Journal of Ambient Intelligence and Smart Environments 16 (2024) 541–566
DOI 10.3233/AIS-230193
IOS Press

Methods for volume inference of non-medical objects from images: A short review

Baticté Nabitchita [a], Norberto Jorge Gonçalves [b], Paulo Jorge Coelho [c,d], Luís Pimenta [b], Eftim Zdravevski [e], Petre Lameski [e], Mónica Costa [a], Paulo Alexandre Neves [f] and Ivan Miguel Pires [g,*]
[a] *R&D Unit in Digital Services, Applications, and Content, Polytechnic Institute of Castelo Branco, 6000-767 Castelo Branco, Portugal*
E-mails: tiquenquilin@outlook.com, monicac@ipcb.pt
[b] *Escola de Ciências e Tecnologia, University of Trás-os-Montes e Alto Douro, Quinta de Prados, 5001-801 Vila Real, Portugal*
E-mails: njg@utad.pt, al70827@alunos.utad.pt
[c] *Institute for Systems Engineering and Computers at Coimbra (INESC Coimbra), DEEC, Pólo II, 3030-290 Coimbra, Portugal*
E-mail: paulo.coelho@ipleiria.pt
[d] *Polytechnic of Leiria, 2411-901 Leiria, Portugal*
[e] *Faculty of Computer Science and Engineering, University Ss Cyril and Methodius, 1000 Skopje, Macedonia*
E-mails: eftim.zdravevski@finki.ukim.mk, petre.lameski@finki.ukim.mk
[f] *Superior School of Technology, Polytechnic Institute of Castelo Branco, 6000-767 Castelo Branco, Portugal*
E-mail: pneves@ipcb.pt
[g] *Instituto de Telecomunicações, Escola Superior de Tecnologia e Gestão de Águeda, Universidade de Aveiro, 3750-127 Águeda, Portugal*
E-mail: ivan.pires@lx.it.pt

Received 16 May 2023
Revised 19 November 2023

Abstract. Nowadays, the object's volume is essential for monitoring any scene. Technological equipment is evolving, and mobile devices and other devices embed high-resolution cameras. The high-resolution cameras open a window for different research studies, where the volume measurement is vital for different areas. This study aims to identify image processing techniques for measuring the object's volume. Thus, a systematic review was performed with a Natural Language Processing (NLP)-based framework for identifying studies between 2010 and 2023 related to the measurement of object volume. As a result of this search, this paper reviewed and analyzed 25 studies, verifying that different computer vision methods accurately handle object recognition. Additionally, an evaluation of the databases presented by the studies above is performed to consider further the design of a new approach to infer the volume of objects from an image.

Keywords: Volume, measurement, image processing, objects, systematic review

1. Introduction

In recent years, technology has greatly evolved in several areas, including the image processing field [8,62,69]. This expansion has been present in several disciplines, such as medicine [17,18,35,44], transmission and encoding

*Corresponding author. E-mail: ivan.pires@lx.it.pt.

1876-1364/$35.00 © 2024 – IOS Press. All rights reserved.

[23,56], pattern recognition [24,41,52], object recognition [55,64], and others. Image processing is a technique for applying various procedures to an image to improve or extract some relevant information [1]. It may be considered signal processing where the input is an image, and the output can be another image, features, or characteristics related to that image [72].

Several options exist to gather, process, and make the data available in terms of hardware to perform image acquisition [4,51,57]. For the scene capture, some conditions should be considered, such as the illumination setup (providing uniform, diffuse distribution of light), the camera quality (whose lens, resolution, and acquisition speed are adapted to the conditions required), the data processing location (that could be onboard or a server-based approach), and if a single camera or a multi-camera setup is more suitable [4,51,57].

Images captured from different devices can be used for various purposes, including measuring volume, height, width, perimeter, area, and other parameters [30,70]. Different techniques can be applied to extract these various features, including image segmentation, canny edge detection, AKAZE feature-point detection and matching, graph-based discriminant analysis, and spatial-spectral multiple manifold discriminant analysis [60,61]. The quality of the images allows the measurement of these parameters with high-quality [43]. Also, with the evolution of technology, these methods can be used to improve different measures, including color identification [59], the volume of box-type objects [34], volume and surface area of ellipsoidal agricultural products [58], and others. It can be helpful in different areas, including Ambient Assisted Living (meals, body parts, etc.) [21,47,53], automotive [5,27], remote monitoring (objects) [28,45], and others.

This study aims to perform a systematic review based on the search of scientific publications related to the measurement of object volume, published between January 2010 and October 2023 in the following scientific databases: IEEE Xplore, Elsevier, Springer, MDPI, and PubMed. It aims to review the existing approaches (methods) for assessing novel computational techniques for precisely estimating the volume of different non-medical objects from their images solely. Additionally, it seeks to investigate and assess the machine learning and advanced image processing algorithms that are used to estimate the volume of objects from various image types, as well as to collect information regarding frameworks that can be used to estimate the volume of a wide range of objects in a variety of settings and conditions. Also aims to evaluate the reported accuracy and dependability of the volume estimation methods and identifies potential applications as well as limitations of these techniques in real-world scenarios. The analysis of the studies' datasets is also performed to determine the datasets previously used (the objects considered, their dimensions, etc.) and evaluate their use to develop a new method to extract volume from 2D images for further application in the medical field to help doctors in remote monitoring, surgeries, and other purposes eventually.

The main contributions of this systematic review are clarifying which methods are applied to infer the objects' volume. Automatic volume measurement using images is a non-contact method that accurately determines the volume of objects in various industries. It offers non-contact measurement, speed, accuracy, cost-effectiveness, versatility, data integration, and enhanced capabilities with Artificial Intelligence. It is particularly useful for measuring fragile, soft, or hazardous materials. Image-based volume measurement systems can be adapted to measure a wide range of objects, making them versatile tools in various applications. Determining the various technological approaches is also critical to developing a new methodology for analyzing the results, of which this paper presents the most popular approaches and features extracted. Finally, an overview of future perspectives is presented.

2. Methods

2.1. Research questions

The main research questions of this systematic review are: (RQ1) Which methods can be used to measure the object's volume from images? (RQ2) Which are the types of objects previously used to measure the volume? (RQ3) What is the applicability of the size of the objects' volume with images?

2.2. Inclusion criteria

Different criteria conditions were defined for analyzing the studies on measuring objects' volume. These were: (1) Studies that focus on object volume estimation from images; (2) Studies using a mobile application and image

processing to object volume estimation; (3) Studies that present segmentation methods to identify objects in images; (4) Studies that are original research studies; (5) Studies that are only focused on images; (6) Studies that are not related to medical research studies; (7) Studies that are not related to robotic developments; (8) Studies that were published between 2010 and 2023; (9) Studies written in English.

2.3. Exclusion criteria

Articles are also excluded based on the following criteria: (1) studies that didn't report the object volume estimation with images; (2) studies that refer to the image processing techniques used; (3) studies that are literature reviews or surveys; (4) studies that are related to robotic developments; (5) studies that are related to the medical subject.

2.4. Search strategy

To find the different studies included in this review, we used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology to identify and process the studies related to object volume estimation from images published between 2010 and 2023. The Natural Language Processing (NLP) toolkit described in [79] was used to perform automatic searches in several electronic databases, including IEEE Xplore, Elsevier, Springer, MDPI, and PubMed. The keywords used for the search were: "object volume estimation from images" and "object volume estimation with mobile devices".

After the search, each study was analyzed by the authors for the evaluation of suitability with the defined criteria in Section 2.2, previously specified by the consent of all authors. The research was performed on 29 March 2023.

2.5. Extraction of study characteristics

After the criteriums analysis of the different studies, the extracted data were mapped in Table 1: year of publication, objects analyzed, purpose, features, methods, and outcomes. When the studies did not present some data, we contacted the corresponding authors to ask for the needed information. All the studies were analyzed to identify the methods for object volume estimation from images using mobile devices. The outcomes are composed of the following metrics available and the different studies.

2.5.1. Accuracy

The accuracy formula (equation (1)) is used to calculate the percentage of a model's total predictions that are accurate. It's especially important in classification issues when we want to know how often the model properly predicts discrete categories as the results.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \tag{1}$$

2.5.2. Precision

The precision formula (equation (2)) measures how well the model predicts the positive outcomes. It indicates the percentage of affirmative identifications that were true. This measure is especially crucial when a false positive comes at a high cost.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{2}$$

2.5.3. R^2

The R^2 formula (equation (3)) is a statistical metric that shows how much of the variation of a dependent variable in a regression model can be accounted for by one or more independent variables. In essence, it offers a measure of the model's fit quality. A higher R^2 value shows an improved match between the model and the data.

$$R^2 = 1 - \frac{Total\ Sum\ of\ Squares}{Sum\ of\ Squares\ of\ Residuals} \tag{3}$$

Table 1
Study analysis

Paper	Year of publication	Population	Purpose of the study	Features	Methods	Outcomes
Dalai *et al*. [13]	2023	NYUv2 [63]	Volume estimation of a rigid object from a single-view image	• Edge • Key points • Shape	• VGG-ResNet for depth analysis • Volume estimation through Hybrid3DU-GNet	• Accuracy (98.59%) • Precision (98.21%) • R-squared (98.2%) • Mean Absolute Percentage Error (6.1%) • Root Mean Squared Error (0.93)
Wang *et al*. [73]	2023	KITTI [22]	Monocular 3D object detection using Depth from Motion approach	• Depth maps after object detection	• Depth from Motion • Modified ResNet-34 with spatial pyramid pooling and feature upsampling	• KITTI average precision for 3D object detection evaluation ranging from 17.46 up to 29.27
Jeon and Heo [29]	2022	Scene Flow [39] KITTI-2015 [40]	Efficient resource-limited mobile MSFFNet for stereo matching, generation of multi-scale cost volumes, performing interlaced concatenation method to generate final cost volume	• Width • Height • Scale	• Multi-scale sequential feature fusion network (MDFFNet) • Adaptative cost volume filtering loss function	• End Point Error for Scene Flow dataset (1.01) • All KITTI benchmark (bg – 2.53%, fg – 4.99%, and All – 2.94%)
Yang *et al*. [77]	2021	Proprietary images	To calculate the volume of food on a plate using a single 2D image based on reference images with known volume and fine-tuning estimation	• Scale • Width • Height	• Two-step AI system • 1st step: Select a reference volume that best matches the image • 2nd step: Fine-tuned volumetric estimation	• Mean Relative Volumetric Error (mRVE) ranging from 11.6% to 20.1%
Kalpitha *et al.* [32]	2021	127,915 CAD models [75]	Handling occlusion in 3D object recognition	• Edges • Corners	• Model-based matching • Canny edge detector • Image Reflection • Region-based segmentation • 3D matching • 3D reconstruction	• Sigmoid activation function: 98.66% (accuracy) • ReLU activation function: 99.50% (accuracy) • Tanh activation function: 97.43% (accuracy)
Poply *et al.* [54]	2021	UNIMIB 2O16 [11]	Predicting the calorie contents of multiple-dish food items by taking their top-view images.	• Mass • Width • Height	• Convolutional Neural Networks • Object Detection • Semantic Segmentation	• 89.3% (mean average precision)

Table 1
(Continued)

Paper	Year of publication	Population	Purpose of the study	Features	Methods	Outcomes
Wittstruck *et al.* [74]	2021	Proprietary images	Methodology for high-resolution RGB data from Unmanned aerial vehicle	• RGB data	• Binary random forest • Gini Index	• 99% (correlation)
Brándi *et al.* [3]	2020	Proprietary dataset	Estimating object volume	• AKAZE features [2]	• AKAZE feature-point detection and matching • Bundle adjustment • Patch-based multi-view stereo • Poisson surface reconstruction	• >90% (accuracy)
Hadi *et al.* [26]	2020	• Cityscapes Dataset [12] • Driving Dataset [71] • WildDash [80]	3D Traffic Scene Reconstruction (3DTSR)	• RGB data	• Instance semantic segmentation method	• AP50: 43.9% (accuracy)
Suzuki *et al.* [66]	2020	Proprietary dataset	Estimating food volume with a point cloud processing method	• Semantic information • Shape • Width • Height	• LMedS method • Graph cut method	• 10.78% (average error rate)
Tomescu [68]	2020	Proprietary images	Food volume measurement	• Angles • Width • Height • Depth maps	• Convolutional neural networks • Model reconstruction • Point cloud fusion (Iterative Closest Point) • Point cloud to voxel grid • Volume calculation	• 70.3% (accuracy)
Gao *et al.* [20]	2019	UEC Food-256 dataset [33]	A novel method, named MUSEFood, for food volume estimation	• Width • Height	• Segmentation method	• 5.11% (relative error)
Lou *et al.* [38]	2019	• SUN RGB-D dataset [65] • NYUv2 RGB-D dataset [63]	Amodal 3D object detection	• Angles • Width • Height	• 3D-SSD	• SUN RGB-D: 50.7% (accuracy) • NYUv2 RGB-D: 39.7% (accuracy)

Table 1
(Continued)

Paper	Year of publication	Population	Purpose of the study	Features	Methods	Outcomes
Yogaswara *et al.* [78]	2019	Proprietary images	Using a computer vision approach, calculate the number of food calories automatically	• RGB data	• Mask Region-based Convolutional Neural Network (R-CNN)	• 89.41% (accuracy)
Dalai *et al.* [14]	2018	Proprietary images	Physical characteristics calculation	• Shape • Width • Height	• K-means clustering • Image segmentation	• 10.33% (average error)
Gao *et al.* [19]	2018	Proprietary images	Food volume measurement	• Width • Height	• Statistical Outlier Removal Filter • Point Cloud Completion • Convex Hull with Boundary Shrinking Property	• 15.21% (average error)
Lo *et al.* [36]	2018	Image set from Yale–CMU–Berkeley object set [6,7]	Food volume measurement	• Angles • Width • Height • Camera coordinates	• Neural Network architecture	• 6.875% (average error)
Parihar *et al.* [49]	2017	NHANES 1999–2000 dataset [46,50]	Estimate the dimensions of an object in a 2D image	• Haar features	• Haar Cascades algorithm • Volume estimation • 3D shape models • Regression analysis	• 93.69% (accuracy)
Yang *et al.* [76]	2017	• BR [67] dataset • UWAOR [41] dataset • UWA3M [42] dataset	Feature description for the 3D local shape in the presence of noise, varying mesh resolutions, clutter, and occlusion	• Triple orthogonal local depth images (TOLDI) representation	• 3D matching	• BR: 99.9% (accuracy) • UWAOR: 19.7% (accuracy) • UWA3M: 17.1% (accuracy)
Fang *et al.* [16]	2015	TADA project [15]	Food portion estimation	• Radius • Height • Width	• The Cylinder Model • Prism Model • Direct Linear Transform (DLT) method	• <6% (average error)
Grum *et al.* [25]	2014	Proprietary images	Identifying and correcting errors in the reconstructed 3D scene	• Shape • Width • Height • RGB data	• Voxel carving • Initial RBF • Updated RBF	• >75% (accuracy)

Table 1
(Continued)

Paper	Year of publication	Population	Purpose of the study	Features	Methods	Outcomes
Chen *et al.* [10]	2012	Proprietary images	Estimating food volume from a single-view 2D image	• RGB data	• Otsu's thresholding • 3D/2D model-to-image registration • Find contours	• Cuboid: 1.27% (average error) • Sphere: 3.38% (average error) • Half-sphere: 4.02% (average error) • Cylinder: 0.52% (average error)
Jia *et al.* [31]	2012	Proprietary images	Estimate the 3D location of a circular feature from a 2D image	• Shape • Camera coordinates • Distance • Length • Width • Height • Diameter	• "Point-clicking" method • "Wireframe-fitting" method	• 5.16% (average error)
Patz *et al.* [48]	2012	Proprietary images	Quantify the influence of the gray-value uncertainty	• RGB data	• Extension of the random walker segmentation	• <5% (average error)
Campbell *et al.* [9]	2010	Proprietary images	Segmentation of a rigid object in a sequence of images	• Shape • Width • Height • RGB data	• Volumetric graph-cut • Gaussian Mixture Model (GMM) • Volume estimation • Boundary estimation	• <5% (average error)

2.5.4. Mean absolute percentage error

Mean Absolute Percentage Error (equation (4)) is a metric used to evaluate the precision of a forecast or prediction. It divides the absolute value of the difference for each observation between the actual and projected values by the actual value. Next, the total number of observations (n) is divided by the sum of this result across all the observations. To turn the final amount into a percentage, multiply it by 100%. It represents the average percentage of forecast inaccuracy.

$$\textit{Mean Absolute Percentage Error} = \frac{100\%}{n} \times \sum_{i=1}^{n} \left| \frac{A_i - F_i}{A_i} \right| \tag{4}$$

2.5.5. Root mean squared error

The Root Mean Squared Error (equation (5)) is a commonly used metric to quantify the discrepancies between values predicted by a model and observed values. For each observation, the square of the difference between the expected and actual values is calculated first. After that, the total of these squared differences for every observation should be done. The mean of these squared differences is then calculated by dividing this amount by the total number of observations (n). Lastly, to get the RMSE, take the square root of this mean. With the advantage that greater mistakes are given proportionally more weight because of the squaring of the errors, the RMSE provides a measure of the error's size.

$$\textit{Root Mean Squared Error} = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} (P_i - O_i)^2} \tag{5}$$

2.5.6. End point error

The End Point Error (equation (6)) is used in domains such as computer vision, especially in optical flow estimates. It gauges the accuracy of the motion vector estimates compared to the ground reality. The difference between the estimated and actual values for the motion vector's horizontal and vertical components is first calculated. These discrepancies are then squared and added together, and the square root of this total is then calculated. The Euclidean distance between the estimated and genuine motion vectors is represented by a single scalar number that the EPE offers, which is an effective motion estimation accuracy measure.

$$\textit{End Point Error} = \sqrt{(u - \hat{u})^2 + (v - \hat{v})^2} \tag{6}$$

2.5.7. Mean relative volumetric error

Mean Relative Volumetric Error (equation (7)) is used in 3D modeling or medical imaging applications to evaluate the precision of volume measurements. It compares an object's measured volume using an imaging method or model to its actual or known volume. The absolute difference between the measured and real volumes is divided by the true volumes to get the relative error in volume measurement for each observation. These relative errors are then averaged across all data. The average departure of the measured volumes from their real values is reflected in this statistic, which gives a mean % error in volume measurements.

$$\textit{Mean Relative Volumetric Error} = \frac{1}{N} \times \sum_{i=1}^{n} \left| \frac{V_{measured,i} - V_{true,i}}{V_{true,i}} \right| \tag{7}$$

2.5.8. Average precision

Average Precision (equation (8)) is used in object identification and information retrieval to assess the precision of an object detection model or the caliber of results that a search algorithm returns. When working with unbalanced datasets, it is especially important to keep in mind that the positive class is much less common than the negative class. Essentially, it adds up the accuracy values at every rank where a pertinent document is found, and then it divides

this amount by the overall count of relevant documents. A greater Average Precision denotes better performance, and it measures the quality of the retrieval or detection across all ranks.

$$Average\ Precision = \frac{\sum_{i=1}^{n}(Precision(k) \times \mathrm{rel}(k))}{Number\ of\ relevant\ documents} \tag{8}$$

2.5.9. Mean average precision

In several domains, including computer vision, object identification, and information retrieval, the Average Precision measure is extended to provide Mean Average Precision (equation (9)). The accuracy across many queries or classes is summarized in a single chart. Mean Average Precision is the average of the Average Precision determined for every object class in the context of object detection.

$$Mean\ Average\ Precision = \frac{\sum_{q=1}^{Q}\left(\frac{\sum_{i=1}^{n}(Precision(k)\times \mathrm{rel}(k))}{Number\ of\ relevant\ documents}\right)_q}{Q} \tag{9}$$

2.5.10. Average error rate

The Average Error Rate (equation (10)) is used to evaluate a model or system's performance in classification tasks. The computation involves averaging the error rates among several classes or instances. The percentage of incorrectly categorized cases inside each class is known as the error rate, which is determined for each class separately and averaged across all classes. Since it guarantees that the error rates of smaller classes are given equal weight to those of bigger classes in the overall evaluation, this metric is especially helpful in situations where the classes have unequal distributions.

$$Average\ Error\ Rate = \frac{\sum_{i=1}^{N}\frac{Number\ of\ errors\ in\ class\ i}{Total\ Number\ of\ instances\ in\ class\ i}}{N} \tag{10}$$

2.5.11. Relative error

A measurement or estimate's accuracy about the real or true value is determined by calculating the Relative Error (equation (11)). The relative error indicates the extent of the inaccuracy concerning the real value and may be given as a fraction or a percentage. One divides the absolute value of the difference between the estimated and actual values by the true value. After performing this computation, a dimensionless number is produced that may be multiplied by 100% to get a percentage. This metric is very helpful for expressing measurement and prediction accuracy in disciplines like statistics, engineering, and physics.

$$Relative\ Error = \frac{|True\ Value - Estimated\ Value|}{True\ Value} \tag{11}$$

3. Results

As represented in Fig. 1, this study identified 12548 papers, including 800 duplicated in different searched databases, which were removed. An automation tool was used for the search, and 3803 studies were marked as ineligible by the device during the first filtering, which cannot be correctly accessed. Next, the tool performed the filtering by keywords in the main text, abstract, and title, resulting in the exclusion of 5602 papers. The remaining 2343 studies were filtered by the study type, where 81 review studies were not considered. The studies not directly related to the volume measurement were also removed, excluding 343 papers. A significant part of the studies was associated with the medical subject, which was not considered for this review, and 1775 studies were discarded. The next stage of the exclusion was associated with using video and other sensors to measure the volume that is out of the scope of this review, resulting in the exclusion of 119 studies. The full-text analysis was performed in the remaining 25 papers, included in the qualitative and quantitative syntheses.

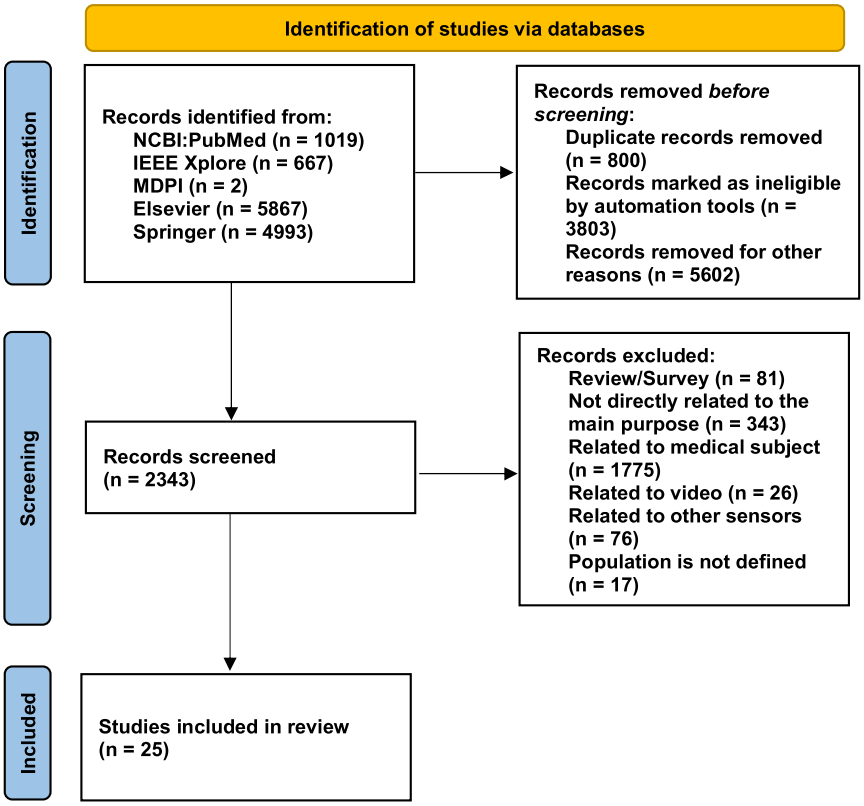

Fig. 1. Flow diagram of identification and inclusion of papers.

This review only summarizes the findings related to the volume measurement of objects in the different studies. The reader must follow the original research to get more relevant information. Based on Table 1, the analyzed studies were published between 2010 and 2023, distributed by two studies in 2023 (8%), one study in 2022 (4%), four studies in 2021 (16%), four studies in 2020 (16%), three studies in 2019 (12%), three studies in 2018 (12%), two studies in 2017 (8%), one study in 2015 (4%), one study in 2014 (4%), three studies in 2012 (12%), and one study in 2010 (4%). Regarding the datasets used, fourteen studies (56%) used publicly available datasets. Following the features extracted from the images, where fourteen studies extracted width (56%), fourteen studies extracted height (56%), seven studies extracted RGB data (28%), six studies extracted shapes (24%), three studies extracted angles (12%), two studies extracted camera coordinates (8%), two studies extracted edges (8%), two studies extracted scale (8%), two studies extracted depth maps (8%), one study extracted corners (4%), one study extracted mass (4%), one study extracted AKAZE features (4%), one study extracted semantic information (4%), one study extracted Haar features (4%), one study extracted Triple orthogonal local depth images (TOLDI) representation (4%), one study extracted radius (4%), one study extracted distance (4%), one study extracted length (4%), one study extracted key points (4%), and one study extracted diameter (4%). The methods used differ from the studies, and the results are reported below.

Table 2 depicts which databases are used for the evaluated works and some of their characteristics, such as their constitution and area of intervention/topic.

Based on the categories of the subjects, the following sections present the description of the studies

3.1. Detection of objects

Kalpitha *et al.* [32] developed a method for handling occlusion in a monochromatic image for 3D object detection. Left and right images of an image scene were captured using a binocular stereo vision system to infer 3D

Table 2
Evaluation of the databases

Database	Subject category	Subjects	Number of images
KITTI [22]	Objects	Car, vans, trucks, trams, pedestrian, cyclists	12,000 images with 40'000 objects
KITTI-2015 [40]	Objects	Objects found on the street while driving	400 highly dynamic scenes
Scene Flow [39]	Objects	Everyday objects, objects from the animation movie Monkaa, dynamic street scene from the viewpoint of a driving car	35,454 training and 4,370 test image pairs
Unnamed (private) [77]	Food	Food	14,892 (VFDL) and 13,694 (VFDS) images, and 1,500 images (GRFD) and 416 (IRFD) images
ModelNet [75]	Objects	Airplane; bathtub; bed; bench; bookshelf; bottle; bowl; car; chair; cone; cup; curtain; desk; door; dresser; flowerpot; glass box; guitar; keyboard; lamp; laptop; mantel; monitor; nightstand; person; piano; plant; radio; range hood; sink; sofa; stairs; stool; tent; table; toilet; tent; TV stand; vase; wardrobe; Xbox	127,915 images
UNIMIB 2O16 [11]	Food	Food	1,027 images
Unnamed (private) [74]	Food	Food	106 images
Unnamed (private) [3]	Objects	Celery root; box; angel statue	n/d
Cityscapes dataset [12]	Objects	Road; sidewalk; parking; rail track; person; rider; car; truck; bus; on rails; motorcycle; bicycle; caravan; trailer; building; wall; fence; guard rail; bridge; tunnel; pole; pole group; traffic sign; traffic light; vegetation; terrain; sky	25,000 images
Driving dataset [71]	Objects	Road; sidewalk; parking; rail track; person; rider; car; truck; bus; on rails; motorcycle; bicycle; caravan; trailer; building; wall; fence; guard rail; bridge; tunnel; pole; pole group; traffic sign; traffic light; vegetation; terrain; sky	10,004 images
WildDash [80]	Objects	Road; sidewalk; parking; rail track; person; rider; car; truck; bus; on rails; motorcycle; bicycle; caravan; trailer; building; wall; fence; guard rail; bridge; tunnel; pole; pole group; traffic sign; traffic light; vegetation; terrain; sky	n/d
Unnamed (private) [66]	Food	Food	n/d
Unnamed (private) [68]	Food	Food	80,000 images
UEC FOOD 256 [33]	Food	Food	31,395 images
SUN RGB-D dataset [65]	Objects	Bathroom; classroom; office; furniture store; bedroom; computer room; lecture theatre; library; study space; home office; discussion area; dining area; conference room; lab; corridor; kitchen; living room; bedroom	10,335 images
NYUv2 dataset [63]	Objects	Bedrooms; home offices; bathrooms; kitchens; offices; bookstore; libraries; playrooms; café; living rooms; reception rooms; dining rooms; study rooms; furniture stores	1,449 aligned depth and RGB images, with 464 new images from 3 different cities
Unnamed (private) [78]	Food	Food	2,500 images
Unnamed (private) [14]	Objects	Brick; PVC; alloy block	n/d
Unnamed (private) [19]	Food	Food	n/d

Table 2
(Continued)

Database	Subject category	Subjects	Number of images
Yale–CMU–Berkeley object set [6,7]	Objects	Chips can; master chef can; cracker box; sugar box; tomato soup can; mustard bottle; tuna fish can; pudding; box gelatin box; potted meat can; banana; strawberry; apple; lemon; peach; pear; orange; plum; pitcher base; bleach cleanser; windex bottle; wine glass; bowl; mug; sponge; skillet; skillet lid; plate; fork; spoon; knife; spatula; power drill; wood block; scissors; padlock; key; marker; adjustable wrench; Phillips screwdriver; flat screwdriver; plastic bolt; plastic nut; hammer; clamp; mini soccer ball; softball; baseball; tennis ball; racquetball; golf ball; chain; foam brick; dice; marbles; cups; colored wood blocks; nine hole peg test; toy airplane; lego duplo; timer; rubik's cube	99,400 images
NHANES 1999–2000 [46,50]	Food	Food	n/d
BR [67]	Objects/shapes	Models	n/d
UWAOR [41]	Objects/shapes	Models	n/d
UWA3M [42]	Objects/shapes	Models	75 images
TADA project [15]	Food	Food	998 images
Unnamed (private) [25]	Objects/shapes	Models	n/d
Unnamed (private) [10]	Food	Food	n/d
Unnamed (private) [31]	Food	Food	240 images
Unnamed (private) [48]	Objects/shapes	Models	75 images
Unnamed (private) [9]	Objects	Models	n/d

information. The comparator models for the computed 3D poses have been built using a virtual camera, and a feature detector has been suggested to find the local features of the object model. The projected picture unknown points were filled using the bilinear interpolation method throughout the image correction process. During the recognition phase, the regions of the rectified images are compared to the comparator models to identify the items.

Brándi *et al*. [3] demonstrate using an image-based volume estimation method. The camera positions are first estimated via bundle adjustment based on the feature points found in the images. The camera positions generate a dense point cloud, into which a convex hull and Poisson surface are fitted. Next, a better surface intersection is created. Finally, the finished surface's volume is calculated and converted to metric measurements using a reference item. The software system's implementation consists of a server and an Android client. The server oversees image processing, while the Android client application provides the user interface for taking photographs and creating a three-dimensional reconstruction of the scanned object.

The reconstruction of a 3D traffic scene was the subject of video footage captured by a dash camera in a moving vehicle [26], where 3DTSR attempts to generate a 3D traffic scene. The 3D traffic scene provides a new platform for applications, including self-driving cars, driving pattern analysis, and traffic accident analysis. The authors' approach used passive sensing, which identifies objects and their positions by using visual data. The COCO dataset and the Mask R-CNN model were utilized to conduct the studies. The investigations showed that it is possible to extract the shape and appearance of objects in the road scene using the suggested segmentation method.

The authors of [38] offered a faster and more accurate remedy to the amodal 3D object detection issue for indoor environments. The solution is obtained via a novel neural network structure that receives two RGB-D images as input and generates orientated 3D bounding boxes. A group of 3D anchor boxes is linked to each place on the prediction layers to address the shape variance of various objects explicitly. The sizes of these boxes vary. The category scores for 3D anchor boxes are generated during testing with multiple placements, sizes, and orientations, leading to final detections using non-maximum suppression. SUN RGB-D and NYUV2 datasets have both been utilized. The results showed that the suggested algorithm is the first 3D detector that works on challenging datasets in almost real-time with comparable performance to other methods.

In [14], the authors employed a k-means clustering technique. The weight was approximated by multiplying the volume calculated by the density after the volume was determined using the model-based area-volume mapping. The segmentation approach, which is based on GDIS, is very accurate and precise. Using a segmentation method based on a grid structure, the ROI (Region of Interest) has been located, and the object's volume has been calculated using this region. The object's volume has been calculated using the approximate height, width, and length that were measured using 3D-based modeling. The study has precision for various things thanks to its design performance.

In [36] work make it possible to estimate the volume of food by obtaining a single-depth photograph from any practical viewing angle. The scientists developed a view synthesis technique based on deep learning to reconstruct 3D point clouds of food items and estimate the volume from a single depth shot. A separate neural network is built to predict a depth image from the opposite viewing angle using a depth image from one viewing angle as a predictor. The whole 3D point cloud map is recreated using the suggested point cloud completion and Iterative Closest Point (ICP) methods by combining the initial data points with the synthesized points of the object items. A database of depth photographs of food object items captured from different viewing angles was created using an image rendering technique, and it was utilized to assess the proposed neural network. The authors estimated the volume of the synthesized 3D point cloud to be equal to the actual volume of the individual object components, allowing them to evaluate the procedure.

In [9] the authors suggested a method for automatically identifying a rigid object's segmentation from a collection of photographs calibrated for camera posture and intrinsic characteristics. Instead of requiring human interaction, the strategy depends on the camera being fixed on interest throughout the process. A graph-cut optimization produces the segmentation of the 3D space that is globally optimal. This segmentation yields a more accurate color model, extracted and repeated until convergence. It suffices to indicate what needs to be segmented and to start an autonomous segmentation process by applying the fixation constraint, which requires that the object of interest be centered in the image. Then, it was found that compared to independent 2D segmentations, a single 3D segmentation significantly improves silhouette quality.

3.2. Detection of objects and shapes

The work presented in [29] focuses on deploying an approach in resource-limited mobile environments through a multi-scale stereo-matching network. The authors present the MSFFNet, a multi-scale stereo matching network that generates a cost volume by connecting multi-scale SFF modules. Then, an adaptative cost-volume-filtering (ACVF) loss function that directly supervises the cost volume is applied. This second step performs two kinds of filtering using the probability distribution generated by the ground truth disparity map and the one estimated from the teacher network. Several results are given with the datasets Scene Flow and KITTI (both 2012 and 2015 versions), comparing with several approaches such as PSMNet, GANet-15, and DispNetC, among others. The End Point Error (EPE) evaluation metric was used for Scene Flow, achieving 1.01. In contrast, the KITTI benchmark achieved 22.53%, 4.99%, and 2.94% for all benchmarks in the background, foreground, and object sections.

Effective volume estimation on single-view objects through hybrid U-Net methodology is the focus of [13]. Using a mean-median filtering approach, the single-view object images are pre-processed. The feature extraction phase focuses on depth and shape analysis. 3D image reconstruction is performed using a hybrid 3DU-GNet framework, leading to volume estimation that can be applied to regular and irregular objects. The authors explain the operation of the hybrid network and mathematical formulations. Performance is analyzed against different existing approaches showing promising results, such as an accuracy of 98.59%, a precision of 98.21%, a Mean Absolute Percentage Error of 6.1%, and a Root Mean Squared Error of 0.93.

Authors of [73] present an integrated framework capable of estimating depth and detection of 3D objects from consecutive-frame images. The work uses the KITTI test set and metrics to evaluate, together with detailed analysis with ablation studies. Using a geometry-aware cost volume to establish the stereo correspondence, the Depth from Motion framework lifts 2D image features to the 3D space, enabling the detection of objects. To achieve the desired 3D detection, the input e closed-loop regularization images are fed to a stereo matching and view transformation. This step performs 2D detection with a ResNet-34 neural network, employs a 2.5D backbone, and performs depth estimation. The result is fed to a final step – 3D backbone and 3D head – that enables the final 3D detection of objects through a voxel-based approach. The paper clearly presents the results as superior to others using the same

KITTI benchmark and test set, with a KITTI average precision for 3D object detection evaluation ranging from 17.46 to 29.27.

For local shape description, the authors of [76] employed a triple orthogonal local depth image technique. It combines a triple orthogonal local depth image (TOLDI) format with a local reference frame (LRF). The proposed LRF differs from many earlier ones in that the z-axis is derived using the key normal points, and the x-axis is calculated by averaging the weighted projection vectors of the radius neighbors. Then, to provide TOLDI feature descriptors, three local depth images (LDI) obtained from three orthogonal view planes in the LRF are concatenated into feature vectors. Using publicly accessible datasets that cover three critical surface matching scenarios – shape retrieval, object recognition, and 3D registration – the performance of the TOLDI approach is carefully evaluated. The studies revealed the efficacy of the suggested approach.

Grum *et al*. [25] developed a technique for modeling a scene with several 3D objects. The suggested method's initialization is a rough 3D model of the scene created from the provided set of multi-view images. The first approach adjusts the positioning of the 3D patches in a location after determining the discrepancy between two scene projections in photos. Next, the technique "shape-from-contours" identifies differences between 3D object projections and their corresponding contours segmented from photographs. Object outlines are determined by using both supervised and unsupervised segmentation.

The study presented by [48] aimed to determine how random walker segmentation gray-value uncertainty affected the result. The edge weights in a weighted network created by random walker segmentation are based on the gradient of the image between the pixels. The developed technique identifies regions where uncertain pixel values substantially impact the segmentation result. As a result, it evaluates the segmentation accuracy and allows the estimation of the probability density function for the volume of the segmented object.

3.3. Detection of food

The work presented in [77] presents a human-mimetic estimation of the volume of food on a plate using a single RGB image. The authors present an AI system to mimic dietitians' reasoning that typically uses common-use objects, such as spoons and cups, among others, to quantify food. As a result, the authors defined several volumes of known value and created a virtual dataset with computer simulation – VFDL and VFDS. They define several classes based on the pre-defined volume and wrap the food part with food. The dataset is used for training and testing, with a second dataset of real images for validation – GRFD and IRFD. While GRFD is created through image capture on a controlled scenario with a fixed camera, IRFD comprises images taken with a mobile phone. The AI system comprises two steps – the first one outputs a vector of probabilities of the food volume concerning the predefined volumes. In contrast, the second implements an inner product of the probability vector and the reference volume vector. Results seem promising, considering that volume is calculated from a single 2D RGB image, with a Mean Relative Volumetric Error (mRVE) ranging from 11.6% to 20.1%.

In [54], the authors suggested utilizing computer vision and deep learning to determine the calorie count of each dish using top-view images of several dishes. The system performs an advanced image segmentation process that replicates instance segmentation, in which each pixel is identified from the instance of objects for each object detected in an image. This is done using convolutional neural networks (CNNs). A calorie table lookup is used to get the estimated calories for food goods using the estimated volume, mass, and other previously known characteristics. At the time of evaluation, the system had a mean average precision (mAP) of 89.30% for object detection and a percentage accuracy of 93.06% for calorie prediction.

The authors of [74] created an uncrewed aerial vehicle (UAV) data analysis system, which was applied to a field of Hokkaido pumpkins in northwest Germany. The methodology, implemented in Python, involved a few steps, such as image pre-processing, pixel-based image classification, classification post-processing for single fruit detection, and fruit size and weight quantification. According to the results, the field sample had a 95% detection rate, a 5% error rate, and a reliable prediction of volume and weight with Pearson's correlation values of 0.83 and 0.84, respectively.

In [66], a method of processing a point cloud has been put forth for more precise estimation. Point clouds, collections of coordinate points on objects, can be used to analyze three-dimensional things. The authors have developed a point cloud method to build the dish space and identify dishes on the dining table. There is yet another way to figure out how much food is in the dish space. The results found low error rates, ranging from 4% to 16%.

Tomescu *et al*. [68] suggested a method for identifying different food items and their mass on mobile devices using only the phone camera, called FoRConvD (Food Recognition using Convolutional Neural Networks and Depth Maps). Volume estimation and food type identification are the two main components of the approach. The type of food is determined using EfficientNet, a convolutional neural network model suitable for mobile devices. A method for measuring the volume called "depth map fusion" involves creating a 3D model of the object using various images collected from multiple angles. The experimental results empirically show that the method proposed for estimating the amount of food is reliable and accurate, with a volume overestimation of 0% to 10% depending on the object's shape.

In [20], the authors recommended using the MUSEFood method to determine food volume. MUSEFood uses the camera to take pictures of the food; it does not require training images to understand how much of it there is, nor does it require a background reference item for the photos to be shot. Using the microphone and speaker in a noisy environment, MUSEFood accurately estimates the vertical distance from the camera to the meal and scales the food in the image to reflect its actual size. MUSEFood outperforms other methods and dramatically speeds up the assessment of food volume, according to trials on real foods.

Yogaswara *et al*. [78] created a system that employs a computer vision method and the Deep Learning Mask Region-based Convolutional Neural Network (R-CNN) technology to automatically calculate the number of calories in meals based on the size of the food volume. The segmentation technique uses the instance-aware semantic segmentation approach to identify each pixel from an instance of an object for each object recognized in a food image. This model will quickly identify each unique food object to calculate the precise quantity of calories for each food object inside one class. The results of this study are likely to help people learn about the number of dietary calories concerning their bodies' calorie requirements, with a mean average precision (mAP) level of 89.4% and a percentage accuracy in calories estimated at 97.48%.

The authors of [19] presented a way for measuring food volume to determine the consumers' daily nutrient consumption accurately. The method relies on a simultaneous localization and mapping technique, a modified convex hull algorithm, and a 3D mesh object reconstruction method (SLAM). The feasibility of wearing monocular SLAM algorithms for continuous meal volume monitoring was examined in this study. A sparse map is produced using SLAM after the camera has taken photographs of the food, and a 3D mesh object has been created using the multiple convex hull approach. The mesh object can then be used to determine the volume of the target item. It has been evaluated in studies to assess the viability and accuracy of the suggested algorithm for calculating meal volume.

Parihar *et al*. [49] demonstrated a technique for calculating an object's dimensions from a user-provided 2D image. It performs three tasks: object recognition using the Haar Cascades technique, dimension computation using a reference object and a conversion factor, volume estimation using conventional 3D shape models for ordinary things, and a human body prediction model based on regression analysis. An accuracy of 93.69% was attained in the final value comparison.

To determine how much energy was consumed during a meal, the authors of [16] provided a method for estimating the amount of a serving of food based on a single-view food image. This work has established a system for estimating meal quantities without requiring manual parameter modification. Even though single-view 3D scene reconstruction is typically an ill-posed problem, the introduction of geometric models, such as the shape of the container, can aid in partially recovering the 3D properties of the food items in the image. The estimated 3D properties of each food item and a scene reference object can be used to determine the volume of each food item in the picture. The weight of each food can then be estimated using the food's density. Assuming precise segmentation and food classification, the studies could calculate the energy in an image of a meal with an error of less than 6%.

In [10], the authors devised a 3D/2D model-to-image registration technique for estimating meal volume from a single-view 2D image containing a reference object. Otsu's thresholding and morphological techniques are used first to distinguish the food from the backdrop image. The user-selected 3D shape model is then used to calculate the meal volume. The model's placement, orientation, and scale are all optimized using a model-to-image registration process. The circular plate of the image is fitted, and its spatial information provides a set of restrictions for resolving the registration problem. Experimental results using uniform test objects and accurately designed food models with known volumes showed how effective the strategies were.

The authors of [31] revealed a technique for determining a circular feature's 3D location from a 2D image. In the past, a meal container, such as a circular plate, was used as a necessary reference rather than a general reference, like

a checkerboard card. In this study, a mathematical model for the system – which consists of a camera and a circular item in three dimensions – is constructed, and the food volume is calculated using this model. The tests showed that in 224 out of 240 photographs of various real objects and food replicas, the relative inaccuracy of the image-based volume prediction was less than 10%.

3.4. An overview of evaluation metrics

All research work presented in this document provides validation through evaluation metrics. Moreover, it is also possible to find a comparison with different previous approaches to highlight the contribution. Some works also present ablation analysis, focusing on the performance of a given aspect or feature of the network. Nevertheless, matching results between the different research works presented is typically very difficult since they tend to utilize different evaluation metrics. One of the efforts to homogenize comes in the form of a dataset. The KITTI dataset is just not a dataset. It has a collection of benchmarks that some works [29,73] take advantage to perform the evaluation.

KITTI [22] benchmarks in the 2012 version evaluate percentages of erroneous pixels and average end point errors for both non-occluded (Nocc) and all (All) pixels reported. For the KITTI-2015 [40] version, the disparity outliers are evaluated for background (bg), foreground (fg), and all pixels on both non-occluded and all pixels.

Nevertheless, the focus is topically given to some loss or matching evaluation metric that analyzes the difference between a given object's calculated volume and the real volume. The EPE (End Point Error) is one of such metrics that measures the average disparity error in pixels of the estimated disparity map.

Presenting the evaluation metrics mathematical expressions here is out of this paper's scope, so the reader is invited to find such information in the referred research works.

4. Discussion

4.1. Interpretation of the results

Measuring the objects' volume may power the development of different systems in different life areas. The identification of objects can handle the identification of the objects used by people and the size of their objects. However, currently, there are different studies related to lifestyles, where identifying the number of aliments is essential to promote different kinds of diets. Only one of the analyzed studies worked with images related to body parts. A set of possibilities is opened with the correct measurement of the volume.

From the analysis of the twenty-five studies, it is observed that a total of thirty-two different databases are used, of which thirteen (about 41%) are fully proprietary and private [3,9,10,14,19,25,31,48,66,68,74,77,78]. There are also two works that, in their composition, apply their databases together with other publicly available [38,49], corresponding to about 6% of the databases. The remaining are exclusively from public databases, used in thirteen studies [13,16,20,26,29,32,36,38,49,54,66,68,73,76], and correspond to about 41% of the databases.

From another perspective and considering the similarity of themes in the databases, it is observed that about 44% of these are applied to food-related studies. In general, these studies [10,16,19,20,31,49,54,66,68,74,77,78] aim for health monitoring and management (e.g., calorie intake) through measurable information such as food volume. In a significant number, with about 61% of the databases involving nineteen studies [3,9,13,14,25,26,29,32,36,38, 48,49,73,76], some seek to identify and reconstruct objects or different shapes. Finally, a database related to the measurements of human parts [49] is applied together with another [46,50], also for measuring objects.

In terms of the description of the databases, twelve (about 39%) are not described, or it was impossible to determine their dimensions and constitution [3,9,10,14,19,25,38,41,49,66,67,80]. Of the remaining nineteen databases, about 47% of these have less than 2,500 images, about 16% are constituted by more or equal to 80,000 images [6,7,68,75], and the remaining 26% have numbers between the previously mentioned ranges [12,12,22,33,39,65,71,77].

4.2. Validity and reliability

Analyzing the studies that present quantitative metrics, the analysis that can be carried out is shown below, considering that this is somewhat abusive in the sense that possible performance comparisons between different studies should be carried out on the same databases to obtain a univocal evaluation of the methods. Other reviews have been performed, but they are mainly related to only one type of object. This paper intends to present a multivariate review of different kinds of objects. In contrast, the authors of [37] only reviewed the measurement of the volume of dietary objects.

Considering the previous issues, it was verified that the authors of [54] and [78] have the same objective to perform calorie count from food volume. The comparison metrics are the same: mean average precision (mAP) and accuracy for calorie prediction. In Poply *et al*. [54], the mAP value is 89.30%, and the calorie prediction percentage is 93.06%, against the 89.40% and 97.48%, respectively, presented by Yogaswara *et al*. [78].

Considering other measurements, the study [74] presents a 5% error rate (with a 95% detection rate) for field samples for UAV images. Suzuki *et al*. [66], for point cloud detection of dishes on the dining table, presented error rates ranging from 4% to 16%. Finally, and also for determining the circular features of the plates, in [31], the authors revealed that the image-based volume inaccuracy prediction was less than 10%.

As previously mentioned, although the formal comparison is not entirely scientifically sound, a brief analysis can infer from the magnitude of the metrics presented the relative quality of the studies.

4.3. Comparison of the different studies analyzed

For the comparison of the studies, Table 3 presents the primary outcomes and limitations of the different studies to further develop a new system for accurately measuring the object's volume. Finally, a new method combining the best results from previous studies is urged.

4.4. Final remarks

This systematic review presented an overview of the solutions for measuring the volume of objects based on image processing techniques. It is intended to develop an application for automatically measuring the object volume based on a set of pictures related to a scene. This research is a preliminary stage about the volume of objects.

The main findings from the 25 studies identified by this systematic review are as follows. Concerning RQ1, "Which methods can be used to measure the object's volume with images?", we conclude that the approaches are broad. Typically for object detection and localization on the scene, various techniques can be followed, and methods based on classical computer-vision approaches, such as model matching, feature point extraction, image segmentation, and others. In terms of 3D reconstruction and 3D matching, most studies are based on CNNs, and Deep Learning approaches for data, such as RGB or RGB-D images and point clouds.

Regarding RQ2, "Which are the types of objects previously used for the measurement of the volume?", most studies focus on finding and measuring food-related shapes. From these shapes, the food's volume in the scene is inferred so that food or dietary assessment regarding calorie intake, etc., can be produced. From another perspective, the studies that are not food-related present case studies for different types of objects and different proposes.

Finally, related to RQ3, "What is the applicability of the measurement of the objects' volume with images?", we identified that these approaches are helpful for the identification of the quantity of ingested meals, monitoring of objects (including their positioning), and the measurement of body parts. There are still many challenges in this process, so that those methods will be improved with the performance of more research studies. The creation of new datasets will promote the comparison of the different results.

5. Conclusions

This article has systematically reviewed measurements of different object volumes from 2D images. A total of 25 studies were considered relevant based on the inclusion criteria, meaning this area is appealing to research. Other

Table 3
Study outcomes and limitations

Paper	Main outcomes	Limitations
Dalai *et al.* [13]	The authors created a method for volume estimation of a rigid object from a single view object using deep learning based on hybrid U-Net	N/D
Wang *et al.* [73]	The authors proposed a framework for monocular 3D detection from videos	Stereo estimation of moving objects Complex framework and lack of generalization
Jeon and Heo [29]	The study presents the development of a network that can be used in resource-limited mobile environments for stereo-matching	Accuracy is relatively lower when compared to some heavy 3D convolution-based networks
Yang *et al.* [77]	The authors created several datasets for training, test and validation. Also, the proposed method calculates the food volume from a single RGB image, being able to normalize the plate size.	The results are inferior to other approaches that use more information, namely multiple-view approaches.
Kalpitha *et al.* [32]	The authors present a method that considers occlusion when recognizing 3D objects. Binocular stereo matching has been used for 3D object recognition to differentiate between items at various distances and to enhance depth perception. During the object recognition process, comparator models in different 3D poses are employed to match the instances of the objects. Bilinear interpolation was utilized to fix the image and anticipate the concealed or occluded locations. A region-splitting technique has been introduced for the region segmentation procedure. Pyramid modeling has effectively matched the repaired photos with the comparator models. A 3D reconstruction of the recognized output was used for validation.	N/D
Poply *et al.* [54]	The study provides a system that uses recent advances in deep learning-based object identification and semantic segmentation to implement our own "improved" pseudo-semantic segmentation procedure. Since the segmentations created during the procedure are used to carry out the calorie predictions, semantic segmentation proves to be helpful in this situation.	Calculating dietary calories is a difficult task. Even the best computer vision system, in this case, would not be able to capture what is inside a food item. However, information can be partially captured by any computer vision system. Depending on their specific ingredients, complex cuisine products like wraps, burgers, burritos, etc., can have a wide range of calories.
Wittstruck *et al.* [74]	The outcomes demonstrated that most of the fruits could be correctly recognized. The picture data might be used to calculate the volumes and weights of the pumpkins with great accuracy, enabling more focused pre-harvest commercialization plans for farmers. Produce producers might strengthen their sales discussions with greater understanding of categorized sales volume as most food merchants need homogenous lots within specified size or weight classes.	If pumpkins are growing vertically in the field, another restriction on the prediction is given since the height of the fruit cannot be determined from the top view image. This could only be seen for a tiny portion of the harvested pumpkins, though.
Brándi *et al.* [3]	Based on modern reconstruction techniques, the model can estimate the volume of certain items with at least 90% accuracy if the pertinent criteria are met. Additionally, a program helps the user capture the photographs, connects to the centralized server, updates the user in real-time on the image processing progress, and then displays the estimated volume next to the rebuilt item.	Algorithms for object recognition can help with segmentation enhancement. It is feasible to locate the camera on a metric scale and exclude the reference item using sensor fusion techniques (Extended Kalman filter, Visual inertial odometry). However, these methods could only be applied if their estimate is accurate enough because the scaling problem is weakly conditioned. The final figures would be significantly impacted by the detection mistake. The reference item may be bypassed in a closed system where numerous calibrated cameras take photographs in predetermined locations. The application may provide a more exciting user experience by re-projecting the reconstructed model using the Aruco-marker.

Table 3
(Continued)

Paper	Main outcomes	Limitations
Hadi *et al.* [26]	The authors demonstrated that the X101-FPN model is the most practical to incorporate into the suggested 3D Traffic Scene Reconstruction job to retrieve the shape and appearance of typical road scene obstacles and execute instance semantic segmentation on Mask R-CNN.	Due to hardware restrictions, the size of training must be limited.
Suzuki *et al.* [66]	By identifying the dish, the authors suggested a point cloud processing technique for accurately measuring the amount of food. The end of a meal can be determined using the proposed approach.	There is still room for improvement in the estimation error rate and stability. Additionally, the technique may be enhanced to estimate the volume of food even when the dining table is cluttered with other items.
Tomescu [68]	The FoRConvD may be used to quickly and easily assess the kind and mass of things connected to food by using the phone's camera. The two essential elements of FoRConvD were food type detection and volume estimation. The foods were categorized into the various categories using the EfficientNet architecture. The depth data were converted directly to point clouds before ICP was used to build the fused model for volume estimate. The volume was then determined by adding the finite surface elements, dividing the result by the appropriate height, and multiplying the result. The research' findings revealed an error in the volume estimates that varied from 0% to +10%.	The volume was calculated using motion structure, combining depth maps and pictures from multiple perspectives to create a 3D model. Methods for model reconstruction and volume computation must consider subsequent tracking optimization. Additionally, a density database must be developed to determine the mass. The technique might also be modified to account for numerous objects at once.
Gao *et al.* [20]	MUSEFood was created to make meal volume estimations using data from several smartphone sensors. MUSEFood uses multi-task learning architectures to exploit FCN and take advantage of information about the geometry of food containers, leading to more accurate and speedy food image segmentation. The authors use the MLS range instead of reference items, which improves user convenience and boosts the precision of meal volume estimation. MUSEFood passes muster in our testing for robustness and flexibility. The shape of the various food containers has no significant impact on the results of the food volume estimation. Because MUSEFood can manage food without traditional shapes, our models are relevant to a more extensive range of foods.	Even though some devices can estimate target distance directly, smartphones with ToF cameras are not very prevalent. If consumers are required to acquire specialist equipment, costs will inevitably increase. Furthermore, these ToF cameras'-built interfaces are not accessible to developers. After gathering the meal volume, the authors may easily estimate the calories and nutrients ingested using a food nutrients database.
Lou *et al.* [38]	For the purpose of detecting amodal 3D objects in indoor settings, the authors introduced an end-to-end neural network. The model is designed to utilize the complementary information in depth and RGB images. As a result, a hierarchical fusion structure conducts 3D bounding box regression and object categorization by combining attributes from diverse input data sources. The composition effectively preserves the finer features and the background information of the scene photographs. Experiments on publicly available datasets show that the method significantly outperforms state-of-the-art approaches in terms of accuracy and processing efficiency. It can enrich the field and application of 3D object identification.	N/D

Table 3
(Continued)

Paper	Main outcomes	Limitations
Yogaswara *et al.* [78]	The authors used the Mask Region-based Convolutional Neural Network to create a computer vision system that can determine food's calorie content based on food volume and mass (R-CNN). Because Mask R-CNN uses ResNet101 as its backbone model, it can compute a pixel-wise mask for every item in the picture. The authors concluded that it could be used to calculate food calories. Furthermore, the algorithm could differentiate between every instance of the same item. Users may more easily determine the number of food calories automatically thanks to the segmentation model, which can be utilized on online apps and Android mobile applications.	Another experiment must be conducted to determine which foods have a convex or concave structure.
Dalai *et al.* [14]	The suggested approach for estimating an object's weight from 2D photos is based on image processing. Getting a more exact estimation of each item's weight enables more precise grading. Fast speed grading and packaging cannot be done at high-speed using conventional weighing techniques. The authors illustrated a method for weight estimate using image analysis that is accurate and quick. The volume of each item is estimated using two perpendicular views, and the weight of each item is then calculated using the object's predetermined density. The proposed weight calculating technique provides results that are 90% accurate.	The recommended system might perform better if it used a more precise shape recognition algorithm that could recognize an object's 3D shape. Therefore, by applying the correct volume calculation technique, the volume may be calculated from a single image, yielding accurate results for both the volume and weight. The recommended method will, however, result in erroneous estimates when the thing is tiny, expensive metal, etc. Because it is hard to identify whether an object is hollow from a 2D representation, such as in an empty soft drink can, it also has problems with hollow things. Configurations involving dynamic cameras and fixed objects, or vice versa, may provide extra difficulties.
Gao *et al.* [19]	The proposed approach demonstrates the viability and accuracy of measuring food consumption continuously. The proposed technique may achieve an overall accuracy of 83% using the statistical outlier filter, point completion method, and multiple convex hull algorithm.	The average percentage error is around 20%.
Lo *et al.* [36]	The authors showed how the suggested network architecture and point cloud completion algorithms could implicitly learn the 3D structures of different shapes and restore the occluded section of food items to enable better volume estimation. The results demonstrate that the suggested approach outperforms other strategies described in earlier research, achieving accuracy in volume estimate of up to 93% using 3D food models from the Yale–CMU–Berkeley object collection. Overall, the authors discovered that combining several methodologies might be one of the viable fixes for tricky nutritional assessment problems. Dietary evaluation using images will undoubtedly be necessary for tracking health.	More research is necessary to assess the effectiveness of the algorithms using real-world scenarios. Even though the suggested model can currently only handle some food items from hidden viewing angles, significant advancements have been made using the current methodology (e.g., the model-based approach). In addition, a more extensive 3D model database is being developed to train the network and maximize the generalization potential of deep learning. The models should subsequently be capable of handling other geometric forms or even unidentified foods that are not included in the training dataset with enough training data.
Parihar *et al.* [49]	The suggested model uses a Haar cascade classifier to identify objects in images to extract key patterns describing the object. Two images of an object were used to determine its real size, with a one-time calibration process for each view. Different techniques for estimating regular and irregular objects' volume and weight are explained.	The suggested system's performance can be enhanced by utilizing a better shape identification technique that can identify an object's actual 3D shape.

Table 3
(Continued)

Paper	Main outcomes	Limitations
Yang *et al.* [76]	The keypoint's normal and the radius neighbors' weighted projection vectors were calculated by the authors to create the LRF. The eigenvectors' sign ambiguity problem is eliminated by the suggested approach. The crucial x-axis is calculated using the suggested approach, which makes use of all locations. To create balanced resilience to noise, different mesh resolutions, clutter, and occlusion, it gives them weights. As a result, a repeatable and reliable LRF is connected to the TOLDI description. TOLDI collects detailed spatial and geometric data from several viewpoints of the immediate surface. Furthermore, there is no need for complicated preprocessing because the TOLDI feature can be retrieved from the originally scanned point clouds. The findings show that our LRF is reliable and reproducible against a range of annoyances.	The authors explicitly describe the local depth information using all the pixel values in one LDI, resulting in a reasonably high-dimensional descriptor. We're excited to come up with a smaller LDI feature representation. The Microsoft Kinect device, stereo sensors, and structure from motion systems are just a few of the new low-cost tools that have been created that can also capture the texture of 3D objects. When the 3D models show low geometric features but abundant photometric signals, integrating RGB information into the TOLDI description might be advantageous. The alternative is to include the suggested LRF and TOLDI description in specific application algorithms, including surface registration and 3D object identification.
Fang *et al.* [16]	The authors suggested a technique for estimating the size of a meal portion from a single-view photograph. The method can automatically calculate volume utilizing the geometric contextual data from the scene instead of depending on manual initialization estimation parameters. Because of the manual setup of parameters, the authors no longer have scaling problems with various cuisines.	For volume estimation, the authors want to employ additional contextual information. The authors can reduce the impact of segmentation and food categorization mistakes (or food portion estimation problems) by creating a more reliable energy estimation system.
Grum *et al.* [25]	The authors provide a way for creating multiple-object 3D scenarios using multiple-view photos. They adopt a two-stage technique to shift the RBF centers to increase the scene coherence with the picture content in textured regions and segmented object outlines. The writers first considered the scene's 3D patches and their projections in the provided picture set. The authors also considered enhancing the consistency between the segmented object outlines and the provided RBF scene model. The texture disparity and the form of the objects are rectified because of moving the RBF centers. Like the suggested technique, other 3D scene representations, such as those based on voxels, parametric models, or meshes, can be used.	Segmenting the provided 3D scene is straightforward since every object is isolated from its surroundings by the RBF surface unless two items are in contact. However, when attempting to represent scenarios with numerous objects, the performance of both space carving and RBF modeling is hindered.
Chen *et al.* [10]	The authors proposed a framework for registering a 3D model of the food item in a single-view 2D image to estimate meal content. Our first testing has demonstrated that, despite the 2D image's absence of detailed volumetric data, our framework can offer an acceptable degree of accuracy. More importantly, this method may be used with any meal if it can be easily put into a standard shape (e.g., an ellipsoid or a wedge).	To further enhance estimation findings, the authors must integrate several models.
Jia *et al.* [31]	Given the 3D food locations, the authors created two model-based methods to estimate meal volume from a single input picture. In our trials, a cuboid and seven meal replicas on a spherical dish were employed to gauge performance. Our findings showed that, when using the "wireframe-fitting" approach with 224 input images, the average volume estimation error was less than 10%.	Due to the bread's modest height and erratic border form, the volumetric inaccuracy is quite substantial. Notably, when the angle between the camera's optical axis and the table is significant, it is challenging to estimate height correctly.

Table 3
(Continued)

Paper	Main outcomes	Limitations
Patz *et al.* [48]	The suggested approach combines the benefits of supervised segmentation with the spread of gray-value uncertainty knowledge. Applications might find this knowledge regarding the impact of gray-value uncertainty applicable. In addition to the medical uses, more information on error propagation might be helpful for other engineering activities and disciplines like quality control or the general extraction of quantitative data.	Because we must store the representation in the tensor product space, the provided model is constrained in terms of the number of RVs and the polynomial order. The approach requires many input samples, or we must utilize the one sample provided to approximate the predicted value to produce an accurate stochastic input picture. Due to the correlation between the noise and potential dependence on the gray value, the model utilized in this work may not be enough for many applications. The authors included stochastic pictures in level-set-based segmentation techniques. We must transmit level sets at an unpredictable speed to utilize them on stochastic images.
Campbell *et al.* [9]	The technique significantly improves autonomous object segmentation. The volumetric method uses the silhouette coherency constraint to segment the item in 3D while concurrently segmenting it across all photos. It enables us to integrate the previously learned color model with a 3D shape to generate a more accurate result. The authors have also demonstrated that it is feasible to use a fixation restriction to initialize an iterative estimating technique to converge to the visual hull of an object observed in several perspectives, therefore eliminating the need for any user input, and automating the entire process.	As seen from the statue sequence, the method is mainly constrained by the color models employed to segment the object. The authors will use more sophisticated picture models to improve the object and background likelihood terms, which should enhance the algorithm's performance.

computer vision techniques allowed the exploration of several approaches to obtain various parameters targeting volume measurements for different objects or shapes.

The selected studies show significant advancements in 3D object recognition and volume estimation, particularly in food analysis. These advancements, particularly in deep learning approaches, are crucial for industries like healthcare, robotics, augmented reality, and smart environments. However, challenges in accuracy and complexity persist, highlighting the need for more efficient algorithms in machine learning and computer vision.

Food-related applications, such as calorie prediction, food amount measurement, and meal volume estimation, have significant implications for nutritional science, dietetics, and consumer technology. However, many papers note limitations in their methods, such as hardware constraints, estimation errors, and model generalization. These limitations highlight areas for future research and development, suggesting a need for more robust, generalizable, and hardware-efficient solutions.

The studies use various datasets to evaluate the proposed measuring methods. However, there needs to be more comparison between the modes, primarily due to the incompatibility of the datasets and the varying sensor types and setups. Most of the studies aimed to compare proposed methods to a certain baseline. This could also be considered a limitation of the studies and, consequently, of this survey as more data was needed to select a single best approach for volume measurement based on imaging processes. However, this review identified the current research trends regarding volume measurements based on imaging processes and which methods yield state-of-the-art predictive and analytical performance.

The information provided in this review suggests several future directions for improving the accuracy and generalization of volume estimation and object recognition methods. These include enhancing algorithms to work better with varying object shapes, sizes, and textures, generalizing models to function effectively across different scenarios and objects, and focusing on hardware optimization. Real-time processing capabilities are essential for robotics, augmented reality, and consumer technology applications, and reducing computational time without sacrificing accuracy is crucial. Complex object interactions are a challenge, and future research should focus on developing algorithms that accurately recognize and estimate object volume in crowded or complex scenes. Integrating volume estimation and object recognition technologies with other technologies like the Internet of Things (IoT), wearable

technology, and machine-to-machine communication can open new avenues for application. Developing comprehensive and diverse datasets is crucial for training effective models, and interdisciplinary collaboration across disciplines like nutrition science, psychology, robotics, and computer vision can lead to more holistic and effective solutions. By addressing these directions, future research can significantly advance the fields of 3D object recognition and volume estimation, leading to more accurate, efficient, and widely applicable solutions across various domains.

Acknowledgements

This work is funded by FCT/MEC through national funds and, when applicable, co-funded by the FEDER-PT2020 partnership agreement under the project **UIDB/50008/2020**. This work is also funded by FCT/MEC through national funds and, when applicable, co-funded by the FEDER-PT2020 partnership agreement under the project **UIDB/00308/2020**.

Conflict of interest

The authors have no conflict of interest to report.

References

[1] K. Adnan and R. Akbar, An analytical study of information extraction from unstructured and multidimensional big data, *Journal of Big Data* **6**(1) (2019), 1–38.

[2] P. Alcantarilla, J. Nuevo and A. Bartoli, Fast explicit diffusion for accelerated features in nonlinear scale spaces, in: *Procedings of the British Machine Vision Conference 2013*, British Machine Vision Association, Bristol, 2013, pp. 13.1–13.11. doi:10.5244/C.27.13.

[3] N. Bándi, R.-B. Tunyogi, Z. Szabó, E. Farkas and C. Sulyok, Image-based volume estimation using stereo vision, in: *2020 IEEE 18th International Symposium on Intelligent Systems and Informatics (SISY)*, 2020, pp. 000055–000060. doi:10.1109/SISY50555.2020.9217089.

[4] A. Bellazzi et al., Virtual reality for assessing visual quality and lighting perception: A systematic review, *Building and Environment* (2021), 108674.

[5] A. Börold, M. Teucke, J. Rust and M. Freitag, Recognition of car parts in automotive supply chains by combining synthetically generated training data with classical and deep learning based image processing, *Procedia CIRP* **93** (2020), 377–382. doi:10.1016/j.procir.2020.03.142.

[6] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel and A.M. Dollar, The YCB object and model set: Towards common benchmarks for manipulation research, in: *2015 International Conference on Advanced Robotics (ICAR)*, IEEE, Istanbul, Turkey, 2015, pp. 510–517. doi:10.1109/ICAR.2015.7251504.

[7] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel and A.M. Dollar, Benchmarking in manipulation research: Using the Yale–CMU–Berkeley object and model set, *IEEE Robot. Automat. Mag.* **22**(3) (2015), 36–52. doi:10.1109/MRA.2015.2448951.

[8] D.B. Camarillo, T.M. Krummel and J.K. Salisbury Jr, Robotic technology in surgery: Past, present, and future, *The American Journal of Surgery* **188**(4) (2004), 2–15. doi:10.1016/j.amjsurg.2004.08.025.

[9] N.D.F. Campbell, G. Vogiatzis, C. Hernandez and R. Cipolla, Automatic 3D object segmentation in multiple views using volumetric graph-cuts, *Image and Vision Computing* **28**(1) (2010), 14–25. Elsevier, Radarweg 29, 1043 NX Amsterdam, Netherlands. doi:10.1016/j.imavis.2008.09.005.

[10] H.-C. Chen, W. Jia, Z. Li, Y.-N. Sun and M. Sun, 3D/2D model-to-image registration for quantitative dietary assessment, in: *2012 38th Annual Northeast Bioengineering Conference (NEBEC)*, IEEE, 345 E 47th st, New York, NY 10017 USA, 2012, pp. 95+.

[11] G. Ciocca, P. Napoletano and R. Schettini, Food recognition: A new dataset, experiments, and results, *IEEE Journal of Biomedical and Health Informatics* **21**(3) (2017), 588–598. IEEE-Inst Electrical Electronics Engineers Inc, 445 Hoes Lane, Piscataway, NJ 08855-4141 USA. doi:10.1109/JBHI.2016.2636441.

[12] M. Cordts et al., The cityscapes dataset for semantic urban scene understanding, in: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 345 E 47th st, New York, NY 10017 USA. 2016, pp. 3213–3223. doi:10.1109/CVPR.2016.350.

[13] R. Dalai, N. Dalai and K.K. Senapati, An accurate volume estimation on single view object images by deep learning based depth map analysis and 3D reconstruction, *Multimed Tools Appl* **82**(18) (2023), 28235–28258. doi:10.1007/s11042-023-14615-7.

[14] R. Dalai and K.K. Senapati, A heuristic grid area based segmentation approach for weight estimation of an object from image, in: *2018 4th International Conference for Convergence in Technology (I2CT)*, IEEE, 345 E 47th st, New York, NY 10017 USA, 2018.

[15] B.L. Daugherty et al., Novel technologies for assessing dietary intake: Evaluating the usability of a mobile telephone food record among adults and adolescents, *Journal of Medical Internet Research* **14**(2) (2012), Jmir Publications, inc, 130 Queens Quay E, Ste 1102, Toronto, On M5A 0P6, Canada. doi:10.2196/jmir.1967.

[16] S. Fang, C. Liu, F. Zhu, E.J. Delp and C.J. Boushey, Single-view food portion estimation based on geometric models, in: *IEEE International Symposium on Multimedia-ISM*, IEEE, 345 E 47th st, New York, NY 10017 USA, 2015, pp. 385–390. doi:10.1109/ISM.2015.67.

[17] F. Ferreira et al., A systematic investigation of models for color image processing in wound size estimation, *Computers* **10**(4) (2021), 43. doi:10.3390/computers10040043.

[18] F. Ferreira, I.M. Pires, V. Ponciano, M. Costa and N.M. Garcia, Approach for the wound area measurement with mobile devices, in: *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, IEEE, Toronto, ON, Canada, 2021, pp. 1–4. doi:10.1109/IEMTRONICS52119.2021.9422661.

[19] A. Gao, F.P.-W. Lo and B. Lo, Food volume estimation for quantifying dietary intake with a wearable camera, in: *2018 IEEE 15th International Conference on Biomedical and Health Informatics (BHI) and the Wearable and Implantable Body Sensor Networks (BSN)*, International Conference on Wearable and Implantable Body Sensor Networks, IEEE, 345 E 47th st, New York, NY 10017 USA, 2018, pp. 110–113.

[20] J. Gao, W. Tan, L. Ma, Y. Wang and W. Tang, MUSEFood: Multi-sensor-based food volume estimation on smartphones, in: *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 2019, pp. 899–906. doi:10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00182.

[21] N.M. Garcia and J.J.P.C. Rodrigues (eds), *Ambient Assisted Living*, CRC Press, 2015. doi:10.1201/b18520.

[22] A. Geiger, P. Lenz and R. Urtasun, Are we ready for autonomous driving? The KITTI vision benchmark suite, in: *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Providence, RI, 2012, pp. 3354–3361. doi:10.1109/CVPR.2012.6248074.

[23] P. Gomes and L.A. da Silva Cruz, Pseudo-sequence light field image scalable encoding with improved random access, in: *2019 8th European Workshop on Visual Information Processing (EUVIP)*, IEEE, 2019, pp. 16–21. doi:10.1109/EUVIP47703.2019.8946268.

[24] M. Grum and A.G. Bors, 3D modeling of multiple-object scenes from sets of images, *Pattern Recognition* **47**(1) (2014), 326–343. Si. Elsevier Sci Ltd, The Boulevard, Langford Lane, Kidlington, Oxford Ox5 1GB, Oxon, England. doi:10.1016/j.patcog.2013.04.020.

[25] M. Grum and A.G. Bors, *Pattern Recognition* **47**(1) (2014), 326–343. Si. Elsevier Sci Ltd, The Boulevard, Langford Lane, Kidlington, Oxford Ox5 1GB, Oxon, England. doi:10.1016/j.patcog.2013.04.020.

[26] S. Hadi, S. Phon-Amnuaisuk and S.-J. Tan, Semantic instance segmentation in a 3D traffic scene reconstruction task, in: *2020 59th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, 2020, pp. 186–191. doi:10.23919/SICE48898.2020.9240300.

[27] T. Hotfilter, F. Kempf, J. Becker, D. Reinhardt and I. Baili, Embedded image processing the European way: A new platform for the future automotive market, in: *2020 IEEE 6th World Forum on Internet of Things (WF-IoT)*, IEEE, 2020, pp. 1–6.

[28] I. Ischuk, A. Dolgov and V. Tyapkin, Verification of a mathematical model for constructing thermal tomograms based on the reduction of a set of different-time visible and infrared images, in: *2020 International Conference on Information Technology and Nanotechnology (ITNT)*, IEEE, Samara, Russia, 2020, pp. 1–5. doi:10.1109/ITNT49337.2020.9253218.

[29] S. Jeon and Y.S. Heo, Efficient multi-scale stereo-matching network using adaptive cost volume filtering, *Sensors* **22**(15) (2022), 5500. doi:10.3390/s22155500.

[30] R.E. Jerome, S.K. Singh and M. Dwivedi, Process analytical technology for bakery industry: A review, *Journal of Food Process Engineering* **42**(5) (2019), e13143. doi:10.1111/jfpe.13143.

[31] W. Jia, Y. Yue, John, D. Fernstrom, Z. Zhang, Y. Yang and M. Sun, 3D localization of circular feature in 2D image and application to food volume estimation, in: *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE Engineering in Medicine and Biology Society Conference Proceedings, IEEE, 345 E 47th st, New York, NY 10017 USA, 2012, pp. 4545–4548.

[32] N. Kalpitha and S. Murali, 3D recognition & automated access control reconstruction from a single image, in: *Materials Today: Proceedings*, 2021. doi:10.1016/j.matpr.2021.04.012.

[33] Y. Kawano and K. Yanai, Automatic expansion of a food image dataset leveraging existing categories with domain adaptation, in: *Computer Vision – ECCV 2014 Workshops, Pt III*, Lecture Notes in Computer Science, Vol. 8927, Springer-Verlag Berlin, Heidelberger Platz 3, D-14197 Berlin, Germany, 2015, pp. 3–17. doi:10.1007/978-3-319-16199-0_1.

[34] G. Küçükyildiz, Image processing based package volume measurement system using kinect sensor, *Sigma J Eng Nat Sci – Sigma Müh Fen Bil Derg* (2022). doi:10.14744/sigma.2022.00003.

[35] G. Litjens et al., A survey on deep learning in medical image analysis, *Medical Image Analysis* **42** (2017), 60–88. doi:10.1016/j.media.2017.07.005.

[36] F.P.-W. Lo, Y. Sun, J. Qiu and B. Lo, Food volume estimation based on deep learning view synthesis from a single depth map, *Nutrients* **10**(12) (2018). MDPI, St Alban-Anlage 66, Ch-4052 Basel, Switzerland. doi:10.3390/nu10122005.

[37] F.P.W. Lo, Y. Sun, J. Qiu and B. Lo, Image-based food classification and volume estimation for dietary assessment: A review, *IEEE J. Biomed. Health Inform.* **24**(7) (2020), 1926–1939. doi:10.1109/JBHI.2020.2987943.

[38] Q. Luo, H. Ma, L. Tang, Y. Wang and R. Xiong, 3D-SSD: Learning hierarchical features from RGB-D images for amodal 3D object detection, in: *Neurocomputing*, Vol. 378, Elsevier, Radarweg 29, 1043 NX Amsterdam, Netherlands, 2020, pp. 364–374. doi:10.1016/j.neucom.2019.10.025.

[39] N. Mayer et al., A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation, in: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4040–4048. doi:10.1109/CVPR.2016.438.

[40] M. Menze and A. Geiger, Object scene flow for autonomous vehicles, in: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Boston, MA, USA, 2015, pp. 3061–3070. doi:10.1109/CVPR.2015.7298925.
[41] A.S. Mian, M. Bennamoun and R. Owens, Three-dimensional model-based object recognition and segmentation in cluttered scenes, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(10) (2006), 1584–1601. doi:10.1109/TPAMI.2006.213.
[42] A.S. Mian, M. Bennamoun and R.A. Owens, A novel representation and feature matching algorithm for automatic pairwise registration of range images, *International Journal of Computer Vision* **66**(1) (2006), 19–40. doi:10.1007/s11263-005-3221-0.
[43] S. Mittal, S. Srivastava and J.P. Jayanth, A survey of deep learning techniques for underwater image classification, *IEEE Transactions on Neural Networks and Learning Systems* (2022).
[44] P.S. Mittapalli and G.B. Kande, Segmentation of optic disk and optic cup from digital fundus images for the assessment of glaucoma, *Biomedical Signal Processing and Control* **24** (2016), 34–46. doi:10.1016/j.bspc.2015.09.003.
[45] O.Y. Morozova, Analysis of typical electric power facilities requiring remote monitoring, in: *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, IEEE, St. Petersburg and Moscow, Russia, 2020, pp. 1255–1257. doi:10.1109/EIConRus49466.2020.9039389.
[46] N. C. for H. Statistics (US), Plan and operation of the third National Health and Nutrition Examination Survey, 1988–94, *US Department of Health and Human Services, Public Health Service, Centers, ...*, 1994.
[47] P.A. Neves et al., Thought on food: A systematic review of current approaches and challenges for food intake detection, *Sensors* **22**(17) (2022), 6443. doi:10.3390/s22176443.
[48] T. Paetz and T. Preusser, Segmentation of stochastic images with a stochastic random Walker method, *IEEE Transactions on Image Processing* **21**(5) (2012), 2424–2433. IEEE-Inst Electrical Electronics Engineers Inc, 445 Hoes Lane, Piscataway, NJ 08855-4141 USA. doi:10.1109/TIP.2012.2187531.
[49] A.S. Parihar, M. Gupta, V. Sikka and G. Kaur, Dimensional analysis of objects in a 2d image, in: *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, International Conference on Computing Communication and Network Technologies, IEEE, 345 E 47th st, New York, NY 10017 USA, 2017.
[50] J. Parker et al., National Center for Health Statistics data presentation standards for proportions, 2017.
[51] R. Pintus, T.G. Dulecha, I. Ciortan, E. Gobbetti and A. Giachetti, State-of-the-art in multi-light image collections for surface visualization and analysis, in: *Computer Graphics Forum*, Wiley Online Library, 2019, pp. 909–934.
[52] I.M. Pires et al., Pattern recognition techniques for the identification of activities of daily living using a mobile device accelerometer, *Electronics* **9**(3) (2020), 509. doi:10.3390/electronics9030509.
[53] I.M. Pires and N.M. Garcia, Wound area assessment using mobile application, *Biodevices* (2015), 271–282.
[54] P. Poply and J.A.A. Jothi, Refined image segmentation for calorie estimation of multiple-dish food items, in: *2021 IEEE International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, IEEE, 345 E 47th st, New York, NY 10017 USA, 2021, pp. 682–687.
[55] S. Qi et al., Review of multi-view 3D object recognition methods based on deep learning, *Displays* **69** (2021), 102053. doi:10.1016/j.displa.2021.102053.
[56] R. Ravikumar and V. Arulmozhi, Digital image processing – a quick review, *International Journal of Intelligent Computing and Technology (IJICT)* **2**(2) (2019), 11–19.
[57] E. Reinhard, W. Heidrich, P. Debevec, S. Pattanaik, G. Ward and K. Myszkowski, *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*, Morgan Kaufmann, 2010.
[58] C.M. Sabliov, D. Boldor, K.M. Keener and B.E. Farkas, Image processing method to determine surface area and volume of axi-symmetric agricultural products, *Int. J. of Food Properties* **5**(3) (2002), 641–653. doi:10.1081/JFP-120015498.
[59] N. Shabairou, E. Cohen, O. Wagner, D. Malka and Z. Zalevsky, Color image identification and reconstruction using artificial neural networks on multimode fiber images: Towards an all-optical design, *Optics Letters* **43**(22) (2018), 5603–5606. doi:10.1364/OL.43.005603.
[60] C. Shah and Q. Du, Spatial-aware collaboration – competition preserving graph embedding for hyperspectral image classification, *IEEE Geoscience and Remote Sensing Letters* **19** (2022), 1–5. doi:10.1109/LGRS.2021.3074328.
[61] G. Shi, H. Huang, J. Liu, Z. Li and L. Wang, Spatial-spectral multiple manifold discriminant analysis for dimensionality reduction of hyperspectral imagery, *Remote Sensing* **11**(20) (2019), 20. doi:10.3390/rs11202414.
[62] J.P. Shim, M. Warkentin, J.F. Courtney, D.J. Power, R. Sharda and C. Carlsson, Past, present, and future of decision support technology, *Decision support systems* **33**(2) (2002), 111–126.
[63] N. Silberman, D. Hoiem, P. Kohli and R. Fergus, Indoor segmentation and support inference from RGBD images, in: *Computer Vision – ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato and C. Schmid, eds, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2012, pp. 746–760. doi:10.1007/978-3-642-33715-4_54.
[64] R.D. Singh, A. Mittal and R.K. Bhatia, 3D convolutional neural network for object recognition: A review, *Multimedia Tools and Applications* **78**(12) (2019), 15951–15995. doi:10.1007/s11042-018-6912-6.
[65] S. Song, S.P. Lichtenberg and J. Xiao, SUN RGB-D: A RGB-D scene understanding benchmark suite, in: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 567–576. doi:10.1109/CVPR.2015.7298655.
[66] T. Suzuki, K. Futatsuishi, K. Yokoyama and N. Amaki, Point cloud processing method for food volume estimation based on dish space, in: *42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society: Enabling Innovative Technologies for Global Healthcare EMBC'20*, IEEE Engineering in Medicine and Biology Society Conference Proceedings, IEEE, 345 E 47th st, New York, NY 10017 USA, 2020, pp. 5665–5668.
[67] F. Tombari, S. Salti and L. Di Stefano, Performance evaluation of 3D keypoint detectors, *International Journal of Computer Vision* **102**(1) (2013), 198–220. doi:10.1007/s11263-012-0545-4.

[68] V.-I. Tomescu, FoRConvD: An approach for food recognition on mobile devices using convolutional neural networks and depth maps, in: *2020 IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI 2020)*, IEEE, 345 E 47th st, New York, NY 10017 USA, 2020, pp. 129–134.

[69] R. Udendhran, M. Balamurugan, A. Suresh and R. Varatharajan, Enhancing image processing architecture using deep learning for embedded vision systems, *Microprocessors and Microsystems* **76** (2020), 103094. doi:10.1016/j.micpro.2020.103094.

[70] T.L. van den Heuvel, D. de Bruijn, C.L. de Korte and B. van Ginneken, Automated measurement of fetal head circumference using 2D ultrasound images, *PloS one* **13**(8) (2018), e0200412.

[71] G. Varma, A. Subramanian, A. Namboodiri, M. Chandraker and C.V. Jawahar, IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments, in: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE Winter Conference on Applications of Computer Vision, IEEE, 345 E 47th st, New York, NY 10017 USA, 2019, pp. 1743–1751. doi:10.1109/WACV.2019.00190.

[72] J. Wang, Z. Mo, H. Zhang and Q. Miao, A deep learning method for bearing fault diagnosis based on time-frequency image, *IEEE Access* **7** (2019), 42373–42383. doi:10.1109/ACCESS.2019.2907131.

[73] T. Wang, J. Pang and D. Lin, Monocular 3D object detection with depth from motion, arXiv, 1 de março de 2023. Acedido: 24 de outubro de 2023. [Em linha]. Disponível em, http://arxiv.org/abs/2207.12988.

[74] L. Wittstruck, I. Kuehling, D. Trautz, M. Kohlbrecher and T. Jarmer, UAV-based RGB imagery for Hokkaido pumpkin (Cucurbita max.) detection and yield estimation, *Sensors* **21**(1) (2021). MDPI, St Alban-Anlage 66, CH-4052 Basel, Switzerland. doi:10.3390/s21010118.

[75] Z. Wu et al., 3D ShapeNets: A deep representation for volumetric shapes, in: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Boston, MA, USA, 2015, pp. 1912–1920. doi:10.1109/CVPR.2015.7298801.

[76] J. Yang, Q. Zhang, Y. Xiao and Z. Cao, TOLDI: An effective and robust approach for 3D local shape description, in: *Pattern Recognition*, Vol. 65, Elsevier Sci Ltd, The Boulevard, Langford Lane, Kidlington, Oxford, GB, Oxon, England, 2017, pp. 175–187. doi:10.1016/j.patcog.2016.11.019.

[77] Z. Yang et al., Human-mimetic estimation of food volume from a single-view RGB image using an AI system, *Electronics* **10**(13) (2021), 1556. doi:10.3390/electronics10131556.

[78] R.D. Yogaswara, E.M. Yuniarno and A.D. Wibawa, Instance-aware semantic segmentation for food calorie estimation using mask R-CNN, in: *2019 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 2019, pp. 416–421. doi:10.1109/ISITIA.2019.8937129.

[79] E. Zdravevski et al., Automation in systematic, scoping and rapid reviews by an NLP toolkit: A case study in enhanced living environments, in: *Enhanced Living Environments*, I. Ganchev, N.M. Garcia, C. Dobre, C.X. Mavromoustakis and R. Goleva, eds, Lecture Notes in Computer Science, Vol. 11369, Springer International Publishing, Cham, 2019, pp. 1–18. doi:10.1007/978-3-030-10752-9_1.

[80] O. Zendel, K. Honauer, M. Murschitz, D. Steininger and G.F. Dominguez, WildDash – creating hazard-aware benchmarks, in: *Computer Vision – ECCV 2018, PT VI*, Lecture Notes in Computer Science, Vol. 11210, Springer International Publishing AG, Gewerbestrasse 11, Cham, CH-6330, Switzerland, 2018, pp. 407–421. doi:10.1007/978-3-030-01231-1_25.