Вычислительные технологии, 2024, том 29, № 2, с. 79–94. © ФИЦ ИВТ, 2024 ISSN 1560-7534
Computational Technologies, 2024, vol. 29, no. 2, pp. 79–94. © FRC ICT, 2024 eISSN 2313-691X

DOI:10.25743/ICT.2024.29.2.007

ML based approach for determining the spatial position and size of objects on images

M. A. Letenkov, E. O. Cherskikh*
St. Petersburg Federal Research Center of the RAS, 199178, St. Petersburg, Russia
*Corresponding author: Ekaterina O. Cherskikh, e-mail: cherskikh.e@iias.spb.su
Received September 05, 2022, revised July 05, 2023, accepted July 20, 2023.

This paper considers an urgent problem of assessing spatial position and geometric characteristics of environmental objects from images. An approach was developed based on combining the results of object detection using the Mask R-CNN model and the reconstruction of depth maps obtained using the RealSense camera. We evaluated the class-averaged values of the relative error in determining the size of objects for test sets of images formed at various levels of scene illumination: 0.1449, 0.3313, 0.6332. Also, within the experiments relative deviation values were obtained when determining the spatial positions of objects: 0.1010, 0.1624, 0.3477.

Keywords: object's size evaluation, spatial position assessing, object detection, depth map reconstruction, Mask R-CNN, Intel RealSense.

Citation: Letenkov M.A., Cherskikh E.O. ML based approach for determining the spatial position and size of objects on images. Computational Technologies. 2024; 29(2):79–94. DOI:10.25743/ICT.2024.29.2.007.

Introduction

Today cyber-physical systems (CPS) based on the use of cloud technologies, Internet of things technologies, artificial intelligence, and machine learning methods are becoming increasingly widespread. CPS [1, 2] include various sensor devices, that collect data on the physical properties of the cyber-physical environment. Interpretation of this data can be used to control the state of the system or expand the functionality of such systems. One of the urgent problems identified in the context of determining the properties of the cyber-physical environment is the problem of identification, determination of spatial coordinates, location and size of environmental objects using various sensor devices based on data from various modalities.

This study was carried out in the context of the CPS, which implements information services to employees of a scientific organization [3]. Current study aims to expand the capabilities of this CPS to interact with objects of the external environment. This system is equipped with a variety of video cameras located throughout its deployment area. Thus, this work is focused on the development of an approach to solving the problem of estimating the positions and geometric characteristics of environmental objects from images recorded by the RealSense D435 depth camera [4] in real time. As environmental objects can be considered furniture, peripherals, office equipment and people. The development of such a solution allows to maintain object records and control changes in various premises. The implementation of the proposed approach can also be used as an independent object tracking system.

1. Related work

Developing an approach to solving the problem of estimating the position and geometric characteristics of environmental objects from images involves solving the following group of main problems: image segmentation; objects classification; estimation of the distances between the camera lens and detected objects.

Most modern methods for classification and detection of object outlines in images are based on the use of artificial neural networks [5, 6], however, training of a neural network from scratch requires a lot of time and a significant set of pre-labeled data. Thus, within this study, pretrained models of neural networks presented on the Tensor Flow project website [7] were analyzed. The Tensor Flow project contains neural network models based on architectures such as Mobilenet [8], Fast R-CNN [9] and Mask R-CNN [6]. According to the results of the analysis, it was decided to use a neural network based on the Mask R-CNN architecture [6], since this neural network is capable to determine not only areas, but also specific boundaries of objects. Furthermore, the speed of this neural network allows it to work in real time even on cameras with low frame rates. One of the fastest-performing implementations of the Mask R-CNN neural network is the Mask R-CNN Inception v2 configuration trained on the COCO dataset [10]. Compared to the ResNet configurations [11], the processing speed of one frame by the neural network in the considered configuration is 5–10 times faster than analogues [6]. A characteristic difference of the Inception configuration is the use of the 3D Inception method for the spatiotemporal analysis of movement signs of visual objects in a video stream [12].

Within current study it is important to note that determining the position and size of an object in global coordinates based on the object's position in image coordinates, requires estimating the distance between camera lens and the detected object. Estimation of such distances can be implemented using specialized hardware, such as LIDAR [13], Kinect [14] and RGB-D cameras [4]. Scanning RGB-D cameras and lidars generate a depth map that allows determining the position and spatial characteristics of captured objects in the scene. Estimating of distances between camera lens and the detected objects can also be implemented using neural network based methods [15–17].

Most known methods for estimating the size and position of objects in space are based on a combination of depth cameras with stereoscopic vision systems, traditional machine vision algorithms or neural network models, heuristic methods and algorithms combining them. So, the authors of [18] explore the capabilities of the Microsoft Kinect v2 sensor using Kinect Fusion algorithms to identify three-dimensional objects and consider the possibility of combining the sensors used. For object detection and estimation, the authors of [19] solve the structured polygon prediction problem and the subsequent depth reconstruction problem by combining depth cameras and machine vision algorithms. A comprehensive solution consisting of deep learning based object detection, image processing, combining RGB-D information and a task management system is proposed in [20]. The position of objects, in the case of [21] — hands, is performed using a single depth camera without markers using a deep neural network based on a synthetic data set. In general, there are many unresolved problems in the field related to the cameras used, data sets and the general focus on statistical characteristics [22].

Within this study, a comparative analysis of the neural network models FCRN-DepthPrediction [23], AlexNet [15], and VGG [16], capable of estimating depth from an image obtained from an ordinary monocular camera, was carried out. The analysis showed that methods for

constructing depth maps based on neural networks show a relatively low level of accuracy for scenes with objects located far enough from the camera lens. It should also be noted that these solutions do not provide an estimate of the distances between the camera lens and detected objects in metric units, which excludes the use of neural network methods to obtain depth maps within current study. Accordingly, it was decided to use the RealSense D435 depth camera [4] since this technology for obtaining a depth map shows high accuracy compared to analogues [24], and also allows estimating the distance between the camera lens and detected objects in low-light scenes [4].

Thus, this study proposes an approach for estimating the spatial position and geometric characteristics of environmental objects from images obtained using a RealSense D435 depth camera [4]. The proposed approach is based on the combined solution of object detection and segmentation and the problem of estimating the distances between the camera lens and detected objects.

2. Development of an approach for determining the spatial position and geometric characteristics of objects from images using machine learning methods

The current study represents the author's method for solving the problem of estimating the spatial position and geometric characteristics of scene objects from images. Within the proposed approach, the source of images is a RealSense D435 camera, mounted vertically on the ceiling of a premises in such a way that the optical axis of the camera lens is in the horizontal plane, and the walls of the room are out of the frame.

The proposed approach involves the implementation of several preliminary steps, performed once during the process of deploying the corresponding solution in the premises under study. As part of this preliminary stage, the distortion effect and perspective distortion are eliminated, the parameters of the scene and video camera are assessed. For this purpose, firstly the initial camera calibration is carried out according to the method [25]: the calibration eliminates the effect of distortion (radial and tangential distortions) and perspective distortions, which leads to an increase in the accuracy of transforming the image coordinates into the coordinates of the real world. The use of this calibration method allows to determine of the main matrix of the camera, containing focal lengths and optical centers, distortion coefficient, rotation and displacement vectors. All this data is essential for the proposed solution to work.

The next step of the preliminary stage is the evaluation of the relevant viewing angles of the video camera along the horizontal (Alpha) and vertical (Betta) image axes. Using camera lens viewing angles as viewing angles becomes impossible since eliminating of the distortion effect leads to loss of the border parts of the image. Thus, the determination of viewing angles should be carried out experimentally, by estimating the size of a visible segment of a flat surface, located at a known distance from the camera lens and oriented perpendicular to the optical axis of the camera.

The last step in the preliminary stage is the estimation of scene parameters. The assessment of the scene under study is carried out experimentally: using the known coordinates of the center of the scene (x_c, y_c), located strictly under the camera, the vertical distance H from the floor to the camera lens is measured. In this case, the coordinates of the center of the scene will be as follows:

$$C_{\text{enter}} = (x_c, y_c, H).$$

The center of the image coincides with the center of the scene up to the z coordinate. Thus, for the center of the image $c_{\text{im}} = (x_{c_{\text{im}}}, y_{c_{\text{im}}})$, the following expression is valid: $x_{c_{\text{im}}} \to x_c$, $y_{c_{\text{im}}} \to y_c$. Thus, the center of the scene under study is displayed on the central pixel of an image obtained from a camera.

Considering the preliminary steps above, the following approach was proposed for assessing the spatial position and geometric characteristics of environmental objects. The developed approach includes the following main stages.

1. Detection of objects in the considered image:
   a) transfer of the image received from the RealSense D435 camera [4] to the Mask R-CNN neural network model [6];
   b) determination the number of objects in the image and their belongness to a certain class of objects;
   c) determination of objects' boxes;
   d) determination of five reference points for each box: vertices and the center of the box.
2. Determination of the spatial areas of the scene associated with the pixels of the image under study.
3. Determination of the height, size and spatial coordinates of objects detected in the image.

The proposed approach contains three key stages, each of them is aimed at solving a separate group of subtasks. Let's consider these stages in more detail.

2.1. Detection of objects in the considered image

At the first stage, the image obtained from the RealSense D435 camera [4] is transferred to the Mask R-CNN neural network model [6], which searches for the coordinates of objects in the image, determines the number of objects and their belongness to certain classes of objects. The inception blocks that are used in this neural network are analogous to those in the Inception-V1 neural network, where the feature map obtained in the previous layer is processed in parallel by four different branches. The first branch of the Inception block is a convolution with a filter (1×1×1), the second and third branches contain two consecutive convolutional layers with filters (1×1×1) and (3×3×3), respectively, and the fourth branch contains two consecutive layers — Max-Pool with a field (3×3×3) and a convolutional layer (1×1×1). Further, the resulting feature map is formed by concatenating the feature maps obtained using all four branches of the Inception module.

The first layers of the Mask R-CNN model are alternating 3D convolutional and pooling layers followed by a sequence of alternating Inception blocks and pooling layers. The output of this neural network is a (1×1×1) 3D convolutional layer. The output feature map of this layer is "stretched" into a vector, the size of which corresponds to the number of classes under consideration.

The neural network of this architecture is capable of processing video fragments of various lengths, each frame is a three-channel colour image of the RGB colour space when processing an image, the neural network model returns the number of objects found and a set of arrays (boxes, classes, masks, degree of correspondence).

Within the proposed approach, it was decided to use boxes to subsequent by determine the key points of objects in the images, since they can be used to obtain all the required information about the objects under study, and furthermore, the use of boxes has a positive

effect on the computational complexity of the proposed approach. Boxes are arrays, each of which consists of four extreme points, corresponding to rectangular areas in which objects are located. Based on the result of box detection, a set of reference points is formed for each box. Such a set consists of the coordinates of the vertices and the geometric center of the associated box. Thus, the result of applying the Mask R-CNN neural network model to a certain image is an array O, which consists of a set of sets of reference points for the detected boxes (Points_j) and of a set of classes (Cls_j), corresponding to the detected objects Ob_j.

2.2. Determination of the spatial areas of the scene associated with the pixels of the image under study obtained from the RealSense D435 camera

Let us consider the developed algorithm for determining the spatial areas of the scene corresponding to the pixels of the considered image. Assume that there is an image of the floor of a certain scene without any objects. The central pixel of the image P_c, located directly under the camera, corresponds in space to a certain square in the center of the scene Fig_c with area S_c. However, if we consider not the central pixel, but some arbitrary pixel of the image P_i, then the situation will be different. In this case, a certain quadrangle Fig_i with area S_i will correspond to pixel P_i in space, and $S_c! = S_i$. This circumstance is due to the fact that parts of the scene, offset from the center of the scene, are observed by the camera at a certain angle, and thus a larger area of space is associated with each pixel of the image. Let us consider in more detail the solution of the problem of associating image pixels and the corresponding spatial areas under the conditions specified above: only the floor is present in the scene image, the camera is fixed perpendicular to the floor plane.

Let the size of the image under study in pixels along the horizontal and vertical axes be equal to $\text{Img}_{\text{pix}_x}$ and $\text{Img}_{\text{pix}_y}$, respectively. In this case, the indices of the central pixel P_c of the image can be determined as follows:

$$P_c(x;y) = \left(\frac{\text{Img}_{\text{pix}_x}}{2}; \frac{\text{Img}_{\text{pix}_y}}{2}\right).$$

As mentioned earlier, the central pixel P_c of the image in space corresponds to a certain square Fig_c in the center of the scene. Each of the sides of a given square is deflected from the normal to the XY plane by some planar angle in the XZ or YZ planes. Accordingly, each of the vertices of this square is deviated from the normal to the XY plane by some planar angles in the XZ and YZ planes. Similarly, angular deviation occurs for an arbitrary pixel P_i with indices along the horizontal and vertical axes i_{pix_x} and i_{pix_y}, the corresponding deviation values can be determined as follows:

$$ang\_dev\_x_{P_i} = \left[-\frac{\text{Alpha}}{\text{img}_{\text{pix}_x}\cdot 2} + \left(i_{\text{pix}_x} - \frac{\text{img}_{\text{pix}_x}}{2}\right)\frac{\text{Alpha}}{\text{img}_{\text{pix}_x}}; \frac{\text{Alpha}}{\text{img}_{\text{pix}_x}\cdot 2} + \left(i_{\text{pix}_x} - \frac{\text{img}_{\text{pix}_x}}{2}\right)\frac{\text{Alpha}}{\text{img}_{\text{pix}_x}}\right],$$

$$ang\_dev\_y_{P_i} = \left[-\frac{\text{Betta}}{\text{img}_{\text{pix}_y}\cdot 2} + \left(i_{\text{pix}_y} - \frac{\text{img}_{\text{pix}_y}}{2}\right)\frac{\text{Betta}}{\text{img}_{\text{pix}_y}}; \frac{\text{Betta}}{\text{img}_{\text{pix}_y}\cdot 2} + \left(i_{\text{pix}_y} - \frac{\text{img}_{\text{pix}_y}}{2}\right)\frac{\text{Betta}}{\text{img}_{\text{pix}_y}}\right].$$

The positions of the vertices of the spatial area of the scene covered by this pixel P_i, respectively, can be determined as follows:

$$1_i: \left(\text{tg}\left[\frac{\text{Alpha}}{\text{img}_{\text{pix}_x}\cdot 2}+\left(i_{\text{pix}_x}-\frac{\text{img}_{\text{pix}_x}}{2}\right)\frac{\text{Alpha}}{\text{img}_{\text{pix}_x}}\right]H; \text{tg}\left[\frac{\text{Betta}}{\text{img}_{\text{pix}_y}\cdot 2}+\left(i_{\text{pix}_y}-\frac{\text{img}_{\text{pix}_y}}{2}\right)\frac{\text{Betta}}{\text{img}_{\text{pix}_y}}\right]H\right),$$

$$2_i: \left(\text{tg}\left[-\frac{\text{Alpha}}{\text{img}_{\text{pix}_x}\cdot 2}+\left(i_{\text{pix}_x}-\frac{\text{img}_{\text{pix}_x}}{2}\right)\frac{\text{Alpha}}{\text{img}_{\text{pix}_x}}\right]H; \text{tg}\left[\frac{\text{Betta}}{\text{img}_{\text{pix}_y}\cdot 2}+\left(i_{\text{pix}_y}-\frac{\text{img}_{\text{pix}_y}}{2}\right)\frac{\text{Betta}}{\text{img}_{\text{pix}_y}}\right]H\right),$$

$$3_i: \left(\text{tg}\left[-\frac{\text{Alpha}}{\text{img}_{\text{pix}_x}\cdot 2}+\left(i_{\text{pix}_x}-\frac{\text{img}_{\text{pix}_x}}{2}\right)\frac{\text{Alpha}}{\text{img}_{\text{pix}_x}}\right]H; \text{tg}\left[-\frac{\text{Betta}}{\text{img}_{\text{pix}_y}\cdot 2}+\left(i_{\text{pix}_y}-\frac{\text{img}_{\text{pix}_y}}{2}\right)\frac{\text{Betta}}{\text{img}_{\text{pix}_y}}\right]H\right),$$

$$4_i: \left(\text{tg}\left[\frac{\text{Alpha}}{\text{img}_{\text{pix}_x}\cdot 2}+\left(i_{\text{pix}_x}-\frac{\text{img}_{\text{pix}_x}}{2}\right)\frac{\text{Alpha}}{\text{img}_{\text{pix}_x}}\right]H; \text{tg}\left[-\frac{\text{Betta}}{\text{img}_{\text{pix}_y}\cdot 2}+\left(i_{\text{pix}_y}-\frac{\text{img}_{\text{pix}_y}}{2}\right)\frac{\text{Betta}}{\text{img}_{\text{pix}_y}}\right]H\right).$$

In accordance with the above expressions, for each pixel P_i, the vertices of the spacial area that correspond to the given pixel can be found. For each such area, its center C_i can be estimated as the arithmetic mean of the coordinates of its vertices:

$$C_i = \left(\frac{1}{4}\sum_{k=1}^{4} k_{i_x}; \frac{1}{4}\sum_{k=1}^{4} k_{i_y}\right).$$

Thus, as a result of this stage, an array of centers of spatial areas of the scene for pixels is formed, corresponding to the reference points of the boxes of objects from the Points_j sets included in the array O.

2.3. Determination of the height, size and spatial coordinates of objects detected in the image

At the final stage of the developed approach, the height, size and spatial coordinates of the objects Ob_j detected in the image are determinated. The input to this stage is the information about the centers of the spatial areas (C_i) obtained in the second stage for pixels P_i, corresponding to the reference points of the boxes of objects from the Points_j sets. For each such center C_i, the distance from the camera lens to a given area of space is calculated in accordance with the following expression:

$$D(C_i)_{\text{floor}} = \sqrt{C_{i_x}^2 + C_{i_y}^2 + H^2}.$$

Thus, we get the expected distances $D(C_i)_{\text{floor}}$, which would be associated with each considered pixel P_i of the image, if the scene were a flat floor, with no objects on it.

Further, for the pixels P_i, corresponding to the reference points of the boxes of objects Ob_j, the distances $D(C_i)_{\text{fact}}$ to the centers of the actual spatial areas of the scene associated with the pixels P_i are determined (Fig. 1, *a*) using the RealSense D435 camera. Now we can subtract from the set of expected distances $D(C)_{\text{floor}}$, the actual distances $D(C)_{\text{fact}}$:

$$D(C)_{\text{delta}} = D(C)_{\text{floor}} - D(C)_{\text{fact}}.$$

Each element of the obtained difference set $D(C)_{\text{delta}}$ represents a certain distance B, equal to the difference between the distance that would be between the camera and the spatial area associated with the corresponding pixel P_i, if a section of the floor were associated with this pixel (distance $A + B$), and the actual distance between the spatial area displayed in this pixel and the camera — distance A (Fig. 1, *b*).

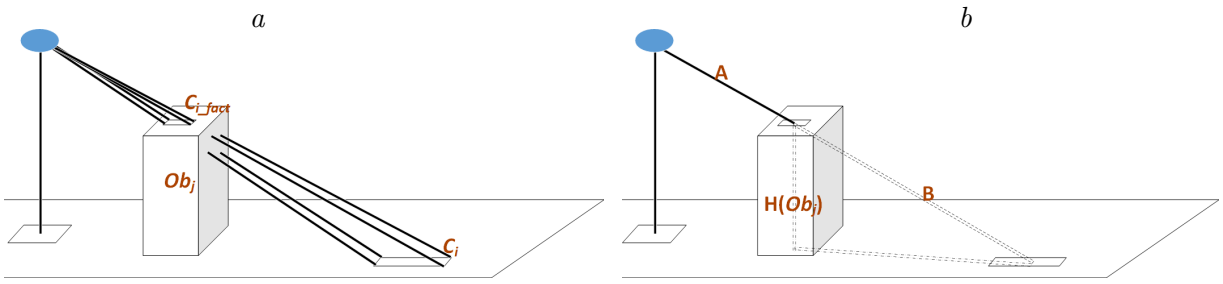

Fig. 1. Illustration of a typical scenes with some arbitrary object (a) and with an indication of the investigated distances (b)

It is important to note that in the absence of errors, all elements of the $D(C)_{\text{delta}}$ set should be greater than or equal to 0. Then, using the values defined above, the heights of objects Ob_j detected on the scene can be determined calculating of the height of each object Ob_j is carried out in two stages. At the first stage, for each target pixel P_{ji} corresponding to one of the reference points (Points_j) from the Points_j set (the set of reference points of the object Ob_j box), the vertical displacement of the spatial area C_{ji} relative to the floor is determined:

$$H(C_{ji}) = \frac{H}{A+B}B = \frac{H}{D(C_{ji})_{\text{floor}}}D(C_{ji})_{\text{delta}}.$$

Thus, based on the number of reference points of the object (Points_j set), a series of estimates of the height of the object Ob_j is formed. At the second stage, the resulting estimate of the object's height $H(\text{Ob}_j)$ is calculated:

$$H(\text{Ob}_j) = \sum_{i=1}^{5} H(C_{ji})w_i,$$

where i is the index of the reference point of the object Ob_j, w are the weight coefficients of the reference points, equal to 0.7 for the reference points associated with the geometric centers of the objects boxes, and 0.075 for other reference points. Based on the results of these calculations, for each object Ob_j detected in the image, a final conclusion about its height is formed.

After the resulting estimate of object's height $H(\text{Ob}_j)$ is obtained for each object Ob_j on the scene, the determination of two other quantities characterizing linear dimentions of objects is carried out as follows:

1. First, the spatial coordinates of the reference points of the object's Ob_j box are determined: $\text{Points}_{ji} = (C_{ji_x}, C_{ji_y}, H(\text{Ob}_j))$, $i = [1, \ldots, m]$), where C_j is an array of spatial areas centers associated with pixels corresponding to the reference points of the object's box (Points_j set), $H(\text{Ob}_j)$ is the final estimate of the object's height, $m = 5$ is the number of reference points of the box of this object.
2. An estimate of the length of an object Ob_j can be defined as the distance between two adjacent reference points. In this work, the length of the object was estimated in accordance with the following expression:
$$L(\text{Ob}_j) = |\text{Points}_{j2} - \text{Points}_{j1}|.$$
3. In a similar way, a width estimation for the object Ob_j can be determined. In this work, the width estimate was determined as follows:
$$W(\text{Ob}_j) = |\text{Points}_{j3} - \text{Points}_{j2}|.$$

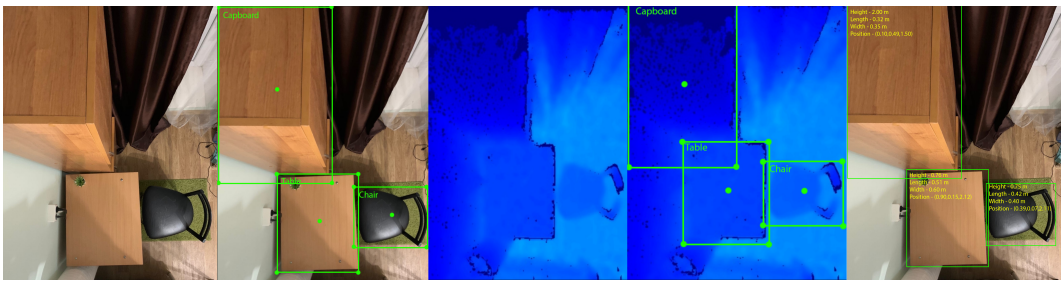

Fig. 2. Step-by-step results of the proposed approach approbation on a certain scene: RGB image of the scene with 2–3 simple objects observed; result of Mask R-CNN operation with highlighted boxes and signed object classes; depth map of the investigated scene obtained using the RealSense D435 camera; depth map of the investigated scene with highlighted boxes and signed object classes; the resulting RGB image with the signed parameters of the objects

To estimate the location of objects, it was decided to take the position of the camera lens as the origin of the reference system. The location of object was approximated by estimates of the spatial positions of the center of these objects. For each object detected in the image, the spatial position of its center was determined in accordance with the following expression:

$$C(\mathrm{Ob}_j) = \left(C_{j5_x}, C_{j5_y}, H - \frac{H(\mathrm{Ob}_j)}{2} \right),$$

where C_{j5} is the position of the center of the spatial area corresponding to the reference point associated with the geometric center of this object's box, H is the distance from the camera lens to the floor of the scene under study along the vertical axis, $H(\mathrm{Ob}_j)$ is the estimate of the height of the object obtained at the previous stage.

Thus, based on the above calculations for each object Ob_j ($\forall \mathrm{Ob}_j \in O$) detected in the image, it is possible to obtain estimated values of the three key characteristics of this object (height, length, and width), as well as obtain an estimate of the location of this object.

Figure 2 below shows the step-by-step results of the approbation of the developed approach to determining the spatial position and geometric characteristics of objects in a certain image.

The first frame of Fig. 2 shows an RGB image of the considering scene. The next one, accordingly to the developed approach, shows the result of objects detection and classification using the Mask R-CNN neural network model. Three objects were detected in the observed image: a chair, a table and a wardrobe. Further, as shown in the third frame depth maps are built for the scene using the RealSense D435 camera. The fourth frame shows the depth map of the scene under study with highlighted boxes and labelled object classes. The results of assessing the spatial position and sizes of the detected objects can be observed in the last frame of the Fig. 2.

Further we will consider the implementation of the proposed approach in terms of quality of estimation the spatial position and size of objects from images.

3. Testing of the Developed Approach for determining the spatial position and geometric characteristics of objects from an image

Approbation and quality assessment of the proposed approach was carried out on a test dataset that includes 12 000 images of various scenes with office premises. This dataset was

generated using the RealSense D435 depth camera [4]. Each image I_i, included in the test dataset corresponds to a certain set of parameter values P_i. These parameters are described below:

A. Target type T. Includes 4 different object classes: cabinets (1), chairs (2), tables (3), plants (4).
B. Size of the target object R. As a parameter characterizing the size of the object, the average value of the object's projections on each of the coordinate axes was chosen — l (m). Thus, each target will be assigned to one of the following categories based on the value of l: 1 — [0,0.5], 2 — [0.5,1], 3 — [1,2].
C. Scene illumination level L. The test dataset includes images generated at various scene illumination levels: 50, 70 and 100 %, where the standard illumination level for office premises is taken as 100 % [17].

The test dataset was divided into three subgroups of images differing in the level of scene illumination: 50, 70, and 100 %. Each selected subgroup of images includes 4000 images containing objects of the following types and sizes: 600 tables of size 2; 150 plants size 1; 50 plants size 2; 8000 chairs in size 1; 4000 chairs of size 2 and 400 cabinets of size 3. As a result of applying the Mask R-CNN neural network model [6] for a subgroup of images with 100 % scene illumination, the following number of objects were detected and correctly classified in Table 1.

Based on the data obtained for all three groups of images, the resulting percentages of detected and correctly classified objects were determined for each selected subgroup with a given level of scene illumination. The result obtained are presented in Table 2.

T a b l e 1. The proportion of detected and correctly classified objects (illumination level — 100 %)

Target type T	Target size R	The percentage of detected objects, %	The share of correctly classified objects from the number of successfully detected, %
Table	2	89	60
Plant	1	88	49
Plant	2	88	48
Chair	1	90	50
Chair	2	89	50
Cabinet	3	90	80

T a b l e 2. The proportion of detected and correctly classified objects using the Mask R-CNN neural network model [6]

Target type T	Target size R	The share of detected and correctly classified objects		
		At 100 % illumination	At 70 % illumination	At 50 % illumination
Table	2	0.53	0.37	0.26
Plant	1	0.43	0.3	0.22
Plant	2	0.42	0.29	0.21
Chair	1	0.45	0.32	0.22
Chair	2	0.45	0.32	0.23
Cabinet	3	0.72	0.51	0.36
Average share		0.5	0.35	0.25

From the data presented above, we can conclude that the accuracy of object detection and classification using the Mask R-CNN neural network model [6] significantly depends on the illumination level of the scene and decreases with decreasing illumination level. It is also important to note that the class of the object affects the quality of the classification, in particular, the calculated indicators have lower values for objects of the following classes: plant and chair.

To directly assess the quality of the developed approach for determining the spatial position and size of objects from images, various quantitative estimates of the accuracy of the proposed solution were formed depending on the scene parameters and the characteristics of the target objects. The corresponding indicators were determined as follows:

1. The relative error was taken as an indicator of the accuracy of determining the sizes of scene objects, which was determined as follows:

$$|a_{\text{real}} - a| \Big/ \left(\frac{a_{\text{real}} + a}{2} \right),$$

where a_{real} is the real value of the object size; a is the value calculated based on the data obtained using the approach considered. The values a_{real} and a were determined directly as average values of the object's projections onto each of the coordinate axes. Thus, the lower the value of the relative error, the higher the accuracy of determining the size of objects using the developed approach.

Figure 3 below shows a diagram of the obtained values of the relative error in determining the sizes for each type of objects for a subgroup of images with a scene illumination equal to 100 %.

The figure above shows a box plot diagram showing the distribution of the obtained values of the relative error for each type of objects. In this chart, the boxes are bound by the first and third quartiles (25^{th} and 75^{th} percentiles, respectively). The average error estimate is shown as a cross in the middle of the boxes, outliers are shown as dots on the graph. The ends of the whiskers are the edges of a statistically significant sample (without outliers), determined in accordance with the following expressions:

$$X_1 = Q_1 - k(Q_3 - Q_1), \quad X_2 = Q_3 + k(Q_3 - Q_1),$$

where X_1 is the lower border of the whiskers; X_2 is the upper border of the whiskers; Q_1 is the first quartile; Q_3 is the third quartile; k is the coefficient, the most used value of which is equal to 1.5.

From the data presented in Fig. 3, we can conclude that the developed approach demonstrates the best results when estimating sizes for objects of the cabinet ($R = 3$) and table ($R = 2$) classes. The mean error values for objects of these classes were 0.1207 and 0.1296, respectively. The developed approach shows a slightly worse quality of determining the size of objects for classes such as chair ($R = 1, 2$), the obtained values of the average error for objects of these classes were 0.1416 and 0.1429, respectively. The developed approach demonstrated the worst results for objects of plant classes ($R = 1, 2$). It is assumed that these results are related to the geometric features of these objects. It is also important to note that for objects of the same class but different sizes, on average, there is no significant change in the values of the relative error. Thus, we can conclude that, up to a certain limit, the actual sizes of objects do not affect the quality of determining the sizes of objects in the context of the developed approach. However, it should also be noted that for objects of

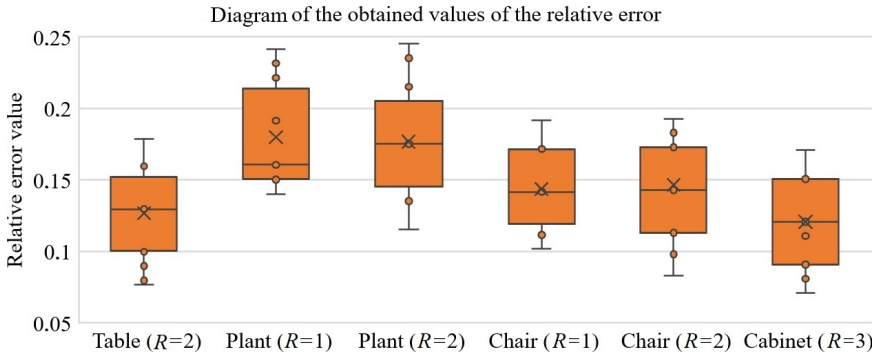

Fig. 3. Diagram of the obtained values of the relative error in determining the sizes for each type of objects

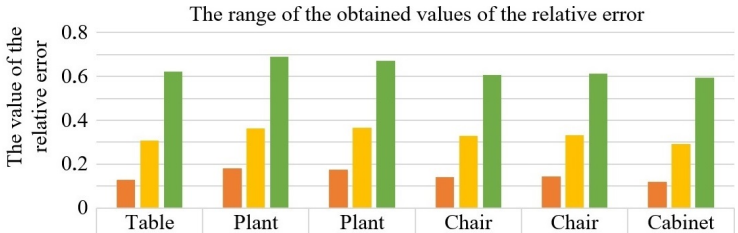

	Table (R=2)	Plant (R=1)	Plant (R=2)	Chair (R=1)	Chair (R=2)	Cabinet (R=3)
Illumination – 100 %	0.1296	0.1809	0.1752	0.1416	0.1429	0.1207
Illumination – 70 %	0.3064	0.3624	0.3658	0.3281	0.3329	0.2923
Illumination – 50 %	0.6214	0.6914	0.6711	0.6083	0.6121	0.5947

Fig. 4. The range of the obtained values of the relative error in determining the object sizes for all levels of illumination

extremely small size, this conclusion will not be valid, since in this case, errors in constructing depth maps, as well as other errors of the developed approach, will play a significantly greater role.

Figure 4 below shows the averaged values of the relative error in determining the size of objects from images at different levels of illumination.

From data presented above it can be concluded that for all types of objects, regardless of their size, there is a significant increase of the relative error in determining the size of object with decreasing scene illumination level. At the same time an increase in error growth rate is observed when going from the illumination level 100 to 70 % and from 70 to 50 %, respectively. In particular, the relative error between illumination levels from 100 to 70 % increases by about 7.5 % for each 1 % decrease in illumination level, and average 9.5 % when illumination levels range from 70 to 50 %. These dependencies are similar for all classes of objects under consideration, regardless of their size. The class-averaged values of the relative error for scenes with different illumination levels are 0.1449, 0.3313, and 0.6332 for illumination levels of 100, 70, and 50 %, respectively.

The results obtained can be explained by the fact that with a significant decrease of illumination level, the accuracy of object detection also decreases. Moreover, additional errors arise when constructing depth maps, which significantly affects the quality of determining the size of objects. This situation is observed when applying the developed approach to determining the spatial position and size for objects in images with an illumination level of 50 %. It is also important to note that a decrease in the illumination level has an extremely

negative impact on the quality of object detection and classification, as a result of which most of the objects presented in the scene remain unnoticed or incorrectly classified.

Thus, it can be concluded that the proposed approach demonstrates acceptable quality of determining the size of objects for images with illumination levels of 100 and 70 %.

2. As an accuracy indicator for determining the spatial position of objects, the value of the relative deviation was chosen, which is determined according to the expression:

$$Err = \frac{|r_{\text{real}} - r|}{l},$$

where r_{real} is the real position of the object; r is the position of the object, calculated using the developed approach; l is the average value of the object projection lengths onto each of the coordinate axes. Thus, the lower the value of the relative deviation, the higher the accuracy of determination the spatial positions of objects.

Figure 5 below shows the range of the obtained relative deviation values while determining the spatial position of objects for a subgroup of images with a scene illumination of 100 %.

According to the data presented in the figure above, the developed approach demonstrates the best average values of the relative deviation for objects of the cabinet ($R = 3$) and table ($R = 2$) classes. The average relative deviation values for objects of these classes were 0.0919 and 0.0932, respectively. Slightly worse results are obtained for objects of the classes chair and plant ($R = 2$), the relative deviation values were 1.1001 and 0.1039, respectively. The worst results in determining the spatial position of objects were demonstrated for objects of the plant and chair classes ($R = 1$), the average value of the relative deviation for these classes was 0.1098 and 0.1071, respectively. It should be noted that the smallest averaged values of the relative deviation are observed for objects of sizes $R = 3$ and $R = 2$. When moving to smaller sizes of target objects ($R = 1$), an increase in the averaged values of the relative deviation is observed. This increase can be explained by a decrease in accuracy of object detection as the size of the object under consideration decreases, as well as a higher specific level of errors arising in the process of reconstructing depth maps. It should also be noted that the variance estimate of the relative deviation for objects of the plant class ($R = 1, 2$) is much higher, and for objects of the table class ($R = 2$) is lower than the average for the samples. This distribution can be explained by the geometric features of these objects: high shape complexity of the plants and the shape simplicity of the table, respectively.

Figure 6 below shows the range of the obtained relative deviation values when determining the spatial position of objects at different levels of illumination.

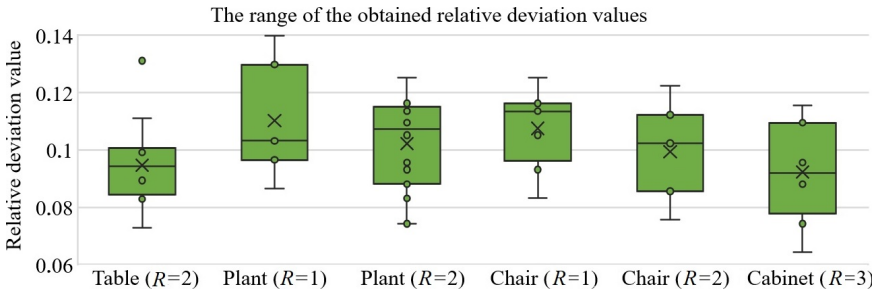

Fig. 5. The range of the obtained relative deviation values while determining the spatial position of objects for a subgroup of images with a scene illumination of 100 %

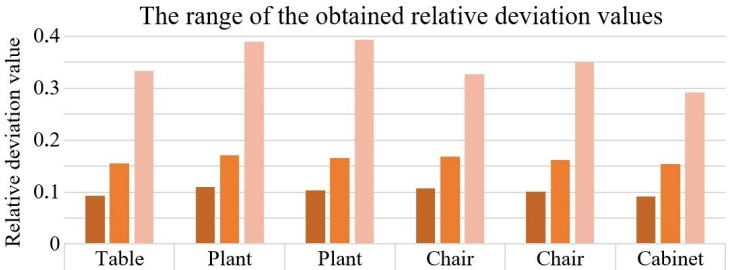

	Table (R=2)	Plant (R=1)	Plant (R=2)	Chair (R=1)	Chair (R=2)	Cabinet (R=3)
Illumination – 100 %	0.0932	0.1098	0.1039	0.1071	0.1001	0.0919
Illumination – 70 %	0.15504	0.17081	0.165205	0.168245	0.161595	0.153805
Illumination – 50 %	0.3332	0.3898	0.3939	0.3271	0.3501	0.2919

Fig. 6. The range of the obtained values of the relative error in determining the size of objects for different levels of illumination of the scene

According to the data presented in the figure above, it can be seen that the proposed approach demonstrates high accuracy in determining the spatial position of objects for images illumination levels of 100 and 70 %. The class-averaged relative deviation values for scenes with illumination levels of 100 and 70 % are 0.1010 and 0.1624, respectively. With a decrease in illumination, a critical decrease in the quality of determining the spatial position of objects is observed, and at 50 % illumination level, the approach becomes unsuitable for practical application. The class-averaged relative deviation for scenes with 50 % illumination is 0.3477. Based on the results obtained, it can be concluded that the developed approach demonstrates a fairly high accuracy in determining the spatial position of objects for images with illumination levels of 100 and 70 %.

Based on the results of the experimental assessment, it can be concluded that the proposed solution allows successfully determining the size and spatial position of objects of various classes from images with illumination levels of 100 and 70 %.

Conclusion

Based on the results of approbation of the proposed approach on a test set of 12 000 images, the developed solution showed an acceptable level of accuracy in determining the dimensions and spatial position of objects for images with illumination levels of 100 and 70 %. The class-averaged values of the relative error in determining the object sizes for the corresponding image sets were 0.1449 and 0.3313, respectively, and the class-averaged values of the relative deviation in determining the spatial position of objects for these image subsets were 0.1010 and 0.1624. To increase the accuracy of the developed approach, the use of object masks should be tested, which can have a positive effect on the accuracy of determining the target parameters. The proposed approach can be used to improve existing systems for detecting and classifying objects or implemented as an independent system in various domestic or industrial premises. In addition, the developed solution may be of interest when combined with existing algorithms for detecting objects and positioning them in space. It should be noted that the Mask R-CNN neural network model [6] used within this approach does not provide an acceptable level of accuracy in detecting and classifying objects. Thus, further work will be aimed at upgrading the proposed solution through the use of more high-precision neural network models, as well as by combining the results of processing series of individual images.

References

[1] **Vatamaniuk I.V., Levonevskiy D.K., Malov D.A., Saveliev A.I.** Models and methods of user interaction with cyberphysical intellectual space. St. Petersburg: Limited Liability Company Publishing House Lan; 2019: 176. (In Russ.)

[2] **Levonevskiy D.K., Vatamaniuk I.V., Saveliev A.I., Denisov A.V.** Corporate information system of user service as a component of cyber-physical intellectual space. Journal of Instrument Engineering. 2016; 59(11):906–912. DOI:10.17586/0021-3454-2016-59-11-906-912. (In Russ.)

[3] **Levonevskiy D.K., Vatamaniuk I.V., Saveliev A.I.** Minos multimodal information and navigation cloud system for the corporate cyber-physical smart space. Software Engineering. 2017; 8(3):120–128. (In Russ.)

[4] **Keselman L., Iselin Woodfill J., Grunnet-Jepsen A., Bhowmik A.** Intel realsense stereoscopic depth cameras. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017: 1–10.

[5] **He K., Gkioxari G., Dollar P., Girshick R.** Mask R-CNN. Proceedings of the Seventh IEEE International Conference on Computer Vision. Venice, Italy. October 22–29, 2017. 2017: 2961–2969. DOI:10.1109/ICCV.2017.322.

[6] **Ammirato P., Berg A.C.** A Mask-RCNN baseline for probabilistic object detection. arXiv Preprint. 2019: 1908.03621.

[7] **Abadi M., Barham P., Chen J., Chen Z., Davis A., Dean J., Devin M., Ghemawat S., Irving G., Isard M., Kudlur M.** Tensorflow: a system for large-scale machine learning. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16). 2016: 265–283.

[8] **Qin Z., Zhang Z., Chen X., Wang C., Peng Y.** Fd-mobilenet: improved mobilenet with a fast downsampling strategy. 25th IEEE International Conference on Image Processing (ICIP). 2018: 1363–1367.

[9] **Girshick R.** Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision. 2015: 1440–1448.

[10] **Lin T.Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., Zitnick C.L.** Microsoft coco: common objects in context. European Conference on Computer Vision. 2014: 740–755.

[11] **Szegedy C., Ioffe S., Vanhoucke V., Alemi A.A.** Inception-v4, inception-resnet and the impact of residual connections on learning. Thirty-First AAAI Conference on Artificial Intelligence. February 4–9, 2017, San Francisco, California USA. DOI:10.1609/aaai.v31i1.11231.

[12] **Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Erhan D., Vanhoucke V., Rabinovich A.** Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015: 1–9. DOI:10.1109/CVPR.2015.7298594.

[13] **Fernald F.G.** Analysis of atmospheric lidar observations: some comments. Applied Optics. 1984; 23(5):652–653.

[14] **Zhang Z.** Microsoft kinect sensor and its effect. IEEE Multimedia. 2012; 9(2):4–10.

[15] **Iandola F.N., Han S., Moskewicz M.W., Ashraf K., Dally W.J., Keutzer K.** SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. arXiv Preprint. 2016: 1602.07360.

[16] **Simonyan K., Zisserman A.** Very deep convolutional networks for large-scale image recognition. arXiv Preprint. 2014: 409.1556. DOI:10.48550/arXiv.1409.1556.

[17] GOST R 55710-2013 Lighting of workplaces inside buildings. Standards and measurement methods: approved and put into effect by Order of the Federal Agency for Technical Regulation and Metrology dated November 8, 2013 No. 1364-st. Available at: `http://docs.cntd.ru/document/1200105707` (accessed April 24, 2020). (In Russ.)

[18] **Andujar D., Dorado J., Fernandez-Quintanilla C., Ribeiro A.** An approach to the use of depth cameras for weed volume estimation. Sensors. 2016; 16(7):972.

[19] **Cai Y., Li B., Jiao Z., Li H., Zeng X., Wang X.** Monocular 3d object detection with decoupled structured polygon estimation and height-guided depth estimation. In Proceedings of the AAAI Conference on Artificial Intelligence. 2020; 34(7):10478–10485.

[20] **Wang D., Li W., Liu X., Li N., Zhang C.** UAV environmental perception and autonomous obstacle avoidance: a deep learning and depth camera combined solution. Computers and Electronics in Agriculture. 2020; (175):105523.

[21] **Mueller F., Davis M., Bernard F., Sotnychenko O., Verschoor M., Otaduy M.A., Casas D., Theobalt C.** Real-time pose and shape reconstruction of two interacting hands with a single depth camera. ACM Transactions on Graphics. 2019; 38(4):1–13.

[22] **Li R., Liu Z., Tan J.** A survey on 3D hand pose estimation: cameras, methods, and datasets. Pattern Recognition. 2019; (93):251–272.

[23] **Laina I., Rupprecht C., Belagiannis V., Tombari F., Navab N.** Deeper depth prediction with fully convolutional residual networks. Fourth International Conference on 3D Vision (3DV). 2016: 239–248.

[24] **Mejia-Trujillo J.D., Castano-Pino Y.J., Navarro A., Arango-Paredes J.D., Rincon D., Valderrama J., Munoz B., Orozco J.L.** Kinect™ and Intel RealSense™ D435 comparison: a preliminary study for motion analysis. IEEE International Conference on Ehealth Networking, Application & Services (HealthCom). 2019; 41(2):607–615.

[25] **Triggs B.** Camera pose and calibration from 4 or 5 known 3d points. Proceedings of the Seventh IEEE International Conference on Computer Vision. 1999; (1):278–284.

Вычислительные технологии, 2024, том 29, № 2, с. 79–94. © ФИЦ ИВТ, 2024 ISSN 1560-7534
Computational Technologies, 2024, vol. 29, no. 2, pp. 79–94. © FRC ICT, 2024 eISSN 2313-691X

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

DOI:10.25743/ICT.2024.29.2.007

Подход к определению пространственного положения и размеров объектов сцены по изображению с использованием методов машинного обучения

М. А. Летенков, Е. О. Черских*

Санкт-Петербургский Федеральный исследовательский центр РАН, 199178, Санкт-Петербург, Россия
*Контактный автор: Черских Екатерина Олеговна, e-mail: `cherskikh.e@iias.spb.su`

Поступила 05 сентября 2022 г., доработана 05 июля 2023 г., принята в печать 20 июля 2023 г.

Аннотация

В контексте определения характеристик киберфизической среды одной из актуальных проблем является определение пространственных координат, местоположения и размеров объектов внешнего окружения. В рамках настоящего исследования для решения указанной проблемы был предложен подход к оценке пространственного положения и геометрических характеристик объектов окружающей среды по изображениям, основанный на объединении

результатов детектирования объектов с использованием нейросетевой модели Mask R-CNN, а также результатов восстановления карт глубины наблюдаемой сцены с помощью камеры RealSense D435.

В работе представлены авторские алгоритмы анализа геометрических свойств наблюдаемых на изображении областей сцены, а также комплементарные алгоритмы оценки размеров и положения объектов. Оценка качества разработанного подхода проводилась на основе тестового набора данных, включающего 12 000 изображений, сформированных при трех различных уровнях освещенности сцены: 100, 70 и 50 %. В рамках эксперимента найдены усредненные по классам значения относительной ошибки вычисления размеров объектов для соответствующих наборов изображений: 0.1449, 0.3313, 0.6332. Значения относительного отклонения при определении пространственного положения объектов по результатам тестирования составили 0.1010, 0.1624, 0.3477. Таким образом, предлагаемое решение позволяет успешно определять размеры и пространственное положение объектов различных классов по изображениям сцены с уровнями освещенности 100 и 70 %.

Ключевые слова: оценка размера объекта, оценка пространственного положения, обнаружение объекта, реконструкция карты глубины, R-CNN-маска, Intel RealSense.

Цитирование: Летенков М.А., Черских Е.О. Подход к определению пространственного положения и размеров объектов сцены по изображению с использованием методов машинного обучения. Вычислительные технологии. 2024; 29(2):79–94. DOI:10.25743/ICT.2024.29.2.007. (на английском)