Extended Target Tracking in Human–Robot Coexisting Environments via Multisensor Information Fusion: A Heteroscedastic Gaussian Process Regression-Based Approach

Pritam Paral, *Member, IEEE*, Amitava Chatterjee, *Senior Member, IEEE*, Anjan Rakshit, and Sankar K. Pal, *Life Fellow, IEEE*

***Abstract*—In this article, a new systematic approach to sensor fusion and state estimation is proposed for extended target tracking in human–robot coexisting environments. The developed method, called *human feature-based extended target tracking via multisensor information fusion* (HFBETT-MSIF), can assimilate information from the onboard camera and sonar sensor of a mobile robot in a unified way, during tracking of a pair of human shoes. A novel generalized measurement model containing the complete information of the human target is formulated for both sensors, thus rendering the tracking system potentially robust to the failure of any one sensor. The study illustrates how *heteroscedastic Gaussian process* (HGP) regression can be used to derive the measurement model. It also develops an advanced HGP model, called *bias-minimized most likely HGP*, to interpret the real-world shoe-contour data subjected to heteroscedastic noise. Performance evaluations conducted for real-life shoe tracking demonstrate the supremacy of the HFBETT-MSIF.**

***Index Terms*—Extended target tracking, heteroscedastic Gaussian process (HGP) regression, human–robot coexisting environment, multisensor fusion, recursive state estimation.**

I. Introduction

IN RECENT years, it has increasingly been found that various human-following or tracking approaches, with a view to overcoming the limitations of single-sensor systems [1], [2], [3], are making use of a sensor fusion mechanism that involves multiple distinct sensing units [4], [5], [6], with each sensing module being designated to perform a specific job. The candidate sensors utilized for such applications are vision sensors, sonar sensors, laser range finders (LRFs) etc., and the fusion approaches predominantly use probabilistic methods, such as extended Kalman filter [4], covariance intersection filter [5], and particle filter [6]. Mostly, such sensor fusion mechanisms require integration of vision sensing with range sensing, and, in most such schemes, it is not possible to separate the integration of visual data and range information, since either one of them furnishes only partial or incomplete information about the target. As a result, if any one sensing unit malfunctions, the tracking system stops working. A new framework of sensor fusion, integrating an LRF and a monocular camera, is proposed in [6] for human tracking, where both sensors can work independently as well as in a cooperative manner, thus, providing a potentially robust solution in the face of the failure of any one sensor. However, a noteworthy limitation of this approach is that although it can track target poses, it is not suitable for tracking the shape of the target, which is especially crucial in the context of practical, comparatively short-ranged tracking applications.

The vast majority of the algorithms for detecting and tracking dynamic objects (e.g., cars, pedestrians, bicycles, etc.) consider the target object as a *point mass* that generates at most one measurement at a single sensor scan. However, in practice, especially where the size of the object cannot be ignored due to the relatively close proximity of the object to the sensors having fair resolutions, the *point-mass assumption* does not hold true. This gives rise to the *extended target tracking* problem [7], [8] where, at a single scan, multiple measurements are acquired whose distribution relies not only on the kinematic states but also on the shape of the target object. Situations get more complicated where we consider targets of arbitrary shapes or where no a priori information about such shapes is available. *Gaussian process* (GP)-based methods [7], [8], [9], [10] are growingly utilized to solve such nonparameteric regression problems at hand.

In a classical *GP regression* (GPR) model [11], noise level is considered constant throughout the input space. However, in many practical problems (e.g., [12] and [13]), the variability in observation is heavily reliant on the input. Over the past couple of decades, different *heteroscedastic GP* (HGP) [14],

Manuscript received 14 October 2022; accepted 22 December 2022. Date of publication 17 January 2023; date of current version 24 July 2023. The work of Pritam Paral was supported by the Ministry of Electronics and IT, Government of India, through the "Visvesvaraya PhD Scheme for Electronics and IT". Paper no. TII-22-4286. *(Corresponding author: Pritam Paral.)*

Pritam Paral and Sankar K. Pal are with the Center for Soft Computing Research, Indian Statistical Institute, Kolkata 700108, India (e-mail: callinpritam@gmail.com; sankar@isical.ac.in).

Amitava Chatterjee is with the Department of Electrical Engineering, Jadavpur University, Kolkata 700032, India (e-mail: cha_ami@yahoo.co.in).

Anjan Rakshit was with the Department of Electrical Engineering, Jadavpur University, Kolkata 700032, India (e-mail: anjanrakshit1951@gmail.com).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TII.2022.3232765.

Digital Object Identifier 10.1109/TII.2022.3232765

[15], [16], [17], [18], [19] algorithms have been developed, where the traditional constant-noise assumption is relaxed and the noise level is considered to be varying in the input domain. Compared with the standard (homoscedastic) GP, the HGP is demonstrated to better approximate various sources of uncertainty. Typically, in the HGP, two GPs are involved, the first one for modeling the underlying function and the second one for learning the heteroscedastic noise, together generating a joint posterior distribution, that is analytically intractable and non-Gaussian.

The numerical solutions can be achieved using various approximate inference methods, such as Markov chain Monte Carlo (MCMC) samplings [14], expectation propagation [15], variational inference [16], Laplace approximation [17], and the most likely noise (MLN) approaches [18], [19]. The most accurate MCMC sampling is heavily time-consuming when dealing with large-scale datasets. Although the expectation propagation approximations are significantly faster compared with MCMC methods, they remain highly expensive in the case of large-scale regression problems. The variational inference and its variants (e.g., variational HGP (VHGP) [16]) offer a reasonable tradeoff between computational efficiency and accuracy. For the Laplace approximation, nonelliptical skewed posterior distributions pose a challenge and serious efforts would be needed to improve the method in this regard. The MLN approaches are simple and computationally appealing in handling regression problems involving heteroscedastic noise, but they may risk overfitting and suffer from drawbacks, such as numerical instability and inaccuracy. For instance, the limitations of the *most likely HGP* (MLHGP) [18] include no guarantee for convergence and possible oscillations because of empirical estimation of the input-dependent noise (IDN) levels. Afterward, the *maximum a posteriori HGP* (MAPHGP) [19], although tried to address these problems, it showed a tendency to overfit significantly, especially when there are numerous latent noise variables to learn.

Considering the abovementioned aspects, a new, efficient approach to multisensor fusion and state estimation is proposed in this study for extended target tracking in human–robot coexisting environments. The developed method, referred to as *human feature-based extended target tracking via multisensor information fusion* (HFBETT-MSIF), can systematically assimilate information from multiple sensors mounted on a mobile robot in a unified way, during human tracking. This inexpensive, low-power robotic system comprises a monocular camera and an ultrasonic (US) sensor and can be particularly beneficial in developing countries. This research considers both shoes of a target person as the *extended targets* and there are several reasons behind selecting shoes [6]. Moreover, in this study, the shape of a target shoe is characterized by its projected 2-D contour, and the boundary of a *star-convex* set is proposed to be used to model the shoe shape. In general, the *radial function* is used to describe arbitrary star-convex shapes [7], and in this work, the function is approximated by a finite set of radial distances sampled at a predefined set of angles.

First, to precisely represent an extended target's states, the present article defines an augmented state vector comprising the shape parameters and pose states. A novel *generalized* measurement model is then developed for both the camera and sonar sensor based on an advanced HGP regression approach. Finally, the derived measurement models are utilized for Bayesian inference for which the popular unscented Kalman filter (UKF) [20] is employed to predict and correct conditional probability density functions. The major contributions of this research work are as follows.

1) Given the relatively short-ranged sensing system involving the sonar sensor and camera, the extents of the target shoes are large enough so that multiple measurements are acquired, when a sensor scan is performed. In the proposed framework, both the sensing modules are endowed with the capability of extracting the range and bearing measurements of various points on the shoes' surfaces at a single scan independently, during the tracking process. Such a strategy enables deriving respective measurement models for the sonar sensor and camera, each containing the complete information of the human target and being involved in the estimation of the shape states of the target in conjunction with its kinematic states. This renders the tracking system potentially robust against the breakdown of any single sensing unit.
2) Since no a priori knowledge about the target shape is available here, the problem of predicting unknown radial distances for shape state estimation can be formulated as a *nonparameteric regression* problem. However, the experimental study reveals that the variability in the measurement of radial distances is dependent on the angle inputs. Hence, the noise level is considered to be varying across the input space (for details, see Section S-I of the Supplementary Material). To deal with the aforementioned heteroscedastic regression problem, an advanced HGP model, called *bias-minimized MLHGP* (BMMLHGP), is developed here, inspired by the concepts of improved MLHGP presented in [21]. This provides an almost unbiased estimate of the heteroscedastic noise level by using the *method of moments* [22] for Bayesian residuals. The algorithm achieves excellent numerical accuracy and stability, aside from a superior computational efficiency. Based on the BMMLHGP, a novel generalized measurement model is derived for both the camera and sonar sensor. To the best of our knowledge and belief, this is the first work to demonstrate how an HGP configuration can be successfully utilized for shape state estimation in an extended target tracking problem.

The rest of this article is organized as follows. In Section II, the state transition model of a target shoe is developed. Sections III and IV, respectively, describe the methodologies for shoe localization based on camera and US sensor. Section V presents the BMMLHGP model and illustrates how regression via BMMLHGP can be utilized to derive the generalized measurement model. In Section VI, the shoe tracking problem at hand is mathematically formulated and the proposed multisensor fusion-based tracking approach is presented. Comprehensive real-life performance evaluations are presented in Section VII. Finally, Section VIII concludes this article.

II. State Transition Model of the Extended Target Object

A. Definition of the State Vector

The extent of a target shoe is modeled here using a *star-convex* shape. Each extended target is assumed to move in a 2-D space

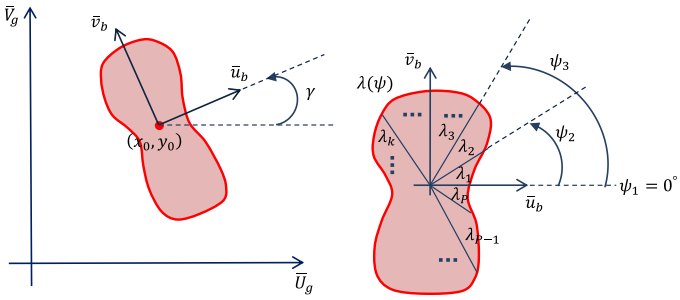

Fig. 1. (a) Illustration of a typical shoe in the GCF $(\overline{U}_g - \overline{V}_g)$ and BAF $(\overline{u}_b - \overline{v}_b)$. (b) Visual representation of the shape state for a shoe.

and its shape is characterized by its projected 2-D contour [7]. Now, to define the state vector, the notions of two reference frames, named *body-attached frame* (BAF) and *global coordinate frame* (GCF), are here presented. A representative example for the projected 2-D contour of a human shoe is shown in Fig. 1(a), where $\overline{U}_g$-$\overline{V}_g$ and $\overline{u}_b$-$\overline{v}_b$, respectively, represent the GCF and BAF, $(x_0,\ y_0)$ denotes the coordinate of the BAF's origin w.r.t. the GCF, and the angle γ symbolizes the heading of the BAF. In general, the pose state at sampling instant t_i $(i = 1,\ 2, \ldots)$, indicated by $\boldsymbol{X}_{\text{po}}(i) \in \mathbb{R}^p$, is defined as $\boldsymbol{X}_{\text{po}}(i) = [x_0(i),\ y_0(i),\ \boldsymbol{X}'_{\text{po}}(i)^T]^T$ [7], where $\boldsymbol{X}'_{\text{po}}(i) \in \mathbb{R}^{p'}$ comprises auxiliary state variables, for example, velocity, angular rate, acceleration, heading etc. In this research, the well-known *constant velocity motion model* is adopted, i.e., $\boldsymbol{X}'_{\text{po}} = [\dot{x}_0,\ \dot{y}_0]^T \in \mathbb{R}^2$, and therefore, $\boldsymbol{X}_{\text{po}}(i) = [x_0(i),\ y_0(i),\ \dot{x}_0(i),\ \dot{y}_0(i)]^T$.

The contour of a star-convex shape in polar coordinates is given by a radial function $\lambda = \lambda(\psi)$ that maps an angle ψ in the BAF to the corresponding radial distance λ [see Fig. 1(b)]. The observed radial function λ is considered as a noise-corrupted version of the *actual* radial function $\tilde{\lambda}$. For practical applications, $\lambda(\psi)$ is approximated by a finite set of radial distances $\{\lambda_k\}_{k=1}^P$ sampled at a predefined set of angles $\{\psi_k\}_{k=1}^P$ [7]. Then, the shape state at sampling instant t_i, indicated by $\boldsymbol{X}_{\text{sh}}(i) \in \mathbb{R}^P$, is given as $\boldsymbol{X}_{\text{sh}}(i) = [\lambda_1(i),\ \lambda_2(i), \ldots,\ \lambda_P(i)]^T$. Finally, the complete state vector of a target at t_i, denoted by $\boldsymbol{X}(i) \in \mathbb{R}^{p+P}$, is given as [7]

$$\boldsymbol{X}(i) = \begin{bmatrix} \boldsymbol{X}_{\text{po}}(i)^T & \boldsymbol{X}_{\text{sh}}(i)^T \end{bmatrix}^T. \tag{1}$$

B. Description of the State-Transition Model

The state-transition model for this extended object, i.e., a human shoe can be described as [7]

$$\boldsymbol{X}(i+1) = \boldsymbol{F}\left(\boldsymbol{X}(i),\ \boldsymbol{w}(i)\right) \tag{2}$$

or alternatively

$$\begin{bmatrix} \boldsymbol{X}_{\text{po}}(i+1) \\ \boldsymbol{X}_{\text{sh}}(i+1) \end{bmatrix} = \begin{bmatrix} \boldsymbol{F}_{\text{po}}\left(\boldsymbol{X}_{\text{po}}(i),\ \boldsymbol{w}_{\text{po}}(i)\right) \\ \boldsymbol{F}_{\text{sh}}\left(\boldsymbol{X}_{\text{sh}}(i),\ \boldsymbol{w}_{\text{sh}}(i)\right) \end{bmatrix} \tag{3}$$

where $\boldsymbol{w}(i) \in \mathbb{R}^q$ is a random noise with mean zero and covariance $\boldsymbol{Q}(i) \in \mathbb{R}^{q\times q}$, and $\boldsymbol{F} : \mathbb{R}^{p+P} \times \mathbb{R}^q \to \mathbb{R}^{p+P}$ denotes the *state-transition function*. $\boldsymbol{F}$ can be decomposed into two components as $\boldsymbol{F} = [\boldsymbol{F}_{\text{po}}^T\ \boldsymbol{F}_{\text{sh}}^T]^T$, where $\boldsymbol{F}_{\text{po}}$ and $\boldsymbol{F}_{\text{sh}}$ represent the dynamical behaviors shown by the pose and shape states, respectively, over time. Similarly, $\boldsymbol{w}(i)$ is expressed as $\boldsymbol{w}(i) = [\boldsymbol{w}_{\text{po}}(i)^T\ \boldsymbol{w}_{\text{sh}}(i)^T]^T$, where $\boldsymbol{w}_{\text{po}}(i)$ and $\boldsymbol{w}_{\text{sh}}(i)$ capture the uncertainties in individual dynamical behaviors with covariances

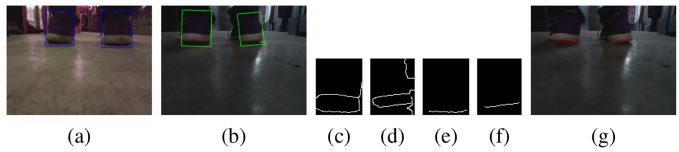

Fig. 2. (a) Shoe templates determined from the reference frame. (b) PSAs computed for a representative frame captured. Shoe edges identified in the PSAs. (c) Shoe: left. (d) Shoe: right. Contact edges between the shoes and the ground. (e) Shoe: left. (f) Shoe: right. (g) CPs detected in the frame.

$\boldsymbol{Q}_{\text{po}}(i)$ and $\boldsymbol{Q}_{\text{sh}}(i)$ $(\boldsymbol{Q}(i) = \text{diag}(\boldsymbol{Q}_{\text{po}}(i),\ \boldsymbol{Q}_{\text{sh}}(i)))$, respectively. In this study, the constant velocity motion model is utilized to represent $\boldsymbol{F}_{\text{po}}$ as [7]

$$\begin{aligned} \boldsymbol{X}_{\text{po}}(i+1) &= \boldsymbol{F}_{\text{po}}\left(\boldsymbol{X}_{\text{po}}(i),\ \boldsymbol{w}_{\text{po}}(i)\right) \\ &= \begin{bmatrix} 1 & 0 & \tau & 0 \\ 0 & 1 & 0 & \tau \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \boldsymbol{X}_{\text{po}}(i) + \begin{bmatrix} \tau^2/2 & 0 \\ 0 & \tau^2/2 \\ \tau & 0 \\ 0 & \tau \end{bmatrix} \boldsymbol{w}_{\text{po}}(i) \end{aligned}$$

where τ is the sampling interval. The heading γ of a target is assumed to be aligned with its velocity, that is $\gamma = \tan^{-1}(\dot{y}_0/\dot{x}_0)$. This research considers time invariant shapes, leading to the use of an *identity function* for $\boldsymbol{F}_{\text{sh}}$ as [7]

$$\boldsymbol{X}_{\text{sh}}(i+1) = \boldsymbol{F}_{\text{sh}}\left(\boldsymbol{X}_{\text{sh}}(i),\ \boldsymbol{w}_{\text{sh}}(i)\right) = \boldsymbol{X}_{\text{sh}}(i) + \boldsymbol{w}_{\text{sh}}(i).$$

III. Vision Sensor-Based Shoe Location Detection (SLD)

During human tracking, from each successive frame fr_i captured at a sampling instant t_i $(i = 0,\ 1, \ldots)$, we extract a pair of quadrilateral image blocks (referred to herein as the left and right *potential subareas* (PSAs) [23], respectively) containing the positions of the target shoes and subsequently determine the target shoes' real locations w.r.t. the GCF at t_i.

A. Visual Features Extraction (VFE) Module

In this article, we develop a VFE module based on *photometric-invariant CFAsT-Match* (PICFAsT-Match) approach proposed in [23] where, first, an intelligent algorithm [23] detects the shoes of the target person in a *reference frame* fr_{ref}, and then, two template images of rectangular shape, one containing mainly the left shoe $(I_{ST(\text{L})})$ and the other mainly the right shoe $(I_{ST(\text{R})})$, are acquired from fr_{ref}. Details of each individual shoe detection (i.e., PSA computation) in a succeeding frame (I_{TR}) captured during pursuit, using template matching, are given in [23]. In this work, similar to [23], the fitness of a single affine transformation Θ in a specific grid G_σ (σ being a precision parameter) is assessed via a *sublinear* approximation of the *Adapted-Colour-Sum-of-Absolute-Differences* (ACSAD) distance function $\xi_\Theta(I_{ST(k)},\ I_{TR})$, $k = \text{L}$ or R. The mathematical definition of ACSAD distance $\xi_\Theta(I_{ST(k)},\ I_{TR})$ is given in detail in Section S-II of the Supplementary Material.

B. SLD Module

After implementing VFE, we apply *bilateral filtering* [24] to smooth the PSAs and then *Canny edge detection* [25] to obtain the target shoe's edges in the smoothed PSAs. Finally, using *Shi–Tomasi corner detector* [25], along with an *empirical geometric model*, we identify multiple points of contact (CPs) of the shoes

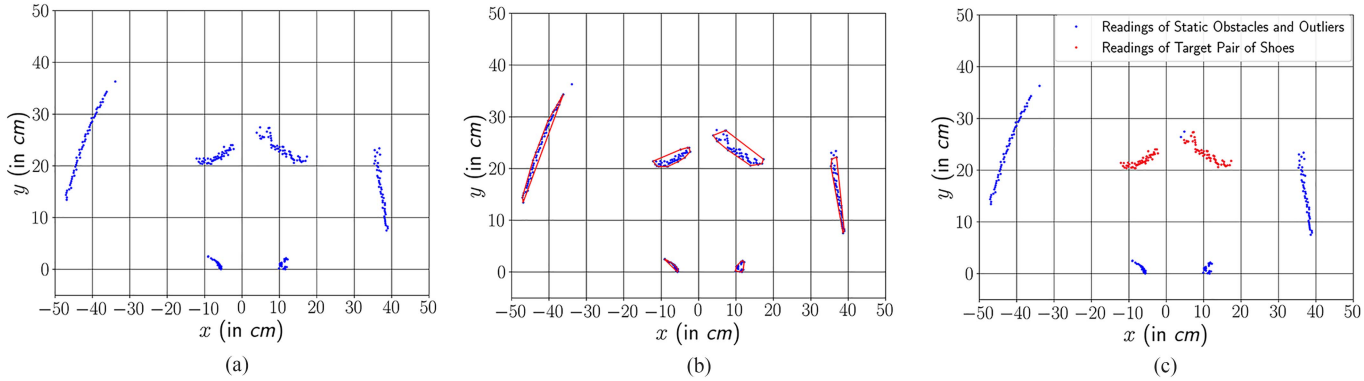

Fig. 3. (a) Readings in a sample sonar scan. (b) Clustering sonar readings with DENCLUE (a *red* contour: a spatial cluster). (c) Finding sonar readings representing the target shoes based on EFBLR, resulting in *filtered* scan.

with the ground. The schemes for computation of PSAs and CPs are shown in Fig. 2 with a real-life test case.

Considering a visual frame fr_{ex} captured at time instant t_{ex} during tracking, let us assume that the number of CPs identified for a target shoe [referred to as *shoe CPs* (SCPs)] is denoted by NV_{ex}. Then, let $A(\text{ex}) = \{[p_m^{(\text{ex})}, q_m^{(\text{ex})}]^T\}_{m=1}^{\text{NV}_{\text{ex}}}$ indicate the set of the image plane coordinates of the SCPs w.r.t. the center of the image (i.e., frame). Also, let the set of the visual ranges and bearings of the SCPs w.r.t. the onboard camera be denoted by $V(\text{ex}) = \{[r_{v_m}^{(\text{ex})}, \theta_{v_m}^{(\text{ex})}]^T\}_{m=1}^{\text{NV}_{\text{ex}}}$. Once we obtain the SCPs in the image plane, we adopt the geometric setting and definitions associated with the *visual observation model* module, described in [6], to determine $V(\text{ex})$, and accordingly the set of the corresponding 2-D positions of the SCPs in the GCF, denoted by $Z_{(\text{vis})}(\text{ex}) = \{\underline{\mathbf{z}}_{(\text{vis}),m}^{(\text{ex})}\}_{m=1}^{\text{NV}_{\text{ex}}}$, with $\underline{\mathbf{z}}_{(\text{vis}),k}^{(\text{ex})} = [\underline{z}_{x(\text{vis}),k}^{(\text{ex})}, \underline{z}_{y(\text{vis}),k}^{(\text{ex})}]^T$ (where $\underline{z}_{x(\text{vis}),k}^{(\text{ex})}$ and $\underline{z}_{y(\text{vis}),k}^{(\text{ex})}$ represent the coordinates of the kth SCP in the GCF).

IV. Sonar-Based Target Shoe Localization

In this work, the sonar-based shoe localization is performed based on the human leg localization approach developed in [26], utilizing *DENsity-based CLUstEring* (DENCLUE) and *Edge Feature Based Leg Recognition* (EFBLR) algorithms [26]. The shoe localization scheme is shown in Fig. 3 with a real-life test case. Let, for a sonar scan Sc_{ex} acquired w.r.t. a robot pose frame rp_{ex} [26] at time instant t_{ex} during tracking, NS_{ex} be the number of sonar readings in *filtered* Sc_{ex} characterizing a target shoe's extent in width. Assuming that $S(\text{ex}) = \{[r_{s_n}^{(\text{ex})}, \theta_{s_n}^{(\text{ex})}]^T\}_{n=1}^{\text{NS}_{\text{ex}}}$ denotes the set of sonar ranges and bearings of the corresponding sonar readings w.r.t. the onboard US sensor at its original heading [refer to Fig. 5(a)], the set of 2-D positions in the GCF that corresponds to $S(\text{ex})$ is given as $Z_{(\text{rng})}(\text{ex}) = \{\underline{\mathbf{z}}_{(\text{rng}),n}^{(\text{ex})}\}_{n=1}^{\text{NS}_{\text{ex}}}$, with $\underline{\mathbf{z}}_{(\text{rng}),k}^{(\text{ex})} = [\underline{z}_{x(\text{rng}),k}^{(\text{ex})}, \underline{z}_{y(\text{rng}),k}^{(\text{ex})}]^T$ (where $\underline{z}_{x(\text{rng}),k}^{(\text{ex})}$ and $\underline{z}_{y(\text{rng}),k}^{(\text{ex})}$ are the coordinates of the kth shoe-reading in the GCF).

V. Generalized Measurement Model for Camera and Sonar: Heteroscedastic GP Regression

Let the set of readings obtained with an onboard sensor for a single target shoe at t_i be $Z(i) = \{\underline{\mathbf{z}}_1^{(i)}, \underline{\mathbf{z}}_2^{(i)} \ldots, \underline{\mathbf{z}}_{\text{NC}_i}^{(i)}\}$,

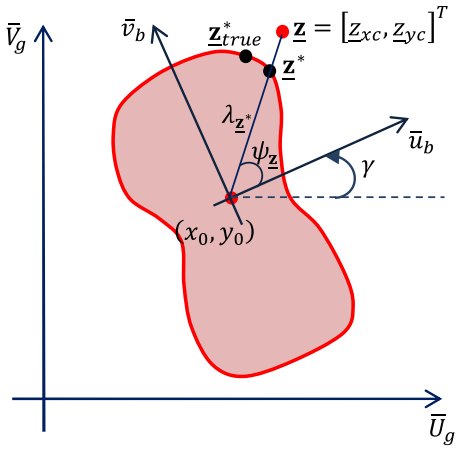

Fig. 4. Positions of a noisy measurement $\underline{\mathbf{z}}$ and the corresponding measurement generating point source $\underline{\mathbf{z}}^*$ in the GCF.

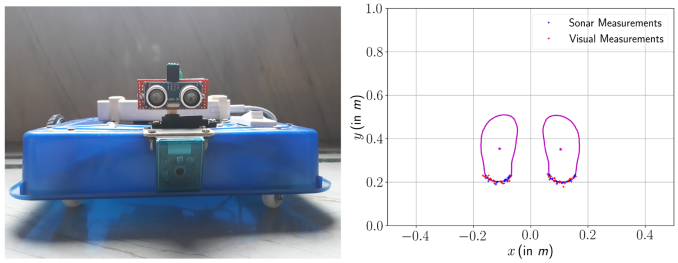

Fig. 5. (a) Wheeled mobile robot utilized and (b) a pair of shoes tracked by the onboard sensors of the robot. *Magenta* "⋆": center of a shoe.

where $\underline{\mathbf{z}}_k^{(i)} \in \mathbb{R}^2$, $k = 1, 2 \ldots, \text{NC}_i$ represent individual readings and NC_i indicates the number of the readings. If we assume these measurements to be conditionally independent of each other given the state $\boldsymbol{X}$, the joint likelihood function $p(\underline{\mathbf{z}}_1^{(i)}, \underline{\mathbf{z}}_2^{(i)} \ldots, \underline{\mathbf{z}}_{\text{NC}_i}^{(i)} | \boldsymbol{X})$ can be expressed as the product of the individual likelihoods of all measurements as follows:

$$p(\underline{\mathbf{z}}_1^{(i)}, \underline{\mathbf{z}}_2^{(i)}, \ldots, \underline{\mathbf{z}}_{\text{NC}_i}^{(i)} | \boldsymbol{X}) = \prod_{k=1}^{\text{NC}_i} p(\underline{\mathbf{z}}_k^{(i)} | \boldsymbol{X}). \tag{4}$$

We define a point source $\underline{\mathbf{z}}^* \in \mathbb{R}^2$ on the boundary of the target that yields the measurement $\underline{\mathbf{z}}$. However, in reality, the estimator does not know the actual point (referred to as $\underline{\mathbf{z}}^*_{true}$ in Fig. 4) that generates the measurement $\underline{\mathbf{z}}$. Therefore, following a similar approximation approach as in [7], we assume that $\underline{\mathbf{z}}^*$ is located on the boundary along the direction indicated by $\psi_{\underline{\mathbf{z}}}$, where $\psi_{\underline{\mathbf{z}}}$ is the angle made by $\underline{\mathbf{z}}$ w.r.t. the BAF (given in Fig. 4). Then, we can consider the measurement $\underline{\mathbf{z}}$ as a noise-contaminated representation of $\underline{\mathbf{z}}^*$ as [7]

$$\underline{\mathbf{z}} = \underline{\mathbf{z}}^* + \boldsymbol{\Delta} \tag{5}$$

where $\boldsymbol{\Delta} \in \mathbb{R}^2$ denotes a zero-mean random noise having covariance as $\mathbf{R}_{\boldsymbol{\Delta}} \in \mathbb{R}^{2\times 2}$. As given in Fig. 4, the point source $\underline{\mathbf{z}}^* = [\underline{z}^*_{xc}, \underline{z}^*_{yc}]^T$ can be characterized by two parameters, namely the radial distance $\lambda_{\underline{\mathbf{z}}^*}$ and the angle $\psi_{\underline{\mathbf{z}}}$, as [7]

$$\underline{\mathbf{z}}^* = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + \begin{bmatrix} \cos(\psi_{\underline{\mathbf{z}}} + \gamma) \\ \sin(\psi_{\underline{\mathbf{z}}} + \gamma) \end{bmatrix} \lambda_{\underline{\mathbf{z}}^*}. \tag{6}$$

The angle $\psi_{\underline{\mathbf{z}}}$ from Fig. 4 is given as follows:

$$\psi_{\underline{\mathbf{z}}} = \tan^{-1}\left(\frac{\underline{z}_{yc} - y_0}{\underline{z}_{xc} - x_0}\right) - \gamma. \tag{7}$$

The radial distance $\lambda_{\underline{\mathbf{z}}^*}$ is generally unknown, since $\psi_{\underline{\mathbf{z}}}$ usually does not coincide with the predefined ψ_k, $k = 1, 2, \ldots, P$, and the problem of predicting $\lambda_{\underline{\mathbf{z}}^*}$ can be formulated as a regression problem [7]. Since the object shape is considered arbitrary, it necessitates the implementation of *nonparameteric* regression. In this research, we develop a *method of moments* [22] based MLHGP model, inspired by the concepts described in [21], for the purpose of solving our regression subproblem. The mathematical fundamentals of HGP and a detailed discussion on MLHGP are presented in Sections S-III and S-IV of the Supplementary Material, respectively.

A. Regression Via BMMLHGP

The MLHGP, one of the most computationally appealing approximation to solve regression problems involving IDN, functions on the principle of simply replacing the noise posterior with a point estimate derived at the most likely value in such a way that the posterior predictive distribution over the test output can be dealt with analytically [18]. Although MLHGP is much simpler and offers better computational performance than other analytical approximations, it is not guaranteed that the algorithm will eventually converge, as the empirical noise estimate is demonstrably biased for majority of local noise cases, mainly when the noise level varies in the input domain [21].

In this article, we develop an enhanced version of the existing MLHGP, inspired by [21], called BMMLHGP, in which an almost unbiased noise estimate is elicited from regression residuals via the *method of moments*. A regression residual Ω_j in GPR, commonly referred to as a Bayesian residual, is represented by the difference between the observation u_j and the posterior mean corresponding to u_j, denoted as m_{u_j}, at training location $\boldsymbol{v}_j$

$$\Omega_j = u_j - m_{u_j}. \tag{8}$$

This can be further written as follows:

$$\Omega_j = (u_j - h_j) + (h_j - m_{u_j}) = \kappa_j + (h_j - m_{h_j}) \tag{9}$$

with h_j being the value of the underlying function $h(\cdot)$ at training point $\boldsymbol{v}_j$, m_{h_j} being the posterior mean corresponding to the true function value h_j at $\boldsymbol{v}_j$, and based on the characteristics of GPR, $m_{u_j} = m_{h_j}$. The items $(h_j - m_{h_j})$ and κ_j in (9), respectively, denote the modeling error (deterministic in nature) and observation error (random in nature). Therefore, each residual Ω_j is a random variable as well, which follows normal distribution given by [21]:

$$\Omega_j \sim \mathcal{N}\left(m_{\Omega_j},\, s^2_{\Omega_j}\right) \tag{10}$$

with mean $m_{\Omega_j} = (h_j - m_{h_j})$ and variance $s^2_{\Omega_j} = e^2(\boldsymbol{v}_j) = e^2_j$, where $e^2(\boldsymbol{v}_j) = e^2_j$ denotes the local noise variance at training input $\boldsymbol{v}_j$. The expected function value m_{h_j} and Bayesian residual Ω_j are accessible from the first GP (denoted as G_1) involved in the developed HGP configuration, which is a standard, homoscedastic GP designated to model the latent function $h(\cdot)$ from Q observed data points $DS = \{(\boldsymbol{v}_j,\, u_j)\}_{j=1}^{Q}$, while one needs to estimate the noise variance e^2_j and actual function value h_j from the available data.

In heteroscedastic regression, the dispersion profile of the residual series is governed by the variable noise level, and on that account, regression residuals can be employed for estimating IDN. To identify the dispersion pattern revealed by the residuals (r-function), regression methods are generally carried out on the transformed residuals $r_j = T(\Omega_j)$ instead of the raw residuals Ω_j [21]. Yet, the realized r-function by fitting a curve for r_j may not offer a bias-free estimate for the noise function $e(\boldsymbol{v}_j)$, based on the transformation function T undertaken. Hence, one needs to calibrate the resulting r-function and render it unbiased for IDN.

In statistics, a popular practice for parameter estimation is the method of moments [22] with which one can obtain an unbiased estimate for the parameters under consideration. For IDN, one can estimate its local levels from various statistical moments, e.g., raw moments, central moments, raw absolute moments (RAMs), or central absolute moments of Bayesian residuals. Here, the RAMs of the residuals are preferred, since each order of the RAMs of the residuals carries information regarding the noise power.

With the moment order ω, the RAM of the residual Ω_j at training input $\boldsymbol{v}_j$ is computed as [21]

$$E\{|\Omega_j|^\omega\} = s^\omega_{\Omega_j}\beta(\omega) = e^\omega_j \beta(\omega) \tag{11}$$

where $\beta(\omega)$ denotes a correcting factor that is a function of the moment order ω and can be expressed as follows:

$$\beta(\omega) = (\Gamma(\omega+1)/\sqrt{2\pi})\exp(-m^2_{\Omega_j}/(4s^2_{\Omega_j})) \cdot \left[D_{(-\omega-1)}\left(m_{\Omega_j}/s_{\Omega_j}\right) + D_{(-\omega-1)}\left(-m_{\Omega_j}/s_{\Omega_j}\right)\right] \tag{12}$$

where $\Gamma(\cdot)$ is the *gamma function* and $D_{(\cdot)}(\cdot)$ is the *parabolic cylinder function* [27].

We can neglect the modeling error $m_{\Omega_j} = h_j - m_{h_j}$, especially when the first GP G_1 utilized for learning the latent function value is precisely defined. Under this condition, we have $D_{(-\omega-1)}(0) = 2^{-(\omega+1)/2} \cdot (\sqrt{\pi}/\Gamma(\omega/2+1))$. Then, we can approximate the correction factor $\beta(\omega)$ as follows:

$$\beta(\omega) \approx (2^{\omega/2}/\sqrt{\pi}) \cdot \Gamma\left((\omega+1)/2\right). \tag{13}$$

Now, from (11), we obtain the local noise level e^ω_j as follows:

$$e^\omega_j = (1/\beta(\omega)) \cdot E\{|\Omega_j|^\omega\}. \tag{14}$$

The expression $(1/\beta(\omega)) \cdot E\{|\Omega_j|^\omega\}$ in (14) represents an unbiased estimate for the local noise level e_j^ω at $\boldsymbol{v}_j$.

We present two representative cases where, corresponding to the first ($\omega = 1$) and second ($\omega = 2$) RAMs of the residual Ω_j, we, respectively, obtain the almost unbiased estimates for the noise standard deviation e_j and variance e_j^2 at $\boldsymbol{v}_j$ as follows:

$$\begin{aligned} e_j &= (1/\beta(1)) \cdot E\{|\Omega_j|\} \approx \sqrt{\pi/2} E\{|\Omega_j|\} \\ e_j^2 &= (1/\beta(2)) \cdot E\{|\Omega_j|^2\} \approx E\{\Omega_j^2\}. \end{aligned} \tag{15}$$

Consequently, with $r_j = |\Omega_j|^\omega$, a new dataset $DS' = \{(\boldsymbol{v}_j, r_j)\}_{j=1}^Q$ can be formed to train another classical GP G_2 for estimating the MLN levels (MLNLs) $\hat{e}_j^\omega = (1/\beta(\omega))m_{r_j}$ at training input $\boldsymbol{v}_j$ and $\hat{e}_\star^\omega = (1/\beta(\omega))m_{r_\star}$ at test input $\boldsymbol{v}_\star$ (where m_{r_j} and $m_{r_\star}$ are the posterior means of the IDN levels at $\boldsymbol{v}_j$ and $\boldsymbol{v}_\star$, respectively) [21]. However, the transformed residuals $r_j = |\Omega_j|^\omega$ are generally non-Gaussian as well as nonnegative, and in G_2, we have used a Gaussian approximation to r_j. It is necessary to refine the MLNLs to $\hat{e}_j^\omega = \max(0,\ (1/\beta(\omega))m_{r_j})$ or $\hat{e}_\star^\omega = \max(0,\ (1/\beta(\omega))m_{r_\star})$ so that a nonnegative noise level is ensured [21]. Apart from performing a better estimation of the IDN level, the BMMLHGP also avoids the computationally heavy expectation maximization (EM)-like procedure employed in the existing MLHGP for the purpose of iterative learning.

B. Proposed Measurement Model

In our unknown radial distance prediction problem, we learn a machine learning model from an observed dataset $DS = \{(\psi_k, \lambda_k)\}_{k=1}^P$, where $\lambda_k \in \mathbb{R}$ is a noisy observation of the true radial function $\tilde{\lambda} : \mathbb{R} \to \mathbb{R}$ at the angle $\psi_k \in \mathbb{R}$ such that

$$\lambda_k = \tilde{\lambda}(\psi_k) + \eta_k, \quad \eta_k \sim \mathcal{N}(0,\ e^2(\psi_k)). \tag{16}$$

The noise variance $e^2(\psi_k)$ is observed to be varying across the input space, thus posing a *heteroscedastic* regression problem.

We apply the BMMLHGP to solve this heteroscedastic regression problem, where two operations are carried out.

1) First, an HGP model is built in which two GPs are involved. The first GP (denoted as GP_1) trained on $DS = \{(\psi_k, \lambda_k)\}_{k=1}^P$ recovers the noisy radial function λ. The second GP (denoted as GP_2) trained on $DS' = \{(\psi_k, r_k)\}_{k=1}^P$ with IDN levels $r_k = |\Omega_k|^\omega$ finds an almost unbiased estimate of the MLNLs $\hat{e}_k^\omega = \max(0,\ (1/\beta(\omega))m_{r_k})$ at angles ψ_k, with $\omega = 1$ or 2.
2) The HGP model predicts the target radial distance $\lambda_{\underline{\boldsymbol{z}}^*} = \lambda(\psi_{\underline{\boldsymbol{z}}})$ for a query angle $\psi_{\underline{\boldsymbol{z}}}$ (given in Fig. 4), given $DS = \{(\psi_k, \lambda_k)\}_{k=1}^P$, $\hat{\boldsymbol{e}}^\omega = [\hat{e}_1^\omega, \ldots, \hat{e}_P^\omega]^T$ at $\boldsymbol{\Psi} = [\psi_1, \ldots, \psi_P]^T$, and $\hat{e}_{\underline{\boldsymbol{z}}}^\omega = \max(0,\ (1/\beta(\omega))m_{r_{\underline{\boldsymbol{z}}}})$ at $\psi_{\underline{\boldsymbol{z}}}$, where $m_{r_{\underline{\boldsymbol{z}}}}$ is the posterior mean of the IDN level at $\psi_{\underline{\boldsymbol{z}}}$.

GP_1 is built with a periodic covariance function $c(\psi, \psi')$ that corresponds to the SE kernel $c_{SE}(\psi, \psi') = \rho_0^2 \exp[-|\psi - \psi'|^2/(2\iota_0^2)]$ parameterized by $\boldsymbol{\Phi}_{\tilde{\lambda}} = \{\rho_0, \iota_0\}$

$$c(\psi, \psi') = \rho_0^2 \exp\left[-\frac{2\sin^2\left(\dfrac{|\psi - \psi'|}{2}\right)}{\iota_0^2}\right] \tag{17}$$

with (ψ, ψ') being all possible pairs in the angle domain.

GP_2 is formed with a different SE kernel $c_r(\psi, \psi')$ parameterized by $\boldsymbol{\Phi}_r = \{\rho_r, \iota_r\}$. Now, by defining $\boldsymbol{X}_{\text{sh}} = \boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_P]^T$, a multivariate *Gaussian distribution* can be built following similar approach as in (S-8) of the Supplementary Material:

$$\begin{bmatrix} \boldsymbol{X}_{\text{sh}} \\ \tilde{\lambda}_{\underline{\boldsymbol{z}}^*} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{0} \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{C} + \boldsymbol{N} & \boldsymbol{C}_{\psi_{\underline{\boldsymbol{z}}}} \\ \boldsymbol{C}_{\psi_{\underline{\boldsymbol{z}}}}^T & c(\psi_{\underline{\boldsymbol{z}}}, \psi_{\underline{\boldsymbol{z}}}) \end{bmatrix}\right) \tag{18}$$

where $\tilde{\lambda}_{\underline{\boldsymbol{z}}^*}$ is the actual radial distance at query angle $\psi_{\underline{\boldsymbol{z}}}$

$$\begin{aligned} \boldsymbol{C} &= \begin{bmatrix} c(\psi_1, \psi_1) & \cdots & c(\psi_1, \psi_P) \\ \vdots & \ddots & \vdots \\ c(\psi_P, \psi_1) & \cdots & c(\psi_P, \psi_P) \end{bmatrix} \\ \boldsymbol{C}_{\psi_{\underline{\boldsymbol{z}}}} &= \begin{bmatrix} c(\psi_{\underline{\boldsymbol{z}}}, \psi_1) & c(\psi_{\underline{\boldsymbol{z}}}, \psi_2) & \cdots & c(\psi_{\underline{\boldsymbol{z}}}, \psi_P) \end{bmatrix}^T \end{aligned} \tag{19}$$

and $\boldsymbol{N}$ represents the diagonal matrix of the MLN variances with elements $[\boldsymbol{N}]_{kk} = \hat{e}_k^2$.

Then, similar to the Supplementary Material, the posterior distribution over the target radial distance $\lambda_{\underline{\boldsymbol{z}}^*}$ at $\psi_{\underline{\boldsymbol{z}}}$ can be obtained as follows:

$$p\left(\lambda_{\underline{\boldsymbol{z}}^*} | \psi_{\underline{\boldsymbol{z}}}, \boldsymbol{\Phi}_{\tilde{\lambda}}, \hat{\boldsymbol{e}}, \hat{e}_{\underline{\boldsymbol{z}}}, DS\right) \sim \mathcal{N}\left(m_{\lambda_{\underline{\boldsymbol{z}}^*}}, s_{\lambda_{\underline{\boldsymbol{z}}^*}}^2\right) \tag{20}$$

where $m_{\lambda_{\underline{\boldsymbol{z}}^*}}$, the posterior mean of $\lambda_{\underline{\boldsymbol{z}}^*}$ at $\psi_{\underline{\boldsymbol{z}}}$, is expressed as follows:

$$m_{\lambda_{\underline{\boldsymbol{z}}^*}} = \boldsymbol{C}_{\psi_{\underline{\boldsymbol{z}}}}^T (\boldsymbol{C} + \boldsymbol{N})^{-1} \boldsymbol{X}_{\text{sh}} \tag{21}$$

and $s_{\lambda_{\underline{\boldsymbol{z}}^*}}^2$, the posterior variance of $\lambda_{\underline{\boldsymbol{z}}^*}$ at $\psi_{\underline{\boldsymbol{z}}}$, is expressed as follows:

$$s_{\lambda_{\underline{\boldsymbol{z}}^*}}^2 = c(\psi_{\underline{\boldsymbol{z}}}, \psi_{\underline{\boldsymbol{z}}}) - \boldsymbol{C}_{\psi_{\underline{\boldsymbol{z}}}}^T (\boldsymbol{C} + \boldsymbol{N})^{-1} \boldsymbol{C}_{\psi_{\underline{\boldsymbol{z}}}} + \hat{e}_{\underline{\boldsymbol{z}}}^2. \tag{22}$$

The conditional distribution in (20) signifies that $\lambda_{\underline{\boldsymbol{z}}^*}$ is linearly related to $\boldsymbol{X}_{\text{sh}}$ as follows:

$$\lambda_{\underline{\boldsymbol{z}}^*} = \boldsymbol{C}_{\psi_{\underline{\boldsymbol{z}}}}^T (\boldsymbol{C} + \boldsymbol{N})^{-1} \boldsymbol{X}_{\text{sh}} + \alpha \tag{23}$$

where α represents a Gaussian noise with mean zero and variance $s_{\lambda_{\underline{\boldsymbol{z}}^*}}^2$ defined in (22).

Now, we can derive the measurement model based on (5), (6), and (23) as follows:

$$\begin{aligned} \underline{\boldsymbol{z}} &= \underline{\boldsymbol{z}}^* + \boldsymbol{\Delta} \\ &= \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + \begin{bmatrix} \cos(\psi_{\underline{\boldsymbol{z}}} + \gamma) \\ \sin(\psi_{\underline{\boldsymbol{z}}} + \gamma) \end{bmatrix} \lambda_{\underline{\boldsymbol{z}}^*} + \boldsymbol{\Delta} \\ &= \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + \begin{bmatrix} \cos(\psi_{\underline{\boldsymbol{z}}} + \gamma) \\ \sin(\psi_{\underline{\boldsymbol{z}}} + \gamma) \end{bmatrix} \boldsymbol{C}_{\psi_{\underline{\boldsymbol{z}}}}^T (\boldsymbol{C} + \boldsymbol{N})^{-1} \boldsymbol{X}_{\text{sh}} \\ &\quad + \begin{bmatrix} \cos(\psi_{\underline{\boldsymbol{z}}} + \gamma) \\ \sin(\psi_{\underline{\boldsymbol{z}}} + \gamma) \end{bmatrix} \alpha + \boldsymbol{\Delta} \\ &= \boldsymbol{H}(\boldsymbol{X}) + \boldsymbol{\zeta} \end{aligned} \tag{24}$$

where

$$\begin{aligned} \boldsymbol{H}(\boldsymbol{X}) &= \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + \begin{bmatrix} \cos(\psi_{\underline{\boldsymbol{z}}} + \gamma) \\ \sin(\psi_{\underline{\boldsymbol{z}}} + \gamma) \end{bmatrix} \boldsymbol{C}_{\psi_{\underline{\boldsymbol{z}}}}^T (\boldsymbol{C} + \boldsymbol{N})^{-1} \boldsymbol{X}_{\text{sh}} \\ \boldsymbol{\zeta} &= \begin{bmatrix} \cos(\psi_{\underline{\boldsymbol{z}}} + \gamma) \\ \sin(\psi_{\underline{\boldsymbol{z}}} + \gamma) \end{bmatrix} \alpha + \boldsymbol{\Delta}. \end{aligned} \tag{25}$$

Algorithm 1: BMMLHGP for Solving the Problem of Unknown Radial Distance Prediction.

1: Train a standard homoscedastic GP GP_1 on the observed dataset $DS = \{(\psi_k, \lambda_k)\}_{k=1}^{P}$ and estimate the posterior distribution over observed radial distances $p(\lambda_k|\psi_k, \boldsymbol{\Phi}_\lambda, DS) \sim \mathcal{N}(m_{\lambda_k}, s^2_{\lambda_k})$.
2: Compute transformed residuals $r_k = |\Omega_k|^\omega$ with $\Omega_k = \lambda_k - m_{\lambda_k}$, forming a new dataset $DS' = \{(\psi_k, r_k)\}_{k=1}^{P}$.
3: Train a second classical GP GP_2 on DS' and estimate the IDN levels $p(r_k|\psi_k, \boldsymbol{\Phi}_r, DS') \sim \mathcal{N}(m_{r_k}, s^2_{r_k})$.
4: Update the MLNLs

$$\hat{e}^{\omega}_{k} = \max\left(0, \frac{m_{r_k}}{\beta(\omega)}\right) \text{ with } \beta(\omega) \approx \frac{2^{\omega/2}}{\sqrt{\pi}} \cdot \Gamma\left(\frac{\omega+1}{2}\right)$$

5: Make Prediction on unknown radial distances

$$p\left(\lambda_{\underline{\mathbf{z}}^*}|\psi_{\underline{\mathbf{z}}}, \boldsymbol{\Phi}_{\tilde{\lambda}}, \hat{\boldsymbol{e}}^{\omega}, \hat{e}^{\omega}_{\underline{\mathbf{z}}}, DS\right) \sim \mathcal{N}\left(m_{\lambda_{\underline{\mathbf{z}}^*}}, s^2_{\lambda_{\underline{\mathbf{z}}^*}}\right)$$

Here, $\boldsymbol{H}(\cdot)$ can be considered as a *nonlinear* function of the state $\boldsymbol{X}$ and $\boldsymbol{\zeta}$ represents a zero-mean random noise whose covariance $\mathbf{R}_{\boldsymbol{\zeta}} \in \mathbb{R}^{2\times 2}$ is

$$\mathbf{R}_{\boldsymbol{\zeta}} = \mathbf{R}_{\boldsymbol{\Delta}} + \begin{bmatrix}\cos(\psi_{\underline{\mathbf{z}}}+\gamma)\\ \sin(\psi_{\underline{\mathbf{z}}}+\gamma)\end{bmatrix} s^2_{\lambda_{\underline{\mathbf{z}}^*}} \begin{bmatrix}\cos(\psi_{\underline{\mathbf{z}}}+\gamma)\\ \sin(\psi_{\underline{\mathbf{z}}}+\gamma)\end{bmatrix}^T. \quad (26)$$

Thus, the derivation of the measurement model is completed through (24)–(26), and this model will be utilized in state estimation for the assimilation of visual and sonar information. Algorithm 1 summarizes the approach to solve our problem of predicting unknown radial distances via BMMLHGP.

VI. Proposed Multisensor Fusion-Based Approach to Human Shoe Tracking

Based on the discussions in Sections III and IV, the information obtained from the camera and sonar for the left (L) [or right (R)] shoe at sampling instant t_i can be characterized by the set of NL(R)V$_i$ measurements $Z_{\text{L(R)(vis)}}(i) = \{\underline{\mathbf{z}}^{(i)}_{\text{L(R)(vis)},m}\}_{m=1}^{\text{NL(R)V}_i}$ and the set of NL(R)S$_i$ measurements $Z_{\text{L(R)(rng)}}(i) = \{\underline{\mathbf{z}}^{(i)}_{\text{L(R)(rng)},n}\}_{n=1}^{\text{NL(R)S}_i}$, respectively. Representing those groups of $(\text{NLV}_i + \text{NLS}_i)$ and $(\text{NRV}_i + \text{NRS}_i)$ measurements at t_i as $\text{IN}_\text{L}(i)$ and $\text{IN}_\text{R}(i)$, respectively, we can formulate the adopted extended target tracking problem as a *state estimation* problem, where the conditional probability distributions of the left and right shoes' states at t_i, denoted as $\boldsymbol{X}_\text{L}(i)$ and $\boldsymbol{X}_\text{R}(i)$, respectively, need to be computed as [7]

$$p\left(\boldsymbol{X}_\text{L}(i)|\text{IN}_\text{L}(0), \text{IN}_\text{L}(1), \ldots, \text{IN}_\text{L}(i-1), \text{IN}_\text{L}(i)\right) \quad (27)$$

$$p\left(\boldsymbol{X}_\text{R}(i)|\text{IN}_\text{R}(0), \text{IN}_\text{R}(1), \ldots, \text{IN}_\text{R}(i-1), \text{IN}_\text{R}(i)\right). \quad (28)$$

The means of the conditional probability distributions defined in (27) and (28) are designated as the *state estimates* $\overline{\boldsymbol{X}}_\text{L}(i)$ and $\overline{\boldsymbol{X}}_\text{R}(i)$, respectively.

The expressions in (27) and (28) can be evaluated recursively based on Bayes' theorem in which the propagations of the distributions are performed by using the state-transition model in (2), and their corrections are carried out by means of the measurement model in (24). In this work, a popular sample-based estimator, namely the UKF [20], is employed as the recursive state estimator and constraints are posed on the shape states $\lambda_k(i)$, $k = 1, \ldots, P$ of a target shoe such that

$$\varepsilon_{\text{MIN}} < \lambda_k(i) < \varepsilon_{\text{MAX}} \quad (29)$$

where $\varepsilon_{\text{MAX}} > 0$ and $\varepsilon_{\text{MIN}} > 0$ represent, respectively, the maximum and minimum permissible sizes of a target shoe to be followed. The proposed approach to tracking a target shoe is summarized in Algorithm 2. Since Algorithm 2 is independent of shoe index L(R), the index is dropped unless otherwise stated (also in an effort to reduce notational clutter).

Algorithm 2: Proposed Approach to Extended Target Tracking Based on Multisensor Information Fusion.

Step-1: Tracking System Initialization
1) Initial state estimate: $\overline{\boldsymbol{X}}(0)$
2) Noise covariances: $\boldsymbol{Q}_{\text{po}}$, $\boldsymbol{Q}_{\text{sh}}$, $\mathbf{R}_{\boldsymbol{\Delta}(\text{vis})}$, $\mathbf{R}_{\boldsymbol{\Delta}(\text{rng})}$
3) Kernel parameters: $\boldsymbol{\Phi}_{\tilde{\lambda}} = \{\rho_0, \iota_0\}$, $\boldsymbol{\Phi}_r = \{\rho_r, \iota_r\}$
4) Moment order in heteroscedastic regression: ω
Step-2: Recursive State Estimation with UKF
for $i = 1, 2, \ldots$ **do**
1) Prediction using the state transition model
Making use of (2) and taking (29) into account, evaluate $p(\boldsymbol{X}(i)|\text{IN}(0), \ldots, \text{IN}(i-1))$
2) Correction using visual information
for $j = 1 : \text{NV}_i$ **do**
Assimilate $\underline{\mathbf{z}}^{(i)}_{(\text{vis}),j}$ by making use of (24)
end for
3) Correction using sonar data
for $k = 1 : \text{NS}_i$ **do**
Assimilate $\underline{\mathbf{z}}^{(i)}_{(\text{rng}),k}$ by making use of (24)
end for
4) Output
return $p(\boldsymbol{X}(i)|\text{IN}(0), \ldots, \text{IN}(i))$
end for

VII. Experimental Results and Performance Evaluations

We evaluate and compare the predictive performance of four GP models, namely the standard GP, the VHGP, the original MLHGP, and the BMMLHGP developed in this work, on the real-world shoe-contour datasets (for both left and right shoes). Then, we evaluate the human following performance of our proposed HFBETT-MSIF using real-world sensor data obtained from a Raspberry Pi (RPi) 3 model B+ based *differentially driven* wheeled mobile robot, shown in Fig. 5(a), endowed with an RPi camera module v2 and a HC-SR04 US sensor. The human following performance of the BMMLHGP-based HFBETT-MSIF is compared with that of two single-sensor (vision/sonar)-based extended object trackers and three other multisensor fusion-based extended object tracking approaches, including the variants of the HFBETT-MSIF based on the standard GP, VHGP, and MLHGP, which are formulated in this work itself and named as MSIF-HFBETT-GP, MSIF-HFBETT-VHGP, and MSIF-HFBETT-MLHGP, respectively. We carry out a set of real-data experiments in four different photometrically affected as well as characteristically diverse environments:

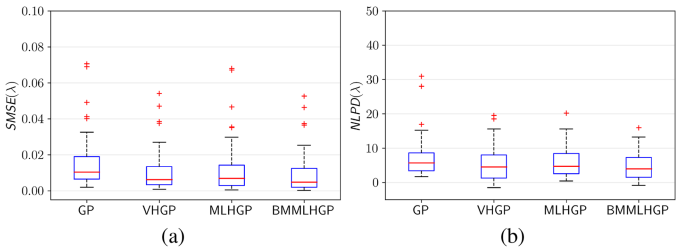

Fig. 6. Predictive performance assessment on real-world shoe-contour datasets over 100 random trials: (a) $SMSE(\lambda)$ on retrieving the noisy radial function λ and (b) $NLPD(\lambda)$ on predicting unknown radial distances.

1) *EN1:* A 10 m $\times$ 3.5 m corridor with rough concrete walls and multiple entrances.
2) *EN2:* An enclosed space of dimension 6 m $\times$ 5.5 m that contains steel furnitures and glass doors, aside from stone walls.
3) *EN3:* A 8.5 m $\times$ 4 m cluttered laboratory containing multiple cardboard boxes, tables, and chairs.
4) *EN4:* Another enclosed space of dimension 9 m $\times$ 4 m comprising concrete walls and structures, combined with wooden walls and doors.

For the purposes of real-life implementation, the complete software is written in a combination of *Opencv-Python* and *Cython*.

A. Comparison of the Predictive Performance of Different GP Models on the Shoe-Contour Datasets

We conduct experiments on two real-world shoe-contour datasets corresponding to a pair of homogeneous shoes (left and right). As quantitative performance measures, we use the *standardized mean squared error* (SMSE) and the *negative log probability density* (NLPD) [21]

$$SMSE(\lambda) = \frac{1}{L}\sum_{j=1}^{L}\frac{\left(m_{\lambda_{\circledast,j}} - \lambda_{\circledast,j}\right)^2}{var(\lambda_{\circledast})} \tag{30}$$

and

$$NLPD(\lambda) = -\frac{1}{L}\sum_{j=1}^{L}\log p\left(\lambda_{\circledast,j}|\psi_{\circledast,j},\, DS\right)$$
$$= \frac{1}{2\,L}\sum_{j=1}^{L}\left[\log\left(2\pi s^2_{\lambda_{\circledast,j}}\right) + \frac{\left(\lambda_{\circledast,j} - m_{\lambda_{\circledast,j}}\right)^2}{s^2_{\lambda_{\circledast,j}}}\right] \tag{31}$$

where $m_{\lambda_{\circledast,j}}$ and $s^2_{\lambda_{\circledast,j}}$, respectively, denote the posterior mean and variance of the predicted radial distance $\lambda_{\circledast,j}$ at test angle $\psi_{\circledast,j}$, $var(\lambda_{\circledast})$ is the variance of the predicted radial distances at all test angles $\boldsymbol{\Psi}_{\circledast} = \{\psi_{\circledast,1}, \ldots, \psi_{\circledast,L}\}$, and L is the number of test angles. For the shoe-contour datasets, the $SMSE(\lambda)$ and $NLPD(\lambda)$ obtained are shown in Fig. 6 , which firmly establish the superiority of the BMMLHGP.

B. Assessment of the Real-Life Shoe Tracking Performance

A series of real people following experiments are conducted to compare six competing tracking methods, where both the kinematic and shape states of two homogeneous target shoes (left and right), each 33 cm $\times$ 13 cm, are estimated based on the real-world camera and sonar data acquired by the robot [see Fig. 5(b)], during human tracking in four experimental environments. In each experiment, the human moved 10 times along different trajectories in each environment. In each trial, the target person walked at an average speed of 0.35 m/s and the tracking process completed in about 22 s.

TABLE I
AVERAGED PERFORMANCE COMPARISON AMONG SIX COMPETING TRACKING ALGORITHMS

Tracker*	RMSE (position) (cm)				RMSE (heading) (°)				mIOU				MET (s)
	EN1	EN2	EN3	EN4	EN1	EN2	EN3	EN4	EN1	EN2	EN3	EN4	
T1	6.38	6.19	6.25	6.07	5.60	5.35	5.43	5.19	0.56	0.59	0.58	0.61	0.56
T2	5.69	5.38	5.53	5.31	5.08	4.83	4.96	4.77	0.61	0.66	0.64	0.67	**0.48**
T3	3.07	2.92	2.99	2.86	3.34	3.19	3.25	3.12	0.79	0.82	0.81	0.83	1.04
T4	2.90	2.77	2.81	2.72	3.09	2.96	3.02	2.88	0.81	0.84	0.83	0.85	1.63
T5	3.53	3.36	3.43	3.28	4.05	3.83	3.92	3.72	0.73	0.76	0.75	0.78	0.91
T6	**2.83**	**2.71**	**2.74**	**2.67**	**2.95**	**2.83**	**2.88**	**2.77**	**0.82**	**0.85**	**0.84**	**0.86**	0.93

*T1: Vision, T2: Sonar, T3: HFBETT-MSIF-MLHGP, T4: HFBETT-MSIF-VHGP, T5: HFBETT-MSIF-GP, and T6: HFBETT-MSIF.
The bold entities represent the *best results* obtained with the competing tracking algorithms corresponding to four performance metrics.

Table I compares average *root-mean-square error* (RMSE) between the estimated and ground-truth positions, RMSE between the estimated and ground-truth heading angles, and *mean Intersection over Union* (mIoU) between the estimated and ground-truth shapes, in each environment, along with *Mean Execution Time* (MET) of the filtering process (involving UKF) at each iteration. In terms of the accuracy in estimating the kinematic and shape parameters, the multisensor fusion approaches perform considerably better compared with the single-sensor-based methods, and further, the BMMLHGP-based HFBETT-MSIF presents overall superior performance among all six competing trackers. This can be attributed to the fact that BMMLHGP provides an almost unbiased estimate of the IDN level, resulting in improved numerical accuracy and stability, as well as minimized risks of global overestimation or underestimation of the noise function $e(\psi_k)$.

For comparing qualitative performances, the study considers four representative runs along curvilinear trajectories, viz. `RUNI`, `RUNII`, `RUNIII`, and `RUNIV`, carried out in environments EN1, EN2, EN3, and EN4, respectively. The corresponding estimation results are shown in Fig. 7(a)–(d), respectively. In an effort to reduce visual clutter, the estimation results provided by the HFBETT-MSIF and two single sensor-based methods are depicted in Fig. 7(a) and (c), whereas the results for the HFBETT-MSIF, its variants based on the original MLHGP and standard homoscedastic GP (i.e., HFBETT-MSIF-MLHGP and HFBETT-MSIF-GP, respectively), are shown in Fig. 7(b) and (d). Referring to Fig. 7(a)–(d), the same parameters are used to initialize the trackers, viz. two identical ellipses with the widths, heights, and initial headings of the left and right shoes, respectively, are utilized as the initial shapes. As evidenced by Fig. 7, the proposed HFBETT-MSIF significantly outperforms the single-sensor (vision/sonar) trackers and also shows better results than the HFBETT-MSIF-MLHGP and HFBETT-MSIF-GP, when it comes to estimating the target shoes' pose and shape states. Due to space limitations, additional state estimation results for `RUNI`, `RUNII`, `RUNIII`, and `RUNIV` are depicted in Section S-V of the Supplementary Material.

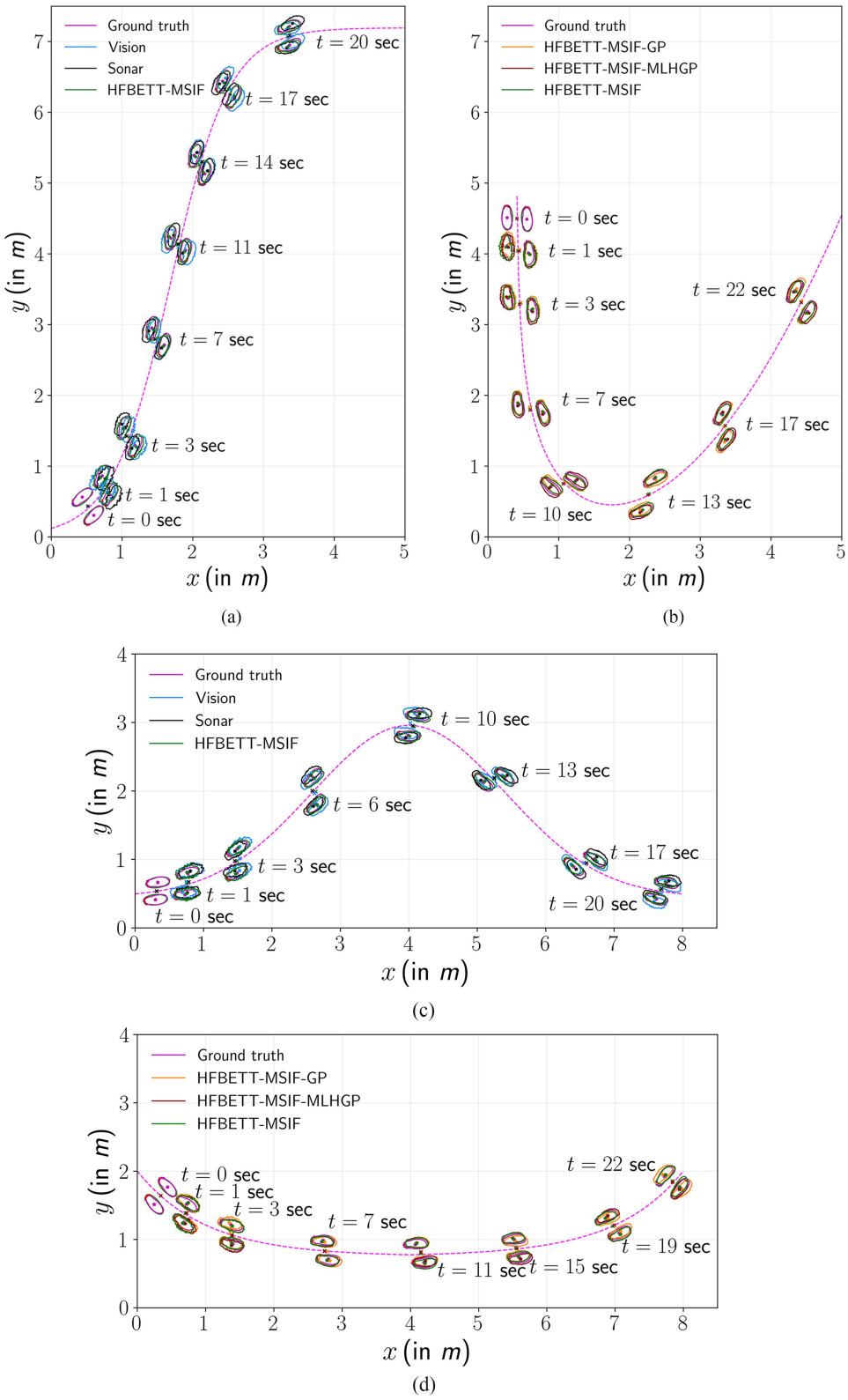

Fig. 7. Ground truth as well as estimated positions and shapes of the target shoes for (a) RUN I, (b) RUN II, (c) RUN III, and (d) RUN IV. (*Blue* and *red* "+": measurements from US sensor and camera, respectively. "⋆": center (x_0, y_0) of a shoe. "×": projected center of gravity of the human target).

VIII. Conclusion

This article proposed a new solution for human feature-based extended object tracking, using fusion of information from camera and US sensor of a mobile robot. The work showed how the onboard sensing modules can function successfully in both independent and cooperative manners. The study also illustrated how HGP regression can be utilized for human shoe tracking involving heteroscedastic noises. The work demonstrated a successful real-life implementation of a popular most likely noise approach, namely the MLHGP and also developed an advanced variant of the MLHGP, called BMMLHGP model, to achieve further enhanced real-life performances.

Acknowledgment

The author S. K. Pal would like to thank the National Science Chair, Science and Engineering Research Board, Department of Science and Technology (SERB-DST), Government of India.

References

[1] C. Micheloni, G. L. Foresti, C. Piciarelli, and L. Cinque, "An autonomous vehicle for video surveillance of indoor environments," *IEEE Trans. Veh. Technol.*, vol. 56, no. 2, pp. 487–498, Mar. 2007.

[2] W. Chung, H. Kim, Y. Yoo, C.-B. Moon, and J. Park, "The detection and following of human legs through inductive approaches for a mobile robot with a single laser range finder," *IEEE Trans. Ind. Electron.*, vol. 59, no. 8, pp. 3156–3166, Aug. 2012.

[3] H. T. Duong and Y. S. Suh, "Human gait tracking for normal people and walker users using a 2D LiDAR," *IEEE Sens. J.*, vol. 20, no. 11, pp. 6191–6199, Jun. 2020.

[4] M. Wang et al., "Accurate and real-time 3-D tracking for the following robots by fusing vision and ultrasonar information," *IEEE/ASME Trans. Mechatron.*, vol. 23, no. 3, pp. 997–1006, Jun. 2018.

[5] R. C. Luo, N.-W. Chang, S.-C. Lin, and S.-C. Wu, "Human tracking and following using sensor fusion approach for mobile assistive companion robot," in *Proc. IEEE Annu. Conf. Ind. Electron.*, Porto, Portugal, 2009, pp. 2235–2240.

[6] J. Yuan, H. Chen, F. Sun, and Y. Huang, "Multisensor information fusion for people tracking with a mobile robot: A particle filtering approach," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 9, pp. 2427–2442, Sep. 2015.

[7] S. Lee and J. McBride, "Extended object tracking via positive and negative information fusion," *IEEE Trans. Signal Process.*, vol. 67, no. 7, pp. 1812–1823, Apr. 2019.

[8] N. Wahlström and E. özkan, "Extended target tracking using Gaussian processes," *IEEE Trans. Signal Process.*, vol. 63, no. 16, pp. 4165–4178, Aug. 2015.

[9] K. Liu, Y. Li, X. Hu, M. Lucu, and W. D. Widanage, "Gaussian process regression with automatic relevance determination Kernel for calendar aging prediction of lithium-ion batteries," *IEEE Trans. Ind. Informat.*, vol. 16, no. 6, pp. 3767–3777, Jun. 2020.

[10] R. R. Richardson, C. R. Birkl, M. A. Osborne, and D. Howey, "Gaussian process regression for in-situ capacity estimation of lithium-ion batteries," *IEEE Trans. Ind. Informat.*, vol. 15, no. 1, pp. 127–138, Jan. 2019.

[11] C. K. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[12] M. Lázaro-Gredilla, M. K. Titsias, J. Verrelst, and G. Camps-Valls, "Retrieval of biophysical parameters with heteroscedastic Gaussian processes," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 4, pp. 838–842, Apr. 2014.

[13] H. Shi, K. Worden, and E. J. Cross, "A cointegration approach for heteroscedastic data based on a time series decomposition: An application to structural health monitoring," *Mech. Syst. Signal Process.*, vol. 120, pp. 16–31, 2019.

[14] P. W. Goldberg, C. K. I. Williams, and C. M. Bishop, "Regression with input-dependent noise: A Gaussian process treatment," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 10, M. I. Jordan, M. J. Kearns, and S. A. Solla, eds., Hoboken, NJ, USA: MIT Press, 1998, pp. 493–499.

[15] L. Muñoz-González, M. Lázaro-Gredilla, and A. R. Figueiras-Vidal, "Divisive Gaussian processes for nonstationary regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 11, pp. 1991–2003, Nov. 2014.

[16] M. Lázaro-Gredilla and M. K. Titsias, "Variational heteroscedastic Gaussian process regression," in *Proc. 28th Int. Conf. Mach. Learn.*, Bellevue, WA, USA, 2011, pp. 841–848.

[17] M. Hartmann and J. Vanhatalo, "Laplace approximation and natural gradient for Gaussian process regression with heteroscedastic student-t model," *Statist. Comput.*, vol. 29, no. 4, pp. 753–773, Jul. 2019.

[18] K. Kersting, C. Plagemann, P. Pfaff, and W. Burgard, "Most likely heteroscedastic Gaussian process regression," in *Proc. 24th Int. Conf. Mach. Learn.*, Corvallis, OR, USA, 2007, pp. 393–400.

[19] N. Quadrianto, K. Kersting, M. D. Reid, T. S. Caetano, and W. L. Buntine, "Kernel conditional quantile estimation via reduction revisited," in *Proc. IEEE 9th Int. Conf. Data Mining*, Miami, FL, USA, 2009, pp. 938–943.

[20] R. Yildiz, M. Barut, and E. Zerdali, "A comprehensive comparison of extended and unscented Kalman filters for speed-sensorless control applications of induction motors," *IEEE Trans. Ind. Informat.*, vol. 16, no. 10, pp. 6423–6432, Oct. 2020. .

[21] Q.-H. Zhang and Y.-Q. Ni, "Improved most likely heteroscedastic Gaussian process regression via Bayesian residual moment estimator," *IEEE Trans. Signal Process.*, vol. 68, pp. 3450–3460, May 2020.

[22] R. V. Hogg, E. A. Tanis, and D. L. Zimmerman, *Probability and Statistical Inference*, 9th ed. Upper Saddle River, NJ, USA: Pearson, 2015.

[23] P. Paral, A. Chatterjee, and A. Rakshit, "Vision sensor-based shoe detection for human tracking in a human–robot coexisting environment: A photometric invariant approach using DBSCAN algorithm," *IEEE Sens. J.*, vol. 19, no. 12, pp. 4549–4559, Jun. 2019.

[24] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. 6th Int. Conf. Comp. Vis.*, Bombay, India, 1998, pp. 839–846.

[25] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. New York City, NY, USA: Pearson, 2017.

[26] P. Paral, A. Chatterjee, and A. Rakshit, "Human position estimation based on filtered sonar scan matching: A novel localization approach using DENCLUE," *IEEE Sens. J.*, vol. 21, no. 6, pp. 8055–8064, Mar. 2021.

[27] A. Winkelbauer, "Moments and absolute moments of the normal distribution," 2014. [Online]. Available: http://arxiv.org/pdf/1209.4340.pdf