Stereoscopic Depth Perception Through Foliage

Robert Kerschner[1], Rakesh John Amala Arokia Nathan[1], Rafał K. Mantiuk[2], and Oliver Bimber[1,*]

[1]Johannes Kepler University Linz, AT
[2]University of Cambridge, UK
[*]oliver.bimber@jku.at

ABSTRACT

Both humans and computational methods struggle to discriminate the depths of objects hidden beneath foliage. However, such discrimination becomes feasible when we combine computational optical synthetic aperture sensing with the human ability to fuse stereoscopic images. For object identification tasks, as required in search and rescue, wildlife observation, surveillance, and early wildfire detection, depth assists in differentiating true from false findings, such as people, animals, or vehicles vs. sun-heated patches at the ground level or in the tree crowns, or ground fires vs. tree trunks. We used video captured by a drone above dense woodland to test users' ability to discriminate depth. We found that this is impossible when viewing monoscopic video and relying on motion parallax. The same was true with stereoscopic video because of the occlusions caused by foliage. However, when synthetic aperture sensing was used to reduce occlusions and disparity-scaled stereoscopic video was presented, whereas computational (stereoscopic matching) methods were unsuccessful, human observers successfully discriminated depth. This shows the potential of systems which exploit the synergy between computational methods and human vision to perform tasks that neither can perform alone.

Introduction

Occlusions caused by vegetation can severely hinder aerial operations, such as search and rescue, wildfire detection, wildlife observation, security, or surveillance. For example, it is almost impossible to detect a standing person in the thermal drone recording shown in Fig. 1b (in the blue box). One of the most promising solutions to this problem is synthetic aperture sensing[1–17], in which multiple images taken at different positions are computationally combined to simulate an advanced (virtual) sensor of a wider (synthetic) aperture. An example result of such an integral image is shown in Fig. 1b, in which a standing person can be much more easily identified (in the blue box). Here we show the result for thermal imaging, but synthetic aperture sensing is equally applicable to radar[18–20], radio telescopes[21,22], interferometric microscopy[23], sonar[24,25], ultrasound[26,27], LiDAR[28,29], and optical imaging[30–32].

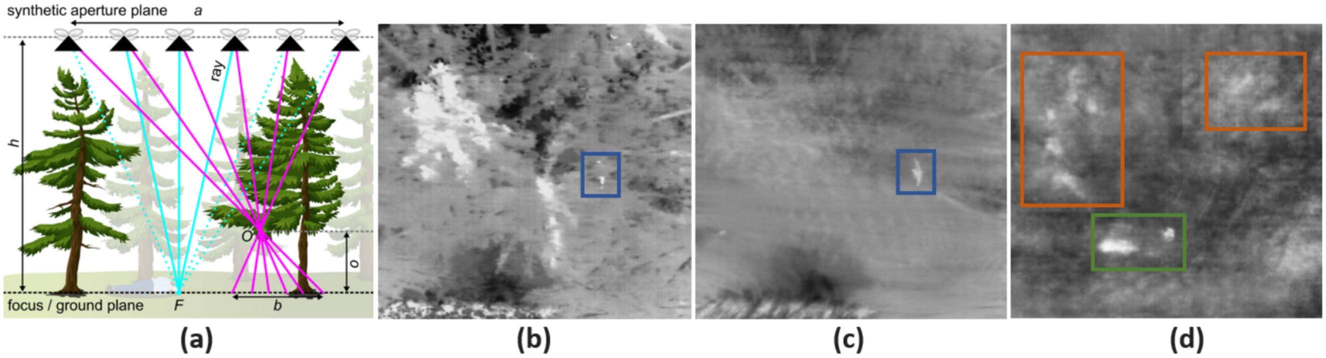


Figure 1. (a) Optical synthetic aperture sensing principle. Registering and integrating multiple images captured along a synthetic aperture of size a while computationally focusing on focal plane F at distance h will defocus occluders O at distance o from F (with a point-spread of b) while focusing targets on F. (b) Conventional thermal aerial image of woodland with an occluded person on the ground (blue box). (c) The same scene as (b) but with suppressed occlusion by integrating 30 thermal images captured along a synthetic aperture of a=14 m at h=26 m AGL. (d) An ambiguous example of an integral image in which true (lying and standing persons in the green box) and false (heated ground patches in red boxes) detections can be made. They cannot be differentiated by other discriminators, such as shape.

An optical synthetic aperture image is formed by superimposing regular images taken with small-aperture optics so that the depth of interest (e.g., the ground level) is brought into focus. This is illustrated in Fig. 1, in which a drone captures multiple images in a fly path over a woodland. The pixels from each camera image are projected onto a hypothetical (virtual) focal plane at distance h from the synthetic aperture's plane (i.e., at the altitude of the flight path) — see the cyan lines projecting on the ground plane in Fig. 1a. Even though the object of interest is occluded in some camera images (dashed lines in Fig. 1a), other views will reveal the object under the foliage. Aligning the focal plane with the forest floor and repeating this for all of its locations results in a shallow depth-of-field integral image of the ground surface (cf. Figs. 1c,d). It approximates the signal of a physically impossible optical lens of the size of the synthetic aperture. The optical signal of out-of-focus occluders, such as the tree crowns, is suppressed (blurred) — see pink lines projecting on the ground plane in Fig. 1a, while focused targets on or near the ground are emphasized. Computation of the integral images can be achieved in real-time and is wavelength-independent. Thus, the method can be applied in the visible range, near-infrared range, or far-infrared range (thermal) to address many different use cases. It has previously been explored in search and rescue with autonomous drones[8, 9], bird census in ornithology[5], and through-foliage tracking for surveillance and wildlife observation[12, 14].

The main limitation of optical synthetic aperture sensing is that its results can be ambiguous if true targets cannot be differentiated from false targets on the basis of clear features such as shape. An example of this is illustrated in Fig. 1d where strong thermal signatures of multiple potential targets near the forest floor are visible. While some of them are the results of sun-heating, only two are the true thermal signatures of people. With two-dimensional information alone, a clear distinction is impossible. However, the height differences between people and the forest floor could serve as an additional cue if it can be preserved in the final image. A computational 3D reconstruction from the sampled multi-view aerial images or the corresponding integral images is currently impossible with state-of-the-art methods in the case of strong occlusion[1], as shown and explained in the *Appendix*. Airborne laser scanning, such as LiDAR, has clear advantages over image-based 3D reconstruction when it comes to partially occluded surfaces, but it also has clear limitations[1]: First, it is not sensitive to the target's emitted or reflected wavelengths. Thus, far-infrared (thermal) signals, for instance, cannot be detected. Second, the point clouds cannot be scanned in high resolution and in real-time due to mechanical laser deflection and high processing requirements. This makes laser scanning unsuitable for many applications that require instant results and high resolutions.

In this article, we explore the synergy between optical synthetic aperture sensing and the human's ability to sense depth in stereoscopic images. We introduce binocular disparity to the optical synthetic aperture images, which then serves as additional cue and discriminator in identification and classification tasks. This enables tasks that cannot be completed with human or computer vision alone. To prove that binocular depth perception is possible for thermal optical synthetic aperture images, which are unnatural for human vision, we test whether human observers can infer depth from such images and complete high-level tasks.

Let us theoretically analyze under what conditions the visual system can fuse and discriminate depth differences between small and occluded targets, such as standing humans (up to 2 m) occluded by tall trees (15–20 m), seen from high altitudes (20–30 m for drones flying above tree level). First, objects that differ much in height (e.g., tree crowns vs. targets on the ground) and are located closely together in the image will result in large disparity gradients (disparity difference divided by the distance between two objects). If the disparity gradient exceeds the limit of human visual perception, diplopia[33] will result, making stereoscopic function impossible. Second, if objects are seen from relatively far distances, and their height difference is small, the disparity difference between them may fall below the stereo acuity limit[34–36], which makes depth discrimination impossible. The latter problem can be addressed by enlarging (scaling) disparities by assuming large viewing baselines (e.g., much larger than a typical inter-ocular distance of 6.5 cm).

Fig. 2 illustrates these two problems for the unoccluded case. Let us consider the some geometric constraints: For a given screen distance v, inter-ocular distance e, and object disparity d, the perceived object distance z is given by [37, ch. 9.2.2]

$$z = \frac{ev}{e-d}. \tag{1}$$

It follows that

$$d = \frac{e(z-v)}{z}. \tag{2}$$

Applying Eqn. 2 to compute the disparity on the focal plane at distance v_f (equals h in Fig. 1a), camera baseline e_f on the synthetic aperture plane, and target distance $z_f = v_f - h_t$ (h_t is the target height) from the synthetic aperture plane; and then scaling the resulting disparity to the display parameters to determine the perceived object distance on the display z_d using Eqn. 1 results in

$$z_d = \frac{e_d v_d}{e_d - \frac{v_d \tan(FOV_d/2) e_f (z_f - v_f)}{v_f \tan(FOV_f/2) z_f}}, \tag{3}$$

where e_d and v_d are the inter-ocular distance and the distance of the display image plane, and FOV_d and FOV_f are the fields of view of the display and camera, respectively.

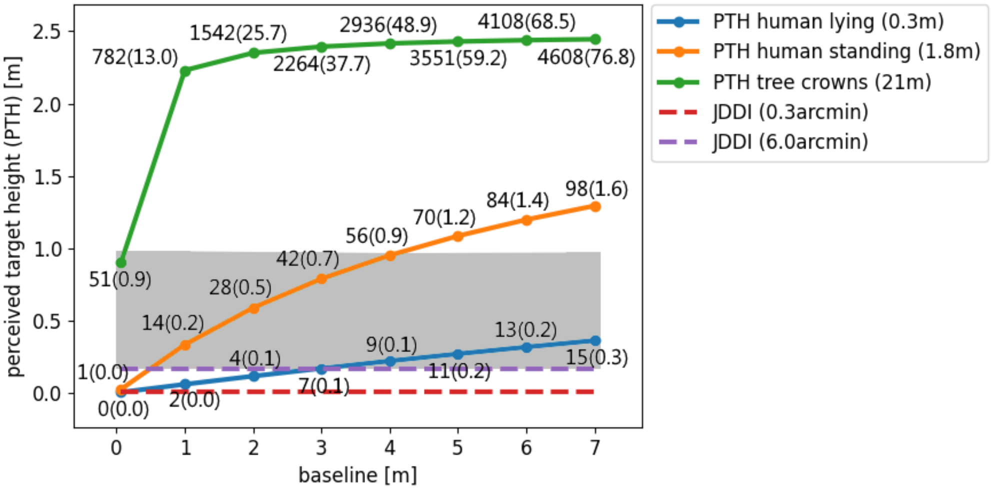


Figure 2. The increase in perceived target height (PTH) with an increasing stereo baseline for three unoccluded objects of different heights (solid plots): tree crowns, lying person, and standing person. Stereo acuity sets the just-detectable depth interval (JDDI) required for perceiving height differences (dashed lines). Both the conservative (0.3 acrmin) and the realistic (6 arcmin) JDDIs are plotted. Disparities, or rather disparity gradients (numbers next to the markers), limit the maximum length of the baseline above which objects cannot be fused due to diplopia. Consequently, the grayed region represents the range in which depth can be perceived (assuming, for example, a disparity gradient limit of 1.0 and a stereo acuity of 6.0 arcmin). Display disparities are given with respect to the ground level and for 60 arcmin object distances. For this plot we assume the capturing and display parameters provided in the *Methods* section.

Consequently, the perceived target height is

$$PTH = v_d - z_d. \tag{4}$$

The just-detectable depth interval is given by[38,39]

$$JDDI = \frac{d_\gamma v_d^2}{ce_d + v_d}, \tag{5}$$

where d_γ is the stereo acuity (in arcmin) and $c = 3437.75$ (1 radian in arcmin).

Now, considering Fig. 2 and the above geometric constraints, it can be seen that the perceived target height (PTH, y-axis) increases with an increased stereo baseline (x-axis). The solid lines show the increase in perceived target height for three different object types: tree crowns at 21 m (green), a lying person at 0.3 m (blue), and a standing person at 1.8 m above the surface (orange). The numbers above the markers indicate the corresponding display disparities and disparity gradients for a given stereoscopic display (assuming minimal object distances of 60 arcmin or 1 deg). The just-detectable depth interval (JDDI) threshold (dashed lines) varies between individuals. With poorer stereo acuity, larger depth intervals are required for perceiving height differences. Under these conditions and assuming an inter-ocular distance of 6.5 cm (the leftmost point in Fig. 2), the height differences between target objects on the ground are unlikely to be detected — even if excellent stereo acuity is assumed. Larger baselines improve the ability to discriminate depth, but they also increase the disparity gradients. If the disparity gradient is excessively large (e.g., 1[33] - 3[40], outside the gray box in Fig. 2), the stereo images cannot be fused.

The third problem is that view-dependent occlusion in the stereo pairs causes binocular rivalry. The rivalry appears when radically different images are presented to each eye, and when it is too strong, it prevents stereoscopic fusion[41,42]. Examples are illustrated in Fig. 3. It has been found that, if partially occluded object fragments are horizontally aligned and match a continuous surface, our visual system tends to extrapolate a coherent surface at an incorrect depth[43]. Horizontally aligned continuous object surfaces, however, are usually not present under realistic occlusion conditions such as ours. Although depth cannot be reconstructed computationally, we show that surface continuity can be reconstructed computationally, which enables human depth perception.

In this approach, we suppress occlusion by means of optical synthetic aperture sensing, as explained above and illustrated in Fig. 1. This also implies that we can compute stereoscopic integral images with suppressed occlusions for a given synthetic aperture of size *a* and for two different viewing positions within *a* and separated by a given baseline *d*. While for the monoscopic case, the center of the synthetic aperture is used as the reference perspective of the resulting integral image, the two baseline-shifted viewing positions are applied for the stereoscopic case. This results in two integral images that reveal a parallax for all

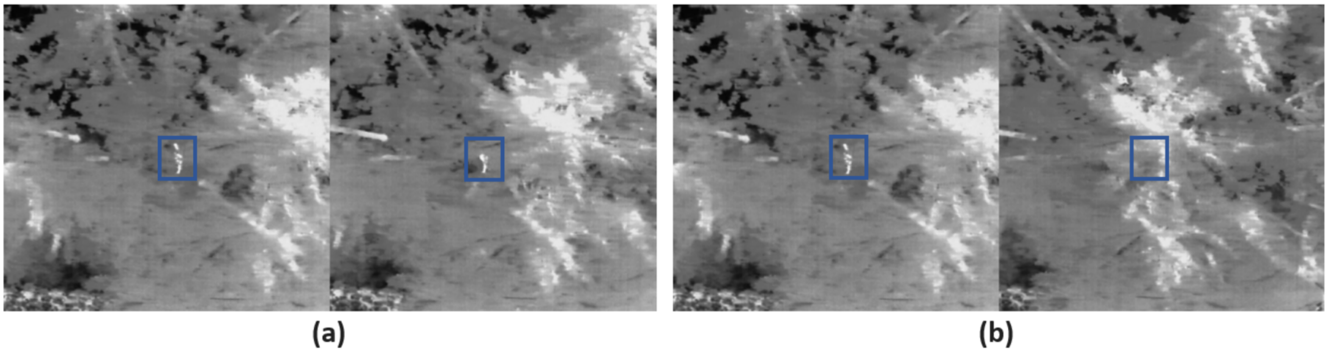


Figure 3. Stereoscopic thermal aerial recordings (left- and right-eye image pairs) of a sparsely occluded, person standing with arms outstretched to the sides (blue box) in woodland. View-dependent partial (a) or full (b) occlusion in stereo pairs cause binocular rivalry and prevent stereo fusion and consequently depth perception.

objects not located on the focal plane. With a large *d* (larger than inter-ocular distance), we upscale disparities so they do not fall below the limits of stereo acuity. The larger *a*, the more occlusion is suppressed, and binocular rivalry and extreme disparity gradients caused by tree crowns can consequently be reduced. However, a wide synthetic aperture also leads to a shallow depth of field and thus to defocus blur and lower contrast. The reduction in contrast and the loss of high spatial frequencies result in degradation of stereo acuity[44]. This is illustrated in Fig. 4.

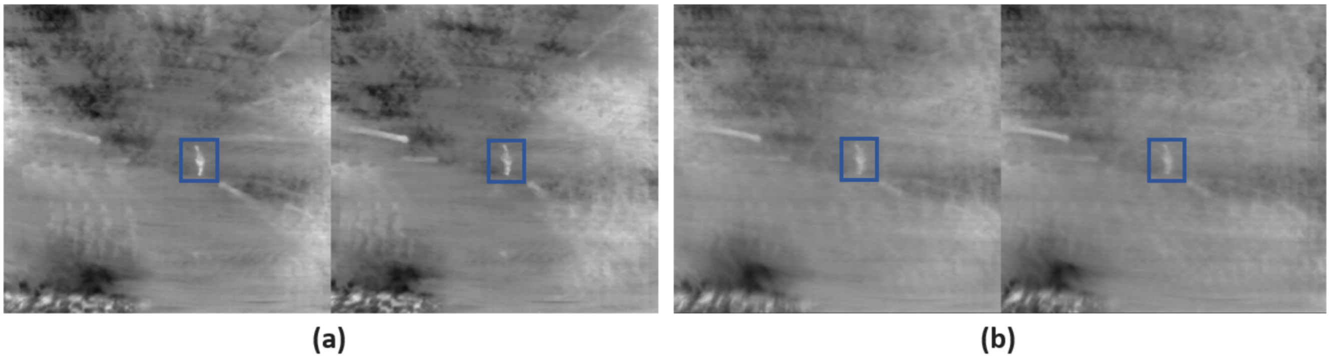


Figure 4. Integral stereo pairs of the scenario shown in Fig. 3, where the synthetic aperture *a* applied is smaller (a) or wider (b). The larger *a*, the more shallow the depth of field. This leads to a reduction in sharpness and contrast.

With the results presented below, we make three main findings: First, occlusion removal in stereoscopic images is of fundamental importance for object identification tasks. Stereoscopic perception alone leads to no significant improvement in the presence of occlusions. In fact, in all test cases with occlusions, observers' performance for stereoscopic images was comparable to that for monoscopic images, and it was not improved by the introduction of motion parallax. Second, while discriminating depth computationally (e.g., using 3D reconstruction from sampled multi-view images) is currently impossible with state-of-the-art methods in the case of strong occlusion, it becomes feasible visually by fusing binocular images with scaled disparities, which can be easily generated with optical synthetic aperture sensing.

Third, the sampling and visualization parameters (best baseline and synthetic aperture size), although restricted by the acuity limits and disparity gradients (refer to Fig. 2), were found to be fairly consistent across all test cases evaluated.

Our findings are discussed in *Summary and Conclusion*. They demonstrate that it is possible to discriminate the depths of objects seen through foliage on the basis of optical synthetic aperture imagery captured with first-person-view (FPV) controlled drones or a manned aircraft. It has the potential to support challenging search and detection tasks in which occlusion caused by vegetation is currently the limiting factor. This includes use cases such as search and rescue, wildfire detection, wildlife observation, security, and surveillance.

Results

We test observers' ability to (a) identify objects and (b) discriminate depth in thermal recordings captured by a drone at 26 m above ground level (AGL), with some of the objects hidden under foliage.

The experiments were conducted for four different scenes (cf. Fig. 5): an open field without vegetation (*scene 1*) with a standing (*object 1*) and a lying (*object 2*) person; a forest (*scene 2*) with one easily detected (based on shape features) person (*object 1*) standing with arms outstretched to the sides; a denser forest (*scene 3*) with a standing (*object 1*) and a lying (*object 2*) person; and a sparser forest (*scene 4*) with a standing person (*object 1*) and a 30 cm high (roughly the height of a lying person) artificial object (*object 2*) of similar shape, footprint, and temperature as the standing person.

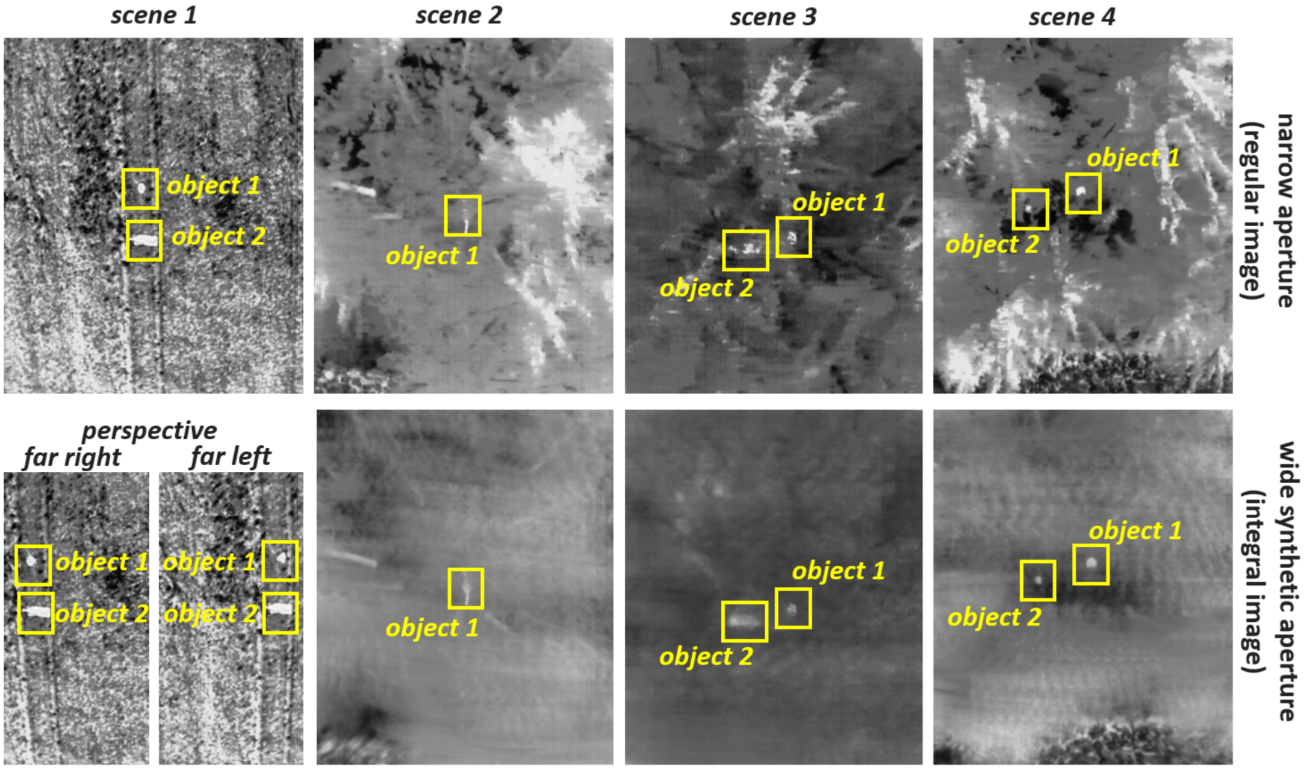


Figure 5. Four test scenes used in the experiments with different occluded (*scene 1*) and unoccluded (*scenes 2-4*) target objects of various heights. A wide synthetic aperture suppresses occlusion in the resulting integral images. In all conditions, users could change the horizontal perspective to find less occluded views and use motion parallax as an additional depth cue.

These recordings were then computationally combined to integral images (as explained in the *Introduction*, cf. Fig. 1a) and presented to 21 observers (11 female, 10 male, average age: 36) via a head-mounted stereoscopic display. At all times, the observers were able to use a game controller to interactively change the horizontal viewing perspective within the limits of the synthetic aperture a covered. This allowed them to find less occluded viewing directions and rely on motion parallax when discriminating depths. We changed visualization parameters, such as aperture size a for integration and (camera) baseline e_f for stereo-pair computation while asking the observers to describe the quality of the perceived images. The synthetic focal plane was set to the depth of the forest floor. Details on how the field and user experiments were carried out are provided in the *Methods* section. Detailed results, the raw and the cleaned data, and the software systems used for our filed and user experiments are available in the supplementary material (see *Code and Data Availability*).

First Experiment: Object Identification

In the first experiment (*object identification*, cf. Fig. 6), we asked the observers to identify (i.e., to detect — not to classify) all objects that appeared to be just above the forest floor in our *scenes 2-4* and to report on how confident they were on a scale from 0 (not confident at all) to 10 (very confident) in their decisions. We asked to detect objects rather than people to avoid any bias in this task. Since the detection of unoccluded objects is trivial, *scene 1* was skipped in this experiment. This task was repeated for regular monoscopic images without disparity (*mono*), for regular stereoscopic images at different baselines e_f (*stereo*), and for stereoscopic integral images at different synthetic apertures a and different baselines e_f (*SA stereo*).

Fig. 6 illustrates how often and with what level of confidence our target objects (*objects 1* and *2*) were detected among all identifications. See the supplementary material for details on which objects were identified. Note that confidence values (the bottom row) were counted as negative if our target objects were not identified. Thus, the negative values indicate misplaced confidence in identifications.

Stereoscopic depth cues (yellow bars in Fig. 6) improved the performance for scene 2, and boosted the confidence across all the scenes. However, when detecting a shorter and more occluded object 2 in scenes 3 and 4, stereoscopic depth cues alone did not improve the performance over monoscopic viewing (blue bars). Only when stereoscopic viewing was combined with synthetic aperture (green bars in Fig. 6, *SA stereo*), did the performance and confidence improve for that object.

This indicates that occlusion removal leads to a measureably better detection of objects and, consequently, to an increase in correct identifications. This method of presentation is still hindered by the lack of distinctive features, occlusions and resulting binocular rivalry. However, it offers marked improvement over monoscopic and stereoscopic presentations.

Since our object identification experiment was a detection and not a classification task, true and false positive rates cannot be determined.

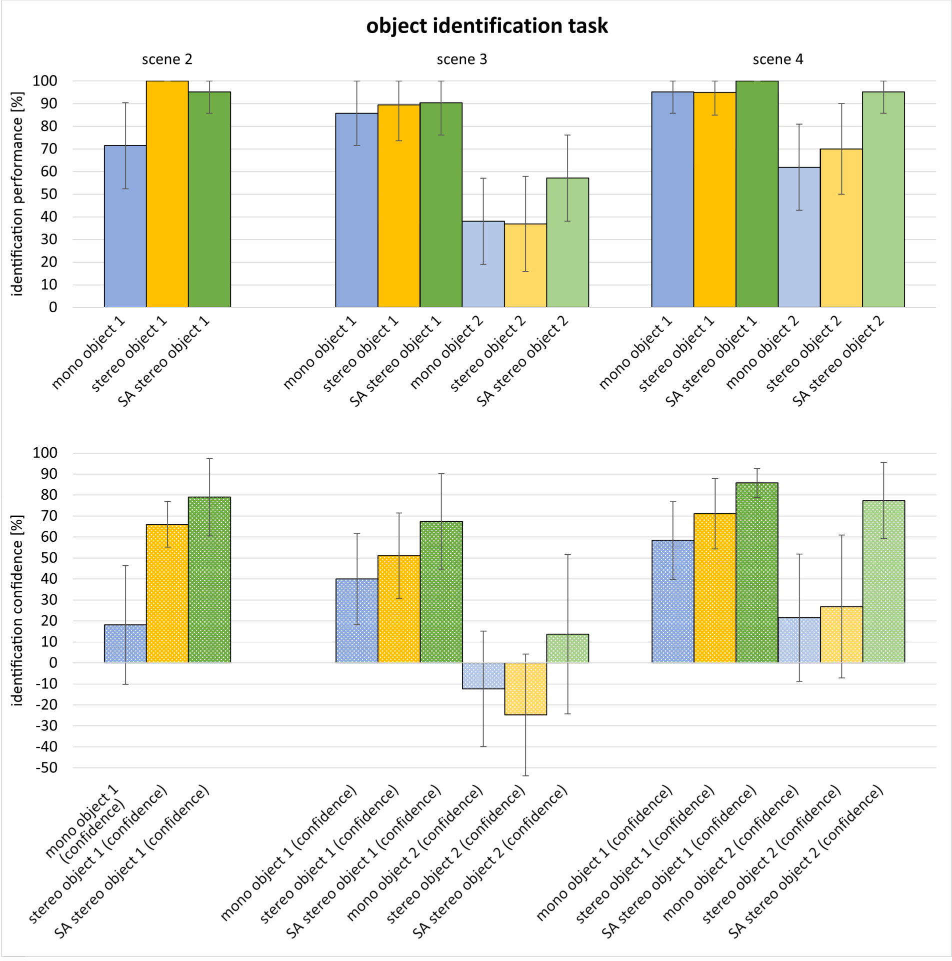
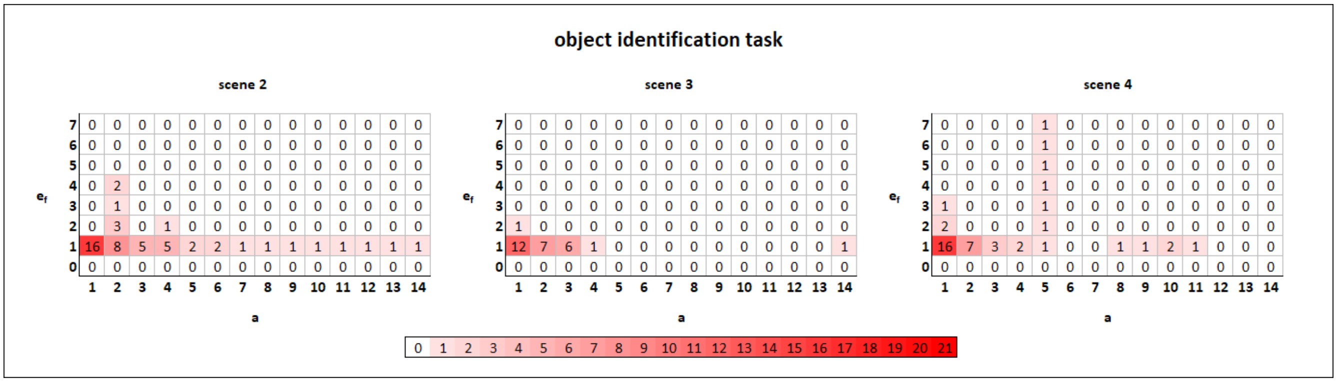


Figure 6. Object identification performance (solid bars, top row) and confidence values (dotted bars, bottom row) for test scenes with occlusion (*scenes 2-4*). We measured how often and with what confidence our target objects (*object 1* and *object 2*) were detected among all identifications and across all observers. The error bars denote 95% confidence intervals (based on the binomial distribution for performance and normal for observers' confidence).

Figure 7 illustrates the ranges of baselines e_f and synthetic apertures a for which our observers performed best. Note that the best parameters varied between the observers and the scenes. Therefore, the color-coded numbers indicate the count of overlapping ranges across all observers (i.e., how often a parameter pair was considered best). Consistently across all three scenes, e_f=1–2 m and a=1–4 m were found to be optimal. Note also that practical GPS positioning was too imprecise for sampling these parameters below 1 m.

Second Experiment: Depth Discrimination

In the second experiment (*depth discrimination*, cf. Fig. 8), we asked observers to indicate the highest object (which was always the standing person, *object 1*, in our experiments) of those identified and again report how confident they were in their



object identification task

scene 2

e_f														
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	2	0	0	0	0	0	0	0	0	0	0	0	0
3	0	1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	3	0	1	0	0	0	0	0	0	0	0	0	0
1	16	8	5	5	2	2	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a	1	2	3	4	5	6	7	8	9	10	11	12	13	14

scene 3

e_f														
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	12	7	6	1	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a	1	2	3	4	5	6	7	8	9	10	11	12	13	14

scene 4

e_f														
7	0	0	0	0	1	0	0	0	0	0	0	0	0	0
6	0	0	0	0	1	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0	0	0	0
4	0	0	0	0	1	0	0	0	0	0	0	0	0	0
3	1	0	0	0	1	0	0	0	0	0	0	0	0	0
2	2	0	0	0	1	0	0	0	0	0	0	0	0	0
1	16	7	3	2	1	0	0	1	1	2	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Figure 7. Ranges of baselines e_f and synthetic apertures a for which our observers performed best in the object identification task. Note that observers performed best over a range of parameters rather than for exactly one pair. The color-coded numbers indicate the count of overlapping ranges across all observers (i.e., how often a parameter pair was considered best). The units of e_f and a are in meters.

decisions. Here, we considered all four test scenes and varied a for the scenes with occlusion (*scenes 2–4*) and e_f for all scenes. The multi-view depth reconstruction results in the *Appendix* reveal that computational depth discrimination is infeasible for our occluded scenes.

Our observers were unable to discriminate depth in monoscopic images, even for the simplest case that is, without occlusion (*scene 1*). Not even motion parallax was sufficient to enable them to determine depth differences. We assume that the monoscopic and parallax depth cues were simply too subtle for aerial viewing conditions (see also the most extreme horizontal perspectives of *scene 1* in Fig. 5). When stereoscopic viewing was enabled (orange bars in Fig. 8), all observers could discriminate depth with good accuracy in the scene with no occlusion (*scene 1*). In the scenes with occlusions (*scenes 2–4*), we observed a consistent improvement in performance when stereoscopic viewing was supported. The persistently low confidence scores (dotted orange bars in Fig. 8), however, underline a remaining strong uncertainty. A significant improvement in both performance and confidence was observed when stereoscopic viewing was used together with synthetic aperture sensing (green bars in Fig. 8).

Fig. 9 illustrates the ranges of baselines e_f and synthetic apertures a for which our observers performed best. As for the results presented in Fig. 7, observers performed best over a range of parameters rather than for exactly one pair. The color-coded numbers indicate the count of overlapping ranges across all observers (i.e., how often a parameter pair was considered best). Consistently across all occluded scenes and in line with the theory explained in Fig. 7, we found e_f=1-2 m and a=1-4 m to be optimal. For the case without occlusion (*scene 1*), however, the average of best baselines was significantly higher than for the cases with occlusion. While larger baselines help in discriminating depths, they also make the left- and right-eye views more different from each other. Therefore, in the cases with occlusions in which the views are likely to vary, the observers did better with smaller baselines.

The occlusion differences between the left and right views are unavoidable even with the synthetic aperture filtering.

Remaining occlusion that appears stronger in one view than in the other leads to binocular rivalry that negatively affects depth perception (cf. Fig. 10). This is not the case for *scene 1* without occlusion and the reason why the observers were able to increase the baseline up to the disparity gradient limit (cf. Fig. 2).

Methods

For our field experiments, we used a real-time kinematics (RTK) enabled DJI Mavic 3T drone with a 640×512@30 Hz thermal camera (61 deg FOV, f/1.0, 5 m-infinity focus). We developed a real-time imaging application using DJI's Mobile SDK 5 that runs on the DJI RC Pro remote controller (Android 10). It supports real-time optical synthetic aperture scanning, occlusion removal, and interactive monoscopic and stereoscopic visualization. Visual results during flight can be presented either live on the remote controller's display (monoscopic) or on a head-mounted display attached to the remote controller's HDMI port (monoscopic or stereoscopic). See *Code and Data Availability* section for how to obtain the imaging application. The data of our four test scenes was recorded at a constant altitude (26 m AGL) by scanning a 1D synthetic aperture over a linear flight path of 14 m and by choosing a sampling distance (0.5 m) that was well above the GPS error. Consequently, a total of 29 thermal images were captured per scan. For a flying speed of 15 m/s and an imaging speed of 30 Hz, this takes approximately 1 s, and the results are instantly displayed. As explained in the *Introduction*, the scanned images are computationally combined to generate stereoscopic integral images, depending on the parameters chosen: synthetic aperture size a (where a=14 m is the

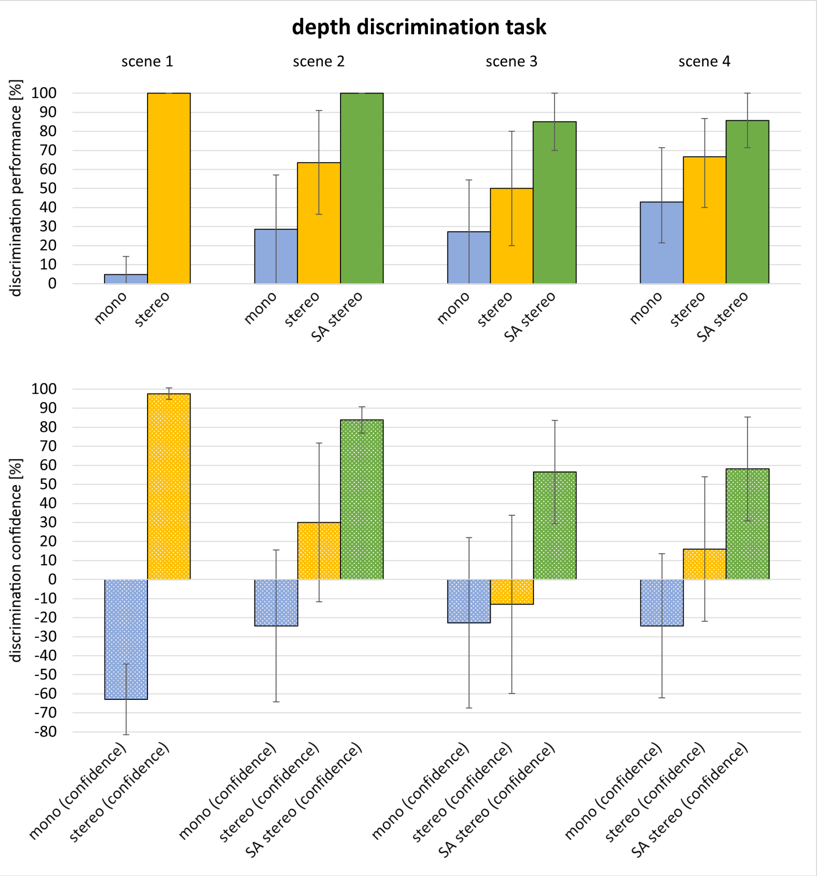
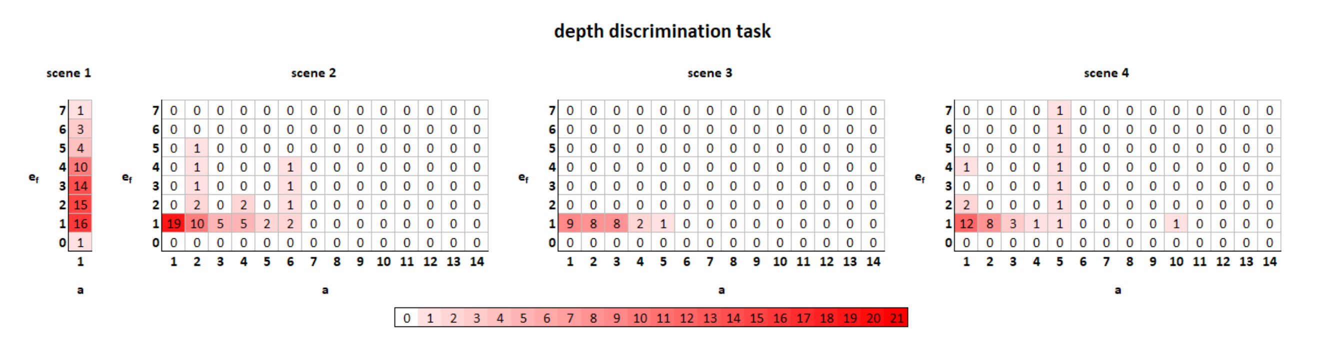


Figure 8. Depth discrimination performance and confidence for all test scenes. We measured how often and with what level of confidence the true highest object (*object 1*) was the highest object identified. The error bars denote 95% confidence intervals (based on the binomial distribution for performance and normal for observers' confidence).

maximum) and baseline e_f (where e_f=14 m-a is the maximum). The synthetic focal plane was always kept on the ground (i.e., h=26 m, cf. Fig. 1a), since focus shifts of the order of the target heights (e.g., 0–2 m AGL) did not significantly change the image content when observed from a large distance of 26 m. Our test sites for *scenes 2–4* were forested to various degrees with conifers and/or a variety of other tree species. The open field site for *scene 1* was a freshly harvested corn field.

All image data captured during the field experiments was recorded by the imaging application on the drone's remote controller and was used later in our offline user experiments. For the experiments, we developed a visualization application that runs on desktop PCs or laptops (Microsoft Windows 11) and that reproduces the same visual experience as the imaging application during flight. It presents the image data to our observers via a head-mounted stereoscopic display without requiring them to be in the field during the actual scans. See *Code and Data Availability* section for how to obtain the visualization application and the data used for our survey. The head-mounted stereoscopic display used was a 1920×1024@60Hz Enmesi E 812 (68 deg diagonal FOV, 2485.2 mm focal distance, 152× magnification, 10 mm eyebox, internally using two 2.1" 1600 x 1600 IPS microdisplays at up to 1058 PPI). Per-observer diopter settings were adjusted on the display before each session. A PowerA Nintendo Switch USB wired game controller was used to change the viewing perspective interactively.

We tested a total of 21 observers (11 female, 10 male, average age: 36, the youngest 14, the oldest 67, recruited through the authors from all social classes – ranging from students over workers through pensioners in Austria). Note that stereoscopic depth perception is considered to be fully developed at the age of 12[45]. As explained in the *Results* section, we first performed the object identification task, then the depth discrimination task. On average, a survey round took a total of 45 minutes per participant. The observers were not informed about study goals. We have intentionally chosen the following order of experiments: *mono*, *stereo*, *SA stereo* – since our hypothesis was that object identification and depth discrimination becomes easier with the introduction of stereoscopic depth queues and occlusion removal. Any other order would have incorrectly biased our results since subjects would have gained knowledge about objects and depths detected under simpler conditions before approaching them under more difficult conditions (e.g., *SA stereo*) before *stereo*, *stereo* before *mono*, or *SA stereo* before *mono*)



depth discrimination task

scene 1

e_f \ a	1
7	1
6	3
5	4
4	10
3	14
2	15
1	16
0	1

scene 2

e_f \ a	1	2	3	4	5	6	7	8	9	10	11	12	13	14
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	1	0	0	0	0	0	0	0	0	0	0	0	0
4	0	1	0	0	0	1	0	0	0	0	0	0	0	0
3	0	1	0	0	0	1	0	0	0	0	0	0	0	0
2	0	2	0	2	0	1	0	0	0	0	0	0	0	0
1	19	10	5	5	2	2	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

scene 3

e_f \ a	1	2	3	4	5	6	7	8	9	10	11	12	13	14
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	9	8	8	2	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

scene 4

e_f \ a	1	2	3	4	5	6	7	8	9	10	11	12	13	14
7	0	0	0	0	1	0	0	0	0	0	0	0	0	0
6	0	0	0	0	1	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0	0	0	0
4	1	0	0	0	1	0	0	0	0	0	0	0	0	0
3	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2	2	0	0	0	1	0	0	0	0	0	0	0	0	0
1	12	8	3	1	1	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 9. Ranges of baselines e_f and synthetic apertures a for which our observers performed best in the depth discrimination task. Note that observers performed best within a range of parameters rather than in exactly one pair. The color-coded numbers indicate the count of overlapping ranges of all observers (i.e., how often a parameter pair was considered best). Note also that for the unoccluded scene (*scene 1*) only e_f is considered as a synthetic aperture for occlusion removal is not required. All units are in meters.

for the same scene. In between the scenes, we displayed a neutral stereo pair to set stereoscopic fusion back to the same initial condition. While feedback from the participants was recorded in a questionnaire, all adjusted parameters were automatically recorded and stored by our application. To explore optimal visualization parameters, we incrementally increased a and e_f (starting with a=0 and e_f=0) while repeating each experimental trial of each task (object identification or depth discrimination) for each scene until the depth perception reported deteriorated. If both parameters were to be changed, we always started with a, followed by e_f. Participants were, at all times, able to interactively change their viewing positions using a game controller, and the time for stereo fusion was always allowed.

Summary and Conclusion

Identification and classification of objects that are strongly occluded by vegetation is aided significantly by the ability to discriminate their depths, which provides important additional information to tell true from false findings, for instance, people, animals and vehicles from sun-heated patches of open ground or the tree crowns, or ground fires from tree trunks. This cannot be accomplished with conventional monoscopic or multi-view aerial images — neither computationally nor by visual inspection of images or video.

While neither human nor computer vision can perform this task on its own, we show that the synergy of both makes it possible. We have demonstrated this based on three main findings: First, occlusion removal in stereoscopic images is necessary to *identify* objects occluded by foliage. This is because binocular images cannot be fused when one of them contains occlusions. Second, the combination of stereoscopic presentation and synthetic aperture imaging gives the highest accuracy of depth judgments when *discriminating depths*. Our observers found the task (of selecting the tallest object) impossible to complete in traditional monoscopic presentation, even in the presence of motion parallax. Stereoscopic presentation made the task straightforward for non-occluded object, but very difficult in the presence of occlusions.

Our third finding is that the relevant sampling and visualization parameters (best camera baseline and synthetic aperture size) are consistent throughout all test cases evaluated and in line with the theory explained in Fig. 2. The ability to discriminate depth is limited by sterescopic acuity, disparity gradient and rivalry. However, the flexibility of synthetic aperture imaging let us choose the reconstruction parameters that navigate those limitations (as discussed in the *Introduction*).

While detecting depth differences computationally (e.g., using advanced multi-view 3D reconstruction) is currently impossible with state-of-the-art methods in the case of strong occlusion[1] (see *Appendix*), we have shown that the human visual system can perform this task robustly. One reason for this might be our visual system's ability to integrate partially occluded surfaces that appear sufficiently continuous in a horizontal viewing direction[43], even in the presence of binocular rivalry. While this continuity is not given in conventional stereo pairs with dense occluders, it is enhanced in stereoscopic integral images with wider synthetic apertures that suppress occlusion. We believe that this is the main reason why depth discrimination improved for stereo integral images, even though binocular rivalry was not fully eliminated in cases of locally varying occlusion densities and too wide baselines.

Wider synthetic apertures reduce occlusions, but they also lower image contrast by blending multiple reprojected images. Therefore, for very large apertures that result in lower contrast, stereo acuity is reduced, making it more difficult to detect small disparities. This correspond to the grayed region shown in Fig. 2 gradually shrinking, with the bottom part shifting towards the

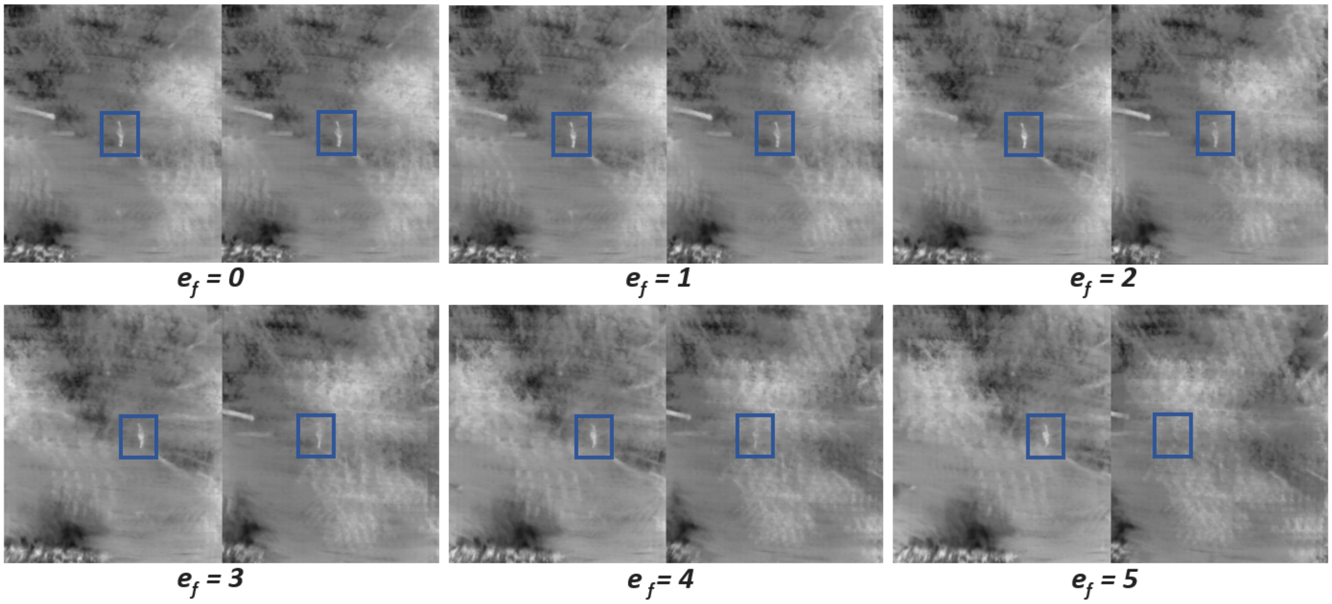


Figure 10. Increasing binocular rivalry in integral stereo pairs with wider baseline due to locally varying occlusion density (a=2, in this example). All units are in meters.

top. This gray region let us deduce the range of effective baselines: a small baseline results in smaller disparities, especially for the objects that are just above the ground level. If the disparity falls below the just-detectable depth interval (JDDI in Fig. 2), no depth difference is seen. But, if the baseline is too large, the disparity gradient may exceed the limit for binocular fusion (upper edge of the gray box in Fig. 2), making the depth judgment also impossible. When all the factors are considered, we can find the range of visualization parameters that provide the most reliable depth cues.

In our experiments, the observers could discriminate disparity of 12 arcmin (resulting from a height difference between a standing and a lying person from 26 m viewing distance and at 1 m baseline, see Fig. 2) seen through foliage. The relationship between depth discrimination precision, occlusion density, imaging, and display parameters has yet to be investigated. Furthermore, the results in Figs. 7 and 9 indicate that it is worthwhile to explore apertures and baselines below 1 m, which could not be investigated in this study because of imprecise GPS positioning.

Deep-learning-based image restoration methods[46] can potentially compensate for the contrast and sharpness loss in the integral images. It may also be possible to retain stereo acuity with more advanced sampling: If a drone were equipped with two cameras at the optimum baseline distance ($e_f \approx 1$ m) and the video was captured in the direction orthogonal to the baseline, the defocus due to the synthetic aperture would affect contrast only in the vertical direction, which is less relevant to binocular fusion. Such avenues should form part of future work.

Our findings demonstrate that human operators can detect depth differences between objects seen through foliage with first-person-view (FPV) drones or manned aircraft, where the thermal images captured are processed in real time with synthetic aperture sensing methods. This has the potential to support challenging search operations where occlusion caused by vegetation is currently the main limiting factor, as is the case for search and rescue, wildfire detection, wildlife observation, security, and surveillance. Although our experiments focused on detecting people, we believe that our findings are equally applicable to other objects, such as occluded vehicles and buildings.

Appendix

Fig. 11 illustrates 3D reconstruction results of our four test scenes, computed with the state-of-the-art structure-from-motion and multi-view stereo pipeline, COLMAP (version 3.9.1)[47,48].

While the occlusion-free *scene 1* can be fully reconstructed from the recorded aerial images, and the height difference between the two targets (standing and lying person) can easily be determined computationally, depth estimation fails for all other scenes. Here, at best, only tree crowns can be partially reconstructed. Using integral images from multiple perspectives on the synthetic aperture does not reconstruct the occluding tree structures, but an extremely noisy ground on which the targets cannot be detected at all.

The reason for this behavior is that without occlusion removal only the unoccluded tree crowns provide consistent and strong image features over a sufficient number of perspectives. The appearance of image features of occluded objects below is too inconsistent to be matched properly. With optical synthetic aperture sensing, in contrast, occlusion caused by the tree crowns is suppressed and is therefore not reconstructed. This comes at the cost of image features of the remaining ground surface loosing contrast and high frequencies (sharpness) in general. They become insufficient for computational stereo-matching - but are obviously sufficient for perceptual stereo fusion.

Note that for each scene all images captured were used and that 3D reconstruction took approx. 15 min on a modern desktop computer.

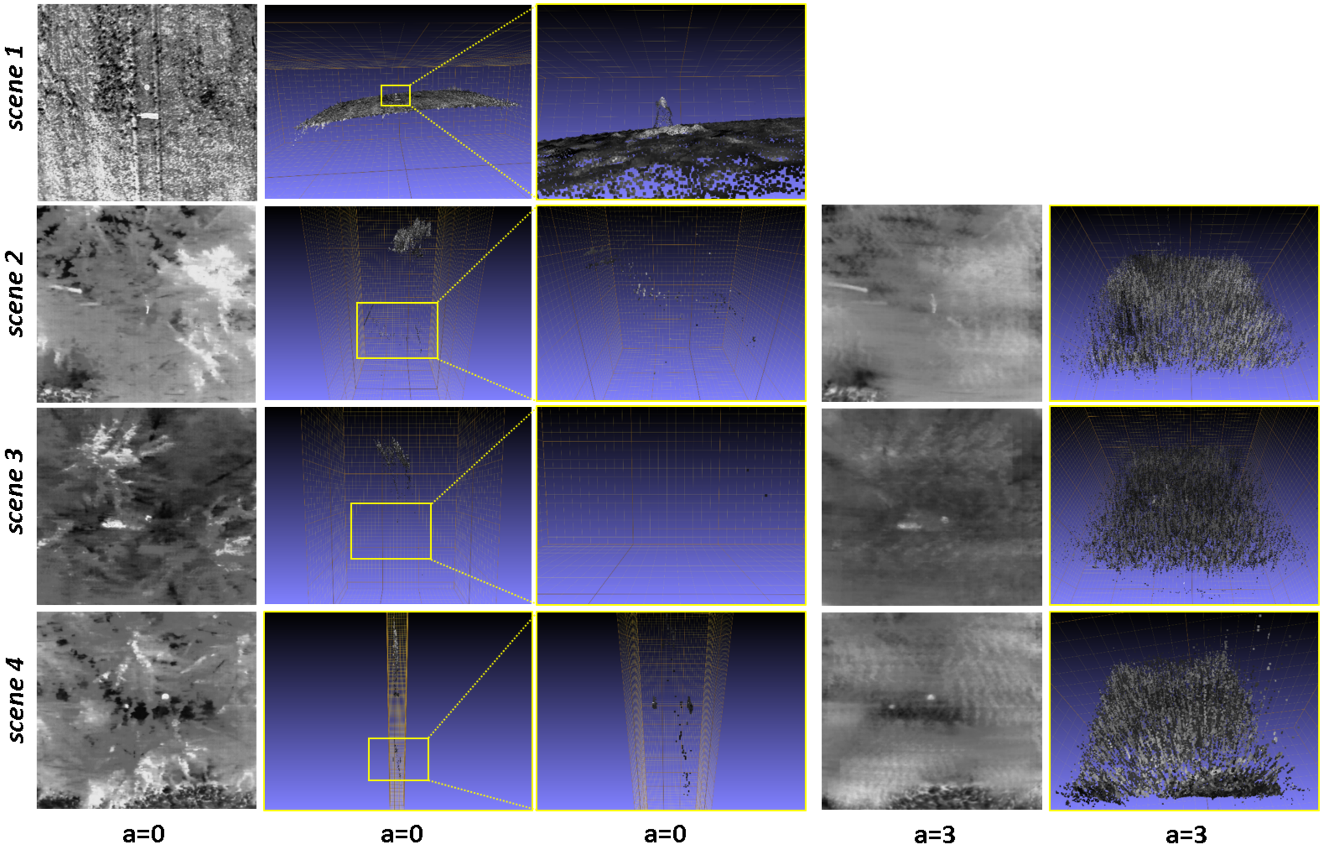


Figure 11. Multi-view 3D reconstruction results of our four test scenes. With regular aerial images ($a = 0$) only the upper depth layer can be reconstructed. For the unoccluded *scene 1*, it contains the two targets on the ground. For all occluded scenes, it contains –at best– the tree crowns. Yellow boxes show close-ups of of the reconstructed ground surface. If occlusion-suppressed integral images (with a=3 in this example) are used instead of regular aerial images, targets remain undetectable in noisy reconstructions of the ground surface. All units are in meters.

References

1. Kurmi, I., Schedl, D. C. & Bimber, O. Airborne optical sectioning. *J. Imaging* **4**, 102 (2018).
2. Bimber, O., Kurmi, I. & Schedl, D. C. Synthetic aperture imaging with drones. *IEEE computer graphics applications* **39**, 8–15 (2019).
3. Kurmi, I., Schedl, D. C. & Bimber, O. A statistical view on synthetic aperture imaging for occlusion removal. *IEEE Sensors J.* **19**, 9374–9383 (2019).
4. Kurmi, I., Schedl, D. C. & Bimber, O. Thermal airborne optical sectioning. *Remote. Sens.* **11**, 1668 (2019).
5. Schedl, D. C., Kurmi, I. & Bimber, O. Airborne optical sectioning for nesting observation. *Sci. reports* **10**, 7254 (2020).
6. Kurmi, I., Schedl, D. C. & Bimber, O. Fast automatic visibility optimization for thermal synthetic aperture visualization. *IEEE Geosci. Remote. Sens. Lett.* **18**, 836–840 (2020).
7. Kurmi, I., Schedl, D. C. & Bimber, O. Pose error reduction for focus enhancement in thermal synthetic aperture visualization. *IEEE Geosci. Remote. Sens. Lett.* **19**, 1–5 (2021).
8. Schedl, D. C., Kurmi, I. & Bimber, O. Search and rescue with airborne optical sectioning. *Nat. Mach. Intell.* **2**, 783–790 (2020).
9. Schedl, D. C., Kurmi, I. & Bimber, O. An autonomous drone for search and rescue in forests using airborne optical sectioning. *Sci. Robotics* **6**, eabg1188 (2021).
10. Ortner, R., Kurmi, I. & Bimber, O. Acceleration-aware path planning with waypoints. *Drones* **5**, 143 (2021).
11. Kurmi, I., Schedl, D. C. & Bimber, O. Combined person classification with airborne optical sectioning. *Sci. reports* **12**, 3804 (2022).
12. Amala Arokia Nathan, R. J., Kurmi, I., Schedl, D. C. & Bimber, O. Through-foliage tracking with airborne optical sectioning. *J. Remote. Sens.* **2022** (2022).
13. Seits, F., Kurmi, I., Nathan, R. J. A. A., Ortner, R. & Bimber, O. On the role of field of view for occlusion removal with airborne optical sectioning. *arXiv preprint arXiv:2204.13371* (2022).
14. Amala Arokia Nathan, R. J., Kurmi, I. & Bimber, O. Inverse airborne optical sectioning. *Drones* **6**, 231 (2022).
15. Seits, F., Kurmi, I. & Bimber, O. Evaluation of color anomaly detection in multispectral images for synthetic aperture sensing. *Eng* **3**, 541–553 (2022).
16. Amala Arokia Nathan, R. J., Kurmi, I. & Bimber, O. Drone swarm strategy for the detection and tracking of occluded targets in complex environments. *Commun. Eng.* **2**, 55 (2023).
17. Amala Arokia Nathan, R. J. & Bimber, O. Synthetic aperture anomaly imaging for through-foliage target detection. *Remote. Sens.* **15**, DOI: 10.3390/rs15184369 (2023).
18. Moreira, A. *et al.* A tutorial on synthetic aperture radar. *IEEE Geosci. remote sensing magazine* **1**, 6–43 (2013).
19. Li, C. J. & Ling, H. Synthetic aperture radar imaging using a small consumer drone. In *2015 IEEE international symposium on antennas and propagation & USNC/URSI national radio science meeting*, 685–686 (IEEE, 2015).
20. Rosen, P. A. *et al.* Synthetic aperture radar interferometry. *Proc. IEEE* **88**, 333–382 (2000).
21. Levanda, R. & Leshem, A. Synthetic aperture radio telescopes. *IEEE Signal Process. Mag.* **27**, 14–29 (2010).
22. Dravins, D., Lagadec, T. & Nuñez, P. D. Optical aperture synthesis with electronically connected telescopes. *Nat. communications* **6**, 6852 (2015).
23. Ralston, T. S., Marks, D. L., Carney, P. S. & Boppart, S. A. Interferometric synthetic aperture microscopy. *Nat. Phys.* **3**, 129–134 (2007).
24. Hayes, M. P. & Gough, P. T. Synthetic aperture sonar: A review of current status. *IEEE journal oceanic engineering* **34**, 207–224 (2009).
25. Hansen, R. E. Introduction to synthetic aperture sonar. In Kolev, N. Z. (ed.) *Sonar Systems*, chap. 1, DOI: 10.5772/23122 (IntechOpen, Rijeka, 2011).
26. Jensen, J. A., Nikolov, S. I., Gammelmark, K. L. & Pedersen, M. H. Synthetic aperture ultrasound imaging. *Ultrasonics* **44**, e5–e15 (2006).
27. Zhang, H. K. *et al.* Synthetic tracked aperture ultrasound imaging: design, simulation, and experimental evaluation. *J. Med. Imaging* **3**, 027001–027001 (2016).

28. Barber, Z. W. & Dahl, J. R. Synthetic aperture ladar imaging demonstrations and information at very low return levels. *Appl. optics* **53**, 5531–5537 (2014).

29. Turbide, S., Marchese, L., Terroux, M. & Bergeron, A. Synthetic aperture lidar as a future tool for earth observation. In *International Conference on Space Optics—ICSO 2014*, vol. 10563, 1115–1122 (SPIE, 2017).

30. Zhang, H., Jin, X. & Dai, Q. Synthetic aperture based on plenoptic camera for seeing through occlusions. In *Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part I 19*, 158–167 (Springer, 2018).

31. Yang, T. *et al.* Kinect based real-time synthetic aperture imaging through occlusion. *Multimed. Tools Appl.* **75**, 6925–6943 (2016).

32. Pei, Z. *et al.* Occluded-object 3d reconstruction using camera array synthetic aperture imaging. *Sensors* **19**, 607 (2019).

33. Burt, P. & Julesz, B. A disparity gradient limit for binocular fusion. *Science* **208**, 615–617, DOI: 10.1126/science.7367885 (1980). https://www.science.org/doi/pdf/10.1126/science.7367885.

34. Deepa, B., Valarmathi, A. & Benita, S. Assessment of stereo acuity levels using random dot stereo acuity chart in college students. *J. family medicine primary care* **8**, 3850–3853 (2019).

35. Filippini, H. R. & Banks, M. S. Limits of stereopsis explained by local cross-correlation. *J. Vis.* **9**, 1–18 (2009).

36. Tyler, C. W. Depth perception in disparity gratings. *Nature* **251**, 140–142 (1974).

37. Hainich, R. R. & Bimber, O. *Displays: fundamentals & applications* (CRC press, 2016).

38. Howard, I. P. *Seeing in depth, Vol. 1: Basic mechanisms.* (University of Toronto Press, 2002).

39. Howard, H. J. A test for the judgment of distance. *Transactions Am. Ophthalmol. Soc.* **17**, 195 (1919).

40. McKee, S. P. & Verghese, P. Stereo transparency and the disparity gradient limit. *Vis. Res.* **42**, 1963–1977, DOI: https://doi.org/10.1016/S0042-6989(02)00073-1 (2002).

41. Treisman, A. Binocular rivalry and stereoscopic depth perception. *Q. J. Exp. Psychol.* **14**, 23–37 (1962).

42. Wilcox, L. M. & Lakra, D. C. Depth from binocular half-occlusions in stereoscopic images of natural scenes. *Perception* **36**, 830–839 (2007).

43. Forte, J., Peirce, J. W. & Lennie, P. Binocular integration of partially occluded surfaces. *Vis. Res.* **42**, 1225–1235, DOI: https://doi.org/10.1016/S0042-6989(02)00053-6 (2002).

44. Halpern, D. L. & Blake, R. R. How contrast affects stereoacuity. *Perception* **17**, 483–495 (1988).

45. Nardini, M., Bedford, R. & Mareschal, D. Fusion of visual cues is not mandatory in children. *Proc. Natl. Acad. Sci.* **107**, 17041–17046 (2010).

46. Zhang, K. *et al.* Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis Mach. Intell.* **44**, 6360–6376 (2021).

47. Schönberger, J. L., Zheng, E., Pollefeys, M. & Frahm, J.-M. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision (ECCV)* (2016).

48. Schönberger, J. L. & Frahm, J.-M. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).

Ethical approval

The ethics committee of the Johannes Kepler University approved the study, all experiments were performed in accordance with relevant guidelines and regulations, and informed consent was obtained from all participants. All experiments have been performed in accordance with the Declaration of Helsinki.

Acknowledgements

This research was funded by the Austrian Science Fund (FWF) and German Research Foundation (DFG) under grant numbers P32185-NBL and I 6046-N, as well as by the State of Upper Austria and the Austrian Federal Ministry of Education, Science and Research via the LIT-Linz Institute of Technology under grant number LIT2019-8-SEE114.

Author Contributions Statement

O.B. developed the concept. R.K. and R.J.A.A.N. implemented the algorithms. R.K., R.J.A.A.N., R.M., and O.B. conceived the experiments, R.K. and R.J.A.A.N. conducted the experiments. R.K. and O.B. and R.M. analyzed the results. O.B., R.K., R.J.A.A.N., and R.M. wrote the paper. All authors reviewed the manuscript.

Additional Information

Code and Data Availability: The python script with the equations in the Introduction which was used for computing Fig. 2, the stereoscopic visualization application and image data used for our user experiments, the raw and cleaned results of our user experiments, and the 3D reconstructions shown the Appendix are available at https://doi.org/10.5281/zenodo.8423145. Since the drone imaging application used for recording our data is a dual use technology, it is available on request from https://github.com/JKU-ICG/AOS/.
Competing interests: None.