LF slice, and even from these views, only parts are required to generate the desired LF slice – some regions of the camera images might even be left unused.

| FOV (degrees) | 27 | 38 | 48 | 59 | 69 | 79 | 89 |
|---|---|---|---|---|---|---|---|
| No. views used | 42 | 44 | 46 | 48 | 50 | 52 | 54 |

Table 1. Number of views used overall for LF synthesis when targeting LF displays with different FOV.

To find out how much we can bound the number of views and pixels to be compressed, we may determine the images and image regions which are actually used during the LF interpolation process, and compress only those for the targeted display. However, assuming receivers with displays with different viewing capabilities makes such an approach impractical, and requires scalability in terms of spatial resolution and FOV. Difference in spatial resolution might be effectively handled by SVC, and is not discussed further here. The differences in FOV however have not been addressed, as studies on the effect of display FOV on the source data used for LF conversion have not been performed so far.

We have performed simulations to see how the FOV of the receiver's LF display affects the way the available captured views are used. We have modeled 7 hyphotetical LF displays, with the FOV ranging between 27º and 89º. Source data with 180 cameras, in a 180º arc setup, with 1 degree angular resolution has been used. Using the tool from [9] and analyzing the pixel usage patterns, we have analyzed how the display's FOV affects the number of views required for synthesizing the whole LF image. This analysis has shown that depending on the FOV of the display, the LF conversion requires 42 to 54 views as input for these sample displays, as seen in Table 1. Please note the actual number depends on the source camera layout (number and FOV of cameras), but the trend is clearly visible.

Looking at the images representing the pixels read from each view also reveals that for most views, only small portions of the view are used, which is especially true for side views. This can be intuitively seen if we consider a 3D display with a wide viewing angle, looking at the screen from a steep angle. In this case, we can only see a narrow image under a small viewing angle – this is also what we need to capture and transmit. This observation suggests that any coding scheme targeting multi-view video on LF displays should be capable of encoding multiple views with different resolution. In case of HPO LF displays, only the horizontal resolution changes. In full parallax setups, both horizontal and vertical resolutions change. Such flexibility is not supported by MVC.

Due to the fact that distributed processing nodes are responsible for different parts of the overall LF, these units require different parts of the incoming views (as seen in Section 2). Thus we may expect that the number of views necessary for one node is lower than for the whole display. Further analyzing pixel usage patterns and separating the parts required by distinct nodes, we can see that this number is indeed lower, however not significantly lower. For example, in case of the 89° FOV display, instead of the 54 views required for the whole LF, one node requires access to 38 views on average, which is still high - decompressing these many full views is a challenge.

As seen previously, not all pixels from these views are necessary to construct the LF. If we look at the patterns showing which regions of the views captured by the cameras are used for the LF conversion process when targeting LF displays with different FOVs, we can see that the area is pointing to the scene center, and is widening with the increased FOV, see Figure 3.

This property may be used to decrease the computational complexity of decoding many views, by decoding only regions of interest for the specific display. H.264 supports dividing the image into regions to distinctly decodable regions using slice groups, however this feature is typically targeted to achieve some level of parallelism in the decoding process. By defining individually decodable slice groups that subdivide the image into

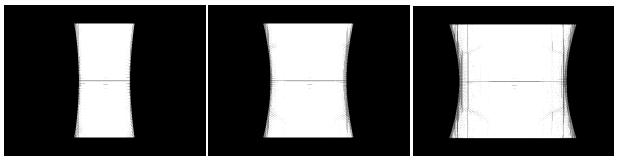


Figure 3. Image regions used from the central camera, by the 27º (left), 59º(center) and 89º (right) LF displays.

vertical regions, and decoding only those required, it is possible to decrease the time required to decode the views. Defining several slice groups would give enough granularity to target a wide range of displays with little overhead.

On the other hand, by separating views into vertical slices, we lose some coding gain due to motion estimation / compensation not going across slice boundaries. Some of this loss might be recovered by using prediction from the center of views to the sides, however such hierarchies are not supported. Exploiting this possibility is an area of future research.

## 6. NONLINEAR CAMERA SETUPS

With the emergence of LF displays with extremely wide FOV, it is more and more apparent that an equidistant linear camera array cannot capture the visual information necessary to represent the scene from all around. A more suitable setup is an arc of cameras, facing the center of the scene. Compressing such captured information with MVC should also be efficient, as the views captured in this manner also bear more similarity than views captured by a linear camera array.

However, the kind of pixel-precise inter-view similarity that MVC implicitly assumes only exist when using parallel cameras on a linear rig, and assuming Lambertian surfaces. It has been shown [10] that the coding gain from inter-view prediction is significantly less for arc cameras than for linear cameras.

Due to the emergence of wide-FOV 3D displays it is expected that non-linear multiview setups will be more significant in the future. Coding tools to support the efficient coding of views rotating around the scene center should be explored, and the similarities inherent in such views exploited for additional coding gains.

## 7. OVERVIEW OF MVC IMPLEMENTATIONS

The features discussed above can be embedded into the systems supporting LF displays if there exists implementations that support real-time operation.

MVC is the compression method of choice for 3D Blu-ray disks, where it is used for encoding the stereoscopic pair more efficiently than simulcasting the two views. Due to this widespread use of the Stereo High Profile of MVC, there are several implementations supporting it. However, support for real-time encoding and decoding of Multiview High Profile with more than two views is very weak, practically nonexistent.

JM 18.6 [11], the latest H.264/AVC reference software supports MVC, but only up to 2 views, which seems to be a hard coded limit. On the other hand it supports the specification of GOP structure explicitly, thus by interleaving frames from multiple views, it is possible to use it for inter-view prediction. It further allows the specification of arbitrary slice groups. Being a reference implementation however, its performance is typically below real-time. When running a single instance of the encoder / decoder, multiple CPU cores are not utilized, however it is possible to run parallel instances of the encoder / decoder during simulcoding, as in this case instances can run independently. Still, due to its low processing speed, this software cannot be utilized in real applications.

JMVC 8.5 [12], the latest H.264/MVC reference software naturally supports MVC with arbitrary number of views. Being a reference implementation, its runtime performance is low, similar to JM. Unlike JM however, depending on setup of inter-view prediction, encoder / decoder instances have to be executed sequentially for each view, and cannot be parallelized, as the dependent views rely on the reconstructed images output by the encoder in previous run. Parallelizing MVC encoding by partially delaying dependent views is possible [8], however this alone does not make JMVC real-time.

x264 [13] the popular, open source implementation of H.264 is considered the fastest pure-software H.264 codec. While it provides real-time encoding and decoding performance for high-resolution 2D videos, it does not support MVC, nor the specification of custom GOP structures to emulate inter-view prediction. Slicing is supported, but only for the purposes of parallel processing – the shape of slice groups cannot be defined externally.

NVENC [14] is a pure-hardware H.264 codec embedded in high-end Nvidia GPUs. It supports faster than real-time 2D video encoding / decoding for very high resolution videos, and it also supports MVC for up to two views. Nvidia does not have plans to extend it to multiple views. Using custom prediction structures and slicing along vertical blocks are not supported.

The DXVA MVC Specification [15] mentions support for the Multiview High Profile, however we have not seen any implementation of this in the latest Windows SDK.

As of commercial H.264 SDKs, we have found only one from MainConcept MVC/3D codec [16], which, according to the publicly available material supports decoding MVC for up to 10 views, but on the encoding side, only Stereo profile is supported.

IP cores (for embedding in hardware codecs in FPGAs or ASICs) have also been announced with MVC support, mostly for Blu-ray decoding. The announcement of the POWERVR VXD392 / VXE382 cores [17] explicitly mentioned Multiview High Profile, the Video Encoder / Decoder fact sheets however reveal that the final products support 2-view MVC.

There have been several attempts towards integrating MVC into open-source H.264 codecs into ffmpeg [18], and x264 [19] (the latter targeted only stereo), however none of these patches made it to the mainline development branch.

## 8. CONCLUSIONS AND FUTURE WORK

Based on the use cases and processing considerations described in this paper, we can formulate at least three aspects that need attention and future research when developing compression methods for LFs. First, we shall add the possibility to encode views having different resolution. Secondly, the ability to decode the required number of views should be supported by the ability to decode views partially, starting from the center of the view, thus decreasing the computing workload by restricting the areas of interest. Third, efficient coding tools for nonlinear (curved) camera setups shall be developed, as we expect to see this kind of acquisition format more in the future.

In the future, we will focus on including many-view MVC encoding / decoding into the x264 codec, which will allow us to exploit the possibilities of MVC (at least partially) in the use cases described. Also, the structure of image data and distributed processing requirements suggest that a novel display-independent representation for LFs should be developed, which gathers the necessary image data into a better localized format, instead of having the image data scattered all around views and compressed as such. We will also explore the SoA of HEVC 3D Extension, and how it can be applied to compress LF data.

## 9. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the PROLIGHT-IAPP Marie Curie Action of the People programme of the European Union's Seventh Framework Programme, REA grant agreement 32449.
The research leading to these results has received funding from the DIVA Marie Curie Action of the People programme of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement 290227.

## 10. REFERENCES

[1] T. Balogh, "The HoloVizio system," Proc. SPIE 6055, *SD&A XIII*, 60550U, 2006

[2] G. Wetzstein, et al, "Tensor Displays: Compressive Light Field Synthesis using Multilayer Displays with Directional Backlighting", In *Proc. SIGGRAPH 2012*

[3] M. Kawakita, et al, "Glasses-free 200-view 3D Video System for Highly Realistic Communication," in *Digital Holography and Three-Dimensional Imaging, OSA Technical Digest*, paper DM2A.1.

[4] 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, pp.626,642, April 2011

[5] H. Schwarz, et al, "3D video coding using advanced prediction, depth modeling, and encoder control methods," *Picture Coding Symposium*, 2012

[6] M. P. Tehrani, et al, "Use Cases and Requirements on Free-viewpoint Television (FTV)", *ISO/IEC JTC1/SC29/WG11 MPEG2013/N14104*, October 2013, Geneva, Switzerland

[7] P. T. Kovács et al, "Requirements of Light-field 3D Video Coding", *ISO/IEC JTC1/SC29/WG11 MPEG2014/M31954*, January 2014, San Jose, US

[8] Y. Chen, et al, "The Emerging MVC Standard for 3D Video Services", *EURASIP Journal on Advances in Signal Processing*, Vol. 2009, No. 1, January 2009.

[9] Adhikarla, V.K.; et al, "Fast and efficient data reduction approach for multi-camera light field display telepresence systems," *3DTV-Con 2013*

[10] K. Wegner, et al, "Compression of FTV video with circular camera arrangement", *ISO/IEC JTC1/SC29/WG11 MPEG2014/M33243*, April 2014, Valencia, Spain

[11] H.264/AVC Software Coordination, JM 18.6, http://iphome.hhi.de/suehring/tml/ (visited 14/06/2014)

[12] H.264/MVC Reference Software, JMVC 8.5, cvs:// garcon.ient.rwthaachen.de (visited 14/06/2014)

[13] x264, http://www.videolan.org/developers/x264.html (visited 28/03/2014)

[14] Nvidia Video Codec SDK, https://developer.nvidia.com /nvidia-video-codec-sdk (visited 14/06/2014)

[15] G. J. Sullivan, Y. Wu, "DirectX Video Acceleration Specification for H.264/MPEG-4 AVC Multiview Video Coding (MVC), Including the Stereo High Profile", http://www.microsoft.com/en-us/download/details.aspx? id=25200 (visited 14/06/2014)

[16] MainConcept Video SDK, http://www.mainconcept.com /eu/products/sdks/video/mvc3d.html (visited 14/06/2014)

[17] Imagination's POWERVR VXD392 and VXE382, http://www.imgtec.com/news/Release/index.asp?img_ccc=1 &NewsID=597 (visited 14/06/2014)

[18] J. Britz, "Optimized implementation of an MVC decoder", MSc thesis at Saarland University, 2013

[19] SoC 2011/Stereo high profile MVC encoding, https://wiki.videolan.org/SoC_2011/Stereo_high_profile_m vc_encoding (visited 14/06/2014)

# X

# ARCHITECTURES AND CODECS FOR REAL-TIME LIGHT FIELD STREAMING

by

P. T. Kovács, A. Zare, T. Balogh, R. Bregovic, A. Gotchev, 2017

Journal of Imaging Science and Technology, 61(1), [010403],
DOI:10.2352/J.ImagingSci.Technol.2017.61.1.010403

# Architectures and Codecs for Real-Time Light Field Streaming

***Péter Tamás Kovács; Tampere University of Technology; Tampere, Finland / Holografika; Budapest; Hungary***
***Alireza Zare; Tampere University of Technology; Tampere, Finland / Nokia Technologies, Tampere, Finland***
***Tibor Balogh; Holografika; Budapest; Hungary***
***Robert Bregović; Tampere University of Technology; Tampere, Finland***
***Atanas Gotchev; Tampere University of Technology; Tampere, Finland***

## Abstract

*Light field 3D displays represent a major step forward in visual realism, providing glasses-free spatial vision of real or virtual scenes. Applications that capture and process live imagery have to process data captured by potentially tens to hundreds of cameras and control tens to hundreds of projection engines making up the human perceivable 3D light field using a distributed processing system. The associated massive data processing is difficult to scale beyond a specific number and resolution of images, limited by the capabilities of the individual computing nodes. We therefore analyze the bottlenecks and data flow of the light field conversion process and identify possibilities to introduce better scalability. Based on this analysis we propose two different architectures for distributed light field processing. To avoid using uncompressed video data all along the processing chain, we also analyze how the operation of the proposed architectures can be supported by existing image / video codecs.*

## Introduction

Three dimensional (3D) displays [1][2] represent a new class of terminal devices, that make a major step forward in realism towards displays that can show imagery indistinguishable from reality. 3D display technologies use different approaches to make the human eyes see spatial information. While the most straightforward approaches use glasses or other head-gear to achieve separation of left and right views, autostereoscopic displays achieve similar effects without the necessity of wearing special glasses. Typical autostereoscopic display technologies include parallax barrier [3], lenticular lens [4], volumetric [5], light field [6][7] and pure holographic [8] displays, most of them available commercially, each with its unique set of associated technical challenges. Projection-based light field 3D displays [9][10][11] represent one of the most practical and scalable approaches to glasses-free 3D visualization, achieving, as of today, a 100 Mpixel total 3D resolution and supporting continuous parallax effect.

Various applications where 3D spatial visualization has added value have been proposed for 3D displays. Some of these applications involve displaying artificial / computer generated content, such as CAD design, service and maintenance training, animated movies and 3D games, driving and flight simulation. Other applications require the capturing, transmission / storage and rendering of 3D imagery, such as 3DTV and 3D video conferencing (3D telepresence). From the second group, applications that require live 3D image transmission are the technically most demanding, as 3D visual information has much more information content compared to its 2D counterpart. In state of the art 3D displays, total pixel count can be 1-2 orders of magnitude higher than in common 2D displays, making such applications extremely data intensive.

We focus on the problems associated with applications that require real-time streaming of live 3D video. We consider projection-based light field 3D displays as the underlying display technology (with tens or hundreds of projection engines) coupled with massive multi-camera arrays for light field capturing (with tens or hundreds of cameras). We explore possibilities that can potentially work in real-time on today's hardware. Please note that the majority of the problems discussed here also apply to other 3D capture and display setups in which the necessary data throughput cannot be handled by a single data processing node, and as such distributing the workload becomes necessary. Our novel contribution presented in this paper lies in: analysis of the typical data flow taking place during converting multi-view content to display specific light field representation; analysis of bottlenecks in a direct (brute force) light field conversion process; presentation of two possible approaches for eliminating such bottlenecks; and a suitability analysis of existing image and video codecs for supporting the proposed approaches.

The paper is organized as follows. First light fields and light field displays, as well as different content generation and rendering methods for light field displays are introduced. Then the proposed architectures for scalable light field decompression and light field conversion are described, followed by a comparative analysis of the discussed architectures and codecs.

## Light Fields, Light Field Displays and Content Creation

The propagation of visible light rays in space can be described by the plenoptic function, which is a 7D continuous function, parameterized by location in 3D space (3 parameters), angles of observation (2 parameters), wavelength and time [12]. In real-world implementations, a light field, which is a simplified 3D or 4D parameterization of the plenoptic function, is used, which allows to represent visual information in terms of rays with their positions in space and directionality [13][14]. In a 4D light field, ray positions are identified by either using two planes and the hit point of the light ray on both planes, or by using a single hit point on a plane and the direction to which the light ray propagates. This 4D light field describes a static light field in a half-space. For a more detailed description, the reader is referred to [15].

Having a light field properly reproduced will provide the user with 3D effects such a binocularity and continuous motion parallax. Today's light field displays can typically reproduce a horizontal-only-parallax light field, which allows the simplification of the light field representation to 3D.

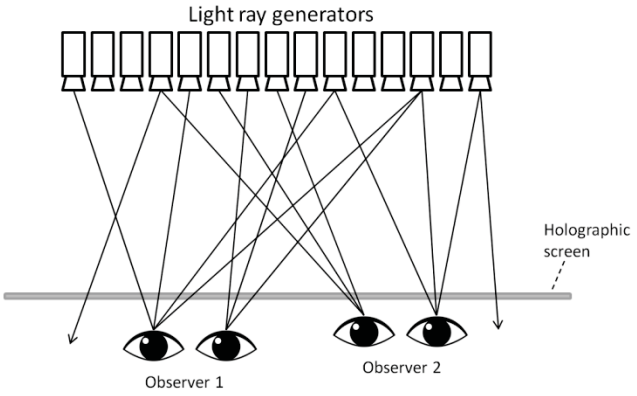



*Figure 1. Basic principle of projection-based light field displays: projection modules project light rays from the back, towards the holographic screen. Object points are visible where two light rays appear to cross each other consistently. Human eyes can see different imagery at different locations due to the direction selective light emission property of the screen.*

### *Optical System and Image Properties*

Projection-based light field displays are based on a massive array of projection modules that project light rays onto a special holographic screen to reproduce an approximation of the continuous light field. The light rays originating from one source (later referred to as light field slices) hit the screen at different positions all over the screen surface and pass through the screen while maintaining their directions. Light rays from other sources will also hit the screen all over the screen surface, but from slightly different directions (see Fig. 1). The light rays that cross the screen at the same screen position but in different directions make up a pixel that has the direction selective light emission property, which is a necessary condition to create glasses-free 3D displays. By showing different but consistent information to slightly different directions, it is possible to show different images to the two eyes, with a specific displacement. Also, as the viewer is moving, eyes will see the other (previously unseen) light rays emitted – causing motion parallax effect. As the angular resolution between light rays is typically below 1 degree, the motion parallax effect is smooth; as the viewing angle can be up to 180 degrees, the image can be walked around. The 3D image can be observed without glasses, and displayed objects can appear floating in front of the screen, or behind the screen surface, like with holograms.

This image formation is achieved by controlling the light rays so that they mimic the light field that would be present if the real object would be placed to the same physical space, up to the angular resolution provided by the projection modules. Such glasses-free display technology has dramatic advantages in application fields like 3DTV, and especially 3D video conferencing, where true eye contact between multiple participants at both sides can only be achieved by using a 3D display [16]. These same applications require real-time processing (capture, transmission, and visualization) of live images, which has special requirements on the rendering system.

Projection-based light field displays from Holografika have been used for the purposes of this work. The interested reader is referred to [7] for a complete description of the display system. Note that other projection-based light field displays based on very similar architectures also exist [10][11][17], thus our findings are relevant for those displays too.

### *Rendering System and Data Flow*

The previously described distributed optical system is served by a parallel rendering system, implemented as a cluster of multi-GPU computers. In the simple case, each GPU output drives a single optical module, multiple outputs of the same GPU correspond to adjacent optical modules, while one computer (having multiple GPUs) corresponds to a bigger group of adjacent modules. As in [18], we define the part of the whole light field that is reproduced by a single optical module a light field slice. One such slice is not visible from a single point of observation, nor does it represent a single 2D view of the scene. The set of light rays contained in such a slice is determined by the optical setup of the display and is calculated during the design process.

The cluster nodes are connected through a network (for example Gigabit Ethernet, Infiniband or 40 Gbit Ethernet), which carries all the data to be rendered / visualized, unless it is pre-computed and stored locally in the rendering nodes. The data received from an external source first needs to travel through the network, in the RAM of the rendering nodes, uploaded to and processed by the GPUs, output on a video interface, and finally displayed by the optical modules. The fastest light field conversion technique currently employed in light field rendering software is interpolation of light rays from the nearest samples. All data that contributes to the calculation of the color of a specific outgoing light ray thus needs to be available in the GPU responsible for that light ray.

### *Content Creation for Light Field Displays*

For the sake of achieving the highest possible light field quality on the display side, we prefer to use a dense set of input views, so that no depth estimation and virtual view synthesis is necessary to perform the light field conversion, thus the rendering process involves light field interpolation only. This approach has been successfully used on many occasions for light field displays resulting in crisp and artifact-free images regardless of scene complexity, material properties or visual effects involved. Therefore this seemingly brute force approach is preferred also for live imagery, so that any type of scene can be handled with no view synthesis artifacts.

While this requires a high number of capture cameras, installing such a rig of low cost cameras is feasible in fixed settings (such as a film studio or a video conferencing room). Installing one for capturing events, while more demanding, is also possible and is done commercially [19] as well as for experimental projects (although most of these target free viewpoint viewing and are not necessarily aware that the captured input can also be used as a light field).

If the captured light field is to be transmitted over longer distances, or stored, it is absolutely necessary to compress it due to the large volume of data. It might even be necessary to compress the video streams captured by the cameras (preferably inside the cameras) to make the data throughput manageable. While compressing tens or even hundreds of views may seem counterintuitive due to reasons of bandwidth requirements, recent results have shown that compressing a 80-view XGA resolution light field stream (while keeping good visual quality) is not much more demanding than compressing a 4K video signal in terms of consumed bandwidth [20]. It has to be noted that the complexity of compressing or even decompressing this many views using the technology discussed in that paper does not allow for real-time operation today.

One of the most well known camera arrays for capturing light fields is the Stanford Multi-Camera Array [21], consisting of 128 video cameras that can be arranged in various layouts, such as a linear array of parallel cameras or a converging array of cameras having horizontal and/or vertical parallax. Numerous other multi-camera setups have been built since then for both research and commercial purposes, e.g. the 100 camera system at Nagoya University [22], the 27-camera system at Holografika [23] (discussed later in this paper) or the 30-camera system from USC Institute for Creative Technologies [24]. These camera systems capture a sufficiently dense (in terms of angular resolution) and wide (in terms of baseline) light field, so that the captured data can be visualized on a light field display without synthesizing additional views beforehand.

One can also use a single moving camera for capturing a static scene [25], or a static camera with a rotating object [26] to obtain a light field. As the last two approaches capture a static light field, they are not in the scope of the discussion presented in this paper, which is about processing light field video. Also please note that plenoptic, range and other single-aperture capture methods are not discussed in this paper, as those cannot capture a scene from a wide viewing angle, unless the captured scene is extremely close (small).

When imagery is derived from a geometrical representation (that is, rendered from a 3D model), the resulting 2D image is a projection of the 3D scene. Synthesizing 2D views from many viewing angles is relatively straightforward in this case, as creating an array of synthetic cameras and rendering additional images from those is a matter of scripting in most modeling tools such as 3ds Max, Maya or Blender. As such, rendering a dense set of views to serve as light field content is only challenged by increased rendering time, and potentially 2D-only visual effects applied in the rendering pipeline which do not work consistently across multiple views. The practicality of this approach has been demonstrated with the Big Buck Bunny light field sequences [27], or the San Miguel sequence [28]. Densely rendered sequences can be used for light field displays without synthesizing additional views. When rendering synthetic scenes, it is also relatively easy to extract ground truth depth information, as depth maps are commonly used during rendering. Thus, depth information is available as a byproduct of the process.

There are attempts to capture and represent LF content with sparser camera views while estimating the geometry (depth) of the scene and augmenting the available views with depth maps to be used for subsequent intermediate view interpolation (synthesis) in order to provide the missing rays. Many techniques have been proposed for rendering intermediate views from a sparse set of views [29], as well as rendering extrapolated views from narrow angle content [30]. Most of these techniques are rooted from stereo matching for depth estimation, and depth based view synthesis by pixel shifting [31] or warping [32], and improving the resulting views by hole filling [33], inpainting [34], and other techniques. As soon as these techniques can provide sufficient quality and work in real-time to generate the rays needed for the display based on live imagery, they will be used in LF displays.

Light fields conversion is possible without explicit geometry / depth information about the scene. The simplest technique involves resampling the light field, which can be implemented by interpolating the 4D light field function from the nearest samples [14]. More involved is the lumigraph approach [13], where a rough geometric model is obtained, which is then used to improve the quality of the rendered lightfield by defining the depth of the object being rendered, thus allowing for better ray space interpolation.

When it comes to live imagery, dense image-based techniques, while requiring a massive amount of images captured or rendered, result in the highest possible light field quality. This dense set of input views is easily converted to display-specific LF slices, however at the expense of getting the large number of input pixels to be available in the display's rendering nodes. The sheer amount of data necessitates novel streaming architectures and codecs that are able to handle such imagery.

### *Light Field Streaming*

When performing light field streaming from an array of cameras, the simplest approach is to have all images captured by the cameras transferred to the GPUs of all rendering nodes to make sure all pixels that might be required during the light field conversion process are immediately available. This brute force approach has previously been implemented and described in [23]. In that system, a linear 27-camera array was used to feed a light field display. The cameras performed MJPEG compression of the images in hardware, which have been transferred to all rendering nodes simultaneously. The rendering nodes decompressed all 27 MJPEG images in parallel, and performed light field conversion on their respective light field slices on GPU. The performance of the whole system reached 15 FPS with 640x480 images, and 10 FPS with 920x720 images back in 2009.

The brute force approach outlined above has several possible bottlenecks. First, the time necessary to decode all the JPEG images (either on CPU or GPU) increases linearly with the number of images (that is, number of cameras), and the resolution of the images. Second, the network bandwidth required to transfer all the compressed images to the rendering nodes increases linearly with the number and resolution of images. Third, the memory required to store the uncompressed images increases linearly with the number and resolution of captured images (both CPU and GPU memory).

If we consider an 80-view input (we used the BBB light field test sequences [27]), each view with 1280x768 resolution and 24 FPS, uncompressed, then the bandwidth required to transmit all pixels is 45.3 Gbit/s, which is difficult to reach with common network technologies.

Considering the same input, using 1:10 ratio JPEG compression, still gives a total bandwidth requirement of 4.5 Gbit/sec and an averaged PSNR of 44.65 dB calculated over the luminance channel. While it is possible to transmit this amount of data trough a 10- or 40-Gigabit Ethernet channel in real time, it is not feasible to decompress and process it. Using the fastest available libjpeg-turbo JPEG decoder library [35], and a 6-core i7-5930K CPU @ 3.5 Ghz, one can decode 80 such views in ~43 ms. If the system performs JPEG decoding only, this results in ~23 FPS decoding speed, and this does not even include any further processing or rendering. Achieving higher resolution or processing a wider or denser set of views is clearly limited by hardware, and thus not scalable.

If we choose H.264 instead of JPEG for compressing the views using the x264 encoder [36] and maintaining a similar PSNR with QP=20 (QP: Quantization Parameter, which determines image quality versus bandwidth in the video encoder), and use ffmpeg [37] to decode the 80 views on the same CPU in parallel, we can reach ~16.94 FPS for decoding all views. It this measurement we used a RAM disk to avoid any significant I/O overhead.

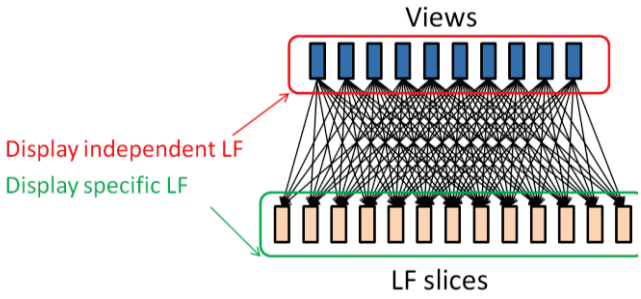



*Figure 2. A set of perspective views depicting the same scene are considered as a display independent light field representation, as it can be used to visualize the light field on any suitable 3D display. The images required by a specific display's projection modules are referred to as light field slices, and the set of light-field slices are therefore referred to as a display specific light field.*

Clearly, such a system having decoding time linear with the number of views and total pixel count cannot scale beyond a specific number of input cameras and resolution, given a specific hardware configuration. While the hardware can be upgraded to some extent, the processing power of a single node cannot be increased indefinitely (nor is it economical).

While increasing the number of cameras and/or the resolution of these cameras increases the quality of the generated light field, the number of rendered light rays remains constant. It seems counterintuitive to use an increasing amount of input data to generate the same amount of outgoing light rays, and, as we will see in the next sections, this can be avoided.

### *Light Field Conversion and Data Access Patterns*

Light field conversion transforms many input 2D views (which can be considered as a display independent light field representation) into light field slices, as required by the optical modules of the targeted light field display. One such light field slice is composed of many outgoing light rays (for example, 1024 x 768 pixels in case of an XGA optical module), which are in turn interpolated from pixels from many input views. Starting from views, we can also see that one input view typically contributes to many light field slices. Fig. 2 shows that views typically contribute to many adjacent light field slices, and that light field slices are typically composed of pixels originating from many adjacent views.

An efficient implementation of the light field conversion process is using weighted look-up tables that map multiple pixels originating from views to an outgoing light ray, with the weighting appropriate for the given light ray. The pixel correspondences stored in the look-up table are calculated based on the geometry of the captured views (intrinsic and extrinsic parameters) and the geometry of the optical modules (display design, calibration data). The light field conversion process only needs to look up and blend the necessary view pixels to generate an outgoing light ray. As such, as long as the capture configuration, display configuration and mapping between them does not change, the look up tables remain constant, the same pixel positions are used while generating the same set of light rays.

*Figure 3. Left: Adjacent rendering nodes consume adjacent, slightly overlapping parts of a source view. Red, green and blue overlays represent the areas of the image used by three rendering nodes that drive adjacent optical modules. Right: Rendering nodes that drive optical modules positioned further away from each other use a disjoint set of pixels from the same source view.*

In the light field conversion algorithm used with HoloVizio displays today, which uses a variant of the light field conversion algorithm described in [14], a maximum of two captured pixels contribute to one outgoing light ray. That is, the maximum number of light rays used by a rendering node cannot be more than twice the number of light rays generated by the same node, regardless of the number of total pixels captured. Take a HoloVizio 722RC display with 72 optical modules as an example, served by 6 rendering nodes, each rendering adjacent light field slices for 12 optical modules. Consider again an 80-view input, each view with 1280x768 resolution. Checking which pixels of a specific view are used by each rendering node, we can realize that rendering nodes that drive adjacent optical modules use slightly overlapping parts of the view (see Fig. 3). However, rendering nodes that are driving optical modules positioned further from each other use a disjoint set of pixels from the input view. None of the other pixels of this view contribute to the final image. These pixels do not have to be transmitted to the renderer. It may even happen that a specific rendering node does not use any pixels of a specific view. In this case, dropping the view on the specific rendering node would not have any effect on the final image.

These properties of the typical data access patterns of light field conversion can be exploited to make the system more scalable, and eventually achieve real-time operation regardless of the number and resolution of the incoming views. Two proposed solutions are described next.

## Streaming Architectures and Codecs

### *Two-Layer Architecture*

One possibility to scale the decompression workload without introducing any new bottlenecks is to introduce two layers in the system: one for decoding the incoming streams in parallel, and a second one to perform the light field conversion (see Fig. 4). The nodes in the first layer decompress the video streams in a way that the decoding workload is distributed equally (so that given N views and M nodes, one node decodes N/M video streams). The number of nodes in the first layer should be chosen so that given the processing power of each node, we have enough nodes to decompress all the individual video streams in real time and transmit them to the second layer.

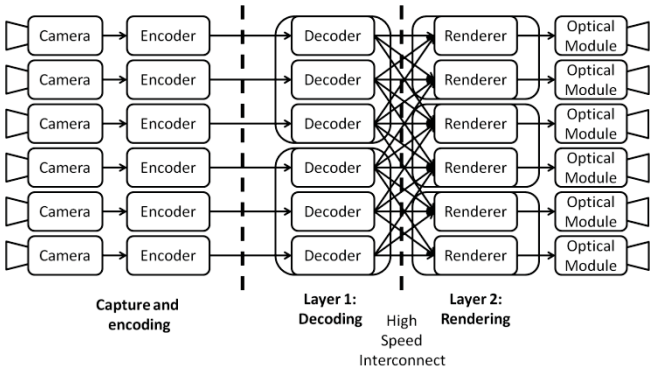



*Figure 4. The two-layer decoder-renderer architecture. The first layer decodes video streams in parallel, while the second layer requests portions of uncompressed video data on demand.*

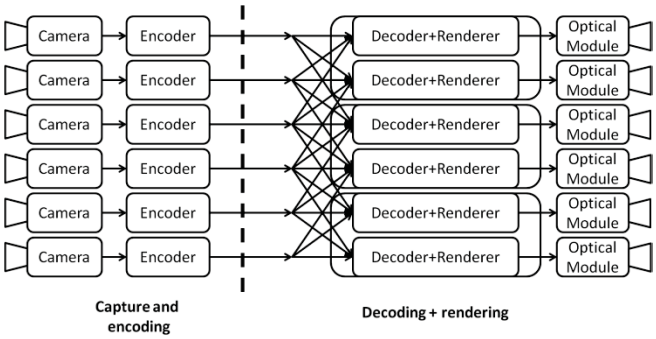



*Figure. 5. The one layer decoder-renderer architecture. Decoders and renderers are running on the same nodes. Decoders decode those parts of the video that the renderer on the same node will need for rendering. No data exchange except frame synchronization takes place between the nodes.*

It is easy to see that in this case the processing power of the individual nodes can be chosen arbitrarily – as long as one can decode at least one video stream, the number and processing power of the nodes can be traded off. The second layer requests uncompressed pixels from the first layer through a high-speed network. As we have seen before, the number of requested pixels can be maximum twice the number of pixels generated by a specific node. In case of driving 12 1280x800 resolution optical modules, this can be up to 1.41 Gbits/second, assuming 24 FPS, which is in the manageable range inside a rendering cluster. To avoid the collection and transfer of single pixels, we can transfer rectangular blocks which form the bounding box of the required pixels. Due to the compact shape of the required image regions, a bounding box does not add too much extra pixels to transmit, but has the benefit of being continuous in memory.

All rendering nodes require a different set of pixels, thus each will request the transmission of different, slightly overlapping regions. As the regions are fixed if the camera settings, display parameters and the mapping between them are fixed, the rendering nodes can subscribe to receive these fixed image regions on each new frame. The bandwidth required for transmitting and receiving the decompressed image regions can be scaled with appropriately choosing the number of processing nodes in both layers.

While the first layer can be connected to the data source using a network necessary to transmit the compressed video, the first and second layers are connected using a high-speed network to be able to carry the uncompressed image regions.

If the selected network architecture can take advantage of transferring large chunks of continuous in-memory blocks more efficiently than doing a large number of small transfers (for example InfiniBand RDMA [38]), then the images can be stored rotated with 90 degrees in memory, so that full height vertical blocks become a continuous memory block, assuming the common line-by-line interleaved image storage format.

In order to reduce the network load between the first and second layer, it is possible to employ a slight, fixed ratio compression method that can be decompressed even when only partially transmitted, such as the S3TC / DXTn texture compression method that is available on virtually all GPUs.

### *One-Layer Architecture with Partial Decoding*

The second proposed architecture does not require two processing layers, as decoding and light field conversion happens on the same processing nodes (see Fig. 5).

In this case however, decoding needs to happen in real time regardless of the number of views to be decoded. To achieve that, we can take advantage of the fact that each renderer is using only portions of the views, as shown before. As we have seen before, the main bottleneck is decompression of all parts of all views, even those that are finally left unused in the specific rendering node. Therefore, assuming that we can save time by skipping the decoding of some views, and also skipping the decoding of image regions that are not needed, we can decode only those image regions which are necessary on the current rendering node. Using this approach, the decoding workload can be reduced with different codecs, if certain conditions are met.

This, however, poses unusual requirements on the image / video codec used for the transmission of the views. In the following subsections, we analyze common image / video codecs from this perspective: JPEG, JPEG2000, H.264 and HEVC. In the case of JPEG and JPEG2000, frames are encoded independently. In the case of H.264 and HEVC, we enable inter frame compression.

In a horizontal only parallax light field display, the image regions are typically vertical. H.264, HEVC and JPEG2000 support vertical slice shapes, while in the case of JPEG, which supports only horizontal subdivision, rotating the image before encoding is necessary.

The configuration used for all tests described below is a 6-core i7-5930K CPU @ 3.5 Ghz PC with a GeForce GTX 960 GPU with 2GB RAM, and all I/O performed on RAM disk.

The experiments were performed on the Big Buck Bunny light field sequences [27], more specifically the Flowers scene. Other content (i.e. Balloons sequence [39] in Fig. 8 and Fig. 9) is used only for illustration purposes.

### *H.264*

H.264 [40] can partition a full frame into regions called slices, which are groups of macroblocks (please note these are different from light field slices). An encoder is commonly configured to use a single slice per frame, but can also use a specific number of slices (see Fig. 6). This configuration may also specify the way macroblocks are partitioned into slices. Slices are self-contained in the sense they can be parsed from the bitstream without having the other slices. Slicing originally serves the purposes of providing error resilience and also to facilitate parallel processing. Slicing introduces some increase in bitrate, but this is minor compared to the overall bitrate, as seen in Fig. 7.

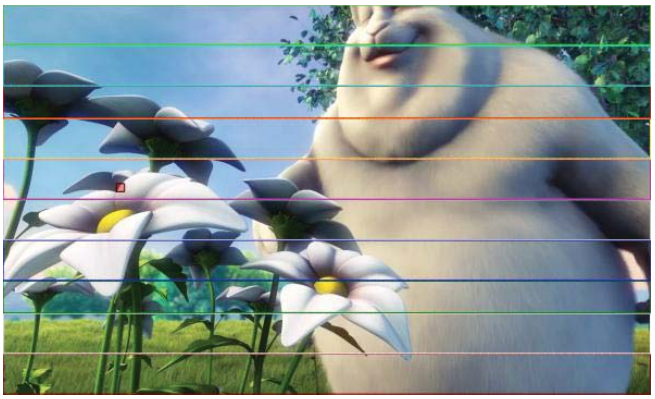


*Figure 6. A frame subdivided into horizontal H.264 slices. In an I-frame, these slices can be decoded independently.*

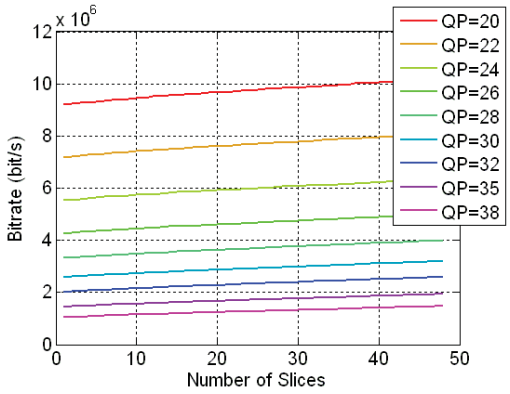



*Figure 7. Bitrate increase caused by slicing a video frame into an increasing number of slices, shown for different QPs.*

To see how H.264's slicing feature can support partial decoding of video streams, we need to consider how it is implemented in the encoder and decoder. In H.264 we can differentiate Intra frames (I-frames) which are encoded independently from other frames in the video sequence, and hence employ prediction only inside the frame. Predicted frames (P-frames) and Bidirectionally predicted frames (B-frames) use image information of previously encoded / decoded frames in the encoder / decoder, exploiting similar blocks of pixels in subsequent frames moving across the image in time, typically representing a moving object.

When using multiple slices, encoders disable intra-frame prediction across slice boundaries when encoding I-frames. Therefore it is possible to decode only parts of an I-frame by decoding the respective slices and skipping the other slices (see Fig. 8). On the other hand, when performing inter-frame prediction in P-frames and B-frames, all encoders we have checked disregard slice boundaries, and perform motion prediction across slice boundaries. This means that the decoding process will also assume that the frame to be used for motion compensation is fully available in the decoder. If, due to partial decoding of the previous frames, this condition is not met, the image regions that correspond to skipped slices will contain bogus color in the Decoded Picture Buffer.

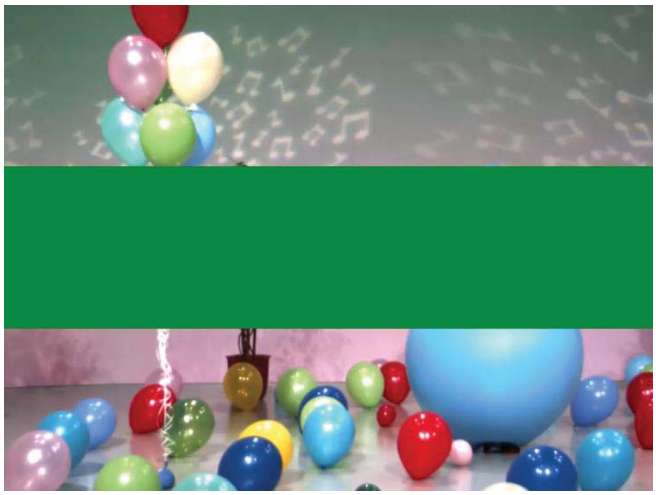


*Figure 8. When using I-frame only encoding in H.264, dropping a slice has no effect on the remaining parts of the image. This image has been subdivided into three slices, and the middle slice dropped prior to decoding.*

*Figure 9. When using P- and B-frames motion vectors pointing out from the undecoded region (middle slice) propagate bogus colors into the slices which we intend to decode.*

Subsequently the motion compensation process will copy bogus colors from the buffer to the image being decoded, whenever a motion vector goes across the boundary of a missing slice (see Fig. 9). When such an erroneous decoded frame is used as a basis of motion compensation, the error further propagates from the undecoded regions as the decoder proceeds further in the stream. Only on the next I-frame the partially decoded image will be correct again, as the decoder does not use any other frames for decoding I-frames.

Therefore skipping the decoding of slices is possible only if the encoder is instructed not to perform motion prediction across slice boundaries (see Fig. 10). As this is not available as an option in any encoder we have evaluated, this functionality has been implemented by modifying the reference encoder JM 18.6 [41]. Our implementation is similar to that of [42], although serving a different purpose.

*Figure 10. Difference of motion vectors in normal encoding and with self-contained slices. Notice that in the normal case (top) motion vectors cross slice / tile boundaries. In the self-contained case (bottom) no motion vectors cross the slice / tile boundaries.*

With restricting the motion vectors, we can achieve truly self-contained slices in the sense that decoding only selected slices becomes possible even for P- and B-frames, while still maintaining a standard confirming bitstream, though with a minor loss in coding efficiency due to partially restricting motion vectors. Implementation details have been described in our previous paper [43].

Once the encoded bitstream with self-contained slices is available at the decoder, there are several options to achieve partial decoding and thus saving decoding time. It is possible to modify the bitstream by dropping NAL units that correspond to the slices we do not wish to decode, like they were dropped by the network. While this results in a corrupt bitstream, some decoders (e.g. ffmpeg) can decode the remaining parts of the image, resulting in sub-linear speedup, as shown in Fig. 11. In this case, error resilience options have been disabled in the decoder to the extent possible, however we suspect that the missing slices in the bitstream still result in some decoding time overhead. Another option is to modify the decoder to explicitly skip the decoding of specific slices, which requires decoder modification (as this functionality was also unavailable in all decoders we have tested).

### *HEVC*

There are many new coding features introduced in HEVC [44] compared to H.264, and even a short summary of these novelties is out of scope in this paper. Therefore we only focus on differences relevant for our use case, and we refer the interested reader to [45] for a summary of other changes and novel features.

Tiles in HEVC are a new concept [46], serving parallelization and packetization purposes. Tiles partition a video frame into a multiple number of rectangular partitions. Compared to slices, tiles provide more flexibility to partitioning and appear to incur less compression penalty since tiles do not contain any header. Furthermore, tiles are independent in terms of entropy coding, as the coder is re-initialized at the beginning of each tile.

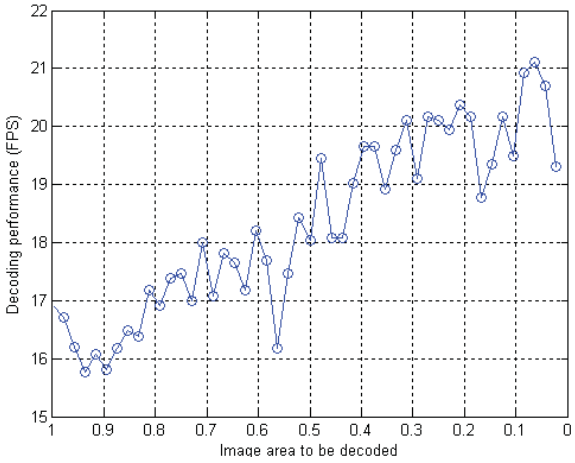



*Figure 11. Speedup of ffmpeg H.264 decoder when performing partial decoding on videos sliced in 48 slices. The vertical axis shows frames per second values for decoding 80 views.*

In addition, intra-frame prediction is limited within tile boundaries. In the HEVC inter-frame prediction scheme, however, the tile boundaries are disregarded. Similar to the H.264 case, motion search has to be restricted to remain inside tile boundaries to ensure partial decoding in P- and B-frames. Tiling is thus suitable for enabling partial decoding, and this has been implemented in the HTM 14.0 reference software [47], alongside a tool for bitstream manipulation that removes selected tiles from the bitstream before decoding [48].

As the coded tiles may be interleaved with other coded data in the bitstream and as parameter sets and headers (e.g. slice segment header) are for the entire bitstream, a dedicated decoding process is defined for decoding particular tiles, while omitting the decoding of other tiles. For that, a tile-based extractor is designed to construct an HEVC full-picture-compliant bitstream corresponding to the desired tiles such that a standard decoder can cope with it.

### *JPEG*

Images stored in JPEG [49] files are typically subdivided into 8x8 blocks (Minimum Coding Unit, MCU). The sequence of these MCUs is basically coded in two parts: the DC component, which is stored as a difference from the previous block, as well as the AC components, which are stored as a sequence of values in a specific ordering, block by block. As all the values and coefficients are Huffman coded, one needs to read all Huffman codes to be able to interpret the ones that are actually needed. Some steps, however can be skipped for the MCUs that are not needed, for example inverse DCT, dequantization, color conversion, etc. Unfortunately, Huffman decoding takes the major part of JPEG decoding time, according to our profiling of libjpeg-turbo. When forcing the decoder to skip all the steps after Huffmann decoding for the unnecessary image portions, one can only gain a rather insignificant speedup, as shown in Fig 12.

There is however an optional feature in JPEG to facilitate separately entropy coded segments, called Restart intervals / Restart markers, which allow resetting the bitstream after every N MCUs, letting the decoder skip N MCUs at a time without decoding anything, by just looking for specific bytes. If the JPEG encoder can be instructed to use a frequent restart interval, this feature can facilitate fast skipping of unnecessary MCUs.

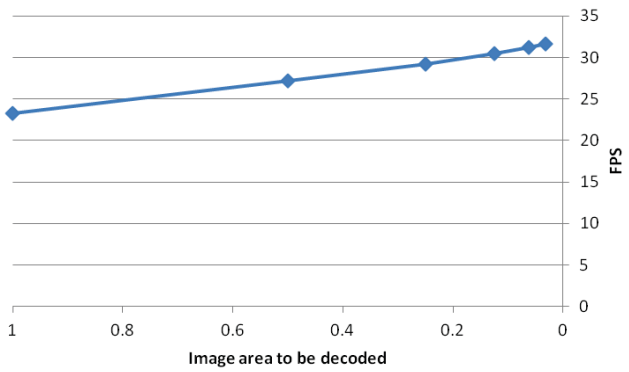



*Figure 12. Speedup of the customized libjpeg-turbo when skipping all steps after Huffman decoding for the unnecessary image parts.*

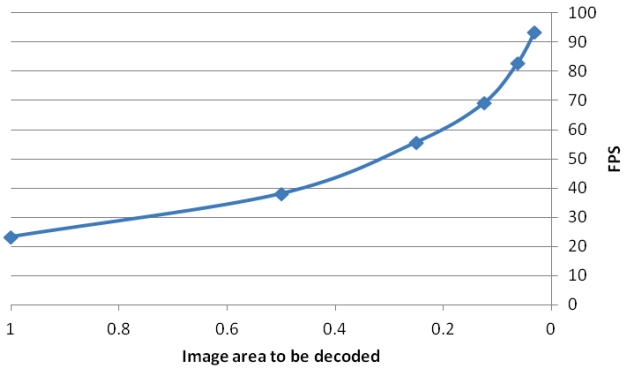



*Figure. 13. Speedup of the customized libjpeg-turbo when skipping through the bitstream using restart markers, also eliminating unnecessary Huffman decoding.*

We have implemented this approach too, by modifying the libjpeg-turbo decoder to skip the unused MCU rows by skipping through restart markers, as well as skipping all other processing steps for those MCUs. The speedup in this case is significant, as shown on Fig 13. For example, by decoding only the quarter of an image, a 2.4x speedup can be gained. While this is still not linear, it is a nice addition to scalability.

The ordering of MCUs in the JPEG stream is fixed, therefore restart markers can be only used to create horizontal strips that can be skipped. If, like in our use case, vertical strips are necessary, then the images need to be rotated 90 degrees before encoding. Also, as mentioned earlier, the encoder needs to be set up to include restart markers at regular intervals, which might not be possible in every JPEG encoder (for example, JPEG capable cameras often have a compression quality settings, but no option for restart markers).

### *JPEG 2000*

JPEG2000 [50] images can also be partially decoded. This functionality has been recently introduced in the proprietary Comprimato JPEG2000 codec [51], the CUDA-based version of which has been used for testing. Decoding 80 views takes 263 ms when decoding entire 1280x768 images, 8 bpp, 4:2:0 subsampling and default compression and decompression parameters, which accounts for 3.8 FPS operation taking only decoding into account.

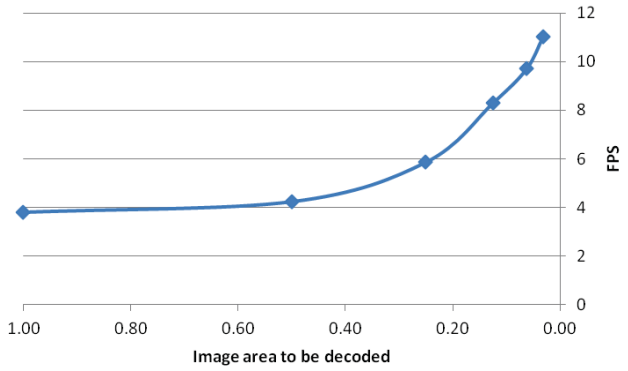



*Figure 14. Speedup of Comprimato JPEG2000 CUDA-based codec when performing partial decoding of a single view. The vertical axis shows frames per second values for decoding 80 views.*

On the other hand, when decoding only selected areas of each image, decoding time decreases significantly, although not strictly proportional with the image area, especially when decoding tiny image areas, as shown on Fig 14. Assuming that the application needs only the quarter of each image, a speedup of 54% can be observed, making decoding speed of 80 views 5.85 FPS. According to Comprimato, the fastest available GPUs at the time of writing can decode ~2.37 times faster, which corresponds to almost 14 FPS. Thus, very soon (after gaining a factor of two speedup over current GPUs) we can expect this solution to work in real time for the benchmark use case.

The main advantage of using JPEG2000 for this purpose is that no special considerations are necessary during encoding, and the partial decompression feature is already built in the decoder as an option.

## Comparative Analysis

### *Architectures*

The presented two-layer architecture is massively scalable, as the number of decoders, and the number of renderers can be chosen according to the performance of the individual nodes, while the bandwidth required by each rendering node is upper bounded by the number of pixels driven by each rendering node. Using this architecture, an arbitrary number of views with arbitrary resolution can be supported. On the other hand, the high speed network necessary between the two layers can make this setup quite costly in a practical case, as the required bandwidth is more than what can be provided by a Gigabit Ethernet network. Any codec can be used that can be decoded by the decoding layer in real time.

The one-layer architecture requires less number of processing nodes (as the separate decoding layer is eliminated) and does not need a high-speed network to connect the two nodes. On the other hand, to achieve real-time performance special considerations have to be taken in the video codec used for feeding the incoming views. The suitability of different codecs for this architecture is analyzed next.

### *Codecs*

Out initial aim was to find a solution for real-time decompression of the (partial) views necessary for light field conversion.

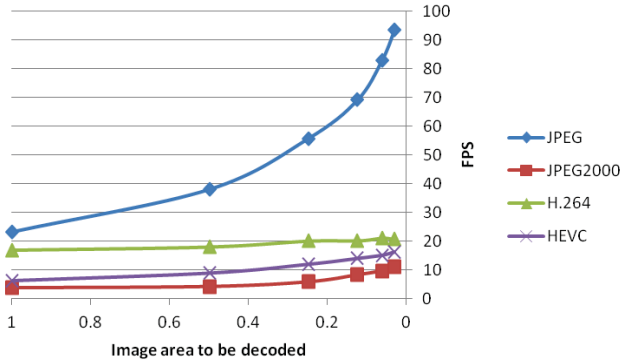



*Figure 15. Comparison of overall speed and speedup of different decoders when decoding partial views. In case of JPEG, restart markers are used. In case of H.264, our custom self-contained slices are used. In case of JPEG2000 no special features are used.*

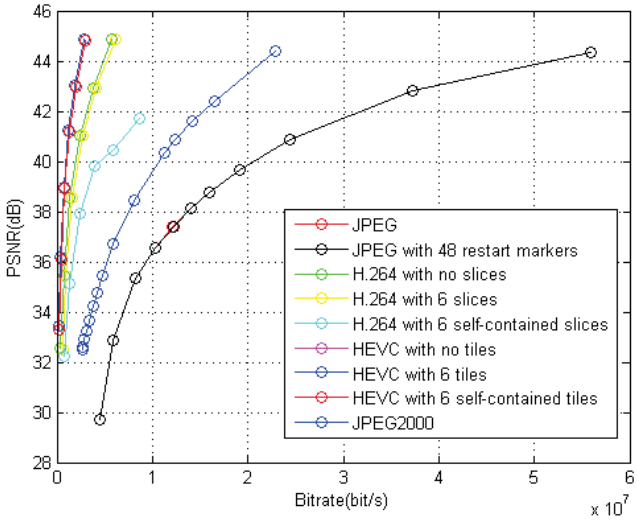



*Figure 16. Comparison of overall quality versus bitrate of the different codecs using the configurations discussed in the paper. Please note reported bitrates are for a single view. JPEG and JPEG with 48 restart markers overlap. The three HEVC curves also overlap.*

Therefore our discussion mainly focused on the runtime performance of decoding, and its improvement when partial decoding is allowed. Here we directly compare the different solutions from this perspective. Also important is the bandwidth required by each codec to achieve same or similar image quality, therefore a brief analysis of bandwidth requirements is presented.

As described in before in detail, none of the evaluated codecs is capable of decoding image portions in a time proportional with the area to be decoded. While JPEG and JPEG 2000 show reasonable speedup, this speedup is not linear, therefore full scalability cannot be guaranteed for any number of views or any resolution, as scalability is only partial. In the case of JPEG, this scalability can only be exploited if the encoder supports it, and is configured appropriately.

It is apparent from Fig. 15 that JPEG is the clear winner when it comes to decoding performance, as even today, >24 FPS decoding can be achieved with using a publicly available codec with a modification to support partial decoding for the targeted use case.

It comes as no surprise that the decoding performance advantage of JPEG comes at a cost in bandwidth (see Fig. 16). The bandwidth required by JPEG to achieve the same quality can easily be 10x higher than, for example H.264. JPEG2000 is in between the two in terms of bandwidth consumption. Please note this comparison of bandwidth requirements is by no means meant as a comprehensive comparison of codec efficiency, but just a rough guideline to see the order of magnitude difference between the codecs under consideration. For a more general comparison of different codecs, the reader is referred to works such as [52][53][54].

In all cases except JPEG2000, partial decoding capability requires some kind of subdivision of the images into independent regions, which increases the necessary bitrate. In case of JPEG, the difference is negligible (<0.2%). In case of H.264 and HEVC, the difference is bigger due to the restriction of motion vectors.

When restricting motion vectors in H.264 and HEVC, the more independent areas are defined, the smaller these areas will become, being even more restrictive in the selection of motion vectors. Smaller independent regions however result in less pixels decoded unnecessarily, and thus a possible higher decoding speedup.

That is, the granularity of the subdivision determines the tradeoff between bitrate increase and the compactness of the bounding rectangle that needs to be decoded to gain access to an arbitrary image region.

### *Alternatives*

To avoid using special codec features and encoder / decoder side modifications, one may also choose a very simple alternative solution: subdivide the source videos into vertical stripes of specific size, and encode them separately, in parallel. This approach has been used in [55], and the authors show good results with this approach for the use case of panoramic video. This however requires even more bitrate than the solutions outlined above, as well as a large number of video streams to be handled synchronously.

Advanced codecs targeting multiview and 3D video, such as MV-HEVC and 3D-HEVC [56] could be used as well, provided real-time decoders would exist, but unfortunately this is not the case at the time of writing.

Scalado RAJPEG [57] is said to enable instant random access / partial decoding to images. However, RAJPEG was not made available for testing.

## Conclusion

We have shown two different approaches to introduce scalability in real-time image-based light field based applications. While the two-layer approach can work with any codec (taking into account the necessary networking load), the one-layer architecture with partial decoding can only be achieved by modifying the way well known codecs typically work (except for the case of JPEG2000). While customization of codecs is possible to suit this requirement, the possibilities of forcing off-the-shelf hardware (for example a camera capable of providing compressed output) to use this custom codec are rather limited.

The special use of video codecs as outlined above indicates that current video compression technology is lacking an important feature. Therefore we have made steps [58][59] to ensure that next generation video technologies have low-overhead random access among their requirements, and this has been accepted in the MPEG

FTV requirements document [60]. This work was done in the hope that next generation video codecs will support this use case natively.

## Acknowledgements

The authors are grateful to Attila Barsi of Holografika for his support in profiling the light-field conversion process.

The research leading to these results has received funding from the PROLIGHT-IAPP Marie Curie Action of the People programme of the European Union's Seventh Framework Programme, REA grant agreement 32449.

## References

[1] D. Ezra, G. J. Woodgate, B. A. Omar, N. S. Holliman, J. Harrold, L. S. Shapiro, "New autostereoscopic display system," in Stereoscopic Displays and Virtual Reality Systems II, Proc. SPIE 2409, San Jose, 1995, pp. 31-40

[2] J. L. Fergason, S. D. Robinson, C. W. McLaughlin, B. Brown, A. Abileah, T. E. Baker, P. J. Green, "An innovative beamsplitter-based stereoscopic/3D display design," in Stereoscopic Displays and Virtual Reality Systems XII, Proc. SPIE 5664, San Jose, 2005, pp. 488-494

[3] D. Sandin, T. Margolis, J. Ge, J. Girado, T. Peterka, T. DeFanti, "The Varrier autostereoscopic virtual reality display," ACM Trans. Graphics, vol. 24, no. 3, pp. 894-903, 2005

[4] C van Berkel, D. W. Parker, A. R. Franklin, "Multiview 3D-LCD," in Stereoscopic Displays and Virtual Reality Systems III, Proc. SPIE 2653, San Jose, 1996, pp. 32-39

[5] G. E. Favalora, J. Napoli, D. M. Hall, R. K. Dorval, M. Giovinco, M. J. Richmond, W. S. Chun, "100 Million-voxel volumetric display," in Cockpit Displays IX: Displays for Defense Applications, Proc. SPIE 4712, Orlando, 2002, pp. 300-312

[6] G. Wetzstein, D. Lanman, M. Hirsch, R. Raskar, "Tensor Displays, Compressive Light Field Synthesis using Multilayer Displays with Directional Backlighting," ACM Trans. Graphics, vol. 31, no. 4, art. 80, Jul. 2012

[7] T. Balogh, "The HoloVizio System," in Stereoscopic displays and virtual reality systems XIII, Proc. SPIE 6055, San Jose, 2006

[8] M. Lucente, "Interactive three-dimensional holographic displays: seeing the future in depth," ACM SIGGRAPH Computer Graphics, vol. 31, no. 2, pp. 63-67, May 1997

[9] Method & apparatus for displaying 3D images. by T. Balogh, U.S. Patent 6,201,565, EP0900501 (1997)

[10] N. Inoue, M. Kawakita, K. Yamamoto, "200-Inch glasses-free 3D display and electronic holography being developed at NICT," in Lasers and Electro-Optics Pacific Rim (CLEO-PR), Kyoto, 2013, pp.1-2

[11] K. Nagano, A. Jones, J. Liu, J. Busch, X. Yu, M. Bolas, P. Debevec, "An autostereoscopic projector array optimized for 3D facial display," in Proc. SIGGRAPH 2013 Emerging Technologies, Anaheim, 2013

[12] E. Adelson and J. Bergen, "The plenoptic function and the elements of early vision," in Computational Models of Visual Processing, M. Landy and J.A. Movshon, eds., MIT, 1991

[13] S. J. Gortler, R. Grzeszczuk, R. Szeliski, M. F. Cohen, "The lumigraph," Proc. SIGGRAPH '96, pp. 43-54, 1996

[14] M. Levoy, P. Hanrahan, "Light field rendering," Proc. SIGGRAPH '96, pp. 31-42, 1996

[15] R. Bregović, P. T. Kovács, A. Gotchev, "Optimization of light field display-camera configuration based on display properties in spectral domain," Opt. Express 24, 3067-3088, 2016

[16] D. Nguyen, J. Canny, "MultiView: spatially faithful group video conferencing," in Proc. SIGCHI Conference on Human Factors in Computing Systems, Oregon, pp. 799-808, 2005

[17] J.-H. Lee, J. Park, D. Nam, S. Y. Choi, D.-S. Park, C. Y. Kim, "Optimal projector configuration design for 300-mpixel light-field 3D display," Optics Express, vol. 21, issue 22, pp. 26820-26835, Nov. 2013

[18] P. T. Kovács, Zs. Nagy, A. Barsi, V. K. Adhikarla, R. Bregović, "Overview of the Applicability of H.264/MVC for Real-Time Light-Field Applications", in Proc. 3DTV Conference 2014, Budapest, 2014

[19] TileSlice Films [Online]. Available: https://vimeo.com/timeslice/videos

[20] A. Dricot, J. Jung, M. Cagnazzo, B. Pesquet, F. Dufaux, P. T. Kovacs, V. Kiran, "Subjective evaluation of Super Multi-View compressed contents on high-end 3D displays," Signal Processing: Image Communication, vol. 39, Part B, November 2015, pp 369–385

[21] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, M. Levoy, "High performance imaging using large camera arrays" ACM Trans. Graphics, vol. 24, no. 3. pp 765-776, July 2005

[22] T. Fujii, K. Mori, K. Takeda, K. Mase, M. Tanimoto and Y. Suenaga, "Multipoint Measuring System for Video and Sound - 100-camera and microphone system," IEEE International Conference on Multimedia and Expo, Toronto, Ont., 2006, pp. 437-440. doi: 10.1109/ICME.2006.262566

[23] T. Balogh, P. T. Kovács,"Real-time 3D light field transmission", in Proc. Real-Time Image and Video Processing, Proc. SPIE 7724, Brussels, 2010

[24] A. Jones, J. Unger, K. Nagano, J. Busch, X. Yu, H.-Y. Peng, O. Alexander, P. Debevec, "Creating a life-sized automultiscopic Morgan Spurlock for CNNs "Inside Man"" Proc. SIGGRAPH '14, no. 2, 2014

[25] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, M. Gross, "Scene reconstruction from high spatio-angular resolution light fields," ACM Trans. Graph, vol. 32, no. 4, art. 73, Jul 2013

[26] A. Jones, I. McDowall, H. Yamada, M. Bolas, P. Debevec, "Rendering for an interactive 360° light field display," ACM Trans. Graphics, vol. 26, no. 3, art. 40, Jul 2007

[27] P. T. Kovács, A. Fekete, K. Lackner, V. K. Adhikarla, A. Zare, T. Balogh, "Big Buck Bunny light-field test sequences," ISO/IEC JTC1/SC29/WG11/M36500, Warsaw, 2015

[28] P. Goorts, M. Javadi, S. Rogmans, G. Lafruit, "San Miguel test images with depth ground truth," ISO/IEC JTC1/SC29/WG11/M33163, Valencia, 2014

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

[30] A. Ouazan, P. T. Kovacs, T. Balogh, A. Barsi, "Rendering multi-view plus depth data on light-field displays", in Proc. 3DTV Conference 2011, Antalya, 2011

[31] C. Fehn, “Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV,” in Stereoscopic Displays and Virtual Reality Systems XI, Proc. SPIE 5291, San Jose, 2004

[32] N. Stefanoski, O. Wang, M. Lang, P. Greisen, S. Heinzle, A. Smolic, "Automatic View Synthesis by Image-Domain-Warping," IEEE Trans. Image Processing, vol.22, no.9, pp.3329-3341, Sept. 2013

[33] M. Solh, G. AlRegib, "Hierarchical Hole-Filling For Depth-Based View Synthesis in FTV and 3D Video," IEEE Journal of Selected Topics in Signal Processing, vol.6, no.5, pp.495-504, Sept. 2012

[34] C. Guillemot, O. Le Meur, "Image Inpainting : Overview and Recent Advances," IEEE Signal Processing Magazine, vol.31, no.1, pp.127-144, Jan. 2014

[35] libjpeg-turbo [Online]. Available: http://libjpeg-turbo.virtualgl.org/

[36] x264 [Online]. Available: http://www.videolan.org/developers/x264.html

[37] FFmpeg [Online]. Available: https://www.ffmpeg.org/

[38] Introduction to InfiniBand™ for End Users – Mellanox [Online]. Available: http://www.mellanox.com/pdf/whitepapers/Intro_to_IB_for_End_Users.pdf

[39] M. Tanimoto, N. Fukushima, T. Fujii, H. Furihata, M. Wildeboer, M. P. Tehrani, “Moving multiview camera test sequences for MPEG-FTV,” ISO/IEC JTC1/SC29/WG11/ M16922, Xian, China, 2009

[40] Information technology -- Coding of audio-visual objects -- Part 10: Advanced Video Coding, ISO/IEC 14496-10, 2003

[41] H.264/AVC Software Coordination [Online]. Available: http://iphome.hhi.de/suehring/tml/

[42] P. Quax, F. Di Fiore, P. Issaris, W. Lamotte, F. Van Reeth, “Practical and Scalable Transmission of Segmented Video Sequences to Multiple Players using H.264,” in Motion in Games 2009 (MIG09), Lecture Notes in Computer Science LNCS series, LNCS 5884, pp. 256-267, 2009

[43] A. Zare, P. T. Kovács, A. Gotchev, “Self-Contained Slices in H.264 for Partial Video Decoding Targeting 3D Light-Field Displays,” in Proc. 3DTV Conference 2015, Lisbon, 2015

[44] Information technology -- High efficiency coding and media delivery in heterogeneous environments -- Part 2: High efficiency video coding, ISO/IEC 23008-2, 2013

[45] G. J. Sullivan, J. Ohm, H. Woo-Jin, T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," IEEE Trans. Circuits and Systems for Video Technology, vol.22, no.12, pp.1649-1668, Dec. 2012

[46] K. M. Misra, C. A. Segall, M. Horowitz, S. Xu, A. Fuldseth, M. Zhou, “An overview of tiles,” IEEE Journal of Selected Topics in Signal Processing, vol. 7, no. 6, pp. 969-977, Dec. 2013.

[47] HEVC Test Model [Online]. Available: https://hevc.hhi.fraunhofer.de/HM-doc/

[48] A. Zare, P. T. Kovács, A. Aminlou, M. M. Hannuksela, A. Gotchev, “Decoding complexity reduction in projection-based light-field 3D displays using self-contained HEVC tiles,” in Proc. 3DTV Conference 2016, Hamburg, 2016

[49] Digital compression and coding of continuous-tone still images, ISO/IEC 10918-1, 1994

[50] JPEG 2000 image coding system, ISO/IEC 15444, 2000

[51] Comprimato JPEG2000 encoder and decoder [Online]. Available: http://www.comprimato.com/

[52] S. Grgic, M. Mrak, M. Grgic, “Comparison of JPEG Image Coders”, in Proc. of the 3rd International Symposium on Video Processing and Multimedia Communications, 2001, pp. 79-85

[53] A. Al, B. P. Rao, S. S. Kudva, S. Babu, D. Sumam and A. V. Rao, "Quality and complexity comparison of H.264 intra mode with JPEG2000 and JPEG," Image Processing, 2004. ICIP '04. 2004 International Conference on, 2004, pp. 525-528 Vol. 1.

[54] M. T. Pourazad, C. Doutre, M. Azimi and P. Nasiopoulos, "HEVC: The New Gold Standard for Video Compression: How Does HEVC Compare with H.264/AVC?," in IEEE Consumer Electronics Magazine, vol. 1, no. 3, pp. 36-46, July 2012

[55] P. Quax, P. Issaris, W. Vanmontfort, W. Lamotte, “Evaluation of distribution of panoramic video sequences in the eXplorative television project,” in Proc. 22nd International Workshop on Network and Operating System Support for Digital Audio and Video, Toronto, 2012, pp. 45-50

[56] G. Tech, Y. Chen, K. Muller, J.-R. Ohm, A. Vetro, and Y.-K. Wang, "Overview of the Multiview and 3D Extensions of High Efficiency Video Coding", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 26, Issue 1, pp. 35-49, Sept. 2015

[57] RAJPEG Technology – Scalado [Online]. Available: http://www.scalado.com/display/en/RAJPEG+Technology

[58] P. T. Kovács, T. Balogh, J. Konieczny, G. Cordara, “Requirements of Light-field 3D Video Coding,” ISO/IEC JTC1/SC29/WG11/M31954, San Jose, 2014

[59] P. T. Kovács, Zs. Nagy, A. Barsi, V. K. Adhikarla, R. Bregovic, “Proposal for additional features and future research to support light-field video compression,” ISO/IEC JTC1/SC29/WG11/M37434, Geneva, Switzerland, 2015

[60] S. Shimizu, G. Bang, T. Senoh, M. P. Tehrani, A. Vetro, K. Wegner, G. Lafruit, M. Tanimoto, “Use Cases and Requirements on Free-viewpoint Television (FTV) v.2,” ISO/IEC JTC1/SC29/WG11/N15732, Geneva, 2015

## Author Biography

*Péter Tamás Kovács received M.S. degree in computer science from the Budapest University of Technology and Economics, Budapest, Hungary in 2004. He is currently pursuing the PhD in signal processing at Tampere University of Technology, Tampere, Finland. From 2006 to 2016 he was with Holografika, where he worked on 3D light-field displays and related capture, compression and rendering technologies. His research interests include 3D displays, light-field in particular, real-time rendering and video compression.*

*Alireza Zare is a master student in the Department of Signal Processing at Tampere University of Technology (TUT). In July 2014, he joined 3D Media Group at TUT as a Research Assistant. Since October 2015, he has been an External Researcher with Nokia Technologies, Tampere, Finland. Currently, his research field is mainly focused on video coding and its application in virtual reality.*

*Tibor Balogh is a graduate of the Budapest Technical University.He is the Founder and CEO of Holografika, and has expertise in the fields of holography, lasers, electro-optical technologies and engineering. Tibor's work has led to his being awarded the Dennis Gabor Prize. He was a World Technology Award finalist in 2006. Tibor holds numerous patents for 3D display, has authored a large number of publications presenting aspects of his display work.*

*Robert Bregović received his MSc in electrical engineering from University of Zagreb (1998) and his Dr.Sc.(Tech) in information technology from Tampere University of Technology (2003). He has been working at Tampere University of Technology since 1998. His research interests include the design and implementation of digital filters and filterbanks, multirate signal processing, and topics related to acquisition, processing/modeling and visualization of 3D content.*

*Atanas Gotchev received the M.Sc. degrees in radio and television engineering (1990) and applied mathematics (1992) and the Ph.D. degree in telecommunications (1996) from the Technical University of Sofia, and the D.Sc.(Tech.) degree in information technologies from the Tampere University of Technology (2003). He is an Associate Professor (Tenure Track) at Tampere University of Technology. His recent work concentrates on algorithms for multisensor 3-D scene capture, transform-domain light-field reconstruction, and Fourier analysis of 3-D displays.*

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-4261-9
ISSN 1459-2045