IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 7, 2020, accepted September 15, 2020, date of publication September 18, 2020,
date of current version September 29, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3025193

Image Object Extraction Based on Semantic Detection and Improved K-Means Algorithm

HANXIAO RONG[1], ALEX RAMIREZ-SERRANO[2], (Member, IEEE), LIANWU GUAN[1], AND YANBIN GAO[1]

[1]College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China
[2]Department of Mechanical Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada

Corresponding author: Lianwu Guan (guanlianwu@hrbeu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61803118, and in part by the Science and Technology Research Program of the Chongqing Municipal Education Commission under Grant KJZD-K201804701.

ABSTRACT Object extraction is an important tool in many applications within the image processing and computer vision communities. You Only Look Once version 3 (YOLOv3) has been extensively applied to many fields as a state-of-the-art technique for object semantic detection. Despite its numerous characteristics, YOLOv3 has to be combined with appropriate image segmentation technologies to achieve effective 2D object extraction in real-time monitoring, robot navigation, and target search. In this article, the K-means algorithm is applied to the segmentation of depth images. Considering the inherent sensitivity to the randomness of the initial cluster center and the uncertainty of cluster number K in the initialization phase of the K-means algorithm, this article proposes a new method that combines the semantic image information with the image depth information. Specifically, this method proposed to pre-classify the center depth of the object to determine the appropriate value of K required in the K-means algorithm. At the same time, the proposed algorithm improves the selection of the initial center via the maximin method. This article introduces a multi-parameter extraction method to enable to correctly identify the object of interest after image segmentation. The technique considers three parameters to achieve this: i) the elements of size, ii) the connected domain, and iii) the diagonal detection. Experiments using open-source datasets demonstrate that the average processing time and the segmentation accuracy of the improved K-means algorithm are 20.36% faster and 3.12% higher than the conventional K-means algorithm, respectively. The extraction accuracy of the proposed method is 6.69% higher than that of the SuperCut extraction method.

INDEX TERMS Semantic detection, K-means, image segmentation, object extraction.

I. INTRODUCTION

Object extraction aims to extract the Region of Interest (RoI) of a given image according to the detected position of an object of interest. As prevalent topics in the field of image processing and computer vision, object extraction has been widely used in diverse fields such as real-time monitoring, robot navigation, and target search [1]. Despite the numerous recent breakthroughs in object extraction, many unsolved problems remain. For example, it is still difficult to guarantee the reliability and robustness of the object model obtained from an image due to the interference caused by different perspectives and scenarios from where the image was captured [2]. Fortunately, the introduction of Convolutional Neural Networks (CNNs) has enabled the development of more intelligent and accurate methodologies to detect and extract objects.

Object semantic detection based on CNN, the premise technology of object extraction, has been intensively researched because of its potential towards enhancing the intelligence and autonomy of robots, an area requiring improved sensing and object perception mechanisms [3]. The Region-based Convolutional Network (R-CNN) method proposed by Girshick *et al.* [4] employs deep CNNs for classifying object proposals to increase object detection accuracy. Despite the fact that object detection accuracy has been increased, the detection speed and the detection of objects with complex geometries needs to be improved. In such direction, Faster R-CNN [5], [6] has been developed giving rise to Mask R-CNN by adding the capability to predict segmentation masks on each Region of Interest (RoI) comprising the image. With such contributions, it is possible to perform

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

parallel object detection and extraction [7]. However, despite the success of the R-CNN framework, the associated complex pipelines of the algorithm limit its widespread application due to the high computing time and prohibitive hardware requirements. To reduce these limitations, the You Only Look Once (YOLO), a faster object detection algorithm based on CNN, was proposed by Redmon *et al.* in 2016 [8]. Compared to state-of-the-art methods, YOLO performs object detection as a single regression problem resulting in effective performance. The improved classifier network and independent logic classifier have been used to further improve the characteristics of the YOLO algorithm. The newly YOLOv3 has been proven to be as accurate as the Faster R-CNN, with the added benefit of running eight times faster [9]. Such features have made YOLOv3 the most popular algorithm in image semantic object detection. For example, Wei *et al.* realized real-time monitoring of railway tracks via YOLOv3 and surveillance video [10]. Peng *et al.* extracted dynamic objects from visual data by developing graph-based image segmentation combined with YOLOv3 [11]. Despite the advantageous characteristics of YOLOv3, the running time and precision of the graph-based image segmentation algorithm are extremely dependent on the quality of the input image. Therefore, there is a need to develop improved image segmentation methods that can be combined with YOLOv3 to extract any RoI semantically while having less dependency on image quality.

In order to achieve effective image segmentation in object extraction, Wu *et al.* proposed the SuperCut method to extract objects from a given bounding box defined within the image of interest by utilizing superpixels and graph cut techniques [12]. In contrast to [12], López-Rubio *et al.* distinguished the foreground from the background by using the Self-Organizing Maps (SOM) cluster method [13]. Allab *et al.* improved the Spectral clustering method by finding the optimal spectral embedding which has better convergence in image data processing [14]. Hou *et al.* utilized the Gaussian Mixture Model (GMM) and the Expectation-Maximization (EM) algorithm to address the problem of image color extraction [15]. Sun *et al.* [16] and Tian *et al.* [17] employed the K-means algorithm to realize the quantization of depth images and the extraction of tomato leaf images, respectively. The fast convergence speed and straightforward logic structure make the K-means algorithm widely used in the color quantization, and image segmentation areas [18]. Typically, the K value required in the K-means algorithm is empirical, and the selection of the initial cluster centers is random. Since the cluster number K and the initial cluster centers have a decisive impact on the precision and processing speed of the algorithm, it is always challenging to properly selected a good K value. To solve the above problems, Xia *et al.* introduced the parallel Canopy K-means algorithm to optimize the initial parameters of the K-means algorithm [19]. Although effective, it requires the experience of the user to determine the threshold needed in the Canopy algorithm. Zhang *et al.* proposed a technique to find the optimal value of K by continuously calculating the Davies Bouldin (DB) index of the segmentation results obtained using diverse segmentation numbers [20]. However, the range of the K value set in advance directly affects the processing time of the method. In [21], Khan *et al.* proposed the Region Splitting (RS) module to pre-classify an RGB image data to determine the initial parameters for the K-means algorithm. While the RS module improves the segmentation quality, it requires a trade-off between computational complexity and the algorithm's performance.

Although the above methods present many effective solutions to existing problems of object extraction and image segmentation, they still do not effectively address the problems of multi-object interference of image segmentation in complex application scenes. For the problem of false extraction caused by the existence of multiple objects, Wei *et al.* pointed out that the mechanism of multi-information and multi-parameter can improve the accuracy and robustness of object extraction [10]. Therefore, an object extraction method combining semantic information and depth information is expected to provide enhanced image extraction results. However, the conventional K-means image segmentation method can not effectively distinguish different objects according to the application scene due to the inappropriate number of segmentation and undeterministic initialization. Under this context, this article proposes an effective object extraction framework to solve these problems. Therefore, an extraction method based on YOLOv3 and the improved K-means image segmentation is proposed in this article to extract objects of interest on RGB and depth images. This approach takes advantage of semantic detection for scene understanding, and depth information for segmentation. The RoI and Region of Uninterest (RoU) of the input image are detected by YOLOv3, followed by the calculation of the center depth of objects with offset parameters. According to the center depth and semantic detection, the value of K is then calculated using a pre-classification technique. The computed K value and the initial centers selected by the maximin method are applied to the depth image segmentation. Finally, an extraction mechanism based on size, connected domain, and diagonal detection of the image is introduced to complete the object extraction algorithm.

The main contributions of this article are: i) an improved K-means image segmentation algorithm which strengthens the adaptability of image segmentation to different application scenes by using a pre-classification and a maximin method; ii) improved accuracy and robustness of object extraction by combining the depth image segmentation with semantic detection and using a multi-parameter extraction mechanism. The remainder part of this article is organized as follows. Section II introduces YOLOv3 and provides the basic evaluation metrics of the conventional K-means algorithm. The improved K-means algorithm and the mechanism of extracting objects are presented in Section III. Section IV compares and analyzes the experimental results of the proposed algorithm with existing approaches,

while the conclusions of the paper are presented in Section V.

II. RELATED METHODS

A. SEMANTIC DETECTION METHOD

YOLOv3, the third version of YOLO [9], has a number of features that are absent in many other advanced methods. First, the basic network of YOLOv3 runs at 45 Frames Per Second (FPS) without batch processing, which makes it extremely fast and even able to detect objects in real-time. Second, unlike the sliding window and region proposal-based techniques, the detection approach used in YOLOv3 infers images globally. With such characteristics, YOLOv3 has the ability to detect all the objects of interest in an image with only one process. Third, based on the generalizable representations of objects, YOLOv3 enables to detect objects steadily when applied to unexpected information or an unknown environment [9].

For each recognized object within a given image, the network of YOLOv3 predicts four coordinates (t_x, t_y, t_w, t_h) and one confidence parameter (t_o) associated with the identified entity. The bounding box obtained from the above mentioned four coordinates is used to describe the location and size of the recognized object within the image. The parameters $\sigma\ (t_x)$ and $\sigma\ (t_y)$ are used to represent any offset from the origin of the image, where the paprmeters c_x and c_y describe the offset from the origin of the cell. The predicted bounding box surrounding the identified object has width and height p_w, p_h. The coordinates and confidence of the corresponding bounding box for a given entity located in cell (x, y) are shown in Fig. 1. Equation (1) is used to compute the center coordinates b_u and b_v within the given cell based on p_w and p_h, the width b_w, the height b_h, and the confidence $\sigma\ (t_o)$ of the bounding box.

$$\begin{aligned} b_u &= \sigma\ (t_x) + c_x \\ b_v &= \sigma\ (t_y) + c_y \\ b_w &= p_w e^{t_w} \\ b_h &= p_h e^{t_h} \\ \sigma\ (t_o) &= \Pr(\text{ object }) * \text{IOU}(b,\ \text{object }) \\ B &= \text{box}\ (b_u, b_v, b_w, b_h) \end{aligned} \tag{1}$$

The $\Pr(\text{ object })$ and $\text{IOU}(b,\ \text{object })$ functions in equation (1)denote the accuracy of the object classification prediction and the Intersection Over Union (IOU) between the predicted box and the ground truth, respectively. The "box" represents the function used to calculate the location and size of the bounding box as defined in [9]. Thus $\sigma\ (t_o)$ represents the confidence that the model accurately provides the box coordinate containing the given object. Such value, $\sigma\ (t_o)$, also includes information on how accurate it thinks that the box is as predicted. After YOLOv3 outputs the bounding box of the detected object, in order to further extract the object from the bounding box, the image segmentation method based on the K-means algorithm is used in the proposed algorithm.

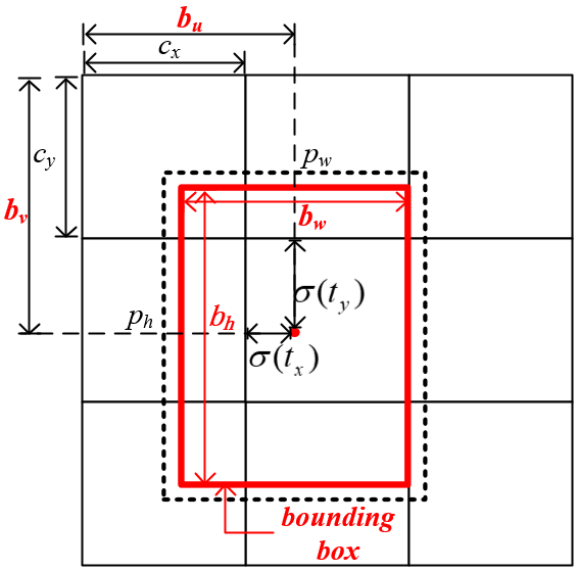


FIGURE 1. Coordinate and size prediction of bounding box.

B. K-MEANS SEGMENTATION ALGORITHM

Image segmentation refers to the decomposition of an image into several non-overlapping meaningful areas with or without the same attributes. Segmenting a given image to include areas representing two objects, "A" and "B", is a typical example. Image segmentation is a critical algorithm in digital image processing. Thus, the accuracy of segmentation directly affects the effectiveness of any application requiring such image information. In conventional image segmentation algorithms, the segmentation process is based on the following three image aspects: 1) the threshold, 2) the edge, and 3) the region [12], [21]–[23]. In traditional image segmentation methods, the algorithms are based on the theory of cluster analysis [24], which humans use when learning to distinguish things by continually adjusting the subliminal clustering pattern. Due to the high efficiency of the classical K-means cluster analysis in large-scale data, the K-means segmentation method is applied in the proposed image segmentation.

The basic idea of the K-means algorithm is to divide the given data into clusters according to the clustering number K, in which the points contained in every cluster are the closest to each other. The K-means algorithm is a typical representation of the clustering method based on the Minkowski Distance function. The following four steps describe the conventional K-means algorithm:

Step 1: Randomly initialize K example cluster centers from the given image data set;

Step 2: Assign all image data points closest to a given example cluster center according to some distance function to create a cluster;

Step 3: Compute the new cluster center for each formed cluster using all data points assigned to the corresponding cluster;

Step 4: Compare the newest cluster center with the corresponding prior cluster center between the present and last cluster center to determine convergence. If the clustering

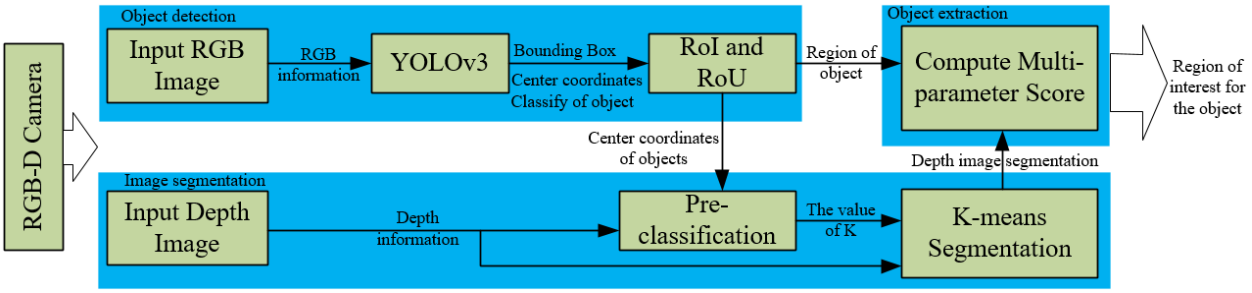


FIGURE 2. Schematic representation of the proposed method.

result does not change from the previous iteration, this represents that the optimization has reached the optimal value and that the clustering process is over; otherwise, repeat Steps 2 and 3.

From the above process, it is clear that the conventional K-means algorithm has some downsides. The main drawback is that the number of clusters, K, needs to be set manually. In many cases, it is difficult to determine the appropriate number of clusters for a given image or set of images in advance. Another drawback is the randomness in defining the initial clustering center. Although the random selection works efficiently when a good set of initial clustering centers is used, but it is undeterministic and unreliable, which leads to inconsistent results or even failure in the image segmentation process. For example, there is no current solution that prevents the chosen initial clustering centers from being too close to each other. Therefore, to remedy the problems of the conventional K-means algorithm, a method of depth image segmentation making full use of the capabilities of YOLOv3 detection is proposed. In order to quantify the accuracy of the image segmentation process and object extraction, there is also a need to chose appropriate evaluation metrics that consider the influence of image data characteristics and depth image noise.

C. EVALUATION METRICS

One of the most effective parameter used to measure the accuracy of image segmentation is the validity index. Typically, the maximum or minimum index value is used to represent that the corresponding segmentation is optimum. Compactness and separation are the main principles for evaluating the effectiveness of segmentation. The compactness parameter is a measure to indicate if the members of the same segmentation are as close as possible to each other. The separation parameter, on the other hand, is a parameter indicating the distance between different segmentation where such value is expected to show that different segmentation is as far as possible from each other. Diverse criteria have been applied in image segmentation, such as the DB index [25], the Calinski-Harabasz (CH) index [26], and the Silhouette Coefficient (SC). The research in [27] points out that SC can work with any distance measure, including the Euclidean and Manhattan distances. SC has the advantage of being insensitive to the information in noise features, which makes it useful in the application of depth image segmentation.

SC is a ratio-type coefficient based on silhouette values for one segmentation, s_i, judging how similar s_i is to its assigned and other segmentation. The silhouette can be calculated by making comparisons between within-segmentation cohesion and the segmentation separation based on the distance to different segmentation. When s_i traverses all the segmentation, it takes the average of the silhouette values to get the SC using (2):

$$SC(k) = \text{avg}(sil(i)) = \frac{1}{p}\sum_{i=1}^{p}\frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{2}$$

where k and p denote the number of the segmentation and the number of total pixels in the input image, respectively. The parameter $a(i)$ represents the average dissimilarity of i ($i \in s_i$) to other segmentation, while $b(i)$ represents the minimized dissimilarity between i and all the segmentation. It must be noted that the range of individual silhouette value $sil(i)$ is $[-1,1]$. A value close to 1 shows that the current pixel is a good match for its assigned segmentation, whereas a $sil(i)$ closer to -1 indicates the opposite. The effectiveness of overall segmentation is computed by taking the average value of all $sil(i)$. In this article, the metric SC is adopted to quantify the effectiveness of the image segmentation.

In the field of object extraction, IOU is usually used as the standard to evaluate the accuracy of the extracted region. For each extracted object, IOU computes the similarity between the extracted region and the real region of the object existing in the given image. This criterion is defined by equation (3):

$$IOU = \frac{OR \bigcap ER}{OR \bigcup ER} \tag{3}$$

where ER and OR represent the extracted region and the reference region, respectively.

III. THE PROPOSED METHOD

The goal of the proposed algorithm is to accurately extract objects of interest from a given RGB-D image. The objects of interest include occluded objects, where the entire object might not be visible in the image as well as unobstructed objects. The proposed method includes the following three aspects: 1) an image semantic object detection process, 2) a K-means depth image segmentation algorithm, and 3) an object extraction tool. The flow of how these aspects are related is illustrated in Fig. 2.

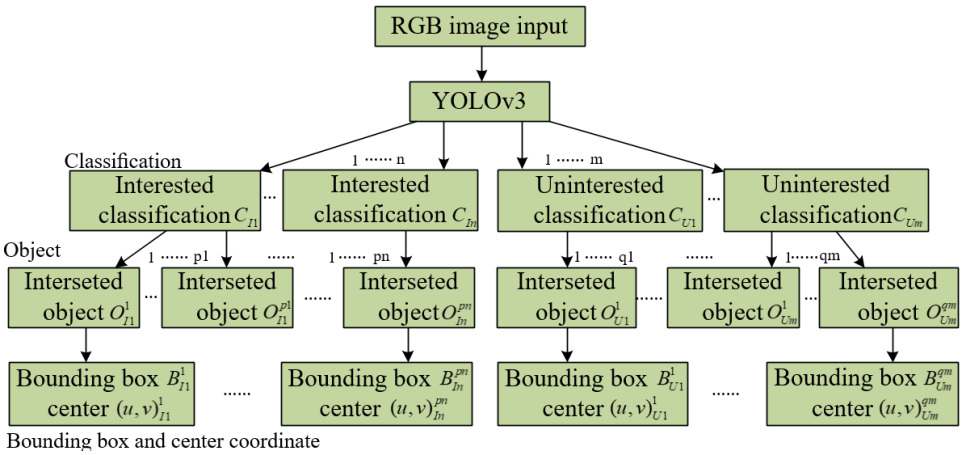


FIGURE 3. Relationship between the parameters of YOLOv3.

A. IMAGE SEMANTIC OBJECT DETECTION

In real world visual information, there may be multiple objects in any given RGB image captured by the available sensors. These objects include items of interest as well as other objects. Although segmentation has the ability to separate different objects, it cannot distinguish the objects of interest from others. Typically, semantic information is used to recognize the objects of interest [28].

The detection results that are typically achieved in YOLOv3 are illustrated in Fig. 3. The identified classification is divided into n classifications of interest, C_{In}, and m classifications of uninterest objects, C_{Um}. Classifications are distinguished by judging whether the detected semantic information belongs to the preset classifications of interest. The bounding boxes coordinate B^{pn}_{In} of the object of interest, O^{pn}_{In}, and uninterested object B^{qm}_{Um} of O^{qm}_{Um} are calculated from (1), where pn and qm in B^{pn}_{In} and B^{qm}_{Um} denote the amount of objects in C_{In} and C_{Um}, respectively. In addition, the center coordinate $(u, v)_O = (b_u, b_v)$ of the corresponding object is computed by (1).

B. DETERMINATION OF K BASED ON SEMANTIC DETECTION

Many methods have been proposed to solve the problem of determining the value of K in the K-means algorithm. However, such approaches are based purely on the exclusive use of 2D images. Similar to [11], the algorithm proposed in this article makes full use of semantic and depth information. The object identification provided by YOLOv3 is based on 2D images. However, due to the fact that the objects in any depth image are more easily distinguished than those in a color 2D image [16], the segmentation based on the depth is employed in this article. Typically, there is a significant difference between the effect of segmentation using the RGB image and depth image as shown in Fig. 4. In addition, different from the traditional image segmentation algorithms based on three color channels, only one depth channel information needs to be processed in the proposed method, which renders the speed of depth image segmentation to be at least three times faster than that of the RGB image segmentation. Due to the global image segmentation of the proposed method, only one operation is needed to achieve effective object segmentation. Thus the proposed mechanism significantly improves the overall efficiency.

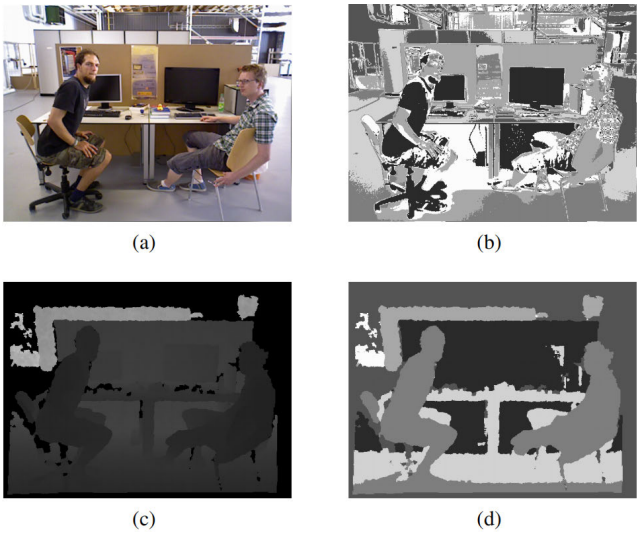


FIGURE 4. Segmentation results using a K = 6 value: (a)The RGB image;(b)The segmentation of the RGB image;(c)The depth image; and (d)The segmentation of the depth image.

In this approach, the parallel mechanism of image semantic detection and segmentation is used. To the best of our knowledge, this is the first time that semantic information has been applied in dealing with the issue of determining the value of K in the K-means algorithm when applied to multi-object segmentation.

1) CALCULATION OF CENTER DEPTH

RGB-D camera collects the color information and the corresponding depth information simultaneously (seen in Fig. 2). The center depth, D_o, of the identified objects are obtained by mapping the detection results to the depth image (Fig. 5(a)). However, considering that the depth information is susceptible to noise, there is a need to use a depth compensation region, described as the average depth of the region around the center of the object. The center depth compensation

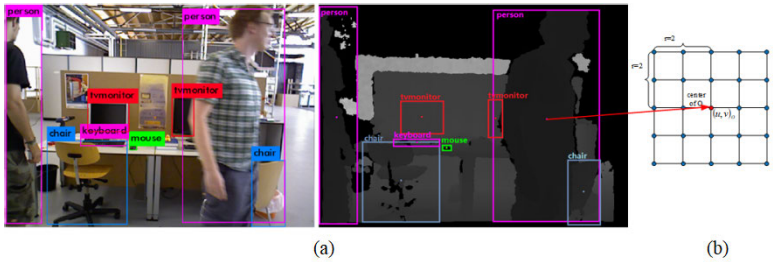


FIGURE 5. Calculation of the center depth: (a)Object detection results and the mapping in depth image, (b)The offset parameter.

region is herein represented by parameter $r \in Z^+$ and described by (4) and Fig. 5(b):

$$r = fix(\min(\sqrt{b_w}, \sqrt{b_h})) \tag{4}$$

where *fix* is the function of taking an integer in the direction of minus infinity [28]. Thus the computed parameter r can be used to adjusts the compensation region adaptively by the size of the object. Subsequently, the center depth can be calculated according to (5):

$$D_o = \frac{1}{(2r+1)^2} \sum_{v=r}^{v+r} \sum_{u=r}^{u+r} d(u, v) \tag{5}$$

in which $d(u, v)$ is the depth of $(u, v)_o$ in image coordinates.

2) CALCULATION OF K

With parameter D_o and the semantic information, the number of image segmentation can then be obtained via the pre-classification of the identified objects. If the center depth difference between objects identified in the same classification is less than a threshold, the corresponding objects can be segmented into the same segmentation. Because each image captured by RGB-D camera consists of at least two parts: foreground and background, the number of segmentation K of each image should be greater than or equal to 2. The K is calculated by the following three steps:

Step 1: set the center depth interval in the interested classification. For the interested classification C_{In}, the center depth of objects in C_{In} is divided into *pn* intervals. According to the depth interval, a new depth set is formed using (6) and (7).

$$D_{In} = \{D_{In\min} \ldots\ldots D_{In\max}\} \tag{6}$$

$$dr\,(C_{In}) = D_{In\max} - D_{In\min} \tag{7}$$

where $dr\,(C_{In})$ is the depth range of C_{In}.

Step 2: calculate the number of segmentation and choose the primary depth set for each classification. The difference between adjacent elements in the depth set D_{In} is computed by (8), where the value of k_{In} is set to 0 in the first iteration. Then count the number of segmentation in each depth set.

$$\Delta d = D_{Inx} - D_{In(x-1)}, \quad (2 \le x \le pn)$$
$$k_{In} = \begin{cases} k_{In} + 1, \Delta d \ge \ell \cdot dr\,(C_{In}) \\ k_{In}, \Delta d < \ell \cdot dr\,(C_{In}) \end{cases} \tag{8}$$

In order to select a more prominent depth interval for segmentation, the benchmark coefficient ℓ is set as 0.7 by experiments. Whenever k_{In} increases by 1, D_{Inx} is recorded as the primary depth $D_{Inp} = \{D_{Inmin}, D_{Inx} \ldots..\}$. Similar to the operation for C_{In}, the primary depth and the corresponding number of segmentation for C_I can be obtained by:

$$\Delta d = D_{Iy} - D_{I(y-1)} (y \ge 2)$$
$$k_I = \begin{cases} k_I + 1, \Delta d \ge \ell \cdot dr\,(C_I) \\ k_I, \Delta d < \ell \cdot dr\,(C_I) \end{cases} \tag{9}$$

It should be noted that the difference between (8) and (9) is the selection of thresholds. The coefficient ℓ is 0.5 in (8), as it has been proven that the lower threshold is conducive to the segmentation accuracy of objects from different classifications [29].

Step 3: traverse all the depth intervals in D_{In} and calculate the total K per (10), where the number of segmentation k_{Um} and k_U for the uninterested classification C_U are obtained from Step 1 and Step 2, respectively.

$$K = 2 + \sum_{p=1}^{n} k_{Ip} + \sum_{q=1}^{m} k_{Uq} + k_I + k_U \tag{10}$$

C. IMPROVED SELECTION OF INITIAL CLUSTER CENTERS

The calculation of K is addressed by pre-classification, then the improved initialization of cluster centers is elaborated in this section. In this article, an alternative selection method, maximin method [30], is explored to random selection of cluster centers. For a given image data $\chi = \{x_1, \ldots, x_N\}$, the first center for the K segments is arbitrarily chosen, and the remaining $K - 1$ centers are selected by the following iterative process:

At each iteration $l (l \in \{2, \ldots, K\})$, the lth center is determined as the point with the greatest minimum distance to the previously selected $l - 1$ centers. The selection of c_l is expressed as:

$$j = \underset{j \in \{1,\ldots,N\}}{\arg\max}\, d\left(x_j, c^{(l-1)}\right) \tag{11}$$

$$j^* = \min \text{ index } (j) \tag{12}$$

$$d\left(x_j, c^{(l-1)}\right) = \min_{h \in \{1,\ldots,l-1\}} d\left(x_j, c_h\right) \tag{13}$$

in which $d\left(x_j, c^{(l-1)}\right)$ denotes the distance between x_j and the center closest to it among the previous centers. Equation (13) shows the tie-breaking strategy adopted in the proposed method when the solution of (11) is not unique. Except for the first center, the choice of the remaining centers is deterministic, which leads to the inherited randomness of the

entire process. The mean of the given data, equation (14), is used as an improved method in (14).

$$c_1 = \frac{1}{N}\sum_{j=1}^{N} x_j \tag{14}$$

The procedure of the improved method is given below.

Step 1: Get the first center c_1 using (14) and set $d_1, \ldots, d_N = \infty$, in which d_j can be calculated by (13). The index of the next center l is set to 2, which is applied in Step 3.

Step 2: Initialize d_{max} between each point and corresponding nearest center to $-\infty$. For point x_j, if $d\left(x_j, c_h\right) < d_j$, update d_j as $d_j \leftarrow d\left(x_j, c_h\right)$. If $d_j > d_{\max}$, d_{max} and the index j^* also needed to be updated by $d_{\max} \leftarrow d_j$ and $j^* \leftarrow j$.

Step 3: Update the current center and l as $c_l \leftarrow x_*$ and $l \leftarrow l + 1$.

Step 4: Repeat Steps 2 and 3 until $i = K$.

D. OBJECT EXTRACTION

After image segmentation, each bounding box includes the segmentation of the corresponding object and the unwanted background. The general extraction methods distinguish object and background based on a single index, such as size, or contour [17], [23]. In this article, the size difference and connected domain analysis are both used to determine the ''most probable'' region for the detected object. Given the segmented image and bounding boxes B_{In}^{pn}, the pre-processing of deleting the negligible region is:

$$S_s(g) = \frac{r(g)}{r\left(B_{In}^{pn}\right)}$$
$$if \;\; S_s(g) \le S_T, \quad g = 0 \tag{15}$$

in which $r(g)$ and $S_s(g)$ denote the size and size score of segmentation g, respectively. S_T is set as 0.2 according to experiments. The parameter $g = 0$ means that segmentation g is too small to be a candidate for the matching object.

Ideally, each detected bounding box for each of the detected objects contains only one main object, so the segmentation with the one connected domain in the box is regarded as the matching object region. However, due to the image noise and partial occlusions, there may be cases where the number of the connected domain does not conform to the above. Different from the method of setting the threshold for the maximum connected domain, the proposed method employs the form of the score as follows:

$$S_c(g) = \frac{1}{cda(g)} \tag{16}$$

where $cda(g)$ is the amount of connected domain of segmentation g. The score $S_c(g)$ represents that the more connected domains, the less possible the region is to be the matching object in the box.

The size score performs highly accurate on the extraction of vertical or horizontal objects. However, as shown in Fig. 6, the results obtained by S_S in the case of the slant object identified is bad and in some cases it fails. Diagonal detection is essential to enhance the robustness of this method. Specifically, take the two highest values in S_D, max1 and max2, and

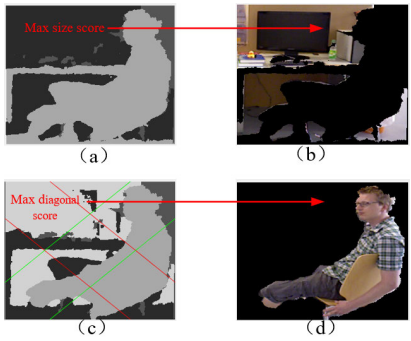


FIGURE 6. Results extracted by size score S_S and diagonal score S_D.

calculate the relative proportion by:

$$tj = \frac{S_s(\max 1)}{S_s(\max 2)}$$
$$tp = 1 - \sqrt{\frac{1}{1+tj}} \tag{17}$$

The diagonal region in the bounding box consists of r_{left} and r_{right}. The two diagonal score S_{Dleft} and S_{Dright} are calculated by the same operation of (16) in r_{left} and r_{right}, respectively. The total score function is defined by (18).

$$S(g) = S_s(g)S_c(g)S_{Dleft}(g)S_{Dright}(g) \tag{18}$$

The matching object for B_{In}^{pn} is then the segmentation that results in the highest score:

$$g = \max S(g) \tag{19}$$

IV. EXPERIMENTAL RESULTS AND DISCUSSION

All experiments used to test and validate the proposed approach were performed using a personal computer with Intel(R) Xeon(R) CPU E5-1630v3 3.70GHZ, 8GB. MATLAB R2019b and Visual Studio 2017 were used as the environment running on a Windows 10 desktop. The proposed method was compared with several popular methods, including one contour detection method and three unsupervised learning methods: SuperCut [12], Self-Organizing Maps (SOM) [13], Spectral method [14], Gaussian Mixture Model (GMM) method [15], and the variants of K-means. The above methods were tested on the TUM RGB-D dataset [31], Princeton Tracking Benchmark [32], and the Cornell Activity Datasets(CAD-60) [33], which provide several sequences of visual data in a variety of environments. Because of the extensive application of human detection and extraction, people in the images are defined as objects of interest.

A. IMAGE SEGMENTATION ALGORITHM RESULTS

In order to verify the effectiveness of determining K via pre-classification, 50 segmentation experiments were respectively carried out on the ten representative images in the dataset. The comparisons between the K_1 calculated by the pre-classification and the highest K_2 selected by the average SC are shown in Table 1. The highest SC values are shown in bold.

TABLE 1. The results of pre-classification and SC under different K_2.

Test image	K_1	K_2=2 SC	K_2=3 SC	K_2=4 SC	K_2=5 SC	K_2=6 SC	K_2=7 SC	K_2=8 SC	K_2=9 SC	K_2=10 SC
1	4	0.8261	0.8560	**0.9057**	0.8692	0.7886	0.7610	0.7406	0.7639	0.7387
2	4	0.8544	0.8194	**0.9214**	0.8891	0.8493	0.7918	0.7646	0.7756	0.7671
3	4	0.6812	0.8977	**0.9784**	0.9567	0.9563	0.9083	0.8692	0.8796	0.8539
4	4	0.7736	0.8458	**0.9427**	0.9077	0.8982	0.8462	0.7835	0.8025	0.7759
5	4	**0.9551**	0.8485	0.9393	0.8953	0.8606	0.8212	0.7467	0.7611	0.7852
6	4	0.6554	0.8686	**0.9522**	0.9023	0.8776	0.8624	0.8306	0.8112	0.7532
7	5	0.7009	0.9105	0.9562	**0.9692**	0.8874	0.8331	0.8043	0.7827	0.7707
8	5	0.6881	0.8934	0.9352	**0.9552**	0.8648	0.8266	0.8011	0.7876	0.7436
9	4	0.9587	0.9473	**0.9698**	0.9108	0.9348	0.8882	0.8671	0.8487	0.8398
10	4	**0.9135**	0.8821	0.9017	0.8542	0.8093	0.7975	0.7896	0.7787	0.7550

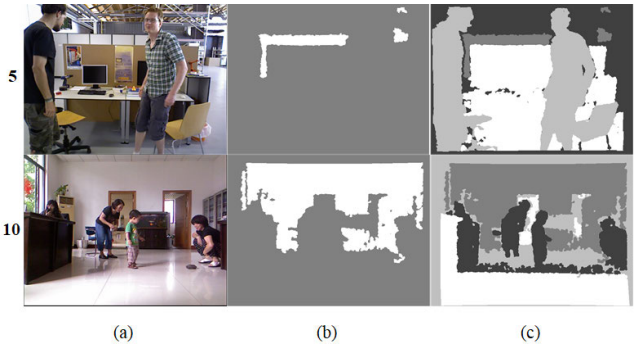


FIGURE 7. The segmentation results on images 5 and 10: (a) original image; (b) the segmentation of K = 2; (c) the segmentation of K = 4.

Since SC tends to decrease when K_2 is greater than 10, Table 1 only lists the validity index under the K_2 from the range of 2 to 10. Overall, the proposed pre-classification method can determine the value of K_1 with the maximum SC. However, the calculated K_1 does not match the optimal K_2 in some cases. For example, the SC of image 10 shows that the optimal segmentation number is 2, while the result of pre-classification is 4. The difference between the two segmentation results is shown in Fig. 7. Although the segmentation in Fig. 7(b) reaches the maximum SC(0.9135), the segmented image cannot completely separate the object from the background. In contrast, the results shown in Fig. 7(c) where a suboptimal SC(0.9017) achieves a more identifiable segmentation result. In other words, the pre-classification method loses 1.29% segmentation accuracy, but the segmented image is more recognizable for the subsequent object extraction. The same situation also happens in image 5. The main reason for this is that the validity index-based method does not consider the existence of multiple objects in the actual scene. In particular, when the object of interest includes similar attributes to other objects (color, depth, etc.), the validity index-based method divides the two objects into the same segment. When compared with the popular validity index-based method, the K value obtained by the proposed pre-classification method has better adaptability and provides better results in a greater range of image scenes.

The segmentation accuracy of the six algorithms (the proposed algorithm, Fuzzy C-means (FCM), conventional K-means, SOM, Spectral method, and GMM method) was also compared on the segmentation of the ten images. Table 2 shows the SC of these six algorithms. The best (highest) SC values and the fastest times are shown in bold.

In Table 2, FCM and the proposed algorithm are generally more effective than the other methods. They have similar effectiveness and, in most cases, FCM is slightly more accurate than the proposed algorithm. However, the processing time of the proposed method is about 4.83% of FCM. When compared with the fastest spectral method, the proposed method is roughly 26.20% slower, but the segmentation accuracy is improved by nearly 21.12%. At the same time, compared with the results of the conventional K-means algorithm, the accuracy and speed of the proposed method are improved by 3.12% and 20.36%, respectively. This demonstrates that the deterministic initialization based on the maximin method is beneficial to the accuracy of image segmentation and dramatically reduces the time processing.

Fig. 8 shows illustrative examples of the obtained results where it can be seen that the K-means based algorithms exhibit similar segmentation effects as the proposed method. In 50 independent runs, the conventional K-means algorithm achieved exactly the same segmentation effect and in some cases, a similar SC value was obtained. The indeterminacy of random initialization leads to the appearance of some poor segmentation with lower SC, which makes the overall performance of the conventional K-means algorithm not as good as the other two methods. In addition, it is evident that SOM is not available for depth image segmentation, and there is an overlapping of objects and background in the spectral method. Therefore, together with the results given in above, the experiments prove that the proposed method and the well-known FCM method produce very similar results both numerically and visually. The proposed method is faster and more stable than the other K-means based method as the maximin is a deterministic initialization approach.

B. OBJECT EXTRACTION RESULTS

When extracting the matching objects in the bounding box, the accuracy of image segmentation has a decisive impact on the overall results. Therefore, only the three most effective methods, FCM, GMM, and the proposed algorithm, are employed in the object extraction experiments. A state-of-the-art algorithm based on graph cuts, SuperCut, was also used to compare the extraction accuracy of different methods. Table 3 shows the average IOU of 10 images under the

TABLE 2. Segmentation effect evaluation of six algorithms.

Test Image	SOM		GMM		Spetral		K-means		FCM		Proposed algorithm	
	SC	Time/ms	SC	Time/ms	SC	Time/ms	SC	Time/ms	SC	Time/ms	SC	Time/ms
1	0.3025	6672	0.8971	2718	0.7458	**184**	0.8663	345	**0.9061**	6700	0.9057	298
2	0.2079	5875	0.9239	2211	0.7691	**200**	0.9020	400	**0.9228**	6727	0.9214	250
3	0.3787	5995	0.9077	1395	0.8054	**135**	0.9462	182	**0.9796**	6677	0.9784	161
4	0.2170	1620	0.9267	859	0.8135	**104**	0.9197	126	**0.9430**	1244	0.9427	117
5	0.3312	6514	0.9376	1690	0.7700	**171**	0.9039	347	**0.9402**	6948	0.9393	208
6	0.2960	6507	0.9448	1311	0.8478	**149**	0.9205	210	**0.9534**	6594	0.9522	189
7	0.2054	6302	0.8428	1862	0.7974	**146**	0.9503	293	**0.9700**	7379	0.9692	225
8	0.2090	6677	0.9336	1793	0.7832	**181**	0.9371	289	0.9204	8390	**0.9552**	210
9	0.1117	6446	0.9567	1488	0.7514	**139**	0.9420	216	**0.9711**	6540	0.9698	157
10	0.2389	6539	0.8962	1967	0.7216	**163**	0.8721	267	**0.9017**	6501	**0.9017**	238

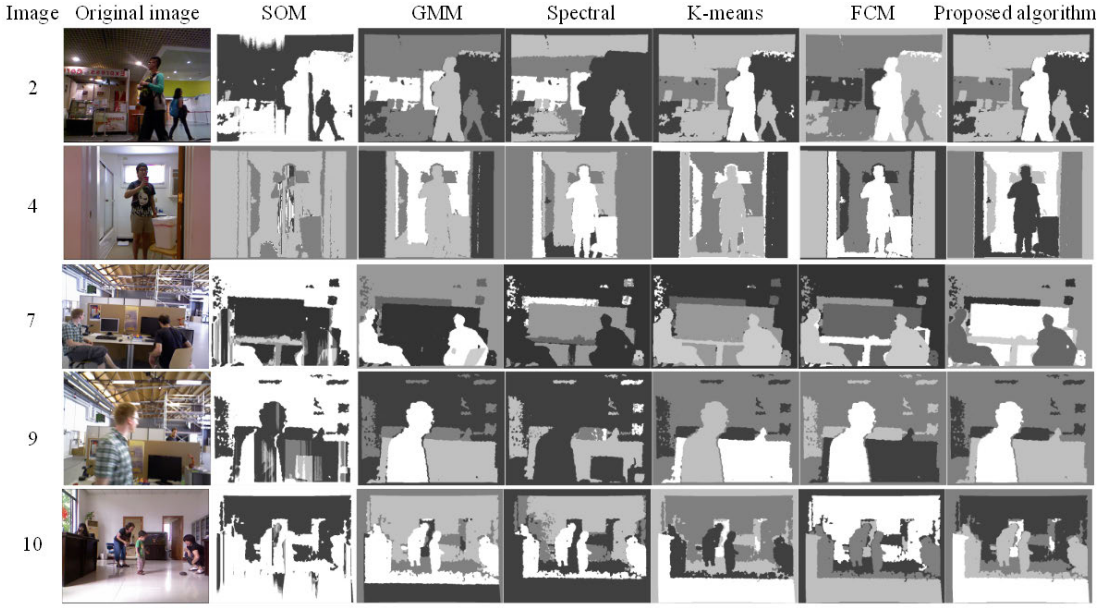


FIGURE 8. Image segmentation results of ten representative images by six algorithms.

TABLE 3. The average IOU of the object extraction results.

Test Image	1	2	3	4	5	6	7	8	9	10
GMM	0.9216	0.9078	0.6482	0.8473	0.9007	**0.9079**	0.7667	0.8893	0.8116	0.8489
FCM	0.9011	**0.9114**	0.9443	**0.8851**	**0.9169**	0.8821	0.8965	0.8767	**0.9126**	0.8590
SuperCut	0.8993	0.8946	0.9521	0.8736	0.8933	0.8684	0.7326	0.8172	0.7667	0.8398
Proposed method	**0.9233**	0.9096	**0.9447**	0.8844	0.9148	0.8949	**0.9142**	**0.8973**	0.9121	**0.8597**

four algorithms. Six representative examples of the obtained results are illustrated in Fig. 9.

Fig. 9 shows that the proposed object extraction method can achieve satisfactory results. In most cases, the IOU differences between FCM and the proposed algorithm are negligible in Table 3. However, the under segmentation and over segmentation of GMM and FCM lead to their inferior performance, which is consistent with the conclusion drawn in the previous section. For example, under segmentation of background caused by the GMM method in image 3 and over segmentation of the left person's face caused by FCM in image 7. The extraction accuracy of the proposed method is 0.71% and 7.26% higher than that of the FCM and GMM algorithm, respectively, showing better extraction performance. Because of the advantage of semantic detection in detecting small objects, some easily ignored objects, such as the right person in image 9 and the left person in image 10, can be extracted completely. Compared with the other three methods, the SuperCut achieves the highest extraction accuracy in images having a simple background, as shown in image 3(Fig. 9). However, in multi-object or complex background images, the average extraction accuracy of the proposed method is approximately 6.69% higher than that of the SuperCut method. One reason is that the use of depth information avoids the negative impact of strong edges on the extracted region. The other reason is that the

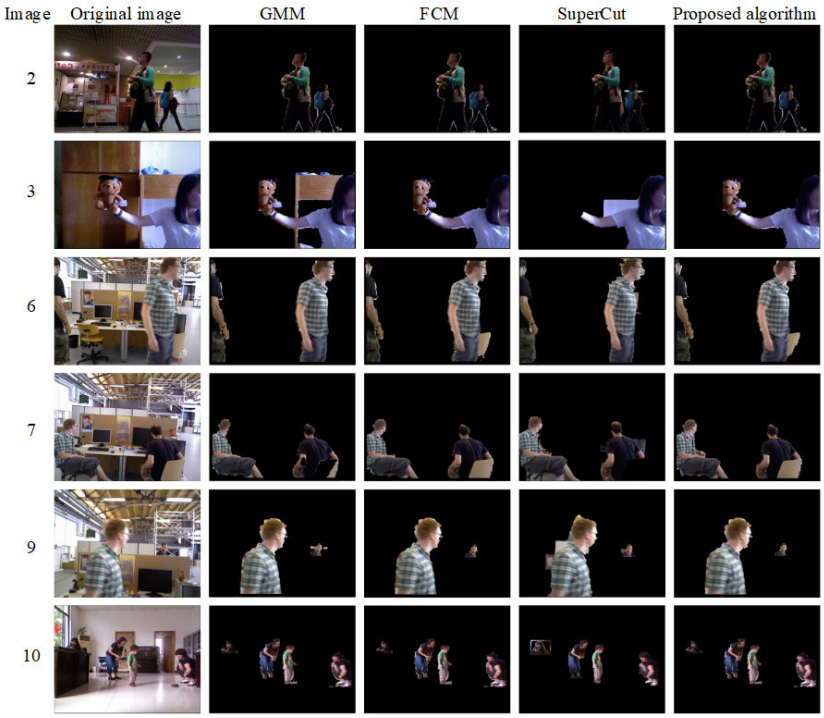


FIGURE 9. Object extraction results under the GMM, FCM, SuperCut and the proposed method.

pre-classification facilitates to improve the recognizability of image segmentation. Furthermore, the matching region is selected in accordance with multiple parameters including the size, the connected domain, and the diagonal detection parameters. In addition, the SuperCut takes more time to process multi-object images than the proposed method. This is due to the fact that the processing time of the SuperCut is proportional to the number of bounding boxes. These experiments demonstrated that the proposed method has better accuracy and stability in both simple and complex scenes.

V. CONCLUSION

In this article, an effective image object extraction method is proposed. The method comprises semantic object detection using YOLOv3 and the object extraction via image segmentation based on an improved K-means algorithm. The determination of the K value based on semantic information and depth information enables the K-means algorithm to determine the appropriate number of segmentation according to the actual scene. Such approach enhances the practicability of image segmentation in real scenes. Meanwhile, the selection of the initial cluster center based on the maximin method improves the determinacy and speed of the image segmentation process. Moreover, the score mechanism considering multiple factors also improves the accuracy and robustness of object extraction. Detailed experiments on test images showed that the proposed method outperforms many well-known methods with respect to image segmentation effect and object extraction accuracy.

In the algorithm proposed in this article, the accuracy of object extraction depends on the performance of YOLOv3 and the accuracy of depth information, which will lead to the vulnerability of the method in extreme cases. On the basis of not affecting the processing time of the algorithm, the way of combining contour detection with the proposed method to improve the accuracy of object extraction will be the focus of future research.

REFERENCES

[1] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[3] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware CNN model," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1134–1142.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[5] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

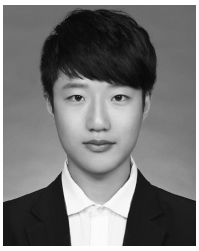
[7] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2980–2988.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[9] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[10] X. Wei, D. Wei, D. Suo, L. Jia, and Y. Li, "Multi-target defect identification for railway track line based on image processing and improved YOLOv3 model," *IEEE Access*, vol. 8, pp. 61973–61988, 2020.

[11] J. Peng, X. Shi, J. Wu, and Z. Xiong, "An object-oriented semantic SLAM system towards dynamic environments for mobile manipulation," in *Proc. IEEE/ASME Int. Conf. Adv. Intell. Mechatronics (AIM)*, Jul. 2019, pp. 199–204.
[12] S. Wu, M. Nakao, and T. Matsuda, "SuperCut: Superpixel based foreground extraction with loose bounding boxes in one cutting," *IEEE Signal Process. Lett.*, vol. 24, no. 12, pp. 1803–1807, Dec. 2017.
[13] F. J. López-Rubio, E. Domínguez, E. J. Palomo, E. López-Rubio, and R. M. Luque-Baena, "Selecting the color space for self-organizing map based foreground detection in video," *Neural Process. Lett.*, vol. 43, no. 2, pp. 345–361, Apr. 2016.
[14] K. Allab, L. Labiod, and M. Nadif, "Simultaneous spectral data embedding and clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6396–6401, Dec. 2018.
[15] X. Hou, T. Zhang, G. Xiong, Z. Lu, and K. Xie, "A novel steganalysis framework of heterogeneous images based on GMM clustering," *Signal Process., Image Commun.*, vol. 29, no. 3, pp. 385–399, Mar. 2014.
[16] Y. Sun, M. Liu, and M. Q.-H. Meng, "Improving RGB-D SLAM in dynamic environments: A motion removal approach," *Robot. Auto. Syst.*, vol. 89, pp. 110–122, Mar. 2017.
[17] K. Tian, J. Li, J. Zeng, A. Evans, and L. Zhang, "Segmentation of tomato leaf images based on adaptive clustering number of K-means algorithm," *Comput. Electron. Agricult.*, vol. 165, Oct. 2019, Art. no. 104962.
[18] K. A. Nazeer and M. P. Sebastian, "Improving the accuracy and efficiency of the k-means clustering algorithm," in *Proc. World Acad. Sci., Eng. Tech.*, 2009, pp. 308–312.
[19] D. Xia, F. Ning, and W. He, "Research on parallel adaptive Canopy-K-Means clustering algorithm for big data mining based on cloud platform," *J. Grid Comput.*, vol. 18, no. 2, pp. 263–273, Jan. 2020.
[20] X. Zheng, Q. Lei, R. Yao, Y. Gong, and Q. Yin, "Image segmentation based on adaptive K-means algorithm," *EURASIP J. Image Video Process.*, vol. 2018, no. 1, p. 68, Dec. 2018.
[21] Z. Khan, J. Yang, and Y. Zheng, "Efficient clustering approach for adaptive unsupervised colour image segmentation," *IET Image Process.*, vol. 13, no. 10, pp. 1763–1772, Aug. 2019.
[22] T. Skyler, C. M. Emre, and B. H. Krizia, "Fast color quantization using MacQueen's k-means algorithm," *J. Real-Time Image Process.* vol. 17, pp. 1609–1624, Oct. 2020, doi: 10.1007/s11554-019-00914-6.
[23] S. Saleh Al-amri, N. V. Kalyankar, and K. S. D., "Image segmentation by using threshold techniques," 2010, *arXiv:1005.4020*. [Online]. Available: http://arxiv.org/abs/1005.4020
[24] J. Fan, D. K. Y. Yau, A. K. Elmagarmid, and W. G. Aref, "Automatic image segmentation by integrating color-edge extraction and seeded region growing," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1454–1466, Oct. 2001.
[25] C. Li, C.-Y. Kao, J. C. Gore, and Z. Ding, "Minimization of region-scalable fitting energy for image segmentation," *IEEE Trans. Image Process.*, vol. 17, no. 10, pp. 1940–1949, Oct. 2008.
[26] L. E. Brito Da Silva, N. M. Melton, and D. C. Wunsch, "Incremental cluster validity indices for online learning of hard partitions: Extensions and comparative study," *IEEE Access*, vol. 8, pp. 22025–22047, 2020.
[27] R. C. de Amorim and C. Hennig, "Recovering the number of clusters in data sets with noise features using feature rescaling factors," *Inf. Sci.*, vol. 324, pp. 126–145, Dec. 2015.
[28] Q.-C. Mao, H.-M. Sun, Y.-B. Liu, and R.-S. Jia, "Mini-YOLOv3: Real-time object detector for embedded applications," *IEEE Access*, vol. 7, pp. 133529–133538, 2019.
[29] T. F. Gonzalez, "Clustering to minimize the maximum intercluster distance," *Theor. Comput. Sci.*, vol. 38, no. 2, pp. 293–306, 1985.
[30] R. J. Hathaway, J. C. Bezdek, and J. M. Huband, "Maximin initialization for cluster analysis," in *Proc. Iberoamerican Congr. Pattern Recognit.*, Iberoamerican, Mexico, 2006, pp. 14–26.
[31] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 573–580.
[32] S. Song and J. Xiao, "Tracking revisited using RGBD camera: Unified benchmark and baselines," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 233–240.
[33] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from RGB-D images," in *Proc. PAIR*, 2011, pp. 1–9.



HANXIAO RONG was born in Shandong, China, in 1992. He received the B.S. degree from Harbin Engineering University (HEU), Harbin, China, in 2016, where he is currently pursuing the Ph.D. degree with the College of Automation. His research interests include automatic underwater vehicles (AUVs), integrated navigation system based on multisensor fusion, and visual navigation.

ALEX RAMIREZ-SERRANO (Member, IEEE) received the B.Sc. degree in mechanical engineering from Mexico, the dual M.Sc. degrees in mechanical and aerospace engineering and computer science/artificial intelligence from USA and Mexico, respectively, and the Ph.D. degree in mechanical and industrial engineering from Canada. He is currently a full time Professor with the University of Calgary, where he has served on diverse roles, including the Former Director of the Manufacturing Program and the Director of the Graduate Program with the Department of Mechanical and Manufacturing Engineering. He is also the Founder and the Director of the Unmanned Vehicle Systems (UVS) Robotarium Research Laboratory, where he performs research and development activities in-ground, aerial, and humanoid robotics. He is also the Founder and the Chief Executive Officer (CEO) at 4Front Robotics, a Canadian-based robotics company developing UVS for operation in confined complex spaces. His research interest includes the design, control, navigation, and modeling of UVS.

LIANWU GUAN received the B.Sc. degree in measurement and control technology and instrumentation and the M.Sc. and Ph.D. degrees in navigation, guidance and control from Harbin Engineering University (HEU), Harbin, China, in 2011 and 2016, respectively. He has been a Lecturer and a Postdoctoral Researcher with the College of Automation, HEU, since 2017. His research interests include inertial measurement unit (IMU), inertial-based integrated navigation and localization systems with their application in pipeline inspection and localization robots, intelligent ski machine, unmanned autonomous vehicle, underwater autonomous vehicle, and other fields.

YANBIN GAO received the B.Sc. and master's degrees from the College of Automation, Harbin Engineering University, in 1985 and 1987, respectively. He is currently a full time Professor with the College of Automation, Harbin Engineering University. He served as the Director of Measurement and Control Technology with the Inertial Navigation Research Institute. His research interests include the development of signal processing, noise suppression, and navigation information conversion technology.