Efficient depth localization of objects in a 3D space using computational integral imaging

Michael Kadosh, Anton Fraiman, Eli Peli, Yitzhak Yitzhaky

Michael Kadosh, Anton Fraiman, Eli Peli, Yitzhak Yitzhaky, "Efficient depth localization of objects in a 3D space using computational integral imaging," Proc. SPIE 12742, Artificial Intelligence for Security and Defence Applications, 1274208 (17 October 2023); doi: 10.1117/12.2683627

Event: SPIE Security + Defence, 2023, Amsterdam, Netherlands

Efficient depth localization of objects in a 3D space using computational integral imaging

Michael Kadosh[a], Anton Fraiman[a], Eli Peli[b], Yitzhak Yitzhaky*[a]

[a]Dept. of Electro-Optical Engineering, School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer Sheva, Israel; [b]Schepens Eye Research Institute of Massachusetts Eye and Ear, Dept. of Ophthalmology, Harvard Medical School, Boston, MA, USA.

ABSTRACT

Accurate localization and recognition of objects in the three dimensional (3D) space can be useful in security and defence applications such as scene monitoring and surveillance. A main challenge in 3D object localization is to find the depth location of objects. We demonstrate here the use of a camera array with computational integral imaging to estimate depth locations of objects detected and classified in a two-dimensional (2D) image. Following an initial 2D object detection in the scene using a pre-trained deep learning model, a computational integral imaging is employed within the detected objects' bounding boxes, and by a straightforward blur measure analysis, we estimate the objects' depth locations.

Keywords: 3D imaging, 3D object localization, computational integral imaging, depth localization.

1. INTRODUCTION

A central component in surveillance operations is detecting and tracking objects in the monitored environment. To address these tasks, the majority of methods were developed for two-dimensional (2D) image and video signals, which are representations of the 3D environment. Such methods do not produce knowledge about the depth locations of the detected objects.

Methods that extract depth location of objects include time-of-flight cameras [1] or structured light imaging [2], but these methods require sources of illumination. A common method for passive depth estimation is stereo imaging, however, it may require complex measurements for the disparity calculations [3]. Recently, monocular images have been used for 3D object detection, mainly through learning procedures [4,5], and the KITTI 3D detection data set [6] has been commonly used as a benchmark for this purpose. These methods, applied largely in scenarios of autonomous driving, depend on labeled examples and their depth accuracies may not be high.

Following a recently developed method [7], in our approach for objects' depth localization we employ computational integral-imaging [8,9] which is a passive imaging technique that can produce information about the depth of objects in the scene by imaging the scene into an array of images slightly shifted from each other. Computational integral imaging calculates reconstructed depth planes (RPs) of the scene from the array of images as follows [9,10]:

$$f^{RP}(x,y,z_0)=\frac{1}{KL}\sum_{k=0}^{K-1}\sum_{l=0}^{L-1}g_{k,l}\left(x+\left(\frac{1}{M_{z_0}}\right)S_x k,\, y+\left(\frac{1}{M_{z_0}}\right)S_y l\right) \quad (1)$$

where $g_{k,l}$ is an elemental image with k and l indices, $K \times L$ are the number of Elemental Images (EIs) in the array, and M is the magnification factor, that is the ratio between the distance from the camera to the reconstructed plane and the camera's focal distance.

* ytshak@bgu.ac.il; phone 972-8-6428618;

Artificial Intelligence for Security and Defence Applications, edited by Henri Bouma, Judith Dijk,
Radhakrishna Prabhu, Robert J. Stokes, Yitzhak Yitzhaky, Proc. of SPIE Vol. 12742,
1274208 · © 2023 SPIE · 0277-786X · doi: 10.1117/12.2683627

S_x and S_y are the distances between the cameras in the x and y directions (x and y define the camera's plane), respectively, and $f^{RP}(x, y, z_0)$ is the 2D reconstructed plane at a distance Z_0 from the camera. In this way, we obtain a set of RPs at different distances Z_i from the camera. In each $RP(Zi)$, the distance (depth) Z_i of an object surface that exists will appear sharp, while objects at other distances from the camera become blurred.

We use this property for the detection of objects' depth locations. The basic approach is to find the depth plane which is the sharpest along the depth axis [11,12]. This sharpness measure can be obtained by calculating the average gradient magnitude of each reconstructed (AGMR) plane, and the depth locations of objects can be obtained according to the depths where peak values in the AGMR appear.

2. METHOD

Recently, we proposed an algorithm for detecting and localizing objects in a 3D scene using computational integral-imaging [7]. We used a custom prototype camera array [13] to obtain an array of 21 images or videos (in a 3 x 7 configuration). The Elemental Videos (EVs), where each image or video observes a slightly different angular perspective of the scene, constitute the array of Elemental Images (EIs) at each time instance. Object detection with instance segmentation is applied to a central elemental image in the array, producing bounding boxes and masks of the detected objects in the 2D image of the 3D scene. We examined several methods for this 2D operation, and present here results using the Mask R-CNN technique [14]. Each of the 2D detected objects at the current video frame goes through a local computational integral imaging process within its bounding box region, forming a reconstructed *depth tube* constructed of local depth planes. All of the detected local objects' tubes go through local AGMR computations that give the depth locations of the 2D detected objects, producing 3D object detections. A schematic description of this procedure is presented in Figure 1.

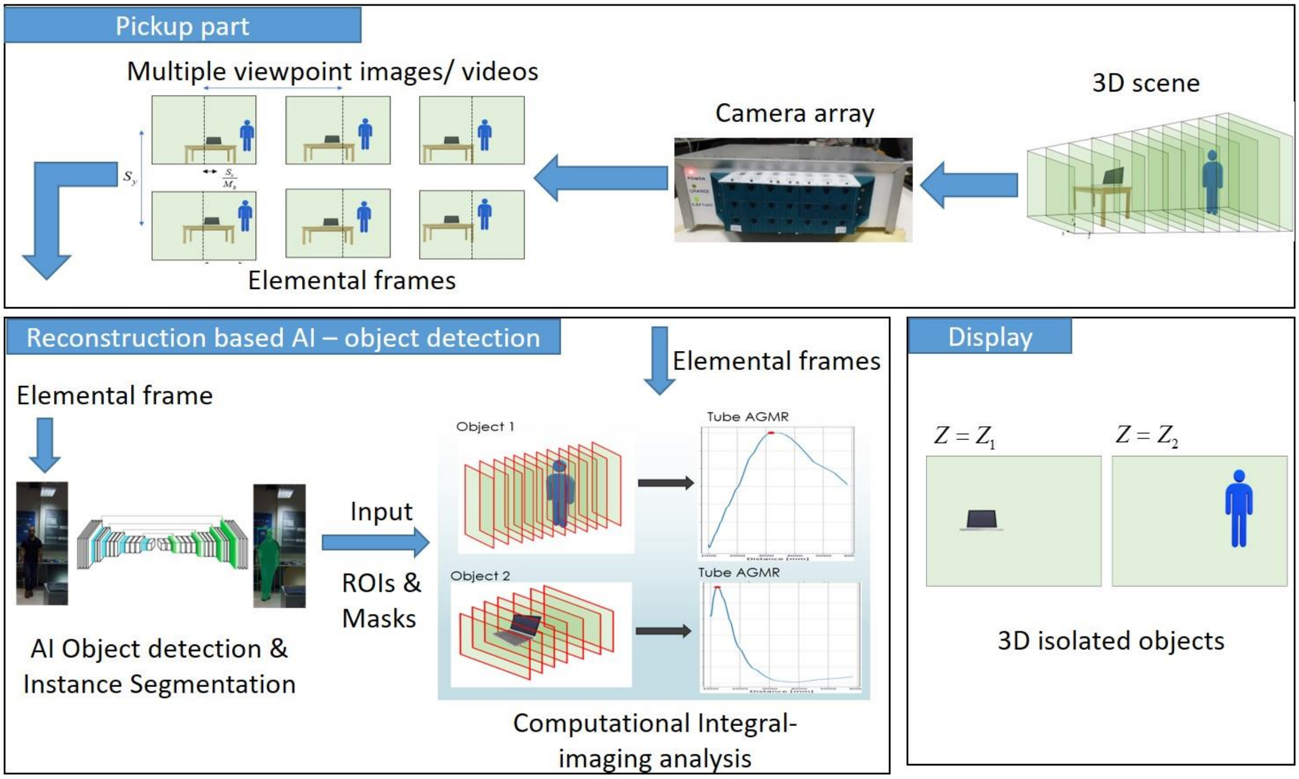

Figure 1. The procedure for object detection and localization in a 3D scene (starting from the upper right side of the figure).

3. RESULTS

Figure 2 presents results of applying the method to two frames of a video that each includes two persons, one walking and one sitting. The figures on the left show the results of applying Mask R-CNN to the frames, detecting in each, the two persons and a chair. The Mask R-CNN was trained on a public MS Coco dataset [15] that has 81 classes. The graphs in the figure show the detected locations (in the depth axis) of the detected persons. The curve in each graph shows the average gradient magnitude of each reconstructed (AGMR) plane calculated at the bounding box region of a detected person. The peaks in the graphs (indicated by red dots) show the detected depth locations of the objects. It can be seen that the detected depth location of the sitting person is about the same in both frames (159cm and 155.5cm), while the detected depth location of the walking person changes according to his location. The small inaccuracies in the detected depths stem likely from the areas in the bounding box that are out of the object mask region (beyond the object edges) that include other objects, that may be moving, as in this case. This can be corrected by applying the computational process only within the object mask region and not within the whole bounding box that may include other objects. The output of the whole process includes the region and mask of each detected object, its category and its estimated distance from the camera.

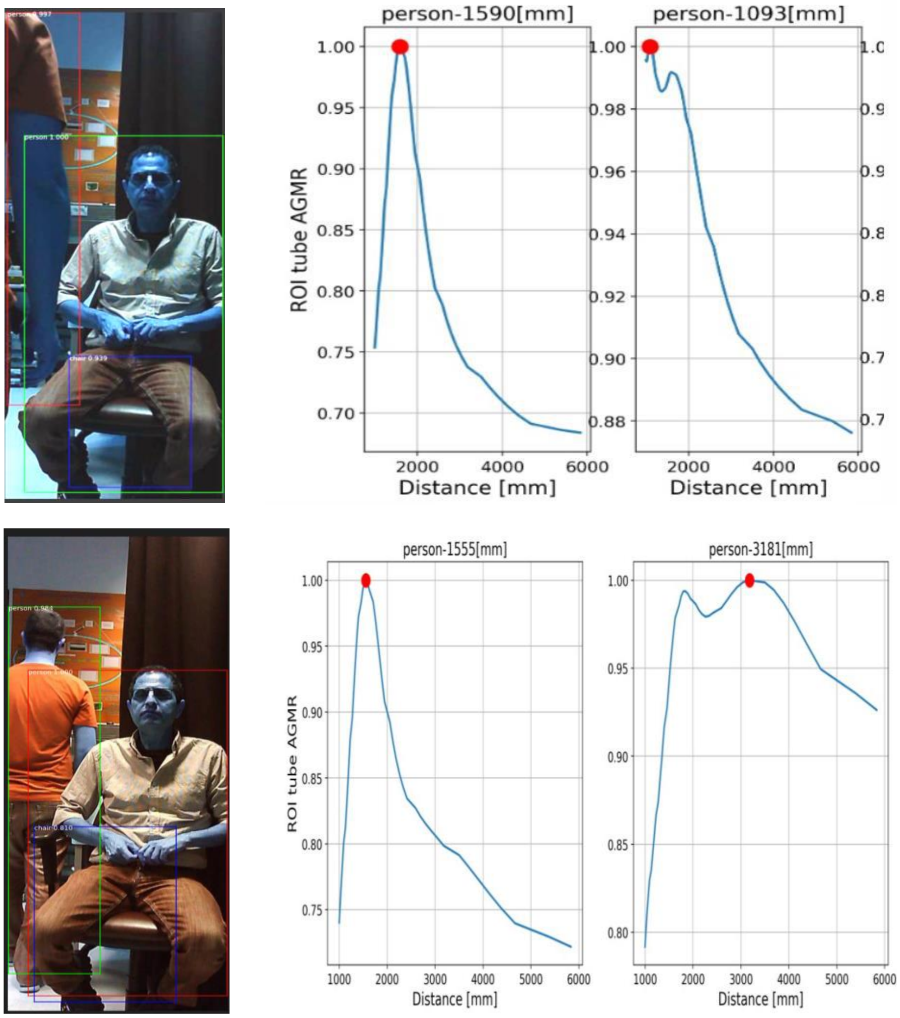

Figure 2. At the left - two video frames (20 (on top) and 70) with two persons each, one sitting and one walking, after applying Mask R-CNN, detecting in each frame, the two persons and a chair (bounding boxes). Note, that in the upper image, only a fraction of the walking person (right arm and shoulder) appears in the frame. Each graph in the figure shows the average gradient magnitude (as a function of the depth axis) of the reconstructed planes (AGMR) calculated at the bounding box region of a detected person. The peaks in the graphs, indicated by red dots, are the detected depth locations of the persons.

4. CONCLUSIONS

In this paper, we demonstrate a method for 3D object localization. The method employs local computational integral-imaging to find depth locations of objects following an object identified via a common 2D object detection with semantic segmentation. The computational integral-imaging is applied only within the detected objects' bounding boxes, making the depth calculation quite accurate and efficient. To create the image array needed for applying the computational integral-imaging, we used a newly developed camera array that simultaneously captures an array of images. For the 2D object detection and segmentation stage, we are currently examining several segmentation methods, which are computationally more efficient than the Mask R-CNN method.

ACKNOWLEDGEMENTS

This study was supported in part by the Israel Science Foundation (Grant No. 1519/20) and the Margaret and Leo Meyer and Hans M. Hirsch Foundation.

REFERENCES

[1] Hansard, M.; Lee, S.; Choi, O.; Horaud, R. *Time-of-Flight Cameras: Principles, Methods and Applications*; SpringerBriefs in Computer Science; Springer: London, UK, 2013; ISBN: 978-1-4471-4657-5.

[2] Geng, J., "Structured-Light 3D Surface Imaging: A Tutorial", *Adv. Opt. Photonics* 2011, *3*, 128–160.

[3] Lang, M., Hornung, A., Wang, O., Poulakos, S., Smolic, A., Gross, M., "Nonlinear Disparity Mapping for Stereoscopic 3D", *ACM Trans. Graph.* **2010**, *29*, 1–10. https://doi.org/10.1145/1778765.1778812.

[4] Xu, B., Chen, Z., "Multi-Level Fusion Based 3D Object Detection From Monocular Images", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2345-2353.

[5] Ding, M., Huo, Y., Yi, H., Wang, Z., Shi, J., Lu, Z., Luo, P. Learning Depth-Guided Convolutions for Monocular 3D Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2020, pp. 1000-1001.

[6] Geiger, A., Lenz, P., Urtasun, R., "Are we ready for autonomous driving? the kitti vision benchmark suite", In CVPR, 2012.

[7] Kadosh, M., Yitzhaky, Y., "3D Object Detection via 2D Segmentation-Based Computational Integral Imaging Applied to a Real Video", Sensors 23 (9), 2023.

[8] Lippmann, G., "La Photographie Integrale. Comptes-Rendus", 1908, 146, 446–451.

[9] Stern, A., Javidi, B., "Three-Dimensional Image Sensing, Visualization, and Processing Using Integral Imaging", Proc. IEEE 2006, 94, 591–607.

[10] Kishk, S., Javidi, B., "Improved Resolution 3D Object Sensing and Recognition Using Time Multiplexed Computational Integral Imaging", Opt. Express 2003, 11, 3528–3541.

[11] Aloni, D., Yitzhaky, Y., "Detection of Object Existence from a Single Reconstructed Plane Obtained by Integral Imaging", IEEE Photonics Technol. Lett. 2014, 26, 726–728.

[12] Aloni, D., Yitzhaky, Y., "Automatic 3D Object Localization and Isolation Using Computational Integral Imaging", Appl. Opt. 2015, 54, 6717. https://doi.org/10.1364/AO.54.006717.

[13] Avraham, D., Samuels, G., Jung, J.-H., Peli, E., Yitzhaky, Y., "Computational Integral Imaging Based on a Novel Miniature Camera Array", Optica Publishing Group: Washington, DC, USA, 2022; p. 3Tu5A–2.

[14] He, K., Gkioxari, G., Dollár, P., Girshick, R., "Mask R-CNN", IEEE: Piscataway, NJ, USA, 2017; pp. 2961–2969.

[15] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., "Microsoft COCO: Common Objects in Context", In *Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 740–755.