A Pretouch Perception Algorithm for Object Material and Structure Mapping to Assist Grasp and Manipulation Using a DMDSM Sensor

Fengzhi Guo, Shuangyu Xie, Di Wang, Cheng Fang, Jun Zou, and Dezhen Song

***Abstract*— We report a new material and structure mapping (MSM) algorithm to assist robotic grasping and manipulation. Building on our new sensor development, the algorithm has four main components: 1) detection of time-of-flight (ToF) durations for the dual modalities of optoacoustic (OA) and pulse-echo ultrasound (US), 2) contour reconstruction by fusing OA and US signals, 3) local noise filtering by checking local consistency of material and structure label (MSL), and 4) medium boundary searching that identifies class boundaries through two-staged clustering and boundary establishment using support vector machine (SVM) hyperplanes. We have implemented our algorithm and tested it with multiple common household items. The experimental results have successfully validated our algorithm design which shows that the average error of contour reconstruction is 0.05 mm and the true positive rate of MSL is over 98%.**

I. Introduction

Considering a robot attempting to grasp/manipulate an optically-transparent plastic bottle half-filled with water, the information on the bottle's material, shape, and water level is significant for the robot to plan the grasping. This is nontrivial because existing sensors and perception algorithms have difficulties in dealing with such challenging objects. As robots move from factory floors to a wide range of domestic environments, the perception capability of unknown objects is essential to achieve effective physical interactions. To enable object material and structure mapping (MSM), our group has devised new dual-modal and dual sensing mechanisms (DMDSM) sensors based on the dual-modal optoacoustic (OA) and pulse-echo ultrasound (US) signals. Time of flights (ToFs) and spectra of signals in both modalities are utilized to perceive the object distance and material/structure, respectively.

Here we report an algorithmic development in MSM with the dual-modal signals acquired by the DMDSM sensors [1]–[6]. Fig. 1 illustrates the performance of our algorithm in contour mapping and material/structure classification for a half-water-filled bottle. Our algorithm has four main components: 1) detection of time-of-flight (ToF) durations for the dual modalities, 2) contour reconstruction by fusing OA and US signals, 3) local noise filtering by checking local consistency of material and structure label (MSL), and 4) medium boundary searching that identifies medium boundaries through two-staged clustering and boundary establishment using support vector machine (SVM) hyperplanes.

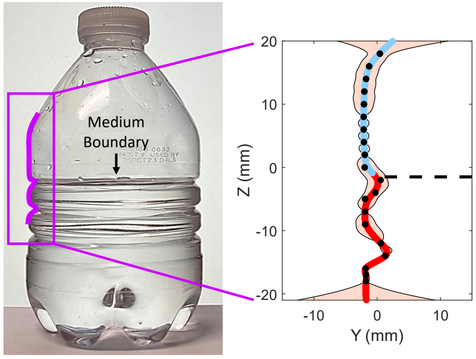


Fig. 1. MSM of a half-filled water bottle. Left: a water bottle with scanning trajectories (purple). Right: the MSM result. Black dots represent the reconstructed scan points while the blue and red curves represent detected plastic surfaces filled with air and filled water, respectively. The orange area represents the 95% confidence interval. The black dashed line represents the medium boundary.

We have implemented the proposed algorithm and tested it in physical experiments. The reconstructed contours indicate that the fusion of OA and US modalities has better positional accuracy than any single modality. The MSM algorithm has been tested with multiple household items, which has effectively provided MSL for all test objects in addition to contour reconstruction. It also shows that the overall true positive rate (TPR) is the highest (over 98%) in all tests.

II. Related Work

Grasping and manipulation of unknown objects is a grand challenge in robotics [7], [8]. For a known object, grasping has been well studied and widely applied on industry floors. When an object is unknown, successful grasping heavily depends on the perception to acquire accurate knowledge of the object's pose, shape, material attributes, and even internal structures. Such information helps estimate the form/force closure and finalize the grasping plan.

Although significant progress has been made in the past few years, unfortunately, existing state-of-the-art sensors still have difficulties in meeting all necessary perception requirements. Photometric stereo [9], lidar [10], and vision-based sensors [11], [12] suffer from the occlusion due to closing-in robot fingers [13] or having a close-range blind zone [14]. Tactile-based sensors [15], [16] and tactile-vision fusion [17], [18] require physical contact with the target object, which may change its poses or damage its surface. This may lead

F. Guo, S. Xie, D. Wang, and D. Song are with CSE Department, Texas A&M University, College Station, TX 77843, USA, Email: dzsong@cs.tamu.edu.

C. Fang and J. Zou are with ECE Department, Texas A&M University, College Station, TX 77843, USA, Email: junzou@tamu.edu.

This work was supported in part by National Science Foundation under IIS-2119549 and NRI-1925037, by Amazon Research Award, and by GM/SAE AutoDrive Challenge.

to either a slow grasping process or a complete grasping failure. Therefore, a compact and non-contact sensing and perception approach is more desirable.

To address the issues of existing sensors, we have developed DMDSM pretouch sensors [1]–[6], which integrate the pulse-echo ultrasound (US) and optoacoustic (OA) modalities together to interrogate the object information without contact. The DMDSM sensor has been iteratively developed to improve its sensing capabilities, cost effectiveness, and portability. The first-generation sensor provides the capabilities of near-distance ranging and material/thickness sensing on regular objects as well as optically and/or acoustically challenging targets [3]. The second-generation version has a much simplified sensor design and configuration with comparable ranging and sensing performances [4]. The most recent third generation version has a self-focused OA/US transceiver which can be integrated with a flat 2D scanning mirror for fast areal mapping and imaging of the object [6]. Although the DMDSM sensor hardware is developed and powerful, the dual modality perception algorithm has not been fully investigated, which is the focus of this paper.

Time-of-flight (ToF) estimation is essential for distance ranging. Our DMDSM sensor presents a unique challenge and opportunity due to dual-response acoustic signal properties. Different methods of ToF estimation have been developed based on the characteristics of the raw signals such as acoustic, optical, and seismic waves. In [19] and [20], the authors survey the commonly used acoustic signal ToF estimation methods. While threshold detection methods like amplitude thresholding and envelop fitting are fast and simple, they are not robust to noise compared to statistical models like cross-correlation and Akaike information criterion (AIC) approaches. While these existing acoustic ToF estimation methods are applicable for handling OA and US signals individually, we are interested in developing a model that can jointly estimate ToFs from both signals to provide better estimation results due to their strong correlation produced by the co-axial and co-registered signal property of the DMDSM sensor.

The fusion of US and OA modalities to obtain a better distance estimation is structurally similar to sensor fusion that combines multiple sensors such as lidars. These approaches often employ Maximum Likelihood Estimation based method [21] to merge the point cloud for mapping or adopt the neural network for object-based lidar point cloud fusion [22]. Unlike other sensors, DMDSM design enables sensor fusion at the device level since there is no perspective difference or synchronization issue across different modalities due to our unique sensor design. This allows us to reconstruct object contour with improved accuracy. Since scan points are often sparse, we adopt the Gaussian process [23] in point interpolation which is widely used to obtain continuous spatial contour with uncertainty characterization [17].

MSM creates labeled point clouds which looks like a clustering problem for point clouds. The problem can be solved by machine learning methods using spatial distance or semantic labels [24]. For lidar or RGB-D scans, point cloud segmentation using deep learning-based methods such as PointNet++ [25], [26] requires a large annotated dataset. However, the point cloud constructed from the DMDSM sensor is sparse and lacks sufficient information and dataset for the deep learning method. Therefore, unsupervised clustering techniques such as k-means [27], k-medoids [28] become our choice of clustering method to build the medium boundary searching algorithm.

III. Problem Formulation

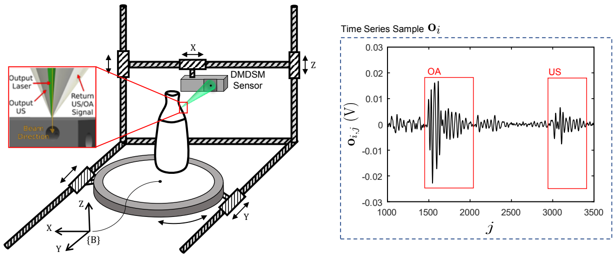


Fig. 2. Schematic illustration of the object scanning system. A representative time series of the received OA and US signals.

A. Background

Before formulating the problem, let us briefly review the design of our new DMDSM sensor and how the data are collected in our scanning system. These results have been presented in our previous publications [1]–[6]. For completeness, we brief the information here. The new fingertip-mounted DMDSM sensor is designed for robot manipulators to sense the distance, material type, and interior structure of an object without physical contact [1]–[4], [6]. Fig. 2 illustrates the object scanning system with the DMDSM sensor [5] and a representative sensor-received time series, where OA and US signals are simultaneously triggered by a single laser pulse. Light and ultrasound signals are coaxially and co-directionally arranged for active object scanning. For the OA modality, upon a laser pulse, OA signal in the form of a sound wave is generated by the induced transitory thermal expansion and received by the transducer embedded in the sensor. For the US modality, the same laser pulse is also incident onto an optoacoustic ultrasound transmitter to send out wideband acoustic waves to the object, and the reflected ultrasound is received by the same transducer. Therefore, triggered by a single laser pulse, the sensor-received time series consists of dual-modality signals. Their ToF signatures are used for distance ranging while the OA and US spectra are used for object material type and interior structure classification.

To further investigate the capability of the DMDSM sensor, we have also developed an object scanning system [5] to collect data from common household items. Fig. 2 illustrates the 4 Degrees of Freedom (DoFs) scanning system which fits an 1-DoF turntable on a 3-DoF linear stage. The hardware development enables us to further develop the perception algorithms that make better use of the dual-modality signals.

B. Problem Definition

1) Assumptions: To focus our attention on the most relevant issues, we have the following assumptions

a.1 The DMDSM sensor is pre-calibrated, and the OA and US signals are co-axial based on the sensor design. The sensor's signal noises are white and follow Gaussian distribution.

a.2 The DMDSM sensor's scanning resolution is much finer than that of material and structural distribution. Therefore, non-boundary points of vicinity share the same material and structural properties. We call this *local consistency* in material and structural class labels.

a.3 Medium boundary is where different material or structure regions meet. We assume the medium boundary only has two different media. This *2-medium* assumption covers the most common case and is the focus of this paper.

2) Inputs: Our algorithm takes in the data collected from the DMDSM sensor. As shown in Fig. 2, the DMDSM sensor is an active point-scanning sensor. When the sensor is at a scan point within its sensing range, the sensor can take voltage time series readings from the acoustic transducer. For each scan point i, the sensor receives consecutive M discrete voltage readings with the sampling frequency f_s to form the time series. Let us define the voltage readings as $\mathbf{o}_i = \{o_{i,j}\}_{j=1:M}$ where $o_{i,j} \in \mathbb{R}$ is the voltage reading from the transducer at discrete time j as shown in the right side of Fig. 2. For all N scan points, let us define $\mathcal{O} := \{\mathbf{o}_i\}_{i=1:N}$ as the transducer inputs to our algorithm.

In addition to the transducer inputs, the platform where the DMDSM sensor is mounted also provides sensor position $\mathbf{p}_i \in \mathbb{R}^3$ and sensor beam direction $\mathbf{v}_i \in \mathbb{S}^2$ at the i-th scan point, where $\mathbb{S}^2$ is the unit 2-sphere in 3D Euclidean coordinate system. Also, $\mathbf{p}_i$'s covariance matrix $\Sigma_{\mathbf{p},i}$ and $\mathbf{v}_i$'s covariance matrix $\Sigma_{\mathbf{v},i}$ are given since they are functions of the mounting platform. Both of which are in frame $\{B\}$, a fixed and right-handed 3D Euclidean system. It can be any other inertial frame of choice. In our scanning system, frame $\{B\}$ represents the object placement table frame. Its origin is at the intersection point of the rotation table's rotation axis and placement plane. Its initial X-, Y-, and Z- axes are parallel to the X, Y, Z directions of the 3D linear stage, respectively. All position variables are defined in $\{B\}$. For all N scan points, we can aggregate sensor position set as $\mathcal{P} := \{\mathbf{p}_i, \Sigma_{\mathbf{p},i}\}_{i=1:N}$ and their corresponding beam direction set as $\mathcal{V} := \{\mathbf{v}_i, \Sigma_{\mathbf{v},i}\}_{i=1:N}$.

3) Outputs: Our algorithm will output the reconstructed object contour points associated with material/structure attribute label. Each contour point is a 3D point in $\{B\}$ and defined as $\mathbf{x}_i \in \mathbb{R}^3$. For all N scan points, the reconstructed contour point set is $\mathcal{X} := \{\mathbf{x}_i\}_{i=1:N}$ and its corresponding covariance matrix set $\mathbf{\Sigma} := \{\Sigma_i\}_{i=1:N}$. In fact, we may generate more than N points in reconstruction through interpolation which results in $N_{\mathrm{E}}, N_{\mathrm{E}} > N$, points. To describe material/structure attribute, let us define label set $\mathcal{L} := \{1, .., H\}$ for H classes of material or structure types. Therefore, we can associate each point with a label as $(\mathbf{x}_i, \Sigma_i, l_i)$ where $l_i \in \mathcal{L}$ is the associated material and structure type label. The problem output is defined as a labeled point cloud set $\mathcal{M} := \{(\mathbf{x}_i, \Sigma_i, l_i)\}_{i=1:N_{\mathrm{E}}}$.

Therefore, the MSM problem can be defined as follows.

Definition 1 (MSM Problem Definition): Given time series set $\mathcal{O}$, sensor's position set $\mathcal{P}$, and sensor's beam direction vector set $\mathcal{V}$, compute an attribute-labeled point cloud $\mathcal{M}$.

IV. Algorithms

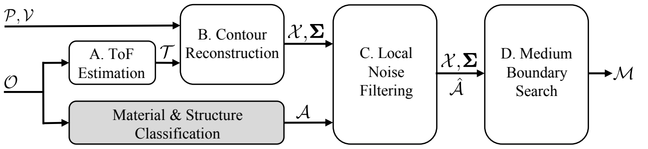

$\mathcal{P}, \mathcal{V}$
$\mathcal{O}$
A. ToF Estimation
$\mathcal{T}$
B. Contour Reconstruction
$\mathcal{X}, \mathbf{\Sigma}$
Material & Structure Classification
$\mathcal{A}$
C. Local Noise Filtering
$\mathcal{X}, \mathbf{\Sigma}$
$\hat{\mathcal{A}}$
D. Medium Boundary Search
$\mathcal{M}$

Fig. 3. Dual modality-based MSM algorithm pipeline.

Fig. 3 shows the overall algorithm pipeline. For each acoustic voltage time series $\mathbf{o}_i$ for the i-th scan point in $\mathcal{O}$, we estimate its ToF. Consequently, we obtain ToF set $\mathcal{T}$. With $\mathcal{T}$, the sensor's position set $\mathcal{P}$, the beam direction set $\mathcal{V}$, we can construct the external contour of the object which is represented as position point set $\mathcal{X}$ with uncertainty characterized by covariance matrix $\mathbf{\Sigma}$. In parallel [5], we apply the BOSS classifier [29] to each $\mathbf{o}_i$ so that it's material or structural label (MSL) distribution is obtained as multinomial probability distribution $\mathbf{a}_i$ which means $\mathbf{a}_i = [a_{i,1}, a_{i,2}, ..., a_{i,H}]^{\mathsf{T}} \in \mathbb{R}^H$, where $a_{i,j} \in [0, 1]$ is the MSL probability that point i is associated with attribute label j and $\sum_{j=1}^{H} a_{i,j} = 1$. Assembling across all scan points, we obtain the initial MSL probability distribution set $\mathcal{A} := \{\mathbf{a}_i\}_{i=1,...,N}$. Utilizing the local structure consistency property and position $\mathcal{X}$, we design a local noise filtering method to reduce detection noise in $\mathcal{A}$ to obtain its denoised version $\hat{\mathcal{A}}$. Finally, we fuse spatial and classification information to search for the medium boundary and build the attribute-labeled point cloud $\mathcal{M}$. We begin with ToF estimation for a detailed explanation.

A. ToF Estimation

For each scan point i, we need to know the sensor-detected distance which is determined by the ToFs of the OA and US signals (see Fig. 4). A typical $\mathbf{o}_i$ has two ToFs: an earlier one for OA and a later one for US. This is because the US signal traverses a round trip (transmitter-target-receiver), which is twice the travel distance of the laser-induced OA signal after a single trip from the target to a receiver, and their ToFs difference is much longer than their durations. Therefore, this allows us to segment out OA and US sub time series with the fixed starting and ending indices from $\mathbf{o}_i$ based on the signal length and measuring distance range. Consequently, we obtain the two sub time series for OA and US as $\mathbf{o}_{i,\phi_{\mathrm{OA}}} := \{o_{i,j}\}_{j \in \phi_{\mathrm{OA}}}$ and $\mathbf{o}_{i,\phi_{\mathrm{US}}} := \{o_{i,j}\}_{j \in \phi_{\mathrm{US}}}$ with time index range ϕ_{OA} and ϕ_{US}, respectively. The dashed red and green frames in Fig. 4 highlight the examples of two sub time series.

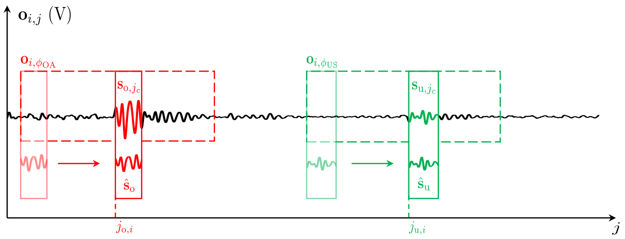

$\mathbf{o}_{i,j}$ (V)
$\mathbf{o}_{i,\phi_{\text{OA}}}$
$\mathbf{s}_{\text{o},j_c}$
$\hat{\mathbf{s}}_{\text{o}}$
$\mathbf{o}_{i,\phi_{\text{US}}}$
$\mathbf{s}_{\text{u},j_c}$
$\hat{\mathbf{s}}_{\text{u}}$
$j_{\text{o},i}$
$j_{\text{u},i}$
j

Fig. 4. Illustration of the ToF estimation algorithm. The red and green colors are corresponding to the ToF estimation processes of OA and US, respectively. The solid boxes are the sliding windows. The dashed boxes demonstrate $\mathbf{o}_{i,\phi_{\text{OA}}}$ and $\mathbf{o}_{i,\phi_{\text{US}}}$, respectively. When the window is sliding, $\mathbf{s}_{\text{o},j_c}$ and $\mathbf{s}_{\text{u},j_c}$ change while their corresponding templates $\hat{\mathbf{s}}_{\text{o}}$ and $\hat{\mathbf{s}}_{\text{u}}$ are fixed.

We apply cross-correlation (CC) [30] method to find the first peak or valley in both $\mathbf{o}_{i,\phi_{\text{OA}}}$ and $\mathbf{o}_{i,\phi_{\text{US}}}$. For the two sub time series, CC aligns each of them to a respective known template by considering phase shift, frequency, and amplitude. The alignment outputs the ToF time indices which are defined as $j_{\text{o},i}$ and $j_{\text{u},i}$ for $\mathbf{o}_{i,\phi_{\text{OA}}}$ and $\mathbf{o}_{i,\phi_{\text{US}}}$, respectively.

For the two known templates, each is a short series with a manually labeled ToF index. Both are pre-selected from historic data due to their clear forms for ToF peak or valley in OA or US signals. The short time series in the lower part of tall solid red and green rectangular frames in Fig. 4 illustrates the two known templates for OA and US modalities, respectively. Applying CC, finding the best matching time series index is to apply sliding window-based linear searching in $\mathbf{o}_{i,\phi_{\text{OA}}}$ and $\mathbf{o}_{i,\phi_{\text{US}}}$. Since the linear search methods in OA and US sub time series are the same, we only detail OA ToF search here. For the known template, we have its vector format as $\hat{\mathbf{s}} = [\hat{o}_{j_s}, ..., \hat{o}_{j_s}]^{\mathsf{T}}$, where j_s and j_e are the starting and ending indices of the OA template. The sliding window starting at $j_c \in \phi_{\text{OA}}$ with length m is denoted as $\mathbf{s}_{j_c} = [o_{i,j_c}, ..., o_{i,j_c+m}]^{\mathsf{T}}, o_{i,j_c} \in \mathbf{o}_{i,\phi_{\text{OA}}}$. Then, we calculate the OA peak/valley index $j_{\text{o},i}$ by maximizing the similarity between $\hat{\mathbf{s}}$ and $\mathbf{s}_{j_c}$

$$j_{\text{o},i} = \arg\max_{j_c} \hat{\mathbf{s}}^{\mathsf{T}} \mathbf{s}_{j_c}. \tag{1}$$

Similarly, we can find the peak/valley index $j_{\text{u},i}$ of US by using the US known template and $\mathbf{o}_{i,\phi_{\text{US}}}$.

Then we can obtain the ToF of OA $t_{\text{o},i} = j_{\text{o},i}/f_s$ and the ToF of US $t_{\text{u},i} = j_{\text{u},i}/(2f_s)$. Sometimes $\mathbf{o}_i$ may contain only one modality due to the material/structure properties of the target. If so, we just calculate the ToF for that existing modality. To summarize, after ToF estimation, we obtain ToF set $\mathcal{T} := \{t_{\text{o},i}\}_{i=1:N} \cup \{t_{\text{u},i}\}_{i=1:N}$.

B. Contour Reconstruction

ToFs from OA and US modalities offer two ways to estimate the distance from the DMDSM sensor and the scan point on the target object. Now let us show how to fuse them to obtain the most accurate point position estimation before we apply interpolation to generate the continuous contour for the object.

1) Scan point position estimation: For the i-th scan point, we use ToF of OA $t_{\text{o},i}$ to show how to obtain its position $\mathbf{x}_{\text{o},i}$ and its covariance $\Sigma_{\text{o},i}$ to characterize its uncertainty. Recall $\mathbf{p}_i$ is the sensor position and $\mathbf{v}_i$ is the scanning vector, Position estimation using OA $\mathbf{x}_{\text{o},i}$ is obtained using sound transmission,

$$\mathbf{x}_{\text{o},i} = \mathbf{p}_i + ct_{\text{o},i}\mathbf{v}_i, \tag{2}$$

where constant $c \in \mathbb{R}$ is the sound speed in air.

Next, let us derive $\Sigma_{\text{o},i}$. From the scanning system, we know the sensor's position and distribution $\mathbf{p}_i \sim \mathcal{N}(\bar{\mathbf{p}}_i, \Sigma_{\mathbf{p},i})$ where $\bar{\mathbf{p}}_i$ is the observed position and $\Sigma_{\mathbf{p},i}$ is its covariance matrix. We assume $\mathbf{v}_i \sim \mathcal{N}(\bar{\mathbf{v}}_i, \Sigma_{\mathbf{v},i})$ follows Gaussian Distribution with covariance $\Sigma_{\mathbf{v},i}$. ToF $t_{\text{o},i} \sim \mathcal{N}\left(\bar{t}_{\text{o},i}, \sigma^2_{\text{o},i}\right)$ also follows normal distribution with variance $\sigma^2_{\text{o},i}$ obtained by $\sigma^2_{\text{o},i} = g(\eta_i)$ where $\eta_{\text{o},i}$ is the signal-to-noise ratio (SNR). SNR is the maximum peak-valley voltage difference over the root of mean noise energy,

$$\eta_{\text{o},i} = \frac{\max_{o_{i,j_1}, o_{i,j_2} \in \mathbf{o}_{i,\phi_{\text{OA}}}} |o_{i,j_1} - o_{i,j_2}|}{\sqrt{\sum_{j=1}^{M_{\text{BG}}} o_{\text{BG},j}/M_{\text{BG}}}}, \tag{3}$$

where $\{o_{\text{BG},j}\}_{j=1:M_{\text{BG}}}$ is the background noise time series with length M_{BG}, and the ratio function $g(\cdot)$ is a staircase function obtained empirically from experiments. It is reasonable to assume $\mathbf{p}_i$, and $t_{\text{o},i}$ are independent.

With the covariance of $\mathbf{p}_i, \mathbf{v}_i$ and variance of $t_{\text{o},i}$, we perform the forward error propagation using (2). For notation simplicity, we denote $\Omega_i := [t_{\text{o},i}, \mathbf{v}_i^{\mathsf{T}}]^{\mathsf{T}}$ and its covariance matrix $\Sigma_{\Omega_i} = \text{diag}(\Sigma_{t_{\text{o},i}}, \Sigma_{\mathbf{v},i})$. Then, the covariance of $\mathbf{x}_{\text{o},i}$ is derived as:

$$\Sigma_{\text{o},i} = \Sigma_{\mathbf{p}_i} + J_{\Omega_i}\Sigma_{\Omega_i}J_{\Omega_i}^{\mathsf{T}}, \tag{4}$$

where J_{Ω_i} is the Jacobian matrix, $J_{\Omega_i} = \frac{\partial(ct_{\text{o},i}\mathbf{v}_i)}{\partial \Omega_i}$.

Similarly, we can obtain the scan point position $\mathbf{x}_{\text{u},i}$ and its covariance $\Sigma_{\text{u},i}$ using US ToF $t_{\text{u},i}$. Although sharing the same $\mathbf{p}_i$ and $\mathbf{v}_i$, its SNR $\eta_{\text{u},i}$ and stair function $g(\cdot)$ need to be established separately based on $\mathbf{o}_{i,\phi_{\text{US}}}$.

We are now ready to fuse the two modalities to obtain a more accurate estimation of the scan points' position by applying the Maximum Likelihood Estimation (MLE),

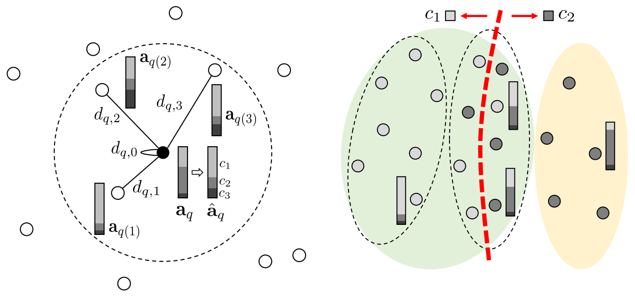
$$\min_{\hat{\mathbf{x}}_i} \left(\|\mathbf{x}_{\text{o},i} - \hat{\mathbf{x}}_i\|^2_{\Sigma_{\text{o},i}} + \|\mathbf{x}_{\text{u},i} - \hat{\mathbf{x}}_i\|^2_{\Sigma_{\text{u},i}} \right), \tag{5}$$

where $\hat{\mathbf{x}}_i$ is the i-th optimized scan point position and its covariance matrix is:

$$\Sigma_{\hat{\mathbf{x}},i} = \left(\Sigma^{-1}_{\text{u},i} + \Sigma^{-1}_{\text{o},i}\right)^{-1}. \tag{6}$$

It is worth noting that we may only have one modality in some cases. If so, we simply skip the fusion steps in (5) and (6) and directly use the results of the remaining modality.

2) Contour reconstruction with point interpolation: For object contour reconstruction, our scan points may not be dense enough. To overcome the issue, we employ Gaussian Process [23] to perform point interpolation. Recall that index set $\{1, ..., N\}$ is for the scan points. We define set $\{1, ..., N_{\text{E}}\}$ as the extended point contour point index set that contains both the scan points and the interpolated points and $N_{\text{E}} > N$ is the total point number for the extended contour point set. To avoid the mean function bending towards the GP mean away, we use the thin plate covariance as the kernel function [31]. Specifically, the uncertainty of the implicit surface for i-th observed point can be calculated by $\sigma_i = \text{tr}\left(\Sigma_{\hat{\mathbf{x}},i}\right)$, where $i \in \{1, 2, ..., N\}$. The covariance Σ_i of the i-th interpolated points is approximated by $\frac{\sigma_i}{3}\mathbf{I}_{3\times3}$, where $i \in \{N+1, N_{\text{E}}\}$ and $\mathbf{I}_{3\times3}$ is the 3×3 identity matrix. Therefore, we obtain the point position $\mathbf{x}_{1:N_{\text{E}}}$ and its covariance $\Sigma_{\mathbf{x}_{1:N_{\text{E}}}}$ in the point cloud $\mathcal{M}$.



(a) (b)

Fig. 5. Conceptual illustration of LNF and MBS. (a) LNF: the middle black solid circle is the denoising target and the hollow circles are other scan points. The grayscale bar represents the MSL probability distribution. The dashed circle shows the neighboring range. (b) MBS: Initial clustering generates the green and yellow shaded clusters. The second stage of clustering further partitions the green cluster into two smaller clusters in dashed eclipses. The last step is medium boundary generation. The class label indices $c_1, c_2 \in \mathcal{L}$.

C. Local Noise Filtering

From Fig. 3, the material and structure classification method [5] outputs label l_i for each scan point. However, the classification may be noisy. According to Assumption a.2 in Sec. III-B.1, we know that labels of points in the vicinity should be the same if they are not boundary or close-to-boundary points. Based on the local consistency assumption and point positions obtained from contour reconstruction, we propose a local noise filtering (LNF) algorithm to de-noise and update the MSL probability distribution $\mathcal{A}$. At this step, we do not differentiate boundary points from non-boundary points which will be dealt with later.

Fig. 5(a) illustrates the LNF algorithm. For a given scan point $q \in \{1:N\}$, we can compute its n nearest neighbor points. For the p-th nearest neighbor point, $0 < p \leq n$, let $d_{p,q} = ||\mathbf{x}_{q(p)} - \mathbf{x}_q||_2$ denote the Euclidean distance between the target point and p-th neighboring point according to the ascending distance where index mapping function $q(p)$ outputs the index of the p-th nearest point of point q in the original index set $\{1:N\}$. Distance to point q itself is $d_{q,0} = 0$. Let vector $\mathbf{a}_{q(p)} \in \mathcal{A}$ be the MSL multinomial probability distribution of the p-th neighboring point.

LNF algorithm is the weighted averaging of the neighboring points' MSL probability distributions where weights should be decreasing function of the Euclidean distance between the target and neighbor. A smaller $d_{p,q}$ means the class correlation between the two points is stronger. To characterize it, we employ a sigmoid function as weight,

$$w_{p,q} := \frac{1}{1+\exp\left(d_{p,q}\right)}. \tag{7}$$

The sigmoid function has two desirable benefits: 1) it is a decreasing function of $d_{p,q}$ which means local consistency is stronger when points are closer, and 2) it significantly reduces the influence of points far from scan point q to ignoble level. Using (7), we calculate the weights for all n nearest points. Let $\mathbf{w}_q := [w_{0,q}, w_{1,q}, ..., w_{n,q}]^{\mathsf{T}}$ be the resulting weight vector. Also, we denote the initial MSL probability corresponding to the n nearest points as a matrix $\mathbf{A}_{q(0:n)} = \left[\mathbf{a}_q, \mathbf{a}_{q(1)}, ..., \mathbf{a}_{q(n)}\right]$ with dimension $H \times (n+1)$. The refined MSL probability distribution vector $\hat{\mathbf{a}}_q$ can be obtained by the weighted average:

$$\hat{\mathbf{a}}_q = \frac{\mathbf{A}_{q(0:n)}\mathbf{w}_q}{\left|\left|\mathbf{A}_{q(0:n)}\mathbf{w}_q\right|\right|_1}, \tag{8}$$

where $||\cdot||_1$ is the L1-norm. Eq. (8) is just the normalization of the MSL probability distribution.

We apply (8) to all points to obtain the updated MSL probability distribution $\hat{\mathcal{A}}$. Although the LNF method can effectively filter the classification noise for interior points of every single medium region, it can also negatively affect the near-boundary points' MSL probability distribution. To avoid the negative effect on the near-boundary points and assign correct labels to points, we propose the medium boundary searching method in the next section.

D. Medium Boundary Searching

The purpose of medium boundary search is to cluster MSL and generate the medium boundary. This is a three-step process (see Fig. 5(b)): initiate clustering, second stage clustering, and boundary generation along with the class label.

1) Initial clustering: In the first step, we apply Partitioning Around Medoids (PAM) [32] method to the scan points to generate k_c clusters based on MSL and point position. For each cluster, we can check if its member points have the same label with the index corresponding to the maximum probability of its MSL distribution in $\hat{\mathcal{A}}$. If not, then the cluster is a boundary-crossing cluster (BCC). The step generates $k_b \leq k_c$ out of k_c BCCs. The green shaded cluster in Fig. 5(b) is a BCC.

2) Second stage clustering: For each BCC, we conduct k_s-stage bisecting PAM to reduce its size. Fig. 5(b) shows that two dashed ellipse enclosed clusters are the outcome of the step. The old BCC is split into a non-BCC and a smaller BCC. At the end of the second step, for each member point in BCC, we replace its MSL probability distribution $\hat{\mathbf{a}}_i$ with $\mathbf{a}_i$ in $\hat{\mathcal{A}}$. This means that we undo LNF algorithm to restore the original MSL probability because we know that LNF does not apply to boundary or close-to-boundary points.

Both the first two steps depend on the PAM clustering algorithm which requires a cost function to evaluate point similarity. In our problem, we consider the similarity of two points with index p, q determined by two terms: the Euclidean distance and the difference of the corresponding refined MSL probability vectors. We propose the new cost function $C_{p,q}$ as:

$$C_{p,q} = ||\mathbf{x}_p - \mathbf{x}_q||_2 + \lambda C_{\text{KL}}(\mathbf{a}_p, \mathbf{a}_q), \tag{9}$$

where $||\cdot||_2$ is L2 norm, $\Delta\mathbf{x} = \mathbf{x}_p - \mathbf{x}_q$, $C_{\text{KL}}(\mathbf{a}_p, \mathbf{a}_q)$ is the KL divergence [33] between MSL probability distribution $\mathbf{a}_p, \mathbf{a}_q \in \hat{\mathcal{A}}$

$$C_{\text{KL}}(\mathbf{a}_p, \mathbf{a}_q) = \sum_{h=1}^{H} a_{p,h} \log(\frac{a_{p,h}}{a_{q,h}}), \tag{10}$$

and $\lambda \in \mathbb{R}$ is a factor to balance the effects of the position and classification results of the scan points. With larger λ, the distance metric has more confidence on $\hat{\mathcal{A}}$.

3) Boundary generation and class labeling: Finally, we estimate the boundary within the cluster by applying an SVM [34] to estimate the hyperplane as boundary separation to determine its final labels l_i. The medium boundary (e.g. the red dashed curve in Fig. 5(b)) is the hyperplane from SVM.

After the first two steps, we label each scan point with index $i \in \{1, ..., N\}$ as

$$l_i = \arg\max_h \hat{\mathbf{a}}_i, \tag{11}$$

where $\hat{\mathbf{a}}_i = [\hat{a}_h]^{\mathsf{T}}_{h=1:H}$.

Using the labels from the scan points, we further build the medium boundaries to separate all the points including scanned and interpolated points to obtain final labels $\{l_i\}_{i=1:N_{\text{E}}}$. To generate the boundaries, we apply standard SVM with the input of scan points' 3D position $\mathbf{x}_i$ and label l_i, $i \in \{1, ..., N\}$, to calculate the hyperplane $\mathbf{k}^{\mathsf{T}}\mathbf{x} = b$ by solving the following optimization

$$\begin{aligned} \min_{\mathbf{k},b} \quad & ||\mathbf{k}||_2^2 \\ \text{s.t.} \quad & l_i(\mathbf{k}^{\mathsf{T}}\mathbf{x}_i - b) \geq 1, \quad \forall i \in \{1, ..., N\} \end{aligned} \tag{12}$$

The hyperplane as the optimization result is used as a medium boundary to separate points with either label c_1 or c_2, where c_1 and c_2 represent the majority label of the scan points in two sub-regions as shown in Fig. 5(b). Then we use the SVM's hyperplane to assign all points (including both scan points and interpolated points) with index $i \in \{1, ..., N_{\text{E}}\}$ by the label from two candidate label indices c_1 and c_2 is listed as follows:

$$l_i := \begin{cases} c_1 & \text{if } \mathbf{k}^{\mathsf{T}}\mathbf{x}_i \geq b \\ c_2 & \text{otherwise} \end{cases} \tag{13}$$

With each point assigned to a label, we create the point cloud $\mathcal{M}$.

V. Experiments

We have implemented the proposed MSM algorithm using Matlab™ and tested both the contour reconstruction and MSL capability which are key outputs of the MSM algorithm.

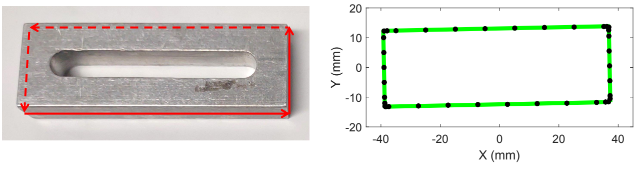
A. Contour Reconstruction

The contour reconstruction experiment validates our ToF-based ranging and dual modality-based scan point estimation in Sec.IV-B. Here we use an aluminum slot shown in Fig. 6 which also has both OA and US responses. The ground truth shape is obtained from caliper measurement with sufficient accuracy. We scan along the red trajectories with $N = 37$ scan points in total. The reconstruction error is the Euclidean distance from point position $\mathbf{x_i}$ to the contour. Tab. I show the average reconstruction error and their standard deviation with the best results in bold font. Since OA only and US only reconstruction results have been presented in our previous work [5], we can directly compare them with results obtained from MLE-based OA and US fusion in (5). It is clear that both the reconstruction error and its standard deviation have been reduced when the MLE-based OA and US fusion approach is used.

TABLE I

CONTOUR RECONSTRUCTION USING DUAL MODALITIES

Modality	Avg. Recons. Err. (mm)	Sta. Dev. (mm)
OA [5]	0.06	0.06
US [5]	0.15	0.11
OA&US Fusion	**0.05**	**0.05**



(a)

(b)

Fig. 6. (a) The aluminum slot used in contour reconstruction test. (b) Reconstruction results: the blacks points are the scan points and the green curve is the reconstructed contour (best viewed in color).

As another experiment, we scan a coke bottle to show that employing multi-scan allows us to perform full 3D reconstruction if needed. The results are in Fig. 7. The yellow curve is the ground-truth medium boundary. Green dots represent the contour interpolation results while red and blue colored dots are the reconstruction of scan points. The 3D contour construction is successful. It is worth noting that the red and blue colored dots are corresponding to the scanning points on the bottle regions filled with air and regions filled with coke, respectively. Those labels are from the initial output of BOSS classifier. They do contain wrongfully classified labels which will be corrected in the next section.

B. Material and Structure Labeling and Mapping

The most unique part of our algorithm is its ability to label and map material and structure. To validate that we continue to use the coke bottle example before we expand the experiment to other common household items.

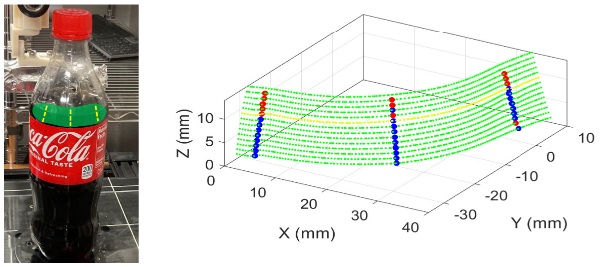
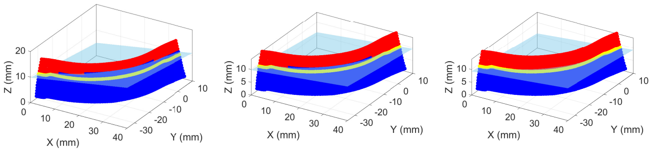


Fig. 7. Mult-scan contour reconstruction of a coke bottle. (a) The coke bottle with the target reconstruction area (green) and the scanning trajectories (yellow). (b) A tri-scan reconstruction result with point interpolation (best viewed in color).

1) Coke bottle example: Fig. 8 illustrates the mapping results for three different setups. Again, red and blue colored regions are corresponding to the bottle region filled with air and the bottle region filled with coke, respectively. The yellow line is the ground truth of the medium boundary. Fig. 8(a) shows the resulting region map with only LNF. Here, we also employ SVM hyperplanes in (13) to partition all points without running the first two MBS steps. Fig. 8(b) shows the map with only MBS. The LNF step is turned off. It is clear that both Fig. 8(a) and Fig. 8(b) have a blue region above the ground truth yellow medium line which are areas of false classifications. Fig. 8(c) employs the complete algorithm pipeline with both LNF and MBS which yields the best results. It is not difficult to see that both LNF and MBS steps are necessary. In the rest of the tests, we will use the full algorithm pipeline.

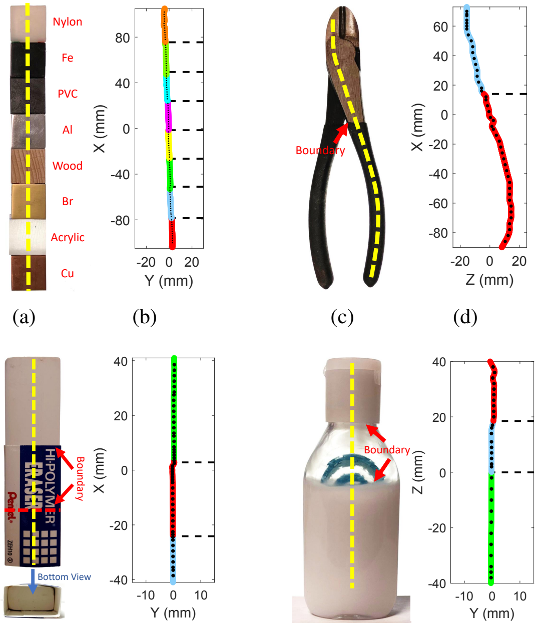


(a) (b) (c)

Fig. 8. Algorithm outputs under different module selection: (a) LNF and SVM only. (b) MBS only. (c) LNF and MBS.

2) MSM experiments with household items: Besides the water bottle in Fig. 1 and the coke bottle, we have also experimented with four more objects: a group of blocks made of different materials, a wire cutter, an eraser halfway out of its paper wrapping, and a conditioner bottle. Fig. 9 shows the MSM results for both contour construction and MSL.

Combined with the two earlier objects, the six test items have different materials (MT) and inner fillings (IF), and, as a result, have different modality responses: some only have US responses while others have both US and OA responses. We further compare algorithm MSL ability using TPR metric. The overall comparison is in Tab. II. In the table, initial TPR refers to the results directly out of the BOSS classifier. LNF TPR counts the TPR after LNF without using MBS. MBS TPR refers to the algorithm with LNF turned off. The last row refers to the full pipeline. It is not surprising the MSM algorithm using full pipeline configuration consistently outperforms its counterparts which shows that our algorithm design is effective.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 9. Material and structure mapping results for 4 different objects. (a, b) multi-material blocks. (c, d) wire cutter, (e, f) eraser, and (g, h) conditioner bottle. The scanning trajectories are shown in the yellow dashed curves. The different colors in (b, d, f, h) are just used to differentiate the different materials or structure types with horizontal dashed black lines indicating the detected medium boundaries.

VI. Conclusion and Future Work

We reported new material and structure mapping algorithm for a fingertip mounted DMDSM sensor designed to assist in grasping/manipulation of unknown objects. The new algorithm enabled our DMDSM sensor to perform contour reconstruction using dual OA and US modalities and provide accurate material and structure labeling for household items. We implemented the algorithm and tested it with common household objects. The experimental results confirmed our design and showed that contour reconstruction accuracy and the true positive rate are 0.05 mm and over 98%, respectively. In the future, we will continue to improve both the sensor design and algorithm design to improve speed and accuracy.

Acknowledgment

We are grateful to S. Yeh, A. Kingery, C. Qian, and Y. Jiang for their inputs and feedback.

References

[1] C. Fang, D. Wang, D. Song, and J. Zou, "Toward fingertip non-contact material recognition and near-distance ranging for robotic grasping," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4967–4974.

[2] ——, "Fingertip non-contact optoacoustic sensor for near-distance ranging and thickness differentiation for robotic grasping," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 894–10 899.

TABLE II

TPR comparison for 6 test objects. The best results are in bold font.

Name	MT & IF Class No.	Modalities	Scan Type	#Scan Pts.	Initial TPR	LNF TPR	MBS TPR	LNF+MBS TPR
Coke bottle	2 (IF)	US	Triple scan	43	95.3%	93.0%	97.7%	**100%**
Water bottle	2 (IF)	US	Single scan	20	90%	95%	**100%**	**100%**
Multi-material blks.	8 (MT)	US & OA	Single scan	98	93.9%	**98.0%**	**98.0%**	**98.0%**
Wire cutter	2 (MT)	US & OA	Single scan	45	93.3%	**100%**	97.8%	**100%**
Conditioner bottle	3 (MT & IF)	US	Single scan	30	96.7%	**100%**	**100%**	**100%**
Half-wrapped eraser	3 (MT & IF)	US	Single scan	49	95.9%	98.0%	**100%**	**100%**

[3] ——, “Fingertip pulse-echo ultrasound and optoacoustic dual-modal and dual sensing mechanisms near-distance sensor for ranging and material sensing in robotic grasping,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 14 105–14 111.

[4] ——, “The second generation (g2) fingertip sensor for near-distance ranging and material sensing in robotic grasping,” in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 1506–1512.

[5] D. Wang, F. Guo, C. Fang, J. Zou, and D. Song, “Design of an object scanning system and a calibration method for a fingertip-mounted dual-modal and dual sensing mechanisms (dmdsm)-based pretouch sensor for grasping,” in *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2022, pp. 341–347.

[6] C. Fang, S. Li, D. Wang, F. Guo, D. Song, and J. Zou, “The third generation (g3) dual-modal and dual sensing mechanisms (dmdsm) pretouch sensor for robotic grasping,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)(Accepted)*. IEEE, 2023.

[7] M. T. Mason, *Mechanics of robotic manipulation*. MIT press, 2001.

[8] M. Ciocarlie, K. Hsiao, E. G. Jones, S. Chitta, R. B. Rusu, and I. A. Şucan, “Towards reliable grasping and manipulation in household environments,” in *Experimental Robotics: The 12th International Symposium on Experimental Robotics*. Springer, 2014, pp. 241–252.

[9] A. Hertzmann and S. M. Seitz, “Example-based photometric stereo: Shape reconstruction with general, varying brdfs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1254–1264, 2005.

[10] Y. Lu, J. Lee, S.-H. Yeh, H.-M. Cheng, B. Chen, and D. Song, “Sharing heterogeneous spatial knowledge: Map fusion between asynchronous monocular vision and lidar or other prior inputs,” in *Robotics Research: The 18th International Symposium ISRR*. Springer, 2020, pp. 727–741.

[11] R. Girdhar, D. F. Fouhey, M. Rodriguez, and A. Gupta, “Learning a predictable and generative vector representation for objects,” in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*. Springer, 2016, pp. 484–499.

[12] H. Cao, G. Chen, Z. Li, Y. Hu, and A. Knoll, “Neurograsp: multimodal neural network with euler region regression for neuromorphic vision-based grasp pose estimation,” *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.

[13] C. E. Smith and N. P. Papanikolopoulos, “Vision-guided robotic grasping: Issues and experiments,” in *Proceedings of IEEE International Conference on Robotics and Automation*, vol. 4. IEEE, 1996, pp. 3203–3208.

[14] M.-C. Amann, T. M. Bosch, M. Lescure, R. A. Myllylae, and M. Rioux, “Laser ranging: a critical review of unusual techniques for distance measurement,” *Optical engineering*, vol. 40, pp. 10–19, 2001.

[15] W. Yuan, S. Dong, and E. H. Adelson, “Gelsight: High-resolution robot tactile sensors for estimating geometry and force,” *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[16] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, and N. F. Lepora, “The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies,” *Soft robotics*, vol. 5, no. 2, pp. 216–227, 2018.

[17] S. Suresh, Z. Si, J. G. Mangelson, W. Yuan, and M. Kaess, “Shapemap 3-d: Efficient shape mapping through dense touch and vision,” in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 7073–7080.

[18] T. Li, X. Shu, K. Yang, C. Wu, and G. Chen, “Robot grasping stability prediction network based on feature-fusion and feature-reconstruction of tactile information,” in *2022 IEEE International Conference on Mechatronics and Automation (ICMA)*. IEEE, 2022, pp. 651–656.

[19] J. C. Jackson, R. Summan, G. I. Dobie, S. M. Whiteley, S. G. Pierce, and G. Hayward, “Time-of-flight measurement techniques for airborne ultrasonic ranging,” *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 60, no. 2, pp. 343–355, 2013.

[20] Z. Qiu, Y. Lu, and Z. Qiu, “Review of ultrasonic ranging methods and their current challenges,” *Micromachines*, vol. 13, no. 4, 2022. [Online]. Available: https://www.mdpi.com/2072-666X/13/4/520

[21] J. Jiao, H. Ye, Y. Zhu, and M. Liu, “Robust odometry and mapping for multi-lidar systems with online extrinsic calibration,” *IEEE Transactions on Robotics*, vol. 38, no. 1, pp. 351–371, 2022.

[22] J. Jiao, P. Yun, L. Tai, and M. Liu, “Mlod: Awareness of extrinsic perturbation in multi-lidar 3d object detection for autonomous driving,” in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 10 556–10 563.

[23] C. E. Rasmussen and C. K. Williams, “Gaussian processes in machine learning,” *Lecture notes in computer science*, vol. 3176, pp. 63–71, 2004.

[24] E. Grilli, F. Menna, and F. Remondino, “A review of point clouds segmentation and classification algorithms,” *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, p. 339, 2017.

[25] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17, Red Hook, NY, USA, 2017, p. 5105–5114.

[26] L. Landrieu and M. Simonovsky, “Large-scale point cloud semantic segmentation with superpoint graphs,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4558–4567.

[27] D. I. Kim and G. S. Sukhatme, “Semantic labeling of 3d point clouds with object affordance for robot manipulation,” in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 5578–5584.

[28] *Partitioning Around Medoids (Program PAM)*. John Wiley Sons, Ltd, 1990, ch. 2, pp. 68–125. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470316801.ch2

[29] P. Schäfer, “The boss is concerned with time series classification in the presence of noise,” *Data Mining and Knowledge Discovery*, vol. 29, pp. 1505–1530, 2015.

[30] S. M. Simkin, “Measurements of velocity dispersions and doppler shifts from digitized optical spectra,” *Astronomy and Astrophysics*, vol. 31, p. 129, 1974.

[31] O. Williams and A. Fitzgibbon, “Gaussian process implicit surfaces,” in *Gaussian Processes in Practice*, 2006.

[32] L. Kaufman and P. J. Rousseeuw, “Partitioning around medoids (program pam),” *Finding groups in data: an introduction to cluster analysis*, vol. 344, pp. 68–125, 1990.

[33] S. Kullback and R. A. Leibler, “On information and sufficiency,” *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[34] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, pp. 273–297, 1995.