Research Article Vol. 1, No. 1 / 15 Jan 2022 / Optics Continuum 103
OPTICS CONTINUUM

Depth from axial differential perspective

ANDREAS FAULHABER,* CLARA KRÄCHAN, AND TOBIAS HAIST

Institute of Applied Optics (ITO), University of Stuttgart, Pfaffenwaldring 9, 70569 Stuttgart, Germany
*faulhaber@ito.uni-stuttgart.de
http://www.ito.uni-stuttgart.de

Abstract: We introduce an imaging-based passive on-axis technique for measuring the distance of individual objects in complex scenes. Two axially separated pupil positions acquire images (can be realized simultaneously or sequentially). Based on the difference in magnification for objects within the images, the distance to the objects can be inferred. The method avoids some of the disadvantages of passive triangulation sensors (e.g., correspondence, shadowing), is easy to implement and offers high lateral resolution. Due to the principle of operation it is especially suited for applications requiring only low to medium axial resolution. Theoretical findings, as well as follow-up experimental measurements, show obtainable resolutions in the range of few centimeters for distances of up to several meters.

1. Introduction

Most technical short range (up to several meters) 3D sensors are currently based on triangulation. This is for good reason, because triangulation offers the possibility to obtain quite accurate distance measurements. Passive sensors (e.g., conventional stereo vision, plenoptic cameras, depth-from-focus, structure from motion) as well as active triangulation sensors using a structured illumination of the scene are used [1–4]. However, there are also disadvantages when using triangulation based sensors, mainly due to the lateral separation of the pupils resulting in potential shadowing, the so-called correspondence problem and a necessary lateral size of the sensor [1].

Although, triangulation plays a major role in distance estimation of human and animal vision - predators highly depend on exact knowledge of prey position - there are other important visual cues, like relative size (perspective) and occlusion, that help to estimate distance [5].

Measuring the image size y' of an object of known size y is a simple, yet powerful technique to estimate the distance to the object and is widely used by e.g., humans. This is possible because the magnification β' is given by [6]:

$$\beta' := \frac{y'}{y} = \frac{a'}{a} \approx \frac{f'}{a} \,. \tag{1}$$

Therefore, measuring y' relatively compared to a given object size y, one can determine the object distance a if the focal length f' of the optical system is given. The focal length given by the objective lens approximates to the image sided distance a' (the distance between the image-sided principal plane and the image plane). However, if the size of the object y is unknown, additional information is necessary.

Current research focuses on finding or estimating the missing information for 3D reconstruction through mainly using machine learning approaches like deep- or convolutional neural networks and (un-)supervised training. Works that employ these algorithmic networks are trained with images and, in the case of supervised learning, with ground truth depth data. The trained network can then be used to estimate depths from unknown 2D images, so called monocular- or single-view depth [7–11].

The standard approach of passive triangulation ("stereo vision") obtains the additional information that is missing in the monocular image by using two separate imaging systems with

#451413 https://doi.org/10.1364/OPTCON.451413
Journal © 2022 Received 16 Dec 2021; revised 16 Dec 2021; accepted 17 Dec 2021; published 13 Jan 2022

OPTICS CONTINUUM

the pupils being laterally separated [1,2]. As already stated, this is a powerful and widely used approach allowing quite accurate depth determination for short distances.

Unfortunately, the lateral separation of the pupils leads to some potential problems. Typically, parts of the scene cannot be measured due to shadowing, i.e., obscuration or occlusion. Also, imaging the objects from different directions might lead to changed appearances and to strong local brightness variations for glossy objects. This makes it difficult to find corresponding pixels in the two images and often leads to a quite large number of pixels for which accurate depth determination is not possible, especially in passive sensors. In addition, the sensor of course needs a certain lateral extent which is given by the triangulation base resulting in many cases in large and bulky devices [1].

In this contribution, we propose to distribute the two pupils *axially* instead of laterally, as shown in Fig. 1. Therefore, occlusions can be avoided and the appearance of objects will not change. By this approach the lateral shift, the so called disparity, of triangulation based sensors is converted to axial into a change of magnification. The difference of the two determined magnifications allows to compute the distance to the object using our method of differential perspective.

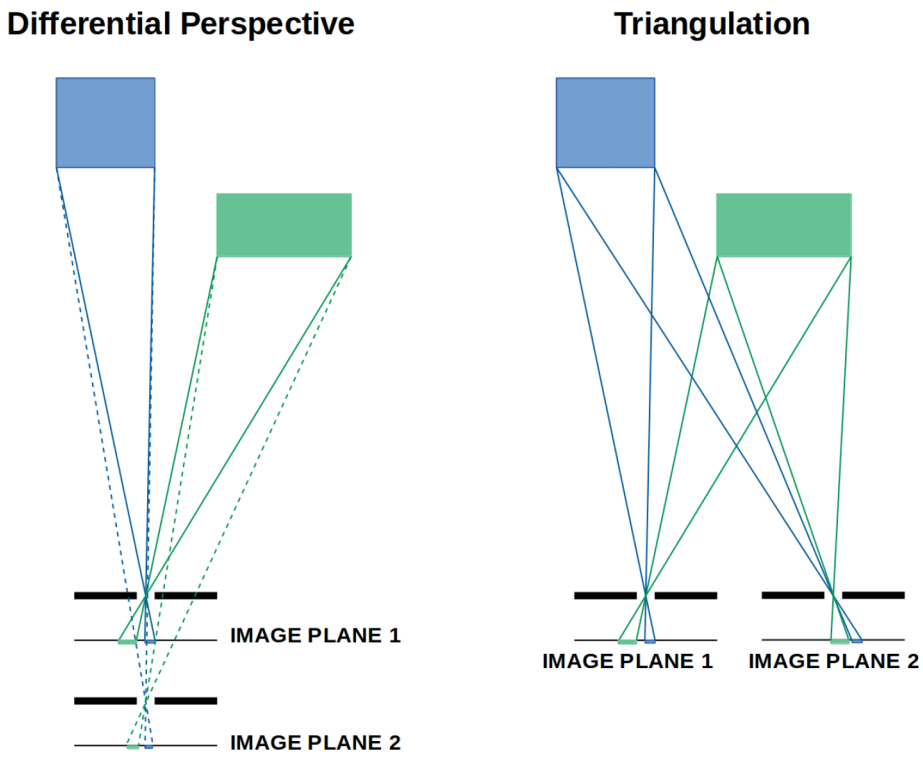

Fig. 1. Principle of differential perspective (left) compared to triangulation (right). The pupils for the two (or more) images are shifted axially instead of laterally. As a result, lateral shift (disparity) is converted into a change of magnification.

Rogers & Bradshaw (1993) already used the term *"differential perspective"* for a certain perspective effect in human vision. They defined it as the varying size ratios of vertical- and horizontal gradients. However, in their definition, the difference in size ratios highly depend on both eyes having a relatively large lateral separation, so again, the result is connected to triangulation (stereo vision) [12]. In contrast, our definition does not rely nor depend on any lateral extent at all.

OPTICS CONTINUUM

Also, Suto et al. (2020), described a related method using two cameras at different positions axially or with different imaging optics (focal length and pupil-size) might seem very similar to our principle approach. Yet, they always utilize at least two imaging optics (and sensors) also with a certain lateral extent, which still can be accounted as a stereo vision method [13].

Another previous publication related to our method is the work by Ma & Olsen (1990). They used a zooming optical lens to generate images at different magnifications through different zooming-factors and calculating the synthetic differences to depths using either gradient-based optical flow or feature matching (standard stereo vision algorithms). Even though our method is to certain extents similar to theirs, they use a mechanically moved lens system with sequential acquisition of images [14].

In our work, we began with a similar sequential method in Sec. 3 to verify our theoretical findings, We then followed with our advanced experimental setup, presented in Sec. 4, of simultaneous acquisition of two camera images with different magnification ratio through an offset. In our method, we first look at the bigger picture of distances of whole objects, which is in most robotic applications more important than the pixel-wise 3D surface depths of objects, as shown by Ma&Olsen. Our depth map can be created of the objects positional depths, which is by far more data efficient through sparse depth data and therefore easier for processing (e.g. object avoidance, SLAM).

2. Theory

In paraxial approximation (here, the employed pinhole camera model) the two different magnifications are given by

$$\beta_1' = \frac{f'}{a} \tag{2}$$

$$\beta_2' = \frac{f'}{a + \Delta a}\,, \tag{3}$$

where we denote the distance of the pinhole to the object and the image plane by a (being negative) and f'. The Δa corresponds to the axial shift of the pupil (pinhole). Of course, for real lenses the distances are to be measured from the corresponding principal planes and the image distance would approximately equal to the focal length of the objective lens.

If we denote the change of magnification by $\gamma = \beta_1'/\beta_2'$, we obtain the following equation for the unknown distance a:

$$\frac{\beta_1'}{\beta_2'} - 1 = -\frac{\Delta a}{a} \quad \rightarrow \quad a = \frac{\Delta a}{\gamma - 1} \tag{4}$$

The determination of $\gamma := \beta_1'/\beta_2'$ is achieved with a particular uncertainty u_γ The uncertainty u_a of the distance position with respect to an uncertainty u_γ is given by

$$u_a = \frac{\partial a(\gamma)}{\partial \gamma}\, u_\gamma = \frac{\Delta a}{(\gamma - 1)^2}\, u_\gamma \tag{5}$$

As one would expect we have a quadratic increase of measurement uncertainty u_a with measurement depth a. This is the same dependence as for stereo vision. However, one has to keep in mind, that the size determination of objects in a scene can be achieved with very high accuracy (see Sec. 3).

In Fig. 2, this distance uncertainty due to realistic uncertainties of magnification differences for particular axial separations of the pupils is shown. For larger separations (e.g., 150 mm) one can achieve uncertainties in the range of centimeters which is sufficient for simple object localization tasks in a wide field of applications, e.g., robotics.

Instead of a pixel-wise working, as it is the case for most stereo vision sensors, our principle is especially suited for distance determination to objects as a whole. Since the appearance of

OPTICS CONTINUUM

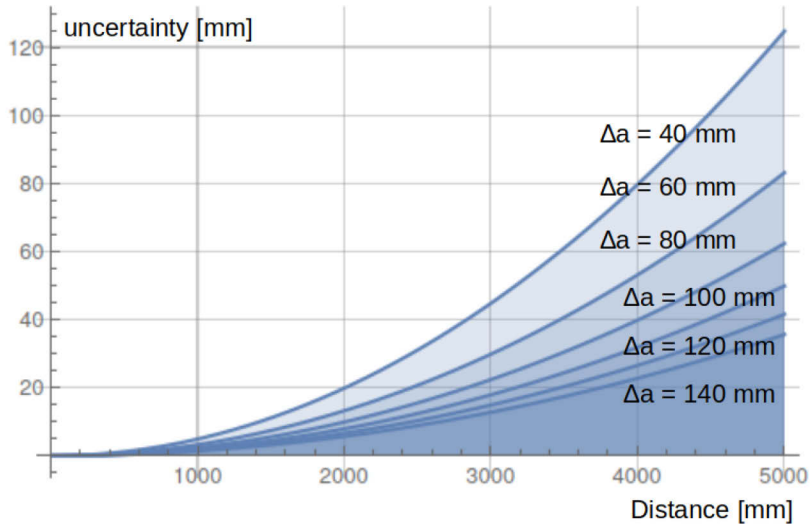

Fig. 2. Uncertainty in distance determination due to an uncertainty of 0.02% of the magnification measurement for different axial separations of the pupils.

the object will not change between the two images it is just necessary that an object or part of it can be segmented. Finding the corresponding object patch in the second image, therefore, is straight-forward (see Sec. 3).

3. Verification of theory

Based on the aforementioned theory, we have built an experimental setup to verify the working principle. The setup, as shown in Fig. 3, consists of an objective lens (SainSonic XR300 HDMC, 35 mm, f/1.7-16) and an image sensor (Ximea XiQ MQ013MG-E2, 1280x1024 pixel, pixel size 5.3 µ) on one side and various different objects in the object field.

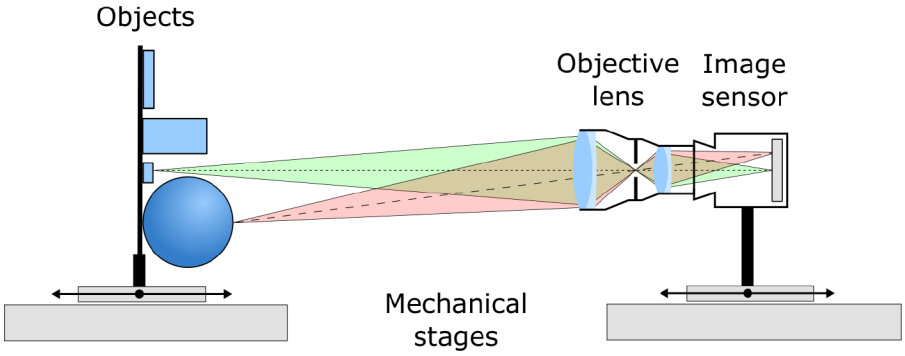

Fig. 3. Schematic of experimental setup with objects and camera (image sensor & objective lens) mounted on movable mechanical stages to vary distances in between both sides.

Some objects are mounted movable on mechanical stages (Walter-Uhl, linear stage LT) to vary their distances to each other. The stage of the camera is moved between two specified positions to simulate two camera positions at distance Δa, that can also be realized e.g. using a beamsplitter with two cameras (see Sec. 4). The movement of some of the objects is used to

OPTICS CONTINUUM

verify the measurement uncertainty of the depth estimation measurement at various distances a relative to the camera(s).

In the experiment, the camera position difference between the simulated cameras was 100 mm. Whereas the distance of the "first" camera (front) to the object varied from 1720–2600 mm, measured from the objective lens to the main object on axis (center of the image). The object's position was stepped through this range in 10 mm and 50 mm steps. The image sensor was connected via USB3.0 to a PC and controlled using the ITOM measurement software [15].

The current evaluation is based on template matching. One region (the "object") in the first image (see Fig. 4) is used as a template in order to search for this corresponding region in the second image. The normalized cross correlation is employed as a means for finding this region. The energy within the correlation spot (6 pixel diameter) is determined. This energy can be used to judge on the similarity of the two regions. If there is a difference in magnification, of course, the correlation spot energy will be reduced. Therefore, the evaluation is repeated for different scaled images (the second image is scaled by Lanczos interpolation). Interpolation and template matching have been done in C++ using OpenCV as the core image processing library [16]. The scaling, that leads to the maximum of the correlation energy, corresponds to the difference in magnification γ.

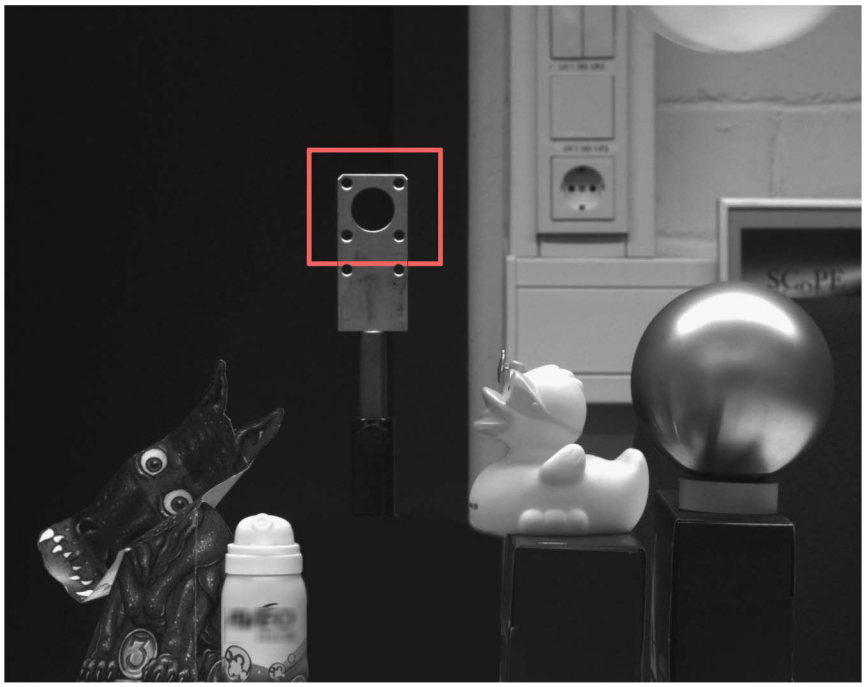

Fig. 4. Acquired scene and the marked area on the movable object for template matching algorithm.

To find this maximum, we determine the cross correlation for 50 different scaled versions of the image patch and fit a fifth-order polynomial through the resulting correlation spot energies. The maximum of the polynomial is found and gives the desired scaling and, therefore, the difference in magnification.

This results in the data plot shown in Fig. 5. We can take a second order polynomial as the calibration curve. The difference of the measured curve to this curve can be used as an estimate for the accuracy of the sensing principle. Over the whole measurement range we have a standard deviation to the calibration curve of 0.87 cm. Furthermore, the maximum deviation is 2.7 cm. At

OPTICS CONTINUUM

larger distances we see a clear increase of uncertainty as expected based on the theory presented in Sec. 2.

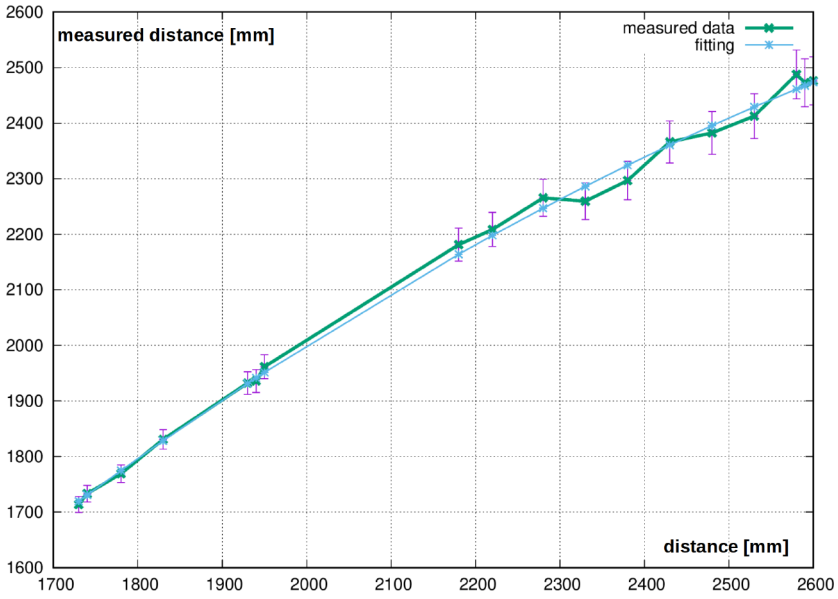

Fig. 5. Measured distance in relation to real distance and calibration function (2^{nd} order polynomial)

4. Advanced experimental setup

Furthermore, we have considered additional experiments since the testing and verification of the theory was successful. The next experiments were concluded with an advanced setup and slightly different arrangement. Also, we have considered taking on simultaneous imaging of two perspectives through placing two cameras with the same objective lenses. Through a beam-splitting device, as shown in Fig. 6, we are able to align the two cameras mechanically to the same optical axis with the second camera 90 degrees off to the side and with a distinct offset distance. The offset enables *differential perspective* with a axis length difference of Δa. This advanced design allows to acquire images simultaneously with both cameras triggered at the same time. The advantages come in handy when the sensor setup or the objects are moving relatively to prevent motion blur or images of different scenes in time, e.g., sequential.

In the further advanced setup, we used the same objective lenses as before (see Sec. 3) in combination with two cameras of the same type (Ximea XiC MC124MG-SY-UB, 12.4Megapixel, 4112x3008 pixels, Sony IMX253 1.1" CMOS, Pixel size 3.45 µ, Monochrome) and a large beamsplitter cube (non-polarizing, 50/50 Transmission, 40mm side length). All detecting optics and devices are mounted fixed to a breadboard to ensure optical axis alignment and rigidity. The distances or offset $\Delta a = 44,5\ mm$ in our setup is limited by the beamsplitter in combination with the camera's Field-of-View (FOV). One has to consider the image sensor size, objective lens focal-length, beamsplitter size diameter and distance between beamsplitter and camera.

Additionally, we selected a new scene with different more complex and specific objects of 3D measurements like metal balls, lamp, wooden hand, textured planes and a movable plane target (see Fig. 7). The latter is movable to place the target, similar to the verification experiment, at specific known positions in the z-depth axis. We also tried other textures of the moving target

OPTICS CONTINUUM

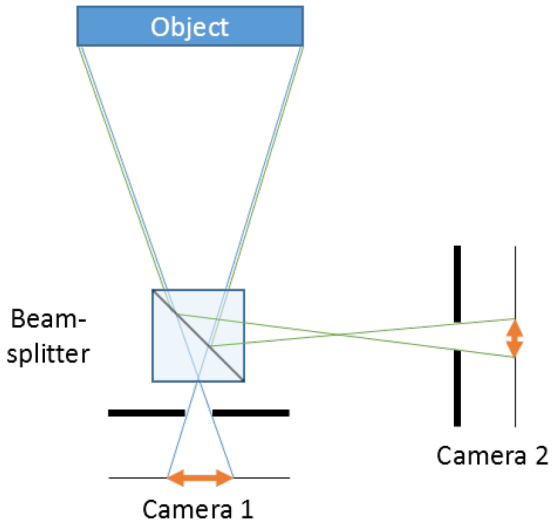

Fig. 6. Scheme of the principle with arrangement for simultaneous image acquisition using a beamsplitter in correspondence with two cameras.

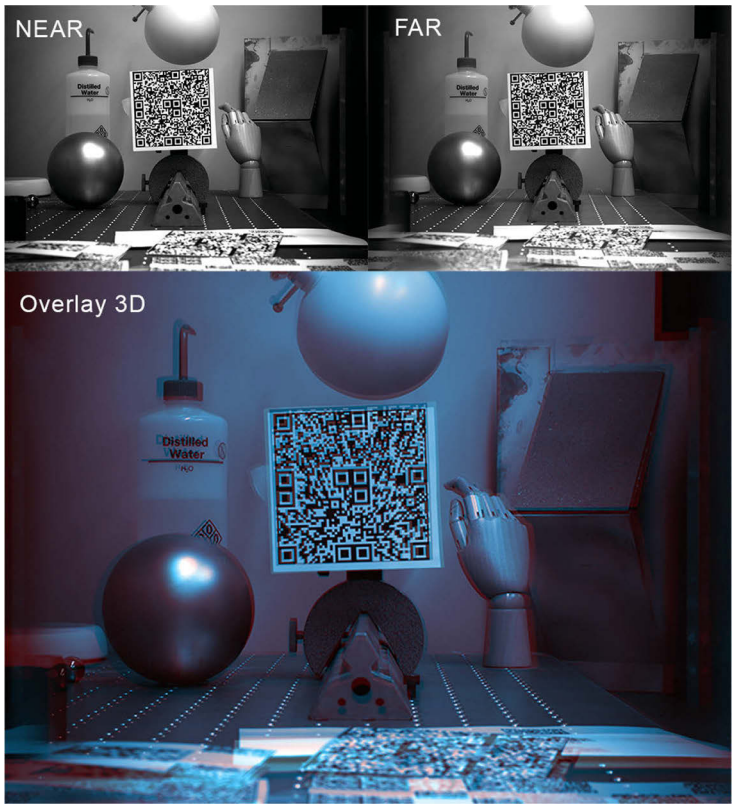

Fig. 7. Images of the scene with various textured objects using camera *NEAR* and *FAR* and a composition of both images superimposed to a 3D anaglyph. The movable object in the image center is textured with four random QR-Code patterns of size 50mm each.

OPTICS CONTINUUM

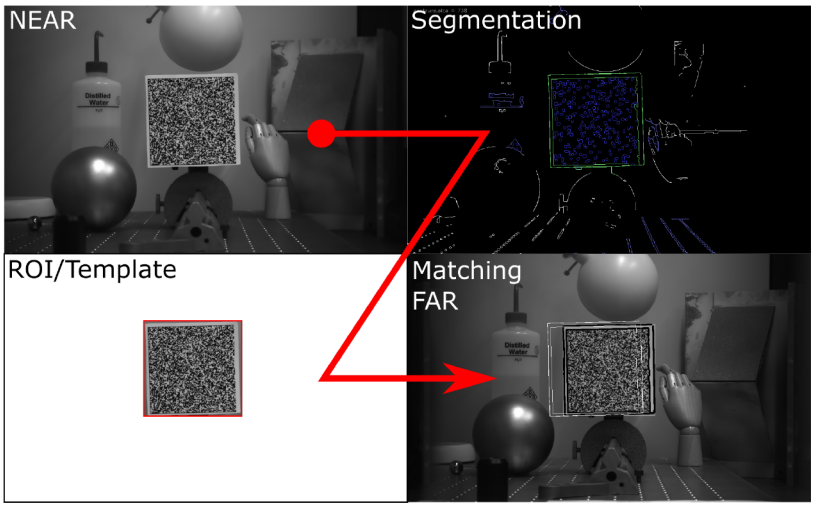

Fig. 8. Images showing the process of the algorithm to calculate object's depth based on segmentation and template-matching.

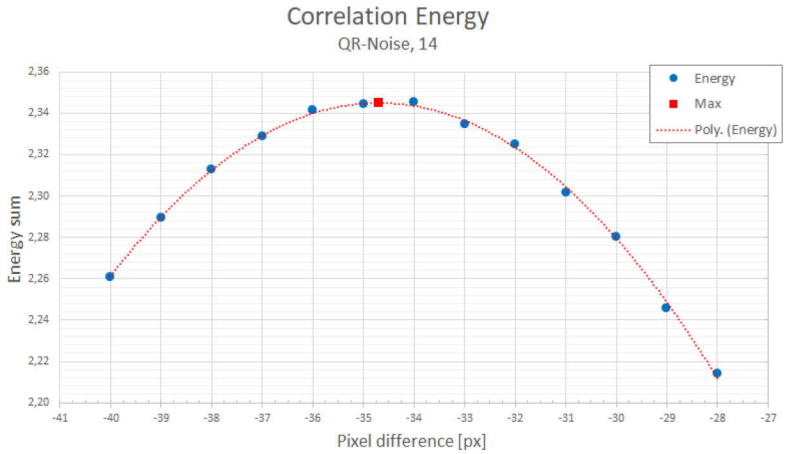

Fig. 9. Correlation energy graph of various template matchings and their pixel differences. Maximum of fitted polynomial yields sub-pixel accurate scaling factor for object depth calculation

through various QR-Code pattern images to vary structure and complexities and to verify their depth calculations by the algorithm.

The algorithm to calculate the depth of a distinct object is currently based on 1^{st} step: segmentation of the object in both images using *Canny-Edge* or *Adaptive Threshold*. Followed by 2^{nd} step: *template matching* using segmented object region-of-interest (ROI), as shown in Fig. 8. We then calculate the magnification ratio in step 3 using the result of the various scaling factors between the ROI (Far) and the tested image (Near). The result is the cross-correlation energy as a polynomial curve, see Fig. 9, where we calculate the maximum as the scaling or the magnification difference ratio γ. In a last step 4, we calculate the object distance utilizing this ratio and the known distances according to Eq. (4).

OPTICS CONTINUUM

5. Results

Beforehand, we have already shown and verified in Sec. 3, that theoretically the new principle works. In the previous section, we showed an advanced setup to further test our hypothesis with different settings and algorithm. The *template-matching* algorithm, as can be seen in Fig. 10, shows both with Canny-Edge and Adaptive-Threshold as segmentation functions good results in the sub-*cm* resolution. Also, there are some more outliers when using *Adaptive Threshold*. The image pairs shown in the graph correlate to the moved target positions with 1cm inter-spacing. With the different QR-Code patterns, results don't differ much from each other, but with less texturing and more white-space the precision slightly decreases.

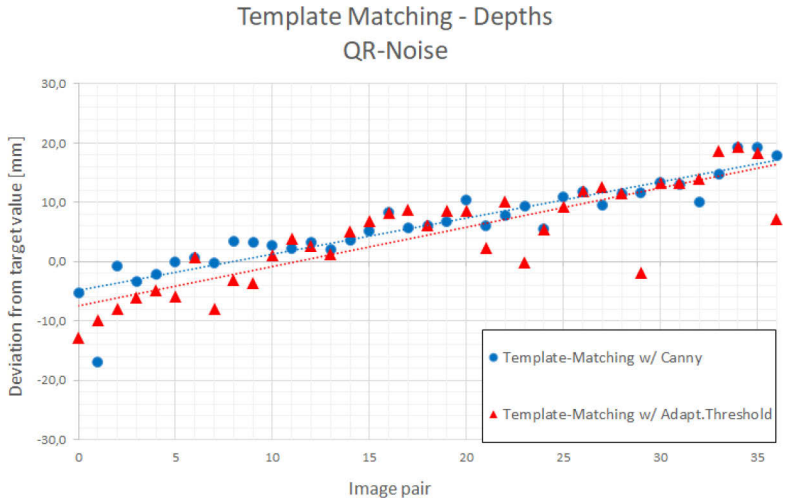

Fig. 10. Result of calculated depths per image pair (moved target) based on the template-matching energy algorithm with different segmentation functions (Canny/ Threshold). Only few outliers and an overall good resolution in the *mm* range and the distinct slope can be out-calibrated.

6. Conclusions

Differential perspective is a new passive distance sensing technique that is able to deliver accuracy in the range of centimeters for distances of several meters.

In the current form the technique is quite simple to implement. Compared to stereo vision shadowing and change of appearance is avoided since the sensor works on-axis. The distance to arbitrary individual objects can therefore be determined in a straight-forward manner, given that the object can be segmented in one of the images. Also, for some applications the reduced lateral extension of the senor might be beneficial. However, the main disadvantage compared to stereo vision is the reduced axial resolution (and accuracy).

The chosen segmentation algorithm itself depends on the objects to be segmented an the background. Here, we largely avoided such complications by using a template matching approach.

Since the main idea is simply to separate the pupil of the two imaging system axially instead of laterally, compared to triangulation based sensors, it is possible to transfer all variations of triangulation to differential perspective. Active illuminations with (single or multiple) fringe or other patterns might be useful and different technical realizations are possible. Furthermore, different algorithms for the processing of the images in order to obtain the difference in magnification are important candidates for future improvements of our method.

Funding. Baden-Württemberg Stiftung.

OPTICS CONTINUUM

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. A. O'Riordan, T. Newe, G. Dooly, and D. Toal, "Stereo vision sensing: Review of existing systems," in *2018 12th International Conference on Sensing Technology (ICST)*, (2018), pp. 178–184.
2. R. Klette, A. Koschan, and K. Schlüns, *Computer vision: Räumliche Informationen aus digitalen Bildern* (Springer, 2013).
3. J. Oliensis, "Exact two-image structure from motion," IEEE Trans. Pattern Anal. Machine Intell. **24**(12), 1618–1633 (2002).
4. J. Weng, T. S. Huang, and N. Ahuja, "Motion and structure from two perspective views: Algorithms, error analysis, and error estimation," IEEE Trans. Pattern Anal. Machine Intell. **11**(5), 451–476 (1989).
5. I. P. Howard and B. J. Rogers, *Seeing in depth, Vol. 2: Depth perception*. (University of Toronto, 2002).
6. F. Pedrotti, L. Pedrotti, W. Bausch, and H. Schmidt, *Optik für Ingenieure* (Springer, 2005).
7. A. Bhoi, "Monocular depth estimation: A survey," arXiv preprint arXiv:1901.09402 (2019).
8. R. Tucker and N. Snavely, "Single-view view synthesis with multiplane images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), pp. 551–560.
9. G. Wang, C. Zhang, H. Wang, J. Wang, Y. Wang, and X. Wang, "Unsupervised learning of depth, optical flow and pose with occlusion from 3d geometry," *IEEE Transactions on Intelligent Transportation Systems* (2020).
10. T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), pp. 1851–1858.
11. K. Konda and R. Memisevic, "Unsupervised learning of depth and motion," arXiv preprint arXiv:1312.3429 (2013).
12. B. J. Rogers and M. F. Bradshaw, "Vertical disparities, differential perspective and binocular stereopsis," Nature **361**(6409), 253–255 (1993).
13. S. Suto, T. Matsumoto, M. Ehara, and S. Ajiki, "Distance measuring camera," U.S. Patent 16636275 (3 December 2020).
14. J. Ma and S. Olsen, "Depth from zooming," J. Opt. Soc. Am. A **7**(10), 1883–1890 (1990).
15. M. Gronle, W. Lyda, M. Wilke, C. Kohler, and W. Osten, "itom: an open source metrology, automation, and data evaluation software," Appl. Opt. **53**(14), 2974–2982 (2014).
16. G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools* (2000).