Single-shot structured light with diffractive optic elements for real-time 3D imaging in collaborative logistic scenarios

Darko Vehar[a], Andreas Hermerschmidt[b], Rico Nestler[a], and Karl-Heinz Franke[a]

[a]ZBS e. V., Werner-von-Siemens-Straße 12, 98693 Ilmenau, Germany
[b]HOLOEYE Photonics AG, Volmerstraße 1, 12489 Berlin, Germany

ABSTRACT

We introduce an innovative concept for 3D imaging that utilizes a structured light principle. While our design is specifically tailored for collaborative scenarios involving mobile transport robots, it is also applicable to similar contexts. Our system pairs a standard camera with a projector that employs a diffractive optical element (DOE) and a collimated laser beam to generate a coded light pattern. This allows a three-dimensional measurement of objects from a single camera shot. The main objective of the 3D-sensor is to facilitate the development of automatic, dynamic and adaptive logistics processes capable of managing diverse and unpredictable events.

The key novelty of our proposed system for triangulation-based 3D reconstruction is the unique coding of the light pattern, ensuring robust and efficient 3D data generation, even within challenging environments such as industrial settings. Our pattern relies on a perfect submap, a matrix featuring pseudorandomly distributed dots, where each submatrix of a fixed size is distinct from the others.

Based on the size of the working space and known geometrical parameters of the optical components, we establish vital design constraints like minimum pattern size, uniqueness window size, and minimum Hamming distance for the design of an optimal pattern. We empirically examine the impact of these pattern constraints on the quality of the 3D data and compare our proposed encoding with some single-shot patterns found in existing literature.

Additionally, we provide detailed explanations on how we addressed several challenges during the fabrication of the DOE, which are crucial in determining the usability of the application. These challenges include reducing the 0th diffraction order, accommodating a large horizontal field of view, achieving high point density, and managing a large number of points. Lastly, we propose a real-time processing pipeline that transforms an image of the captured dot pattern into a high-resolution 3D point cloud using a computationally efficient pattern decoding methodology.

Keywords: single-shot-3D, structured light, diffractive optical element, M-array, perfect submap, pseudo-random arrays, real-time 3D imaging

1. INTRODUCTION

In a stereoscopic depth measurement, an object is captured from different spatial positions with two cameras. The 3D (depth) information is determined by triangulation based on the resulting depth-dependent displacement (disparity) between between corresponding pixel positions of a scene point. The correct assignment of image point correspondences presents a significant challenge. This ill-posed problem doesn't have a general solution since neither the existence nor the uniqueness of correspondences is guaranteed when capturing unknown scenes. The correspondence problem can be simplified by substituting one of the cameras with a projector, thereby projecting a known pattern onto the object surface that is captured by the other camera. The projected pattern should be designed so that there is an unambiguous positional coding throughout the image, ensuring the presence and uniqueness of correspondences. These are called structured light systems for 3D reconstruction.

Further author information: (Send correspondence to D.V.)
D.V.: E-mail: darko.vehar@zbs-ilmenau.de, Telephone: +49 (0)36 7768 976 86
A.H.: E-mail: andreas.hermerschmidt@holoeye.com, Telephone: +49 (0)30 4036 938 27

Nevertheless, even in cases where the projection of a known pattern exhibits existence and uniqueness, an industrial environment presents significant challenges. Factors such as ambient and global natural or artificial illumination being particularly problematic due to its broadband and energetic strength, as well as materials with varying colors, reflectivity, and high contrast textures, can distort or even obliterate the original pattern. Moreover, the dynamic nature of the scenes in industrial environments poses an additional challenge, especially when the 3D-system is part of a mobile robot. In order to always ensure fast and safe actions of robots in environments, 3D data acquisition must take place with minimal latency and thus high temporal availability (3D frame rate).

Various optical 3D technologies are available for enabling mobile robots to detect obstacles and people in real-time, particularly in logistics scenarios. These technologies include stereo vision cameras, structured light cameras, Time-of-Flight (ToF) cameras, and (flash) LiDAR systems, each offering a unique balance of accuracy, range, and cost. In this paper, we focus on structured light technology, presenting a sensor composed of a camera and a DOE projector, capable of capturing 3D images in a single camera shot.

1.1 Novelty of the Proposed Approach

Our approach imposes new conditions on the imaging properties of the camera and projector, as well as their geometric relationship. We propose a front-parallel setup where the camera and projector have the same object-side field angle. Drawing inspiration from stereo camera systems, this configuration has two significant implications for the projected structured light, which we utilize to enhance the efficiency and robustness of our pattern decoding.

- Firstly, in this geometric setup, let us assume both the camera and projector are modeled using a pinhole camera model, with identical object-side field angle. As a result, any spot emitted by the projector onto a surface parallel to the two image planes will map to a spot of the same size in the camera image, regardless of the surface's distance. Each distinct codeword can then be decoded using identical image processing steps, simplifying and streamlining the detection process. Though this principle aligns with the operational logic of stereo vision systems, it is often neglected in structured light systems. This represents a significant departure from traditional methods and increase the the efficiency of the process.
- Secondly, in a configuration where the camera and projector planes are parallel, a codeword from the projected pattern aligns with the corresponding epipolar line—effectively becoming an image row in the case of horizontal alignment (and similarly, an image column for vertical alignment). Ideally, this means the projector pattern only needs to contain a single row or column of pixels, spanning the entire height or width of the projector image. This arrangement facilitates the generation of compact patterns, the length of which depends on the disparity range of the stereo setup, as well as the minimum height and the smallest possible codewords. This methodology allows for detecting finer 3D variations in the scene while maximizing the hamming distance. It ensures that the pattern decoding remains resilient to external factors affecting the scene image capture, such as changes in ambient lighting, sensor noise, or even occlusions. This strategy significantly deviates from conventional published patterns, which typically prioritize maximizing the pattern's width and height and ensuring codeword uniqueness, often overlooking considerations of robustness.

Considering that objects in an industrial scene are not flat planes but encompass a variety of 3D shapes, and acknowledging that a projector image cannot always be rectified as in a stereo camera setup, we outline key decisions concerning the geometrical properties of the pattern. We propose an innovative decoding method for efficient and robust 3D reconstruction of a scene using a camera and projector. This efficient decoding enables real-time processing, facilitating rapid and accurate 3D scene reconstruction.

The paper is further structured as follows: Section 2 reviews the state of the art in structured light systems for 3D reconstruction. Section 3 presents our previous work on systems that utilize single-shot 3D reconstruction. In Section 4, we introduce the proposed pattern, and discuss the design and manufacturing details of the diffractive optical element. Section 5 outlines the processing pipeline for 3D reconstruction using a diffractive optical element based projector and a camera. Section 6 provides a comprehensive discussion of our experimental results. Section

7 summarizes the advantages of our proposed single-shot structured light system. Finally, Section 8 concludes the paper and outlines future directions for our research.

2. STATE OF THE ART

2.1 Triangulation Based 3D Cameras

Our primary focus is on the properties of commercially available systems, as these have demonstrated robust and stable performance across diverse environments. This includes general-purpose triangulation-based 3D imaging sensors, which are suitable for both natural and industrial settings. Although our approach belongs to the category of spatial light coding systems, we also draw parallels with stereo vision systems due to the similar principles applied.

2.1.1 Stereo Vision Cameras

Historically, stereo vision systems such as Pointgrey's Bumblebee and Videre's STOC (Stereo on Chip) were widely used. However, technological advancements have led to the introduction of systems that incorporate additional texture illumination. This technique effectively addresses areas on objects with insufficient texture, where traditional correspondence analysis might struggle. The illuminating pattern is typically a dot pattern, either random or pseudo-random. Established block matching methods are applied alongside disparity optimization to solve the stereo correspondence problem. Given the known camera parameters, these systems can also generate 3D point clouds. Several companies have brought to market single-shot systems that incorporate these advancements, including Intel's Realsense, IDS's Ensenso, ZED by Stereolabs, and the MYNT EYE depth camera. Significant contributions to the domain of multi-shot systems have come from systems like Fraunhofer IOF's Gobo and Cognex's 3D-A5000, formerly known as EnShape. The ongoing development and refinement in this field underscore the potential of stereo vision cameras for sophisticated 3D imaging and depth sensing.

2.1.2 Structured Light Systems Using a Single Camera and 2D-pattern Projector

Spatial light coding In the field of single-shot spatial light coding, only two systems have thus far been robust enough for the practical and commercial use.

The first of these, patented as the "Light Coding" method, was introduced by Primesense Ltd. in 2007. In collaboration with Microsoft, they developed a depth sensor that was integrated into the Kinect for Xbox360 in 2010. The device operates by projecting a fixed dot pattern, generated using an NIR laser and diffractive optics, onto the scene. This pattern is then captured by a camera, generating a depth image with a resolution of 640 x 480 at a rate of 30 frames per second. The patent suggests the potential for a decoding step to be executed in the Fourier transformed image, as evidenced by the almost square-like shape of the pattern. Beyond Kinect, this technology from Primesense was also utilized in products such as ASUS's WAVI Xtion, the Structure Sensor by Occipital Inc., and the Astra by Orbbec. After Primesense was acquired by Apple, the technology is integrated into the iPhones for the FaceID functionality.

The second system, developed by Chiaro Technologies, relies on their patented "Symbolic Light" technology. This system, known as Cloudburst, operates at a rate of 15 frames per second, providing 40,000 3D data points. The row-like structure of their pattern implies that they leverage the epipolar constraint during the decoding step. This technology was commercialized by Cognex, with the release of the 3D-A1000 in 2019.

These two systems showcase the progress in single-shot spatial light coding, and its potential influence on the evolution of depth-sensing and 3D reconstruction technologies.

Temporal light coding The majority of commercially available systems based on the structured light principle belong to this category. They predominantly employ Grey-Code-Sequences and phase-shifting methods to generate high-resolution 3D images of static scenes. Often referred to as 3D scanners, these systems typically use projector technologies such as Digital Micromirror Devices (DMD), Liquid Crystal Displays (LCD), or Liquid Crystal on Silicon (LCoS) to project a sequence of temporally coded patterns. Due to this temporal encoding of each point within the scene, it becomes theoretically feasible to establish pixel correspondences without considering adjacenct pixels.

2.2 Types of Projectors for Structured Light

This subsection explores the three main types of projectors used for structured light applications, including photomask-equipped projectors, programmable projectors, and those realized with diffractive optical elements and a collimated laser beam, highlighting their advantages and disadvantages in terms of energy efficiency, image decoding, and physical size.

A projector equipped with a photomask and a light source e.g. LED, such as the one depicted in Figure 1, represents a cost-effective and simple construction. This configuration has the notable advantage of being able to mimic the exact imaging properties of the camera. This advantage is particularly relevant for our research, as we assume an ideal stereo setup which allows for efficient decoding of the projected pattern.

Programmable projectors employ technologies like DMD, LCD or LCoS. They can project high-resolution monochromatic or color images. Such types of projectors are instrumental for prototyping since they can be programmed to display any pattern type. They are frequently used in commercial systems featuring time-multiplexed light patterns such as phase shifting. One of their key advantages is that, after geometric calibration (5.1), their image can be undistorted before projection, thereby simplifying the decoding and triangulation steps. However, their physical size is a major drawback.

Finally a projector realized with diffractive optical elements and a collimated laser beam can project binary patterns. It has the advantage of being compact enough to fit even into mobile phones and is very energy-efficient, as almost all of the laser energy is used for scene illumination. As a specific laser wavelength is used, disturbing ambient light can be filtered out using a narrow-band filter in front of the camera. This is more challenging to achieve with other projector types due to their use of broadband light sources.

2.3 Patterns Based on M-arrays and Perfect Submaps

Structured light patterns can be designed using various mathematical techniques to create unique and identifiable patterns. One such method uses M-arrays or perfect submaps, which allow for high resolution single-shot 3D imaging.

M-arrays and perfect submaps, often used interchangeably, actually have unique mathematical distinctions. An M-array, or pseudorandom array, is an array that encompasses all permutations of submatrices of constant size, excluding only the zero-filled matrix. Contrarily, a perfect submap[12] is a subset of an M-array and is defined as a matrix where each submatrix of constant size occurs at most once. This makes perfect submaps often more fitting for structured light systems, providing necessary uniqueness and flexibility over M-arrays' typically redundant exhaustive permutations.

Types of M-array Symbols An M-array, as previously discussed, comprises elements derived from an alphabet of k symbols. The choice of symbols largely depends on the projector used. Additionally, the shape and size of the uniqueness window are selected with a specific decoding strategy in mind. Symbols can vary extensively in size, ranging from a single pixel[3,4] up to geometric shapes[5678] that extend across multiple projector pixels. They can utilize monochrome coding, color coding, or a combination of the two.[9]

Strategies that depend solely on color or gray levels for symbol encoding may be sensitive to the spectral properties and texture of the observed surfaces.[10] Binary coding often becomes the preferred choice in many applications aiming to achieve the highest signal-to-noise ratio. This method can be efficiently implemented using a collimated laser beam combined with a DOE.[4]

2.4 Pattern Decoding Strategies

The choice of decoding strategy depends largely on the specific needs of the structured light system. One frequently employed decoding technique is template matching.[5] This method operates by correlating the captured image with a known pattern, facilitating the identification of codewords within the image. However, this approach requires multiple templates for each known pattern to address the potential geometric distortions of the projection.

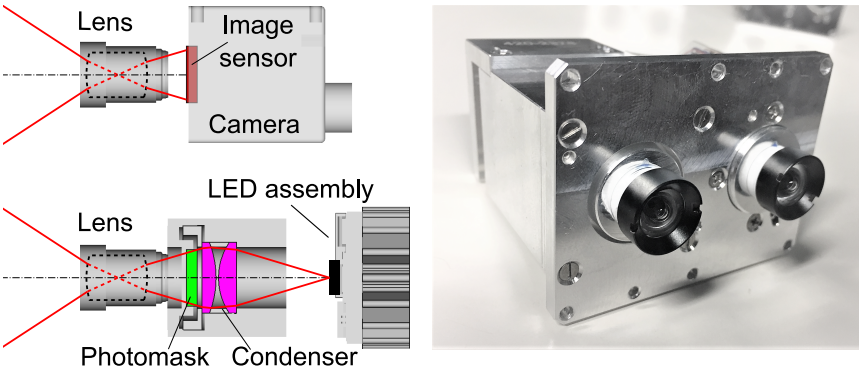

Figure 1. Schematic of the camera projection module (left) and the assembled prototype (right) for 3D surface acquisition in sewer inspection scenarios. The projector is realized through the combination of a white LED, a condenser, and a photomask featuring an M-array pattern.

Tang et al.[8] implement a multi-template strategy to detect the corner points of a binary grid pattern. The white grid cells within this pattern contain geometric symbols. These symbols undergo classification in a subsequent step via a deep neural network, an approach analogous to the task of handwritten digit recognition.

The methodologies outlined in Song et al.[11] and Gu et al.[7] also apply a different grid pattern. Initially, they detect and refine the grid cell corners, followed by the classification of embedded symbols in the grid cells using a deep neural network. Song et al. employ a classic image processing algorithm for grid cell corner detection, while Gu et al. leverage the U-Net architecture[12] for the same task. Although using grid cell corner points as a reference for triangulation-based 3D reconstruction often results in a relatively low lateral resolution, Jia et al.[6] partially mitigate this issue by incorporating distinct feature points of the geometric symbols as reference points.

A higher lateral resolution can be achieved by employing binary patterns with single-pixel-sized symbols, designated either as 0 (black pixel) or 1 (white pixel). As Wijenayake et al.[3] noted, patterns in this category should adhere to the constraint of no two white pixels being adjacent (based on the eight-neighborhood concept). This restriction is crucial for independently identifying each white projected pixel within a captured image.

3. OUR PREVIOUS WORK WITH A STRUCTURED LIGHT-BASED 3D SYSTEM

Our team has expertly designed and constructed structured light single-shot 3D sensors[13] based on perfect submap projection pattern for the inspection of sewer pipe systems. It is engineered to provide a very large field of view, with 76° vertically and 61° horizontally, and a stereo angle of 29°. The prototype shown on the right in Figure 1 is one of six mutually rotated modules for a complete 360-degree 3D capture of sewer surfaces ranging from 200 mm to 400 mm in diameter. The robust 3D-modules, mounted on a mobile carrier, are designed to withstand the harsh conditions within sewer networks while delivering high-resolution 3D imaging capabilities. The lack of scene spectral and textural information in the captured camera images of the pattern is compensated with an additional ambient illumination and texture image acquisition. This capability allows for automatic detection and categorization of defects, enabling efficient and accurate inspections in AI-supported downstream systems.

The structured light projector in each module is composed of three key elements: a white LED, a condenser, and a photomask. The binary pattern utilized, shown in Figure 3, is a perfect submap with a minimum Hamming distance of 3 and a minimum word weight of 4. The camera is programmed to alternate between capturing the texture and the projected patterns as it traverses the sewer pipe, effectively rendering the sewer surfaces in 3D with the level of detail demonstrated in Figure 2.

The employed projector principle incorporating a photomask did not offer optimal energy utilization, and the projection suffered from depth-of-field limitations in the measurement room. Consequently, we decided to explore the use of a DOE paired with a laser, which promises greater energy efficiency and potential for miniaturization.

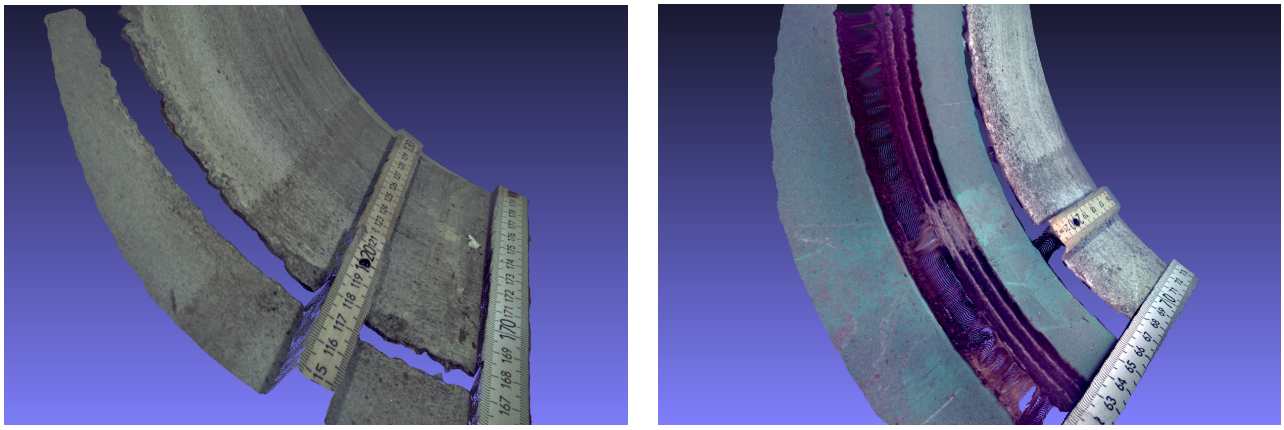

Figure 2. 3D color-textured point cloud of the junction between two concrete pipes (left) and the juncture between a plastic and concrete pipe (right), captured by the structured light module for sewer inspection.

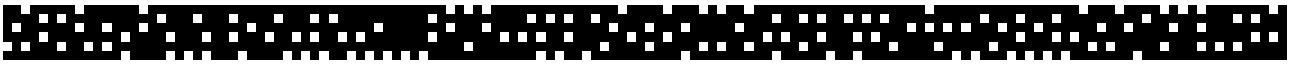

Figure 3. The binary pattern used in the structured light modules for sewer inspection features a uniqueness window size of 6×6 and a minimum Hamming distance of 3. This pattern is concatenated both horizontally and vertically to fill the entire projector image.

4. PROPOSED PATTERN AND DOE DESIGN

The projection pattern consists of points arranged in a grid. It is binary-coded, i.e. it consists of bright points on a dark background. The pattern is composed of the horizontal and vertical concatenation of the so-called base pattern. The base pattern consists of individual overlapping sub-patterns - codewords - of the same size as shown in the Figure 5.

Codeword A codeword is a square-shaped matrix with a side length of W, which should be less than or equal to the height H of the base pattern. This codeword consists of points arranged in a grid under the following conditions: a) two points must not be adjacent in an eight-neighborhood, b) a codeword should not contain

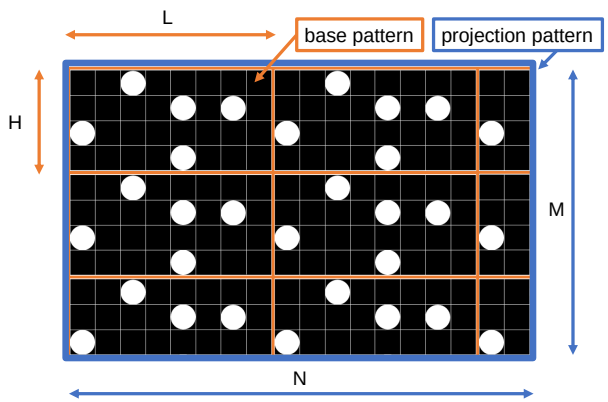

Figure 4. Exemplary representation of the projection pattern (of the projector image), marked as blue rectangle of size M×N consisting of several concatenated base patterns (orange) of size H×L.

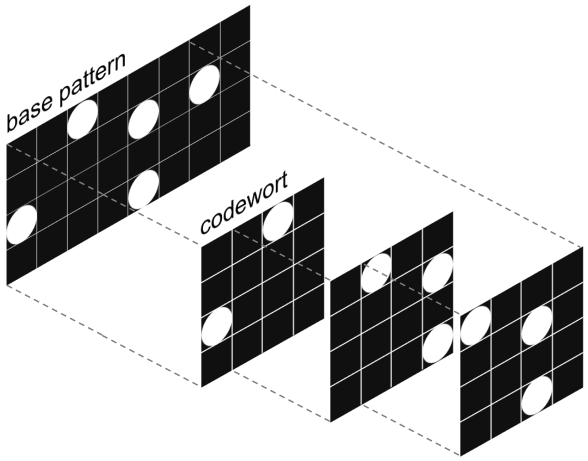

Figure 5. A base pattern of length L=8 and height H=4 consisting of 32 different coded codewords of size 4×4. Three of the 32 codewords are highlighted.

two or more empty columns, and c) the minimum weight of the codeword (number of spots) should exceed one. The first two conditions aim to avoid cases where two adjacent dots (a dash) might be imaged as a single dot on a slanted surface, while the third condition helps to prevent noise-induced single bright spots from being detected as valid codewords. A flat codeword defined as a matrix implies that the object surface to be captured within a codeword must be approximately continuous (continuity constraint). The size of the projected codeword determines the minimum detectable object size, suggesting that codewords should ideally be as small as possible.

Base pattern A base pattern exhibits the properties of the perfect submap. A base pattern of length L and height H comprises exactly L×H overlapping codewords, all distinct from one another. The pattern is designed such that its left and right, as well as upper and lower sides, can be connected, demonstrating the so-called toroidal property. This becomes necessary as the projector image is composed of horizontally and vertically concatenated base patterns. The projector resolution M×N, representing the number of displayable points in the horizontal and vertical directions, thus defines the number of measurable 3D points.

Minimal pattern size The minimum length of the base pattern depends on the geometric arrangement of the camera and projector and on the smallest and largest distance of the object to be measured from the camera or projector. We simplify this by assuming a standard stereo configuration, which makes the equation for disparity more straightforward. The effective disparity range, denoted as Δd, measured in the projector image coordinates, is pivotal for determining the minimum base pattern length, L.

$$\begin{aligned} L = \Delta d = d_{max} - d_{min} &= \frac{f_{proj}B}{Z_{min}} - \frac{f_{proj}B}{Z_{max}} \\ &= f_{proj}B\left(\frac{1}{Z_{min}} - \frac{1}{Z_{max}}\right) \end{aligned} \tag{1}$$

In the equation above, f_{proj} represents the effective focal length of the projector, B is the baseline or the distance between the camera and the projector, Z_{min} and Z_{max} are the minimum and maximum distances of the object from the camera or projector, respectively.

In an ideal situation, the height of a base pattern and the height of a codeword only need to span a single projector pixel. This is not the case in real scenarios, where the image sensor and the DOE are not perfectly aligned or when there is observable radial distortion in both the camera and the projector. An additional height of the base pattern then becomes necessary. This adjustment compensates for misalignment and distortion, ensuring unambiguous assignment of the projected base pattern from the camera image to the corresponding region in the projector image. In situations where the camera and projector are convergent or divergent with respect to each other, the height H of the base pattern must gain additional height to ensure that it encompasses the maximum vertical distance between two epipolar lines in the projector image.

Restrictions posed by the application of DOE The requirements for our use case encompass a large projection angle, high contrast, and effective suppression of the undesired undiffracted laser beam, usually also referred to as the 0th-order beam. As highlighted by Vandenhouten et al.,[4] these criteria can currently only be fulfilled by binary DOEs. The authors also note that these binary DOEs can only generate patterns that maintain central symmetry with respect to the 0th order of diffraction, i.e., the pattern's center. We have accordingly adapted our pattern generation algorithm to accommodate this constraint. Furthermore, when dealing with central symmetric patterns, base pattern concatenation should be executed from the center of the projector image outward towards the edges.

4.1 Deriving Geometrical Properties of the Proposed DOE Binary Pattern

Our proposed system, as shown in Figure 6, is designed to capture objects at distances of up to 3.5 meters. We selected a camera based on the 1/1.2" Sony IMX174 equipped with an 8mm lens, offering a horizontal field of view (FOV) of 69° and a vertical FOV of 47°. The camera features a resolution of 1936 × 1216 pixels. We

Figure 6. An experimental setup featuring a 2.35 MPix camera, collimated laser, and DOE. The setup was geometrically calibrated using a checkerboard target, visible in the background on the right.

Figure 7. Perfect submap with 6×6 uniqnuness window and hamming distance of 3.

determined its intrinsic parameters using our commercial toolkit 3D-EasyCalib,[14] with the effective focal length $f_{cam} = 1411.5px$ being crucial for subsequent computations.

To differentiate between individual spots, we set the projector resolution to half of the camera's resolution. Consequently, we calculated the effective focal length of the projector, f_{proj}, to be $f_{cam}/2 = 705.25px$. This allowed us to compute the baseline B, which is required to achieve a specific depth resolution ΔZ at some depth Z, and can be expressed as:

$$B = \frac{Z^2}{f_{proj}\Delta Z}\Delta d, \tag{2}$$

where Δd specifies the minimum disparity difference still detectable by the decoding algorithm, typically falling within the subpixel range. We set the baseline to $0.1m$.

We defined the target object capture range from $Z_{min} = 0.75m$ to $Z_{max} = \infty$. Applying these values to the pattern length equation (eq.1), we found that the pattern must be at least $L = 94.1px$ wide. Based on previous experience with generating perfect submaps, we set the uniqueness window size to 6×6. We aimed to produce a pattern of the necessary length with the largest possible minimum Hamming distance. The resulting pattern, as depicted in Figure 7, achieves a minimum Hamming distance of 3. The full projector pattern, measuring 968 $\times$ 607 pixels, consists of 100,414 spots.

4.2 Design and manufacturing of the Diffractive Optical Element

For excellent suppression of the undiffracted order in transmission, we decided to realize the DOE as a binary (i.e. 2-level) surface relief microstructure in fused silica, and the required pattern symmetry was taken into account from the start (see section 4).

While it is possible to create arbitrary diffraction angles even with binary DOEs,[15] it is usually sufficient to compute a periodic DOE with spatial periodicity, provided the unit cell is chosen large enough. For this DOE, a rather large unit cell size of $6mm \times 6mm$ was used, leading to rather small deviations between nominal spatial spot positions and spot positions that can be obtained using the spatial frequency grid represented by the harmonic orders of that unit cell.

For a working distance of 1m, the theoretical average position deviations from the nominal positions were as small as 0.04398mm and 0.04005mm in the x- and y-direction, respectively. Even for the large spot number of

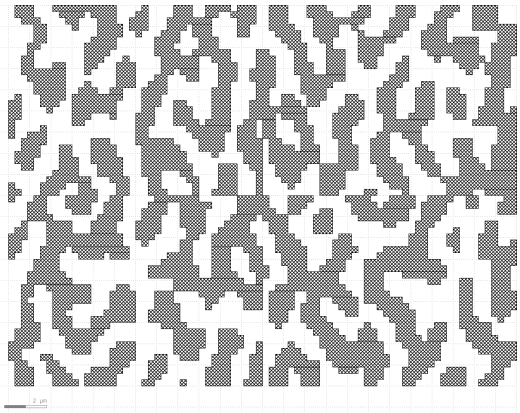

Figure 8. Surface microstructure detail with 300nm pixel size, exhibiting a clearly recognizable anisotropy.

100,414, the maximum deviations were only 3 times that size in either direction. Compared to the minimum, average and maximum spot spacings of 2.60mm, 2.95mm and 4.16mm, respectively, the corresponding average and largest relative position deviations of 1.4% and 4.6% are tolerable. The high precision is due to the high resolution of the spatial frequency grid and its corresponding spatial positions - only 0.1% of the possible positions were used to create the pattern.

Because of the different diffaction angles with respect to the x- and y-axis, the microstructures exhibit a slight anisotropy: the average feature size is smaller along the x-axis compared to the y-axis (see 8). Due to this anisotropy, the zeroth order power is dependent on the polarization direction of the incident light. Therefore, for the design of the microstructure, the polarization has to be taken into account.

The base microstructure was computed using an IFTA[16] algorithm, followed by an optimization based on more precise simulation methods like e.g. RCWA.[17] With the latter method, the polarization dependent effects can be taken into account and the processing parameters can be determined with sufficient precision. The etch depths for suppressing the zeroth order differ by 50nm for the two polarizations. For a nominal etch depth of 1130nm, this amounts to 4.4% difference, which is considerably more than the RIE process depth uncertainty and should not be neglected. The critical dimension of the DOE was 300nm, equal to the pixel size used for the design computations.

The first fabrication step of the DOE was e-beam lithography for the creation of a chromium etch mask on a fused silica substrate. In a subsequent reactive ion etching (RIE) process, the required surface relief profile was obtained. The known fabrication inaccuracies and proximity effects were compensated for, so that a zero order as low as 0.1% was obtained for the target laser wavelength of 850nm and selected polarization.

5. PROPOSED PROCESSING PIPELINE FOR 3D RECONSTRUCTION WITH A DOE BASED PROJECTOR AND A CAMERA

5.1 Geometric Camera and Projector Calibration

The calibration process is crucial to ensure accurate measurements in camera-projector systems that use structured light for computing 3D points through point triangulation. The accuracy of this calibration directly influences the precision of the ensuing 3D measurements. Understanding the importance of calibration is therefore paramount for obtaining reliable results. This section discusses the calibration target and the methods employed for its use with the camera and a projector, which is treated as a inverse camera. Our extensive experience in geometric calibration has enabled us to optimize this process using our own software tools (3D-EasyCalib[14]), yielding optimal outcomes in terms of 3D measurement accuracy.

During the calibration process, we approximate the imaging properties of both the camera and the projector using a pinhole model. To account for the lens distortion observed in the camera, we employ the radial component of a three-term Brown-Conrady distortion model.[18] Real-world setups often exhibit slight deviations from the

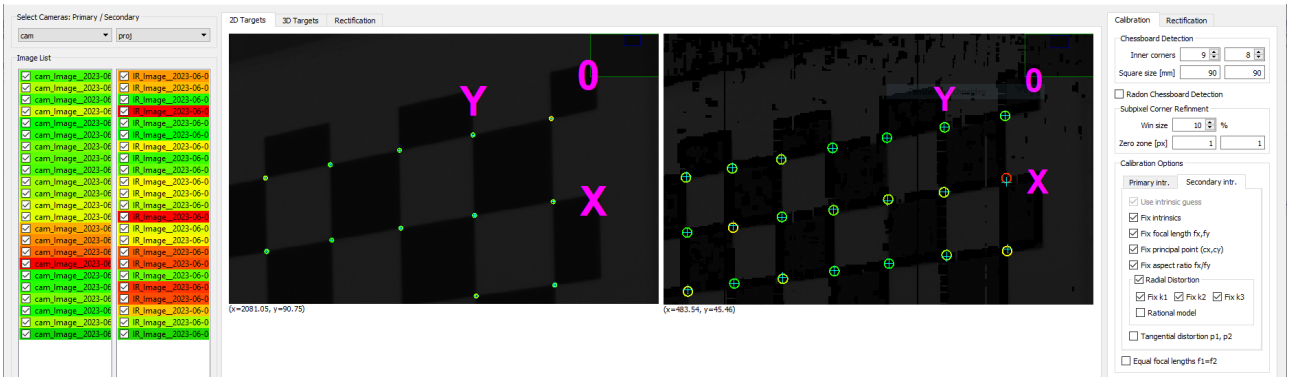

Figure 9. Extrinsic calibration of camera and projector using our commercial toolkit 3D-EasyCalib. The GUI displays calibration points on the target in both the camera image (left) and the projector image (right).

laser's design wavelength, leading to geometrically distorted projected images. The Brown-Conrady model is particularly beneficial for accurately modeling this common occurrence. The calibration parameters are used in the process of geometric image correction, which we perform before 3D-processing.

We capture between 10-20 calibration images for intrinsic camera calibration, projector calibration, and extrinsic calibration, using a target with a checkerboard pattern, as seen in Figure 6. Target detection, sub-pixel processing of calibration points, and parameter optimization were largely automated using the toolkit 3D-EasyCalib.[14]

Additional steps are required for the intrinsic calibration of the projector and the extrinsic calibration of the camera-projector module with a target. For each target pose, we capture a pair of images. The first image is taken with the projected pattern turned off, while the second image is captured with the projected pattern turned on. The first image assists in calibration target detection, while the second image is used for pattern decoding. We apply local homographies around each corner of the checkerboard pattern to accurately transfer calibration points from the camera to the projector images.[19]

The projector calibration matrix, as well as the orientation and pose of the projector in relation to the camera (extrinsics), are estimated through numerical optimization.[14] All these steps are crucial for ensuring the reliability and accuracy of the 3D measurements.

5.2 Pattern Decoding

Decoding the pattern is a fundamental step in structured light imaging. This process involves establishing correspondences between sets of camera pixels and their original codewords—sets of pixels—in the spatially coded pattern of the projector image. Various decoding strategies can be employed (see subsection 2.4), including the use of k-nary search trees or a lookup table with hashing, as implemented in data structures such as an unordered map or dictionary. The choice of decoding strategy can greatly affect the performance and the usability of the structured light 3D system.

Image traversal For each pixel of the image, a block (or window) of size $F \times F$ is spanned around it. We determine the block size by the integer scaling factor $S = f_{cam}/f_{proj} \in \mathbb{N}$ (quotient of the effective focal lengths) and the codeword size $W \times W : F = SW$. If f_{cam}/f_{proj} is not an whole number, we scale the camera image accordingly. Following this, the corresponding homologous point $p_{proj} = (x_{proj}, y_{proj})$ in the projection pattern is determined for each image region I_F of size $F \times F$ with the center coordinates $p_{cam} = (x_{cam}, y_{cam})$.

Spot enhancement We convolve the image section I_F with the mean-free, discrete Gaussian kernel $G_{K\times K}$, $I_G = G_{K\times K} \circ I_F$ to emphasize the spots and suppress the background. The value of σ is adjusted based on the imaged spot size.

Local adaptive binarization Next, the image section I_G from the previous step is converted into a binary representation I_B, using a threshold determination method like that suggested by Niblack.[20] The threshold T of the central pixel (i_F, j_F) of the $F \times F$ block is computed from the mean value m and the standard deviation s of the intensity values within this window: $T\left(i_F, j_F\right) = m + ts$. We choose the parameter t depending on the image contrast. The binary value is then determined as follows:

$$I_B\left(i_F, j_F\right) = \begin{cases} 1, & \text{if } I_G(i_F, j_F) > T(i_F, j_F) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Codeword reading In the binary image region thus pre-processed I_B, we read out the S^2 codewords contained within a regular grid (every S-th place in the column and every S-th place in the row):

$$\begin{gathered} codeword\left(k, l\right) = I_B\left(i_F + kS, j_F + lS\right) \\ k, l \in [1 \ldots W],\ i_F, j_F \in [1 \ldots S] \end{gathered} \tag{4}$$

If there's a large triangulation angle (ratio between the base distance and the distance of the object) or a large scaling factor, we resample the camera image with additional affine-distorted grids.

Correspondence determination with look-up For each codeword read in the image region I_B, we take its position in the base pattern from the coding table (codeword = key) and enter it into the correspondence table. The correspondence table $\mathbb{R}^+ \times \mathbb{R}^+ \rightarrow \left(\mathbb{R}^+ \times \mathbb{R}^+\right)^{S^2} : p_{proj} = (x_{proj}, y_{proj}) \mapsto p_{cam} = (x_{cam}, y_{cam})$ assigns up to S^2 corresponding camera points p_{cam} to each projector point p_{proj}. If the x or y coordinate of the camera pixel exceeds the product of the scaling factor and the base pattern's length $(S \times L)$ or height $(S \times H)$, we adjust the coordinates of the projector point accordingly (see Figure 4),

$$p_{proj} \leftarrow p_{proj} + \left(L\lfloor i_{cam}/SL \rfloor, H\lfloor j_{cam}/SH \rfloor\right). \tag{5}$$

Here, $\lfloor i_{cam}/SL \rfloor$ represents the integer division of the x-coordinate of the camera pixel by the product of the scaling factor S and the length L of the base pattern rounded down, and the expression $\lfloor j_{cam}/SH \rfloor$ is viewed similarly. It is important that we construct a pattern with enough height so that a base pattern can be uniquely identified among all other vertically concatenated base patterns. We enter an "invalid" in the correspondence table if we don't find the codeword in the coding table. For each projector pixel, we compute an outlier-free average across all valid camera pixels.

5.3 3D Point Reconstruction

For 3D-Reconstruction we use the intrinsic parameters K_{cam}, K_{proj} and extrinsic parameters $R, \mathbf{t}$ from the calibration step. Having obtained corresponding points $\mathbf{p}_{cam} = (x_{cam}, y_{cam}, 1)^T$, $\mathbf{p}_{proj} = (x_{proj}, y_{proj}, 1)^T$ in the camera and projector images from the decoding step, we aim to solve these equations to find $\mathbf{X}$. These points are projections of the unknown world point $\mathbf{P}$,

$$\mathbf{p}_{cam} = K_{cam}[I|\mathbf{0}]\mathbf{P}, \quad \mathbf{p}_{proj} = K_{proj}[R|\mathbf{t}]\mathbf{P}. \tag{6}$$

We then combine these two equations according to the linear triangulation method from [21, p. 312] into a form $A\mathbf{X} = \mathbf{0}$ and solve using least squares. The output results in an ordered 3D point cloud: $\{P_{ij} = (X_{ij}, Y_{ij}, Z_{ij})^T \in \mathbb{R}^3, i = 1, \ldots, M, j = 1, \ldots, N\}$, representing the absolute coordinates of the object's surface in the camera coordinate system.

Please note that using a fixed pattern, such as a DOE in a projector, does not allow us to invert the distortion or rectify the camera-projector configuration, as we can do in a stereo camera setup or with a programmable projector. This additional step simplifies and accelerates the execution time of point cloud computation, making it highly recommended for real-time systems, whenever feasible.

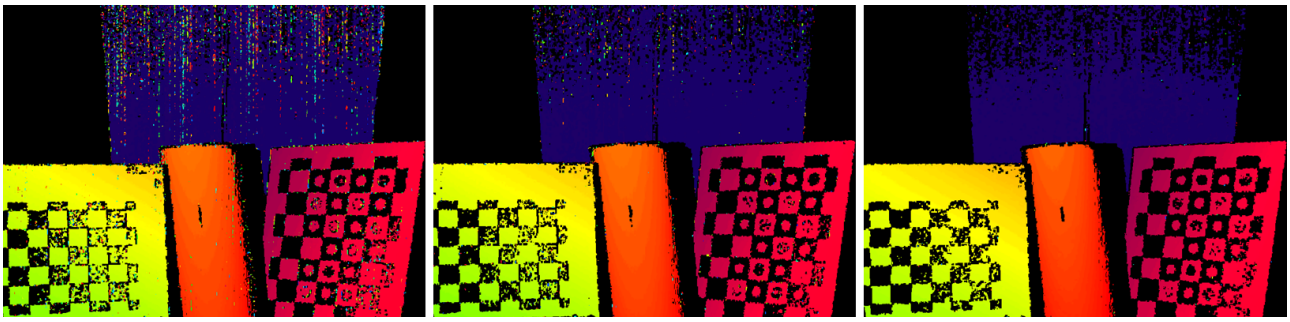

Figure 10. Illustration of the impact of Hamming distance on the decoding process. The leftmost image utilized a pattern with a minimum Hamming distance of 1, the middle image used a minimum Hamming distance of 2, and the rightmost image implemented a minimum Hamming distance of 3. An observable reduction in code collisions is evident as the minimum Hamming distance increases.

6. EXPERIMENTAL RESULTS

To start, we focus on demonstrating the impact of Hamming distance on the resulting decoded image, and consequently, the disparity image. We deployed a DMD projector and assessed a variety of generated binary patterns against a range of real-world scenes. Notably, the only variable manipulated in these tests was the pattern, while the pattern decoding remained constant. The results of these experiments served as valuable constraints for our pattern-generating algorithm.

Of all the variables we examined, we found that Hamming distance exerted the most significant influence on the robustness of the decoded pattern. As shown in Figure 10, we obtained three disparity images by using different minimum Hamming distances: 1, 2, and 3. It's important to note that all other pattern properties, including the uniqueness window size and minimum word weight, were held constant across the experiments.

The effects of varying the minimum Hamming distance were clearly visible in the resulting images. The background of the scene, visualized as a blue plane, has low contrast, and the noise tends to interfere with the projected pattern. In the leftmost image, where the minimum Hamming distance was 1, code collisions occurred frequently, leading to isolated spots of incorrect disparities.

Increasing the minimum Hamming distance proved effective in mitigating these issues. For instance, when we used a minimum Hamming distance of 3, as seen in the rightmost image, the pattern noise was nearly eliminated.

Furthermore, the projected codewords often became significantly distorted in areas of varying contrast, such as on the checkerboard pattern, or when depth discontinuity caused the codewords to break. The increased minimum Hamming distance again proved beneficial in eliminating these code collisions. The first image shows numerous visible code collisions, while the second shows a significant reduction. In the third image, where the minimum Hamming distance was 3, these code collisions were eliminated entirely.

To summarize, our experiments suggest that employing a greater minimum Hamming distance significantly contributes to a more robust decoding process, minimizing code collisions and noise in the decoded patterns.

We further present raw images of the projected pattern (Figure 11), resulting disparity images, and 3D point clouds (Figure 12) generated during the experimental development of our single-shot structured light system. The experimental subjects included everyday objects and people, posing real-world challenges often encountered in structured light applications.

Our system demonstrates a satisfactory detection rate. As with all structured light systems, a certain limitation exists in our method as well - the signal intensity of the projected pattern on distant objects and walls tends to be too low for robust pattern decoding. This problem, common to all structured light systems, arises due to the finite operating range of the equipment.

To demonstrate the depth perception capabilities of our system, we display disparity images focusing primarily on horizontal disparities. This approach, although seemingly simplistic, provides a comprehensive visual representation of scene depth. It is important to note, though, that we do not disregard the significant vertical

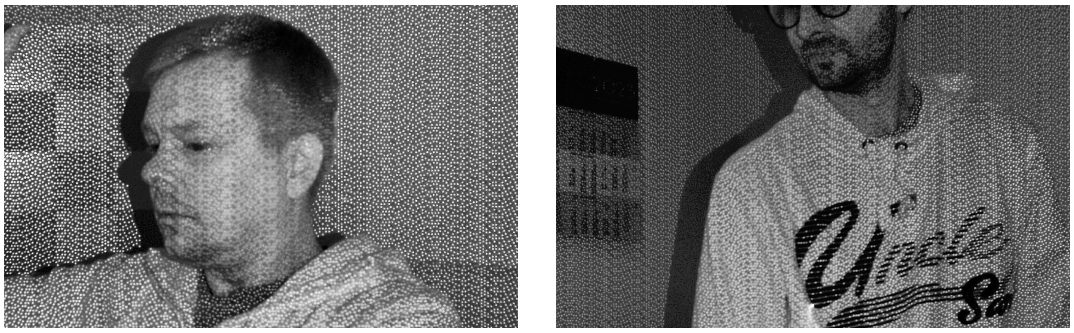

Figure 11. Enlarged image regions showing a projected spot pattern on human skin, walls, and clothes. Contrast has been significantly increased for better visibility. The image on the right displays a 0th diffraction order as a bright spot in the projection center, at the bottom of the cropped image.

disparity. Vertical disparity plays a crucial role during triangulation, providing the necessary information for producing metric measurements. This allows for the preservation of real-world geometry within our results - an achievement significantly more challenging than conventional 'depth sensing' methods, which merely estimate relative depth relationships among objects in the scene.

Our method, therefore, stands out in its ability to capture and represent the scene with a high degree of accuracy in real-time. This accuracy is not just limited to estimating relative distances, but extends to preserving the geometrical integrity of the scene, making it a valuable tool for a variety of applications.

7. ADVANTAGES OF OUR PROPOSED 3D SINGLE-SHOT IMAGING SENSOR

In this section, we gather and highlight the key advantages of our proposed system, designed for triangulation-based 3D reconstruction using a robust unique coding scheme for projecting a light pattern with a DOE-based projector. Our discussion encapsulates both the hardware and software facets of the implementation, considers the requirements specific to industrial applications, and provides a comparative analysis with other stereo and structured light systems.

- Real-time capability for dynamic scenes: Unlike temporally coded methods that require multiple shots to compute depth, our method calculates the depth measurement in a single camera shot. This ability enables us to 3D-image rapid object or camera movements along the optical axis. Notably, the high algorithmic complexity involved in the decoding (correspondence) step often limits the 3D data rate of available systems in the market, which typically achieve only a fraction of the camera frame rate. Our method addresses this limitation, providing a more efficient solution.

- Robustness in non-rectified systems: The lateral expansion of our patterns is realized in both horizontal and vertical directions, which enables decoding of the camera image in both non-rectified images and those affected by optical distortions.

- Reduced computational effort for pattern decoding: Instead of complex cost aggregation and similarity measure calculations for finding homologous points often used in correspondence analysis, we use a lookup table that assigns each codeword its vertical and horizontal position in the projected pattern.

- Ensuring the functional safety of 3D measurement data: Our system is designed such that each image region overlapping a projected codeword corresponds to a single 3D measurement point or is clearly flagged as invalid. This approach draws parallels with signal transmission protocols, similar to the transmission of electrical signals over a wire, and the principles of error detection from coding theory.

- Cost and energy efficient system design: A minimal configuration involving one camera and one projector is more energy-efficient and compact than setups involving, for instance, two cameras and one projector.

Figure 12. Disparity image and point cloud reconstruction of a person waving (top), a person standing (middle) and a plant on a chair (bottom) utilizing our approach and experimental setup.

- Robustness to affine pattern distortions due to the 3D scene: A pattern that uses higher geometrical symbols, like a dash, can appear as a point on a slanted surface. While a dot also undergoes transformation, since we know our pattern only includes dots, this does not confuse the decoding step.
- Maximum 3D measurement point density: While some single-shot methods use abstract, lateral extended symbols, such as circles or lines composed of multiple projector pixels as codewords, our approach ensures that the resolution of the 3D point cloud equals the projector resolution, yielding a denser distribution of 3D measurement points.
- Robust usability even with coloured textured scene objects: Our method employs binary light projection, which can be robustly detected on colored surfaces, unlike color-coded methods that can fail when the colors of the objects in the scene are unknown.

8. CONCLUSIONS AND FURTHER WORK

In this paper, we introduced an innovative approach to 3D snap-shot imaging, leveraging the principles of structured light to create a robust and efficient sensing system. The core of our approach lies in pairing a standard camera with a projector utilizing a diffractive optical element (DOE) and a collimated laser. This combination allows the projection of a unique, specially designed light pattern onto the scene, facilitating effective encoding and decoding for reliable 3D reconstruction.

Our proposed structured light system uses binary patterns with single-pixel-sized symbols. We emphasized the importance of a minimum Hamming distance in pattern generation, which significantly contributes to the robustness of pattern decoding and subsequently improves the quality of the reconstructed 3D scene. Furthermore, we elaborated on the advantages of our system in handling typical industrial scenarios, such as managing depth discontinuities, varying surface textures, and illumination conditions. We showcased the system's performance with real-world experiments, demonstrating the effectiveness of our approach in dealing with these complexities.

The configuration of our 3D sensor is particularly tailored for collaborative scenarios involving mobile transport robots, though its applications are not limited to this. It exhibits great potential for similar environments requiring real-time, reliable 3D data acquisition.

As we move forward, our immediate focus is to evaluate the sensor's performance in a wider range of applications, including but not limited to, industrial quality assurance, patient monitoring in medical environments, and in-vehicle applications. We also plan to explore the potential of DOE-based projectors in different wavelength ranges, aiming to enhance eye safety and pattern contrast, which could potentially extend the sensor's effectiveness to outdoor environments.

By integrating unique pattern coding, robust decoding strategies, and leveraging the capabilities of DOE-based projectors, our proposed system presents a promising contribution to the field of 3D imaging, with a specific emphasis on structured light applications. We anticipate that further refinements and applications of this work will continue to push the boundaries of what is achievable in this exciting field.

ACKNOWLEDGMENTS

The results presented in this article are based on our work within the joint projects AuZuKa "Automatische Zustandsanalyse Kanalnetz durch virtuelle Begehung", grant no. 13N13895 and "Innovative Photonik für Automatische Kollaborative Systeme in dynamischen Waren-Transportprozessen" (AutoKoWaT), grant no. 13N16335. Both projects are funded by the German Federal Ministry of Education and Research (BMBF). AutoKoWaT takes place within the framework of the funding program "Photonik für die digital vernetzte Welt – schnelle optische Kontrolle dynamischer Vorgänge".

REFERENCES

[1] Morano, R. A., Ozturk, C., Conn, R., Dubin, S., Zietz, S., and Nissano, J., "Structured light using pseudorandom codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(3), 322–327 (1998).

[2] Salvi, J., Pagès, J., and Batlle, J., "Pattern codification strategies in structured light systems," *Pattern Recognition* **37**(4), 827–849 (2004).

[3] Wijenayake, U. and Park, S.-Y., "An m-array technique for generating random binary pattern based on a connectivity constraint," in [*Workshop on Image Processing and Image Understanding*], (2012).

[4] Vandenhouten, R., Hermerschmidt, A., and Fiebelkorn, R., "Design and quality metrics of point patterns for coded structured light illumination with diffractive optical elements in optical 3D sensors," *Digital Optical Technologies 2017* **10335**, 1033518 (2017).

[5] Yu, Q. C., Feng, H. M., and Zhang, H. X., "Multi-resolution decoding method of Symbol M array surface structured light," *Proceedings - International Conference on Artificial Intelligence and Computational Intelligence, AICI 2010* **2**, 63–68 (2010).

[6] Jia, X. and Liu, Z., "One-shot m-array pattern based on coded structured light for three-dimensional object reconstruction," *Journal of Control Science and Engineering* **2021**, 1–16 (2021).

[7] Gu, F., Du, H., Wang, S., Su, B., and Song, Z., "High-capacity spatial structured light for robust and accurate reconstruction," *Sensors* **23**(10), 4685 (2023).

[8] Tang, S., Zhang, X., Song, Z., Jiang, H., and Nie, L., "Three-dimensional surface reconstruction via a robust binary shape-coded structured light method," *Optical Engineering* **56**(1), 014102–014102 (2017).

[9] Wijenayake, U. and Park, S.-Y., "Dual pseudorandom array technique for error correction and hole filling of color structured-light three-dimensional scanning," *Optical Engineering* **54**(4), 043109–043109 (2015).

[10] Maurice, X., Graebling, P., and Doignon, C., "Epipolar based structured light pattern design for 3-d reconstruction of moving surfaces," in [*2011 IEEE International Conference on Robotics and Automation*], 5301–5308, IEEE (2011).

[11] Song, Z., Tang, S., Gu, F., Shi, C., and Feng, J., "Doe-based structured-light method for accurate 3d sensing," *Optics and Lasers in Engineering* **120**, 21–30 (2019).

[12] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in [*Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*], 234–241, Springer (2015).

[13] Künzel., J., Vehar., D., Nestler., R., Franke., K., Hilsmann., A., and Eisert., P., "System for 3d acquisition and 3d reconstruction using structured light for sewer line inspection," in [*Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*], 997–1006, INSTICC, SciTePress (2023).

[14] Vehar, D., Nestler, R., and Franke, K.-H., "3D-EasyCalib – toolkit for the geometric calibration of cameras and robots," in [*22. Anwendungsbezogener Workshop zur Erfassung, Modellierung, Verarbeitung und Auswertung von 3D-Daten, 3D-NordOst*], 15–26, GFaI e. V. (Dec. 2018).

[15] Hermerschmidt, A., Krüger, S., and Wernicke, G., "Binary diffractive beam splitters with arbitrary diffraction angles," *Opt. Lett.* **32**(5), 448–450 (2007).

[16] Wyrowski, F. and Bryngdahl, O., "Iterative fourier-transform algorithm applied to computer holography," *J. Opt. Soc. Am. A* **5**(7), 1058–1065 (1988).

[17] Moharam, M. G., Gaylord, T. K., Grann, E. B., and Pommet, D. A., "Formulation for stable and efficient implementation of the rigorous coupled-wave analysis of binary gratings," *J. Opt. Soc. Am. A* **12**, 1068–1076 (May 1995).

[18] Brown, D. C., "Decentering distortion of lenses," *Photometric Engineering* **32**(3) (1966).

[19] Moreno, D. and Taubin, G., "Simple, accurate, and robust projector-camera calibration," *Proceedings - 2nd Joint 3DIM/3DPVT Conference: 3D Imaging, Modeling, Processing, Visualization and Transmission, 3DIMPVT* **2012**, 464–471 (2012).

[20] Niblack, W., [*An Introduction to Digital Image Processing*], Strandberg Publishing Company, Topstykket 17, DK-3460 Birkeroed, Denmark (1985).

[21] Hartley, R. and Zisserman, A., [*Multiple View Geometry in Computer Vision*], Cambridge University Press, New York, NY, USA, 2nd ed. (2003).