

Robot perceptual adaptation to environment changes for long-term human teammate following

The International Journal of
Robotics Research
1–15
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0278364919896625
journals.sagepub.com/home/ijr


Sriram Siva  and Hao Zhang

Abstract

Perception is one of the several fundamental abilities required by robots, and it also poses significant challenges, especially in real-world field applications. Long-term autonomy introduces additional difficulties to robot perception, including short- and long-term changes of the robot operation environment (e.g., lighting changes). In this article, we propose an innovative human-inspired approach named robot perceptual adaptation (ROPA) that is able to calibrate perception according to the environment context, which enables perceptual adaptation in response to environmental variations. ROPA jointly performs feature learning, sensor fusion, and perception calibration under a unified regularized optimization framework. We also implement a new algorithm to solve the formulated optimization problem, which has a theoretical guarantee to converge to the optimal solution. In addition, we collect a large-scale dataset from physical robots in the field, called perceptual adaptation to environment changes (PEAC), with the aim to benchmark methods for robot adaptation to short-term and long-term, and fast and gradual lighting changes for human detection based upon different feature modalities extracted from color and depth sensors. Utilizing the PEAC dataset, we conduct extensive experiments in the application of human recognition and following in various scenarios to evaluate ROPA. Experimental results have validated that the ROPA approach obtains promising performance in terms of accuracy and efficiency, and effectively adapts robot perception to address short-term and long-term lighting changes in human detection and following applications.

Keywords

Robot perceptual adaptation, multisensory fusion, long-term autonomy, human following

1. Introduction

Perception is an essential capability for autonomous robots to perceive the surrounding world and their own states so that they can accomplish other fundamental functionalities, such as navigation and human–robot teaming. For example, robots following humans to perform search and rescue missions and robots working collaboratively in human–robot teaming both need the ability to collect information about their environments to make decisions; robots also need to collect information about themselves and terrain characteristics to perform navigation in unstructured field environments.

While the robot perception research community has made impressive strides over recent years, it remains an unsolved problem for real-world field robotics applications, especially for robots that operate in dynamic and unstructured field environments. Several key challenges in robot perception must be addressed. As modern robotic platforms are often equipped with a variety of sensors it is

necessary to effectively integrate this multisensory data. In addition, there is often limited on-board computing power for robots operating in field environments, and efficient algorithms are required to deliver real-time performance with this limited computational power.

In the past several years, the robotics community has paid more attention towards long-term autonomy: robots that are able to operate for days, months, years, and eventually a lifetime. Long-term autonomy introduces new additional challenges to robot perception. For example, as demonstrated in Figure 1, when robots follow a human teammate to perform search and rescue missions over long

Human-Centered Robotics Lab, Colorado School of Mines, Golden, CO, USA

Corresponding author:

Hao Zhang, Human-Centered Robotics Lab, Colorado School of Mines, Golden, CO 80401, USA.
Email: hzhang@mines.edu

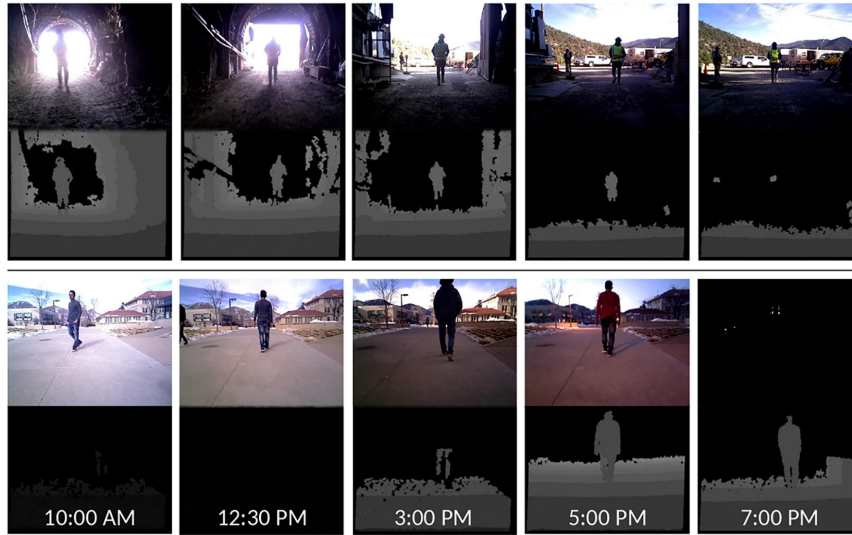


Fig. 1. Motivating examples of robot perceptual adaptation in human teammate following applications. When a robot follows a human during a long-term operation (e.g., search and rescue), the robot requires the capability of perceptual adaptation to adapt to fast changes (e.g., when moving from a dark tunnel to a bright open area shown in the top row) and long-term changes (e.g., different times of the day shown in the bottom row) and in order to avoid perception failures. The problem of robot adaptation to fast and dramatic environment changes has not yet been well addressed. For example, given a stream of color and depth data as the input, existing methods of human detection cannot adaptively choose sensing modalities and often lose the person when the robot travels from a dark mine to a bright open area with fast and dramatic lighting changes.

periods of time, robots often need to operate in indoor and outdoor environments, navigate between dark tunnels and bright open areas, and perform human detection under significant illumination variations at different times of the day. These changing, unstructured field environments often cause failures for robot perception, such as an inability to continue tracking teammates. Thus, addressing the short-term and long-term environment changes is vital to enable long-term autonomy.

Several methods were previously implemented to address the problem of long-term autonomy in dynamic unstructured environments. A widely used paradigm is to learn a unified representation of unstructured environments, which can be applied to various scenarios at different time. For example, in long-term place recognition (also known as loop-closure detection), methods based on holistic layouts (Han et al., 2017a; Wu and Rehg, 2011) or landmarks (Sunderhauf et al., 2015; Yuan et al., 2011) were designed to construct a representation of the environment that is robust to long-term variations over time to find a match with previously visited places. These techniques look past the changes of the environments over discrete time points (e.g., morning versus evening and summer versus winter) to determine the underlying place features that are most representative. However, these techniques learn fixed unified representations that do not adapt or change according to environment dynamics. A conventional robot adaptation paradigm is based upon case-based reasoning (Watson and Marir, 1994), which accomplishes adaptation by switching between multiple perception modalities

depending upon the current context. However, they are limited to cases that have been manually predefined, which makes them impractical for dynamically changing unstructured environments that may have a large number of context cases in a continuous and high-dimensional space. In addition, online learning (Hagras et al., 2004; Liu et al., 2008; Tapus et al., 2010) was also widely studied, where adaptation is achieved by continuously training the model using streams of data in an online fashion. Because online learning models drift and often require many iterations to converge to an optimal model again, they are less effective in scenarios when a robot needs to adapt to fast or repeated changes of the environment.

In this article, we propose a novel approach named *robot perceptual adaptation* (ROPA) that learns a dynamical fusion of multisensory perception data, which is adaptive to continuous short-term and long-term environment changes. ROPA is inspired by the observation that the human eye is able to adapt to a wide range of lighting conditions, and by the psychological findings in perceptual adaptation of humans. Human perceptual adaptation is a fundamental property of perceptual processing to “calibrate perception to current inputs” (Rhodes et al., 2010) and to “maintain the match between visual coding and the visual environment” (Vinas et al., 2012; Webster, 2011). Our application in this article focuses on human detection based upon different types of features extracted from color and depth sensors installed on a mobile robot to perform long-term human teammate following. During real-world human–robot teaming in a field environment, autonomous robots

often need to follow a human teammate to perform certain operations (e.g., following a skier while capturing videos and following a rescuer while carrying equipment as shown in Figure 1) in unstructured and dynamic scenarios that evolve over time.

ROPA is a principled approach that formulates perceptual adaptation as a joint learning problem to simultaneously learn a base perception model to optimally fuse multisensory input data and a calibration term to adapt to environment changes. In order to fuse multisensory inputs, we implement sparsity-inducing norms that enforce the base perception model to learn sparse weights of multisensory input features and apply the weights for data fusion. To achieve perception calibration, we estimate the representativeness of the input feature modalities. Representativeness of a feature modality is referred to as its capability to represent the environment. When the environment changes, the representativeness of each feature modality also changes (e.g., depth can better represent dark environments). Accordingly, by automatically selecting feature modalities that are more representative in a specific environment, our approach provides a calibration of the perception model according to the environmental change. All the above components are mathematically integrated into a joint learning formulation under the unified theoretical framework of regularized optimization. In long-term human following, for each of new data instances ROPA uses the joint base perception model and calibration term to classify humans under environment changes; classification results are applied by a decision-making module to control the robot to navigate and follow the human. In order to evaluate ROPA and benchmark techniques for robot perceptual adaptation in human-following applications, we collect a new large-scale dataset called PEAC. The dataset consists of multisensory perception input data collected from physical mobile robots in real-world field applications under short-term and long-term environment variations. Our experimental results over the PEAC dataset have validated that the proposed approach outperforms previous state-of-the-art methods for human following, obtains real-time performance, and is capable to address long-term and short-term environment changes.

The contributions of this article¹ are as follows.

- We propose a fresh human-inspired idea that addresses a new research problem of robot perceptual adaptation to short- and long-term environment variations through calibrating robot perception to the current context in teammate following applications.
- We introduce ROPA that estimates the importance of heterogeneous sensory data, integrates all information to build a perception model, and, more importantly, calibrates the perception model to adapt to short- and long-term variations of the environment.
- We implement a new algorithm to solve the formulated optimization problem, which possesses a theoretical guarantee to converge to the optimal solution.

- As a practical contribution, we collect a new large-scale dataset from mobile robots, called *perceptual adaptation to environment changes* (PEAC), which includes three representative human-following scenarios of long-term autonomy in field applications to benchmark methods for robot adaptation to short- and long-term environment changes.

The remainder of this article is organized as follows. We review the related work on robot perception and adaptation in Section 2. The proposed adaptation approach is discussed in Section 3, and its optimization solver is described in Section 3.4. After describing the new dataset and our applications, we present and analyze the experimental results in Section 5. Finally, we conclude the article in Section 7.

2. Related work

In this section, we provide a review of related research on long-term autonomy, robot adaptation, and human following. Long-term autonomy has received an increase in attention from the robotics community, because of increasing use of robots in environments presenting long-term dynamics (for example, robot following of humans in field environments throughout long periods of time can experience long-term changes). Robot adaptation is considered one viable solution to enable long-term autonomy, as adaptation enables the robot to cope with the changing environment.

2.1. Long-term autonomy

As robots are leaving factories and entering unstructured and dynamic environments, the ability of robots to reliably operate over long periods of time under dynamically changing conditions needs to be addressed. Several learning-based approaches have been proposed to support robot long-term autonomy.

These approaches can be generally categorized into two paradigms: online learning and representation learning. (1) *Online learning* methods address long-term autonomy by continuously or iteratively updating model parameters during task execution (Kleiner et al., 2002; Thrun and Mitchell, 1995; White et al., 2012). The online learning paradigm is widely applied to a variety of applications with robots operating in dynamic or evolving environments (Leite et al., 2013), including health care, education, and assistive robotics in work Kanda et al. (2010) and home Fernaeus et al. (2010) environments. (2) *Representation learning* aims at learning from data to construct a representation of the robot's surrounding environment, which is robust or insensitive to environment variations. This paradigm is widely used in long-term place recognition (also known as loop-closure detection) to achieve simultaneous localization and mapping (SLAM) in long-term settings (Lowry et al., 2014; Neubert et al., 2013; Rosen et al.,

2016; Sünderhauf et al., 2014). Most of the learning techniques for long-term place recognition focus on creating representations that encode the holistic layout of the environment based upon global feature extraction (Lowry et al., 2016), deep learning (Arroyo et al., 2016; Sunderhauf et al., 2015), multimodal feature integration (Han et al., 2016), and spatio-temporal fusion (Zhang et al., 2016). Several recent methods use landmarks to create long-term environment representations (Sunderhauf et al., 2015; Yuan et al., 2011). However, these methods are specifically implemented for place recognition or robot localization problems, and cannot be applied to appropriate perception of objects of interest under short- and long-term context or environment changes. Tung et al. (2019) evaluated YOLO's (Redmon et al., 2016) performance during long-term changes on detecting an object of interest over long periods of time under various lighting conditions. It was observed that YOLO continuously struggles to detect the same object during sudden changes and during night time. Online learning methods lack the ability to address dramatic and fast changes. Although, they are effective in adapting to environments with slow and gradual changes, they fail under drastic changing conditions as they need time to converge.

2.2. Robot adaptation

Although robot learning (Argall et al., 2009) has been addressed by many researchers, robot adaptation still remains to receive comparatively less attention. This is because many robot systems are designed to be used in very specific domains for a brief period of time (Shibata and Tanie, 2000; Thrun et al., 1999). Early research focused on high-level behavior-based methods to address robot adaptation in general. For example, Parker (2000) developed various architectures to enable teams of heterogeneous robots to dynamically adapt their actions over time. Following a similar direction, case-based reasoning (Watson and Marir, 1994) methods were used by Floyd et al. (2015) and Zhang et al. (2005) for robot behavioral adaptation in evolving environments. Another adaptation scheme was proposed in Dettmann et al. (2014) to control complex robots, which selects a solution from a library of well-performing solutions, given specific tasks and conditions. In these early methods, robot adaptation is generally manually pre-determined, requiring significant domain expertise.

Adaptation based on human intent is studied in the domain of human-robot interaction and collaboration. One of the key components of this adaptation is to be able to recognize human intent and activities (Evrard et al., 2009; Gribovskaya et al., 2011; Kosuge and Kazamura, 1997). For example, a robot may need to recognize human intent and activities based upon visual feedback (Agravante et al., 2014) or audio command (Medina et al., 2012). Another popular learning-based adaptation paradigm is reinforcement learning, which is usually designed for robot behavior adaptation (Jevtić et al., 2018; Mitsunaga et al., 2006;

Ritschel and André, 2017). Recently, several methods (Kruijff-Korbayová et al., 2015; Li et al., 2015; Nikolaidis et al., 2017a,b) studied co-adaptation problems addressing how robots and humans on the same team can collaboratively adapt to each other and complete the joint task effectively. Almost all learning-based methods focus on behavior adaptation. The critical problem of robot perceptual adaptation has not been well understood and studied.

2.3. Human following

A large portion of methods used in human following typically involve detecting humans (Dalal et al., 2006; Redmon et al., 2016) and tracking humans (Nam and Han, 2016; Sminchisescu and Triggs, 2003).

Researchers have tried to solve the problem of human following using different approaches. Most of these methods, in general, involve dividing the query image into several regions and using region proposals to predict the bounding boxes the human might be in (Redmon et al., 2016) and Ren et al. (2015). The core of most of these methods involve both feature extraction (to represent the regions from the bounding box in a different representation space) and then a classifier (to classify if the input feature representation of that region is of a human or not). Feature extraction methods can be further categorized into local features (Bay et al., 2006; Calonder et al., 2010; Mikolajczyk and Schmid, 2001; Rublee et al., 2011) that describe the local information from different regions of interest, and global features (Arroyo et al., 2015; Dalal and Triggs, 2005; Oliva and Torralba, 2006) that generally describe the image as a whole. A global feature vector is generated based on the feature statistics. Dalal and Triggs (2005) is one of the most used global features for whole-body human detection that captures the local shape and edge information of the whole image. Dalal et al. (2006) used optical flow field's internal difference to recognize moving humans. While local and global features can both be used to detect humans, global features have proved to give better results (Wang et al., 2009). Recently, many methods use a combination of different sensors to achieve the capability of human detection in autonomous robots (Keller et al., 2011; Xia et al., 2011). It has also been proven that use of multi-sensory and multi-feature representation can significantly improve performance in long-term settings (Han et al., 2017a; Siva and Zhang, 2018a). Although these methods perform well in most scenarios, they cannot be considered in situations where the robot needs to adapt with the environment. Our approach can take in values from different sensors and calibrate their importance based on the environmental context, allowing robots the ability to adapt with the environment.

3. The ROPA approach

To address multisensory robot perceptual adaptation in long-term autonomy, we propose the ROPA approach to

learn and calibrate a perception model that can adapt to short-term and long-term environment changes.

Notation. Matrices are represented by boldface uppercase letters, and vectors by boldface lowercase letters. Given a matrix $\mathbf{U} = \{u_{ij}\} \in \mathbb{R}^{n \times m}$, we represent the i th row and j th column as $\mathbf{u}^i$ and $\mathbf{u}_j$, respectively. The ℓ_2 -norm of the vector $\mathbf{u}$ is defined as $\|\mathbf{u}\|_2 = \sqrt{\mathbf{u}^\top \mathbf{u}}$. The Frobenius norm of the matrix $\mathbf{U}$ is defined as $\|\mathbf{U}\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^m u_{ij}^2}$.

3.1. Multimodal sensor fusion

Given a collection of n data instances (i.e., data samples in a dataset), the extracted normalized feature vectors are represented as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the feature vector from the i th instance. We assume that the features are extracted from different robot sensors (e.g., color and depth), and various types of features are extracted, with each type named a *modality*. That is, a modality is the set of features that are computed using a feature extraction method from the input of a specific sensor. Then, each heterogeneous feature vector $\mathbf{x}_i \in \mathbb{R}^d$ is assumed to consist of m -modalities of normalized features, such that $d = \sum_{j=1}^m d_j$. The label vector of classes associated with $\mathbf{X}$ is denoted by $\mathbf{Y} = [\mathbf{y}^1; \mathbf{y}^2; \dots; \mathbf{y}^n] \in \mathbb{R}^{n \times c}$, where c is the number of classes. Each element y_{ij} of the matrix $\mathbf{Y}$ indicates how likely the input feature vector $\mathbf{x}_i$ belongs to the j th class and is manually labeled as the ground truth during the training phase.

Then, we formulate the multisensory recognition task as an optimization problem using the objective:

$$\min_{\mathbf{W}} \|\mathbf{Y} - (\mathbf{X}^\top \mathbf{W} + \mathbf{1}_n \mathbf{b}^\top)\|_F^2 \quad (1)$$

where $\mathbf{1}_n \in \mathbb{R}^{n \times 1}$ is the constant vector of all ones, $\mathbf{b} \in \mathbb{R}^{c \times 1}$ is the bias vector that can be calculated by $\mathbf{b} = \mathbf{Y}^\top \mathbf{1}_n / n$. The solution to the optimization problem in (1) is a parameter matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_c] \in \mathbb{R}^{d \times c}$, which consists of the weights $\mathbf{w}_i \in \mathbb{R}^d$ of each element in the feature vector with respect to the i th class.

Different types of features encode different attributes of the environment (e.g., shape, color, and edges). When fusing the features, some modalities are more informative than others depending on the robot operation environment, and it is desirable to estimate the importance of each modality. Inspired by sparse optimization (Han et al., 2017a), to identify discriminative modalities, we design a norm $\mathcal{R}_M(\mathbf{W})$ as a regularizer to (1), which enforces sparsity among the modalities and the grouping effect of the features within the same modality. The $\mathcal{R}_M$ -norm can be expressed as $\mathcal{R}_M(\mathbf{W}) = \sum_{i=1}^c \sum_{j=1}^m \|\mathbf{w}_i^j\|_2$, which applies the ℓ_2 -norm to the weights of feature elements within each modality and the ℓ_1 -norm across different modalities. As $\mathcal{R}_M$ encodes the weight structure among modalities, we call it a modality norm.

As modern robots are usually equipped with a variety of sensors (e.g., color and depth), it is also desirable to estimate the importance of each sensor for multisensory robot perception. For example, a robot operating in dark can benefit more from depth sensors rather than color sensors. To meet this need, we introduce a sensory norm $\mathcal{R}_S$ to identify the discriminative sensors, which is defined as $\mathcal{R}_S(\mathbf{W}) = \sum_{i=1}^c \sum_{k=1}^l \|\mathbf{w}_i^k\|_2$. It applies the ℓ_2 -norm to the weights of the features computed from the same sensor, and applies the ℓ_1 -norm to the weights of features from different sensors.

By applying both modality and sensory norms, we formulate multisensory sensor fusion as a regularized optimization problem with the following objective function (where α is a trade-off hyperparameter):

$$\min_{\mathbf{W}} \|\mathbf{Y} - (\mathbf{X}^\top \mathbf{W} + \mathbf{1}_n \mathbf{b}^\top)\|_F^2 + \alpha(\mathcal{R}_M(\mathbf{W}) + \mathcal{R}_S(\mathbf{W})) \quad (2)$$

3.2. Perception calibration

The key novelty of this article is the introduction of the perception calibration capability, which is inspired by the psychology study on how human perception adapts in a given environmental context. Mathematically, we denote the environmental context as a matrix $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n] \in \mathbb{R}^{s \times n}$, where $\mathbf{e}_i = [e_i^1, e_i^2, \dots, e_i^s]^\top \in \mathbb{R}^s$ is a low-dimensional vector consisting of s environmental context variables (e.g., lighting, fog intensity, and ground traction) obtained along with the i th data instance.

To achieve perception calibration, we estimate the representativeness of each feature to represent the environmental context. Representativeness of a feature is referred to as its capability of representing the environment. When the environmental context changes (e.g., lighting variations), we can estimate the feature representativeness change to still represent the environment under such context changes. Therefore, estimating the feature representativeness change encodes our insight of calibrating the perception model (i.e., dynamically adjusting the weights of the features) according to the context. To achieve our insight, two procedures need to be performed: (1) computing the representativeness of the features to represent the environment in each instance, and (2) associating the computed feature representativeness with the context variables.

We denote the feature representativeness to represent the environment in the instances as $\mathbf{G} = [\mathbf{g}^1, \mathbf{g}^2, \dots, \mathbf{g}^n]^\top \in \mathbb{R}^{n \times d}$. Then, each $\mathbf{g}^i \in \mathbb{R}^d$ denotes the representativeness of the feature vector $\mathbf{x}_i$ to represent the environment, which can be estimated by $\mathbf{g}^i \mathbf{x}_i = 1$. By computing the Moore–Penrose inverse (also known as the pseudo-inverse), we can obtain $\mathbf{g}^i = \mathbf{x}_i^\top (\mathbf{x}_i \mathbf{x}_i^\top)^{-1}$.

We associate the representativeness matrix $\mathbf{G}$ (of the features to represent the environment) with the context variables through designing a novel loss function

Algorithm 1. An iterative algorithm to solve the formulated optimization problem in (4).

Input : Feature matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$, label matrix $\mathbf{Y} \in \mathbb{R}^{n \times c}$, context variables $\mathbf{E} \in \mathbb{R}^{s \times n}$

1. Calculate the representativeness matrix $\mathbf{G}$ from $\mathbf{X}$.
 2. Let $t = 1$. Initialize $\mathbf{W}(t)$ and $\mathbf{V}(t)$ by solving $\min_{\mathbf{W}} \|\mathbf{Y} - (\mathbf{X}^\top \mathbf{W} + \mathbf{1}_n \mathbf{b}^\top)\|_F^2$ and $\min_{\mathbf{V}} \|\mathbf{G} - \mathbf{E}^\top \mathbf{V}\|_F^2$.
 3. **while not converge do**
 4. Calculate the block diagonal matrix $\mathbf{D}^i(t+1) (1 \leq i \leq c)$, where the j th block is $\frac{1}{2\|\mathbf{w}_j(t)\|_2} \mathbf{I}_j$.
 5. Calculate the block diagonal matrix $\widehat{\mathbf{D}}^i(t+1) (1 \leq i \leq c)$, where the k th block is $\frac{1}{2\|\mathbf{w}_i^k(t)\|_2} \mathbf{I}_k$.
 6. Calculate the block diagonal matrix $\widetilde{\mathbf{D}}^j(t+1) (1 \leq j \leq d)$, where the p th block is $\frac{1}{2\|\mathbf{v}_j^p(t)\|_2} \mathbf{I}_p$.
 7. For each $\mathbf{w}_i (1 \leq i \leq c)$, $\mathbf{w}_i(t+1) = (\mathbf{X}\mathbf{X}^\top + \alpha \mathbf{D}^i(t+1) + \alpha \widehat{\mathbf{D}}^i(t+1))^{-1} \mathbf{X}(\mathbf{y}_i - \mathbf{b}_i)$.
 8. For each $\mathbf{v}_j (1 \leq j \leq d)$, $\mathbf{v}_j(t+1) = (\mathbf{E}\mathbf{E}^\top + \beta \widetilde{\mathbf{D}}^j(t+1))^{-1} \mathbf{E}\mathbf{g}_j$.
 9. $t = t + 1$.
-

Output: $\mathbf{W} = \mathbf{W}(t) \in \mathbb{R}^{n \times c}$; $\mathbf{V} = \mathbf{V}(t) \in \mathbb{R}^{s \times d}$

$\mathcal{L}(\mathbf{G}, \mathbf{E}; \mathbf{V})$, with the objective of learning a projection (parameterized by $\mathbf{V}$) from the context variables $\mathbf{E}$ to encode $\mathbf{G}$. Specifically, the loss function $\mathcal{L}$ is defined as

$$\mathcal{L}(\mathbf{G}, \mathbf{E}; \mathbf{V}) = \|\mathbf{G} - \mathbf{E}^\top \mathbf{V}\|_F^2 \quad (3)$$

where the parameter matrix $\mathbf{V} \in \mathbb{R}^{s \times d}$ includes the weights of the environment context variables with respect to the features. That is, $\mathbf{V}$ captures the underlying information of how much each of the elements in the feature vectors should change, given the change in the environmental context.

Similar to the motivation of estimating the importance of features and sensors, it is desirable to learn the importance of the environment context variables to calibrate the perception model. Therefore, we develop the new context norm $\mathcal{R}$ over the parameter matrix $\mathbf{V}$, defined as $\mathcal{R}_C(\mathbf{V}) = \sum_{p=1}^s \|\mathbf{v}^p\|_2$, which enforces the sparsity between the environment context variables with respect to all features.

Therefore, to adapt to short-term and long-term environment changes, we integrate the proposed perceptual calibration capability with multisensory fusion. We formulate multisensory perceptual adaptation as a joint learning problem under the unified regularized optimization framework, with the final objective function:

$$\min_{\mathbf{W}, \mathbf{V}} \|\mathbf{Y} - (\mathbf{X}^\top \mathbf{W} + \mathbf{1}_n \mathbf{b}^\top)\|_F^2 + \|\mathbf{G} - \mathbf{E}^\top \mathbf{V}\|_F^2 + \alpha(\mathcal{R}_M(\mathbf{W}) + \mathcal{R}_S(\mathbf{W})) + \beta \mathcal{R}_C(\mathbf{V}) \quad (4)$$

where β is a trade-off hyperparameter.

3.3 Multisensory robot perceptual adaptation

After solving the formulated regularized optimization problem in (4) during training (using the solver in Algorithm 1), we obtain the optimal $\mathbf{W}^* = [\mathbf{w}_1^*, \mathbf{w}_2^*, \dots, \mathbf{w}_c^*]$ and $\mathbf{V}^*$. Then, during online execution, given a newly

acquired data instance $\mathbf{x}$ and its environmental context variable $\mathbf{e}$, adaptive robot perception to determine the class label $y(\mathbf{x}, \mathbf{e})$ can be performed by

$$y(\mathbf{x}, \mathbf{e}) = \max_i (\mathbf{x}^\top)(\text{diag}(\mathbf{e}^\top \mathbf{V}^*) \mathbf{w}_i^* + \mathbf{w}_i^*) + b_i \quad (5)$$

where $\text{diag}(\cdot)$ denotes a function to covert a vector into a diagonal matrix. Given any vector $\mathbf{a} \in \mathbb{R}^z$, $\text{diag}(\mathbf{a}) = \mathbf{a} \times \mathbf{I}_{z \times z}$, where $\mathbf{I}$ is an identity matrix. The output of (5) provides the decision of whether a human is present or not. Given a query feature vector $\mathbf{x}$, and its associated environmental context $\mathbf{e}$, the term $\text{diag}(\mathbf{e}^\top \mathbf{V}^*) \mathbf{w}_i^*$ provides an estimation of the calibration needed to adjust each feature weight given $\mathbf{e}$ for the query $\mathbf{x}$. This calibration is then added to the feature weights $\mathbf{w}$ to determine the class label.

One of the advantages of our approach is that classification is integrated with feature learning and model calibration under the unified regularized optimization framework, thus eliminating the requirement of using additional classifiers. In addition, our formulation is based on convex linear models, which makes model parameter estimation and online inference highly efficient. Thus, ROPA is able to achieve high-speed onboard processing, which can significantly benefit real-time robotics applications.

3.4 Optimization algorithm

Although our formulation is convex, the objective function in (4) is difficult to solve in general because of the three non-smooth regularizers. Another contribution of this article is that we implement an iterative algorithm to solve the optimization problem, which is presented in Algorithm 1.

To learn the optimal $\mathbf{W}$, we compute the derivative of the objective function in (4) with respect to $\mathbf{w}_i (1 \leq i \leq c)$ and set it to a zero vector, as follows:

$$\mathbf{X}\mathbf{X}^\top \mathbf{w}_i - \mathbf{X}(\mathbf{y}_i - \mathbf{b}_i) + \alpha \mathbf{D}^i \mathbf{w}_i + \alpha \widehat{\mathbf{D}}^i \mathbf{w}_i = \mathbf{0} \quad (6)$$

where $\mathbf{D}^i (1 \leq i \leq c)$ is a block diagonal matrix with j th diagonal block computed by $\frac{1}{2\|\mathbf{w}_i^j\|_2} \mathbf{I}_j$; $\mathbf{w}_i^j$ is the j th segment of $\mathbf{w}_i$ consisting of the weights from the j th feature modality; $\hat{\mathbf{D}}^i$ is a block diagonal matrix with the k th diagonal block computed by $\frac{1}{2\|\mathbf{w}_i^k\|_2} \mathbf{I}_k$; and $\mathbf{w}_i^k$ is the k th segment of $\mathbf{w}_i$ consisting of the weights of features from k th modality. After solving (6), the vector $\mathbf{w}_i$ can be computed by

$$\mathbf{w}_i = (\mathbf{X}\mathbf{X}^\top + \alpha \mathbf{D}^i + \alpha \hat{\mathbf{D}}^i)^{-1} \mathbf{X}(\mathbf{y}_i - \mathbf{b}_i) \quad (7)$$

To compute the optimal value for $\mathbf{V}$, compute the derivative of the objective function in (4) with respect to the columns $\mathbf{v}_j$, ($1 \leq j \leq d$) of $\mathbf{V}$ and set the resulting expression to zero, as follows:

$$\mathbf{E}\mathbf{E}^\top \mathbf{v}_j - \mathbf{E}\mathbf{g}_j + \beta \tilde{\mathbf{D}}^j \mathbf{v}_j = \mathbf{0} \quad (8)$$

where $\tilde{\mathbf{D}}^j$ is a block diagonal matrix with j th block computed by $\frac{1}{2\|\mathbf{v}_j^p\|_2} \mathbf{I}_p$; $\mathbf{v}_j^p$ is the p th segment of $\mathbf{v}_j$ that specifies the weights of the p th environmental context variable with respect to the j th feature modality. Solving the above equation, we can obtain

$$\mathbf{v}_j = (\mathbf{E}\mathbf{E}^\top + \beta \tilde{\mathbf{D}}^j)^{-1} \mathbf{E}\mathbf{g}_j \quad (9)$$

Because $\mathbf{D}^i$ and $\hat{\mathbf{D}}^i$ depend on $\mathbf{W}$, and because $\tilde{\mathbf{D}}^j$ depends on $\mathbf{V}$, $\mathbf{D}^i$, $\hat{\mathbf{D}}^i$, and $\tilde{\mathbf{D}}^j$ are also unknown variables. Therefore, we design and implement an iterative algorithm to solve this optimization problem, which is presented in Algorithm 1. The proposed optimization solver holds a theoretical convergence guarantee to the global optimum, as described by theorem 1, which proves that Algorithm 1 decreases the value of the objective function with each iteration and converges to the global optimal value. First, we present a lemma.

Lemma 1. For any two given vectors $\mathbf{a}$ and $\mathbf{b}$, the following inequality relation holds: $\|\mathbf{b}\|_2 - \frac{\|\mathbf{b}\|_2^2}{2\|\mathbf{a}\|_2} \leq \|\mathbf{a}\|_2 - \frac{\|\mathbf{a}\|_2^2}{2\|\mathbf{a}\|_2}$

Proof. We have

$$-(\|\mathbf{b}\|_2 - \|\mathbf{a}\|_2)^2 \leq 0 \quad (10)$$

$$-\|\mathbf{b}\|_2^2 - \|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2\|\mathbf{a}\|_2 \leq 0 \quad (11)$$

$$2\|\mathbf{b}\|_2\|\mathbf{a}\|_2 - \|\mathbf{b}\|_2^2 \leq \|\mathbf{a}\|_2^2 \quad (12)$$

$$\|\mathbf{b}\|_2 - \frac{\|\mathbf{b}\|_2^2}{2\|\mathbf{a}\|_2} \leq \|\mathbf{a}\|_2 - \frac{\|\mathbf{a}\|_2^2}{2\|\mathbf{a}\|_2} \quad (13)$$

Theorem 1. Algorithm 1 converges to the optimal solution to the optimization problem in (4)

Proof. From Algorithm 1, we know that

$$\begin{aligned} \mathbf{W}(t+1) = \min_{\mathbf{W}} & \|\mathbf{Y} - (\mathbf{X}^\top \mathbf{W} + \mathbf{1}_n \mathbf{b}^\top)\|_F^2 \\ & + \alpha \sum_{i=1}^c \mathbf{w}_i^\top(t+1) \mathbf{D}^i(t+1) \mathbf{w}_i(t+1) \\ & + \alpha \sum_{i=1}^c \mathbf{w}_i^\top(t+1) \hat{\mathbf{D}}^i(t+1) \mathbf{w}_i(t+1) \end{aligned} \quad (14)$$

and

$$\begin{aligned} \mathbf{V}(t+1) = \min_{\mathbf{V}} & \|\mathbf{G} - \mathbf{E}^\top \mathbf{V}\|_F^2 \\ & + \beta \sum_{j=1}^d \mathbf{v}_j^\top(t+1) \tilde{\mathbf{D}}^j(t+1) \mathbf{v}_j(t+1) \end{aligned} \quad (15)$$

Then it can be derived that

$$\begin{aligned} \mathcal{F}(t+1) & + \alpha \sum_{i=1}^c \mathbf{w}_i^\top(t+1) \mathbf{D}^i(t+1) \mathbf{w}_i(t+1) \\ & + \alpha \sum_{i=1}^c \mathbf{w}_i^\top(t+1) \hat{\mathbf{D}}^i(t+1) \mathbf{w}_i(t+1) \\ & \leq \mathcal{F}(t) + \alpha \sum_{i=1}^c \mathbf{w}_i^\top(t) \mathbf{D}^i(t) \mathbf{w}_i(t) + \alpha \sum_{i=1}^c \mathbf{w}_i^\top(t) \hat{\mathbf{D}}^i(t) \mathbf{w}_i(t) \end{aligned} \quad (16)$$

and

$$\begin{aligned} \mathcal{J}(t+1) & + \beta \sum_{j=1}^d \mathbf{v}_j^\top(t+1) \tilde{\mathbf{D}}^j(t+1) \mathbf{v}_j(t+1) \leq \mathcal{J}(t) \\ & + \beta \sum_{j=1}^d \mathbf{v}_j^\top(t) \tilde{\mathbf{D}}^j(t) \mathbf{v}_j(t) \end{aligned} \quad (17)$$

substituting the values of $\mathbf{D}^i$, $\hat{\mathbf{D}}^i$ and $\tilde{\mathbf{D}}^j$, we obtain

$$\begin{aligned} \mathcal{F}(t+1) & + \alpha \sum_{i=1}^c \sum_{j=1}^m \frac{\|\mathbf{w}_i^j(t+1)\|_2^2}{2\|\mathbf{w}_i^j(t)\|_2} \\ & + \alpha \sum_{i=1}^c \sum_{k=1}^l \frac{\|\mathbf{w}_i^k(t+1)\|_2^2}{2\|\mathbf{w}_i^k(t)\|_2} \\ & \leq \mathcal{F}(t) + \alpha \sum_{i=1}^c \sum_{j=1}^m \frac{\|\mathbf{w}_i^j(t)\|_2^2}{2\|\mathbf{w}_i^j(t)\|_2} \\ & + \alpha \sum_{i=1}^c \sum_{k=1}^l \frac{\|\mathbf{w}_i^k(t)\|_2^2}{2\|\mathbf{w}_i^k(t)\|_2} \end{aligned} \quad (18)$$

and,

$$\begin{aligned} \mathcal{J}(t+1) & + \beta \sum_{p=1}^s \sum_{j=1}^d \frac{\|\mathbf{v}_j^p(t+1)\|_2^2}{2\|\mathbf{v}_j^p(t)\|_2} \\ & \leq \mathcal{J}(t) + \beta \sum_{p=1}^s \sum_{j=1}^d \frac{\|\mathbf{v}_j^p(t)\|_2^2}{2\|\mathbf{v}_j^p(t)\|_2} \end{aligned} \quad (19)$$



Fig. 2. Illustration of the scenarios in which the PEAC dataset was collected for the task of human teammate following.

From Lemma 1, we can derive the following equations:

$$\begin{aligned} & \sum_{i=1}^c \sum_{j=1}^m \|\mathbf{w}_i^j\|_2 - \sum_{i=1}^c \sum_{j=1}^m \frac{\|\mathbf{w}_i^j(t+1)\|_2^2}{2 \|\mathbf{w}_i^j(t)\|_2} \\ & \leq \sum_{i=1}^c \sum_{j=1}^m \|\mathbf{w}_i^j\|_2 - \sum_{i=1}^c \sum_{j=1}^m \frac{\|\mathbf{w}_i^j(t)\|_2^2}{2 \|\mathbf{w}_i^j(t)\|_2} \end{aligned} \quad (20)$$

$$\begin{aligned} & \sum_{i=1}^c \sum_{k=1}^l \|\mathbf{w}_i^k\|_2 - \sum_{i=1}^c \sum_{k=1}^l \frac{\|\mathbf{w}_i^k(t+1)\|_2^2}{2 \|\mathbf{w}_i^k(t)\|_2} \\ & \leq \sum_{i=1}^c \sum_{k=1}^l \|\mathbf{w}_i^k\|_2 - \sum_{i=1}^c \sum_{k=1}^l \frac{\|\mathbf{w}_i^k(t)\|_2^2}{2 \|\mathbf{w}_i^k(t)\|_2} \end{aligned} \quad (21)$$

$$\begin{aligned} & \sum_{p=1}^s \sum_{j=1}^d \|\mathbf{v}_j^p\|_2 - \sum_{p=1}^s \sum_{j=1}^d \frac{\|\mathbf{v}_j^p(t+1)\|_2^2}{2 \|\mathbf{v}_j^p(t)\|_2} \\ & \leq \sum_{p=1}^s \sum_{j=1}^d \|\mathbf{v}_j^p\|_2 - \sum_{p=1}^s \sum_{j=1}^d \frac{\|\mathbf{v}_j^p(t)\|_2^2}{2 \|\mathbf{v}_j^p(t)\|_2} \end{aligned} \quad (22)$$

Adding Equations (18)–(22) on both sides we get that

$$\begin{aligned} & \mathcal{F}(t+1) + \mathcal{J}(t+1) + \alpha \sum_{i=1}^c \sum_{j=1}^m \|\mathbf{w}_i^j(t+1)\|_2 \\ & + \alpha \sum_{i=1}^c \sum_{k=1}^l \|\mathbf{w}_i^k(t+1)\|_2 + \sum_{p=1}^s \sum_{j=1}^d \beta \|\mathbf{v}_j^p(t+1)\|_2 \\ & \leq \mathcal{F}(t) + \mathcal{J}(t) + \alpha \sum_{i=1}^c \sum_{j=1}^m \|\mathbf{w}_i^j(t)\|_2 \\ & + \alpha \sum_{i=1}^c \sum_{k=1}^l \|\mathbf{w}_i^k(t)\|_2 + \sum_{p=1}^s \sum_{j=1}^d \beta \|\mathbf{v}_j^p(t)\|_2 \end{aligned} \quad (23)$$

Equation (23) proves that the value of the objective function decreases in each iteration. Because the formulated objective function is convex, Algorithm 1 converges to the optimal solution.

As the optimization problem in (4) is convex, Algorithm 1 converges to the global optimal solution fast. In each iteration of our algorithm, computing steps 4–6 is trivial. We compute steps 7 and 8 by solving a system of linear equations with a quadratic complexity.

4. Dataset and implementation

4.1. The PEAC dataset

A practical contribution of this research is the collection of a new dataset of PEAC. Although long-term autonomy has been recently attracting an increasing attention in robotics, before this work, no dataset is publicly available for benchmarking robot perceptual adaptation, which consists of multisensory perception data collected from physical robots in real-world field applications under short-term and long-term lighting changes. Motivated by this desire, we collected the PEAC dataset. We utilized the Clearpath Husky and Jackal mobile robots (shown in Figure 2) to follow and detect an individual human subject. The robots are equipped with the Asus Xtion PRO structured-light camera without any automatic gains to collect color-depth data, and the Adafruit TSL2561 digital luminosity sensor to collect luminosity data. The color and depth images have a resolution of 640×480 . The luminosity readings is normalized between 0 (no lighting intensity) to 1 (maximum lighting intensity). Both structure-light camera and luminosity sensor run at 30 frames per second.

During data collection, we assume the robot is performing a human-following task, which is a desired capability in robotics applications (e.g., to carry rescue gear for humans in a long-term search and rescue operation). The robots are manually controlled by a separate human operator to follow a single human teammate walking in front of the robot. The dataset is collected in three different scenarios shown in Figure 2.

- **Scenario I (entering–exiting a mine):** A mobile robot follows a human subject entering and exiting two different mine drifts (i.e., horizontal openings made in a mine). One drift is dark (Figure 2), and the other has light bulbs installed in the drift (Figure 3). When the robot travels from the inside to the outside of the mine drift (or vice versa), the environment exhibits significant lighting changes. This testing scenario represents possible situations when robot performs underground search and rescue, for example, in mine, cave, and subway environments.
- **Scenario II (traveling indoor–outdoor):** A robot follows an individual human to travel inside and outside of a

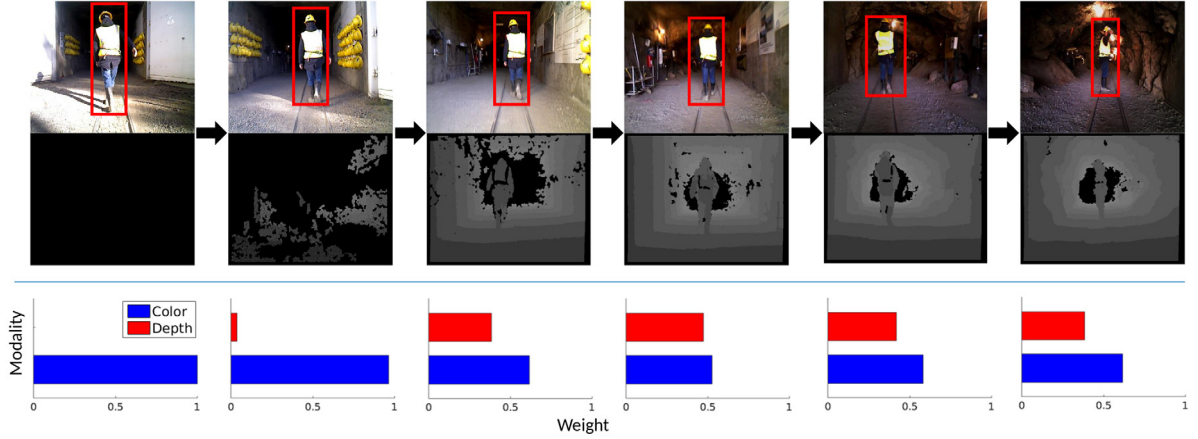


Fig. 3. Qualitative and quantitative results on Scenario I (entering-exiting a mine). The top row depicts an example of the qualitative results from the robot’s viewpoint when it follows a human subject to navigate into a dark mine drift from a bright open area. The bottom row illustrates the importance of color-depth sensor modalities learned and adapted by ROPA.

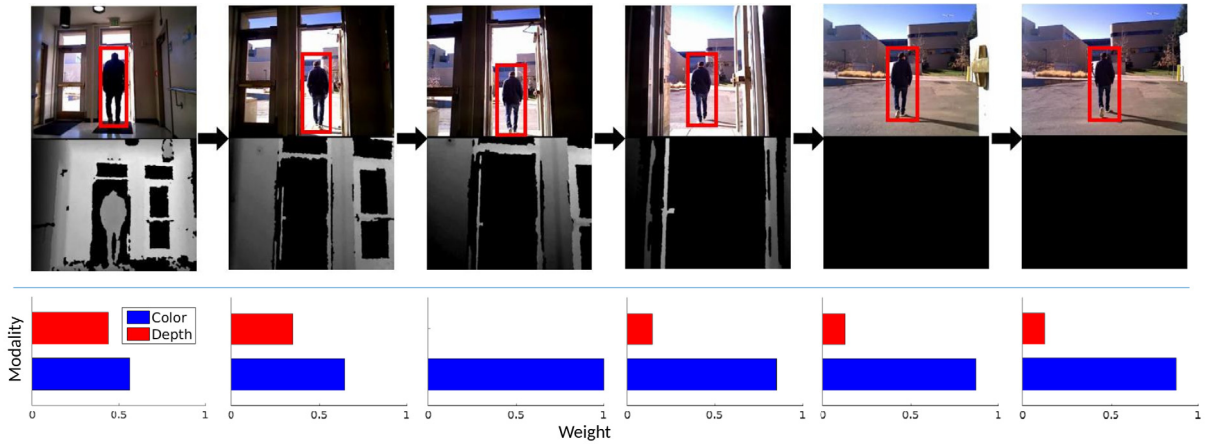


Fig. 4. Qualitative and quantitative results on Scenario II (traveling indoor-outdoor). The top row illustrates an example of the recognition results from the robot’s viewpoint when it follows the human to navigate from the inside to the outside of a building. The bottom row shows the importance of color-depth sensor modalities learned and adapted by our ROPA algorithm.

building. This scenario may happen in robot-assisted daily living and/or urban search and rescue applications. Similar to Scenario I, this scenario contains the challenge of fast illumination changes when the robot enters and leaves the building, as shown in Figures 4.

- Scenario III (following all day): A mobile robot follows a human subject in an outdoor environment in different times of the day from dawn to dusk, and across different months. The environment changes dramatically, especially lighting levels from noon to late evening (Figure 5). In the daytime under strong sunshine, structured-light depth sensors fail. In evenings with poor lighting conditions, color cameras do not work well. This Scenario III includes significant challenges caused by long-term environment changes across a day and seasons.

In each scenario, we collect 20 instances of human following, each consisting of 700–1,000 color-depth images. In addition, we collect the light level data (in lux) using digital luminosity sensors installed on the robots to document lighting variations in the environments. The ground truth of human detection is manually labeled by bounding human subjects in the color-depth scene with a box. The PEAC dataset is publicly available at <http://hcr.mines.edu/code/PEAC.html>.

4.2. Implementation

Three real-time feature extraction methods were applied to compute heterogeneous features from the proposed regions of the color and depth images readings in our experiments. (1) Histogram of oriented gradients (HOG) features (Dalal

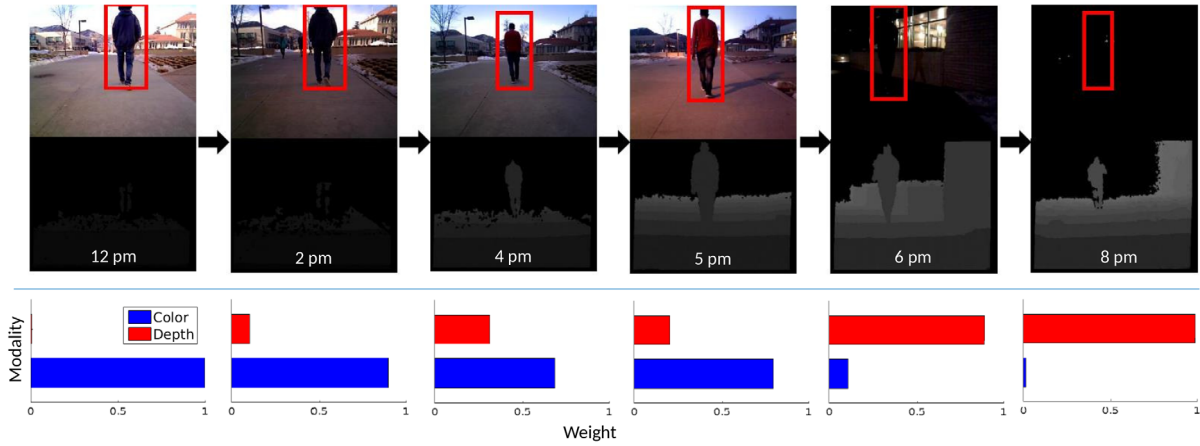


Fig. 5. Qualitative and quantitative results on Scenario III (following all day). The top row illustrates an example of the qualitative results from the robot’s viewpoint when it follows a subject to navigate in a campus environment from noon until late evening with dramatic long-term lighting changes. The bottom row shows the importance of color-depth sensor modalities learned and adapted by ROPA.

and Triggs, 2005) are used to represent shapes by counting occurrences of gradient orientation. (2) GIST features (Oliva and Torralba, 2006) are built from the response of steerable filters at different orientations and scales. (3) Local difference binary patterns (LBP) features (Arroyo et al., 2015) compute binary strings from simple intensity and gradient differences of image grid cells. These three features are selected because they are widely used in previous work on long-term place recognition. Each modality of features from color or depth data is normalized to ensure that features in the modalities have the same value range. To ensure real-time performance, we do not use deep features. However, in principle, any features that can produce a vector-based representation can be applied as an input modality to work with ROPA. The luminosity data is adopted to represent the context, i.e., the lighting variation. Object proposals are provided by a proposal generation approach applied on both color and depth data (Ren et al., 2015) to obtain regions of interest that potentially contain humans. The current implementation of ROPA is programmed using a mixture of unoptimized Matlab and C++ on an onboard Linux computer within the Jackal and Husky robots, which has an i5 2.5 GHz CPU and 8 GB memory, with no GPU.

Given the detection results of humans from ROPA that addresses short-term and long-term environmental changes, human following is performed by a decision-making module for robot navigation control. The module used in this work is implemented using an apprenticeship learning approach in the general framework of a Markov decision process (MDP), which is detailed in our previous work (Han et al., 2017b), which leverages human demonstrations to learn a navigation policy (e.g., move forward or backward, turn left or right, stop, speed up, slow down, etc.) based on human detection results.

5. Experiments

In this section, we present and analyze experimental results. To evaluate the performance of ROPA for multisensory robot perceptual adaptation, we performed extensive experiments on the PEAC dataset in the human-following task.

To estimate ROPA’s parameters, we employed around 4,000 color-depth frames from Scenario I (entering-exiting a mine) *only* for training. Features extracted from color-depth images using different techniques are concatenated into a final vector as the input to ROPA. The luminosity sensor reading is used to encode the context, i.e., environment lighting changes. Throughout our experiments, we set the values of the trade-off hyperparameters α and β to 0.1.

In the testing phase, we applied the trained model to the five instances from Scenario I (entering-exiting a mine), and the same model to instances from Scenario II (traveling indoor-outdoor) and Scenario III (following all day). That is, no instances from Scenarios II and III were used for training. Thus, we can evaluate how the ROPA model can be scaled to previously unseen situations.

To show the advantage of ROPA, we first compared ROPA with baseline techniques. We implemented baselines based upon single sensor and single feature type (i.e., HOG, LBP, or GIST from color or depth sensor). The decision model in (5) without the calibration term was applied along with these features to recognize humans. In addition, we compared ROPA with previous state-of-the-art methods, including (1) multimodal convolutional neural network (mCNN) (Eitel et al., 2015), (2) shared representative appearance learning (SRAL) (Han et al., 2017a), and (3) You Only Look Once (YOLO; Redmon et al., 2016). The used techniques except YOLO were trained on the same set of data from the PEAC dataset. We used the YOLO model

Table 1. Comparison of average accuracy over all robot operation scenarios in the PEAC dataset.

Methods	Scenario I	Scenario II	Scenario III
HOG	89.22%	79.00%	27.74%
LBP	81.40%	71.95%	44.83%
GIST	79.53%	71.12%	55.10%
HOG-D	70.11%	65.16%	46.84%
LBP-D	61.07%	70.10%	47.45%
GIST-D	67.57%	48.98%	51.48%
YOLO	88.13%	77.65%	47.08%
SRAL	92.30%	85.07%	51.13%
mCNN	90.65%	87.24%	49.45%
ROPA	96.91%	89.72%	79.17%

pretrained by the YOLO’s authors under various illumination conditions (Redmon and Farhadi, 2018).

5.1. Scenario I (entering–exiting a mine)

The qualitative results obtained by ROPA are illustrated in Figure 3, which shows that ROPA allows the robot to accurately recognize the human when navigating from a bright outdoor open area into a dark mine drift under dramatic illumination changes.

In addition, we compute numerical results to quantitatively evaluate ROPA. To evaluate how well ROPA can identify humans under lighting changes, accuracy is employed as an evaluation metric. Our ROPA obtains an average accuracy of 96.91% in Scenario I. We also compared ROPA with baseline and previous methods in Scenario I. The comparison results are presented in Table 1. It is observed that techniques based on color features (i.e., HOG, LBP, GIST, and YOLO) generally perform better than methods using depth features (i.e., HOG-D, LBP-D, and GIST-D). In addition, through integrating multisensory multimodal features, sensor fusion approaches (i.e., SRAL and mCNN) can outperform techniques using single types of features. ROPA significantly improves performance and obtains the best accuracy, owing to its capability to immediately calibrate multisensory perception and adapt to environment changes.

In addition to accuracy, we also evaluate how ROPA can adapt to environment changes by analyzing the importance of different sensor modalities, i.e., color and depth in this experiment. The quantitative results are graphically presented in Figure 3. It is observed that, when the robot stays outside of the mine under direct sunshine, it completely relies on color cues to recognize the subject because the structured-light depth sensor on the robots fail under direct sunshine, and cannot provide depth information. ROPA can automatically learn this fact from data without the requirement of hard coding. When the robot follows the human subject into the mine drift with reduced lighting, we observe that ROPA starts using the depth information to

combine with color cues to recognize humans, which demonstrates ROPA’s on-the-fly multisensory perception calibration capability to adapt to environment changes.

5.2. Scenario II (traveling indoor–outdoor)

We further evaluate ROPA’s performance in Scenario II, in which a robot identifies and follows a human subject to travel between the inside and outside of a building. In this scenario, the robot needs to address the challenge of adapting to lighting differences, when it travels from the inside of a building to the outside (or vice versa). In the experiment, the perception model learned in Scenario I is directly applied to Scenario II without re-training or additional training, in order to show the ROPA’s capability of adapting from one scenario to a new, previously unexperienced scenario.

The qualitative results obtained by ROPA over Scenario II are shown in Figure 4, which demonstrates that ROPA can recognize humans while following the subjects traveling from the inside of a building with low lighting levels into a bright outdoor environment. Quantitatively, ROPA obtains an average accuracy of 89.72% in identifying human subjects. In addition, we compare our ROPA approach with the baseline techniques, and present the results in Table 1 of the main article. Consistent with the results observed in Scenario I, methods based on color cues perform better than methods that employ depth features only. In indoor environments with a reduced lighting level (but not completely dark), color cues still contribute to human recognition. When multisensory multimodal features are fused together, SRAL and mCNN obtain an improved accuracy over the baseline techniques based on a single type of features only. Owing to the capability of calibrating multisensory perception, ROPA adapts to environment changes and outperforms the baseline and previous approaches. In addition, Figure 4 graphically illustrates the importance of the color and depth sensor modalities in the experiment. It can be observed that as the robot starts to navigate toward the outside of the building, color cues start to become more critical to recognize the human subject. The changes of the sensor weights match the variations of the surrounding environment, which demonstrates the multisensory perceptual adaptation capability of the robot enabled by ROPA.

5.3 Scenario III (following all day)

To evaluate our approach’s multisensory perception adaptation capability to *long-term* environment variations, we evaluate ROPA in Scenario III, in which a mobile robot follows a human to navigate in outdoor campus environments at different times of the day (e.g., from morning, noon, afternoon, until evening). In this experiment, the environment shows significant variations in different hours of the

day during long-term robot operations. The perception model learned in Scenario I is directly used in this new long-term scenario without re-training, in order to demonstrate the ROPA's ability to calibrate perception automatically.

The qualitative results obtained by ROPA are illustrated in Figure 5, which demonstrates ROPA's capability to recognize humans under long-term lighting changes during the day. Quantitatively, ROPA obtains an average accuracy of 79.17%. With no re-training or incremental training, ROPA can still recognize humans well. However, in general, the accuracy is lower than the scenarios with short-term environment changes. The comparison of our ROPA approach with baseline and previous state-of-the-art methods is listed in Table 1. An interesting observation is that the SRAL and mCNN approaches based on sensor fusion perform worse than some baseline techniques using a single type of features. This phenomenon results from the fact that for most of the time in this scenario, only color or only depth information is available (Figure 5). For example, under the Sun, the depth sensor often fails, while during or after sunset, the color sensor does not operate well. In this scenario, sensor fusion paradigms may not be able to work well.

Adaptation of the importance weights of color and depth modalities is illustrated in Figure 5. The results validate that ROPA continuously calibrates multisensory robot perception to the surrounding environment with long-term changes. When the sunshine is strong (e.g., at the noon time when the depth sensor does not work), ROPA adapts to utilize color cues only for perception. On the other hand, when the environment is very dark (e.g., at 20:00), the color camera does not work, and ROPA can automatically calibrate perception to depend on depth cues for recognition and following. Another interesting observation from the experimental results is that, the weight of the color sensor at 17:00 is greater than the weight at 16:00, although the environment has a decreased *natural* lighting. This occurs because the street lights were turned on right before 17:00, which provides additional artificial lighting to the scene, making the overall lighting on the pedestrian path better than that at 16:00.

6. Discussion

6.1. High-speed processing

Owing to ROPA's ability to integrate feature fusion, perception calibration, and classification in the unified formulation, and the efficiency of our convex objective function, our ROPA approach can achieve high-speed onboard processing. To validate this advantage, we perform additional experiments in Scenario I, using our CPU implementation on a Jackal robot's onboard computer. Without counting the time on extracting HOG, LBP, and GIST features, we obtain a processing rate at around 100 Hz. When counting the time for feature extraction, we obtain a processing rate

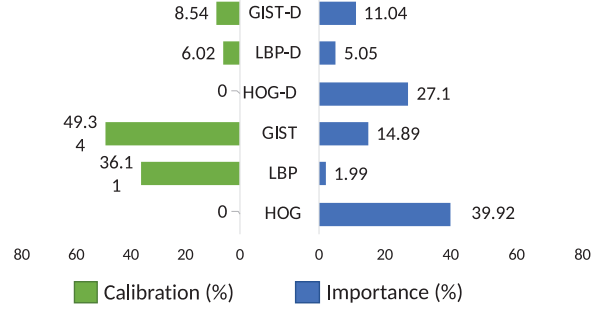


Fig. 6. Experimental results on feature modality analysis during the training phase.

of around 15 Hz. These results indicate the promise of ROPA to be applied in real-time robotics applications. In addition, any feature descriptor can be integrated to the ROPA approach, which provides the flexibility to further improve the overall processing speed when faster, higher-quality features are available.

6.2. Feature modality analysis

The proposed ROPA approach is capable to automatically estimate the importance of each feature modality, and how much each modality should be calibrated (i.e., calibration). We perform such experiments in the training phase using the data instances from Scenario I. The results are illustrated in Figure 6. The importance histogram indicates that, among the feature types used in the experiment, HOG is the most important in sensor fusion, followed by LBP, and GIST is the worst. The calibration graph illustrates that the GIST feature modality from the color data requires the most calibration, whereas the HOG modalities from both color and depth cues do not require calibration, in general. This result indicates that HOG features are relatively insensitive to the lighting changes.

6.3. Hyperparameter selection

The hyperparameters α and β in our formulation (4) are designed to control the strength of regularization terms over feature learning and perception calibration, respectively. Their optimal values can be determined using cross-validation during training. From the result in Figure 7, we observe that when $\alpha = \beta = 0.1$, ROPA statistically obtains the best accuracy. In general, the range of $\alpha, \beta \in (0.01, 1)$ can result in satisfactory accuracy, which also shows that both regularization terms are useful.

7. Conclusion

In this article, we have introduced the novel, bio-inspired approach named ROPA to enable the new robot capability of calibrating multisensory perception, in order for robots to adapt to short-term and long-term environment changes.

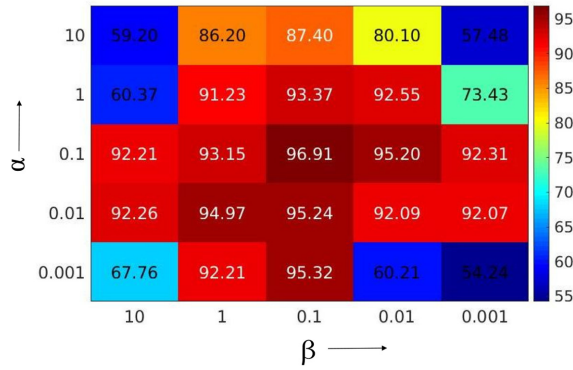


Fig. 7. Experimental results on hyperparameter analysis during the training phase.

Our focused application in this article aims at the task of long-term human following in the field environment, which is essential to real-world human–robot teaming applications.

The ROPA approach has been formulated as a joint learning problem to simultaneously estimate the representativeness of each feature modality, integrate heterogeneous features, and more importantly, calibrate the perception model to adapt multisensory perception with environmental changes. In order to fuse multisensory input data, we have implemented sparsity-inducing norms that enforce the base perception model to learn sparse weights of the multisensory input features and use the learned weights for multisensory fusion. To achieve perception calibration, we have estimated the representativeness of the input features to encode the environment, and provided a calibration of the base model according to the environmental change. All components were integrated under the unified theoretical framework of regularized optimization.

In addition, we have collected the new large-scale PEAC dataset containing multisensory data instances in scenarios of robot following of humans in a wide range of field environments with short-term and long-term environment changes. This open dataset provides a benchmark to evaluate and compare the approaches designed for robot perceptual adaptation to short-term and long-term variations of the robot operation environment in long-term human-following applications. We have conducted extensive experiments to evaluate and analyze ROPA in various scenarios using the PEAC dataset. Experimental results have validated that, through calibrating perception, ROPA is able to effectively adapt to environment changes, obtains promising accuracy and efficiency, and outperforms baseline and previous methods in dynamic and unstructured environments in long-term human-following applications.

The proposed mathematical formulation under regularized optimization for ROPA is general and has a potential to provide a framework that addresses robot adaptation to other environment and context changes (e.g., terrain changes and season changes). One of the challenges that

prevent us from extending this approach to other adaptation scenarios is the lack of data on long-term robot adaptation, which will be our future work. Another future research can focus on integrating prior knowledge of cause and effect of environment changes in the adaptation process.

Funding

This research was partially supported by the Army Research Office (ARO; grant number W911NF-17-1-0447), US Air Force Academy (USAFA; grant number FA7000-18-2-0016), Distributed and Collaborative Intelligent Systems and Technology (DCIST) CRA (grant number W911NF-17-2-0181) and National Science Foundation (NSF; grant number IIS-1849348).

Notes

1. A preliminary non-archived version of the article describing the dataset and initial experimental results was presented as a spotlight talk at the *ICRA Workshop on Robot Teammates Operating in Dynamic, Unstructured Environments* (RT-DUNE) (Siva and Zhang (2018b)).

ORCID iD

Sriram Siva  <https://orcid.org/0000-0003-3457-2085>

References

- Agravante DJ, Cherubini A, Bussy A, Gergondet P and Kheddar A (2014) Collaborative human-humanoid carrying using vision and haptic sensing. In: *IEEE International Conference on Robotics and Automation*.
- Argall BD, Chernova S, Veloso M and Browning B (2009) A survey of robot learning from demonstration. *Robotics and Autonomous Systems* 57(5): 469–483.
- Arroyo R, Alcantarilla PF, Bergasa LM and Romera E (2015) Towards life-long visual localization using an efficient matching of binary sequences from images. In: *IEEE International Conference on Robotics and Automation*.
- Arroyo R, Alcantarilla PF, Bergasa LM and Romera E (2016) Fusion and binarization of CNN features for robust topological localization across seasons. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*.
- Bay H, Tuytelaars T and Van Gool L (2006) Surf: Speeded up robust features. In: *European Conference on Computer Vision*. Berlin: Springer, pp. 404–417.
- Calonder M, Lepetit V, Strecha C and Fua P (2010) Brief: Binary robust independent elementary features. In: *European Conference on Computer Vision*. Berlin: Springer, pp. 778–792.
- Dalal N and Triggs B (2005) Histograms of oriented gradients for human detection. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.
- Dalal N, Triggs B and Schmid C (2006) Human detection using oriented histograms of flow and appearance. In: *European Conference on Computer Vision*. Berlin: Springer, pp. 428–441.
- Dettmann A, Langosz M, von Szadkowski K and Bartsch S (2014) Towards lifelong learning of optimal control for kinematically complex robots. In: *Workshop at IEEE International Conference on Robotics and Automation*.

- Eitel A, Springenberg JT, Spinello L, Riedmiller M and Burgard W (2015) Multimodal deep learning for robust RGB-d object recognition. In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 681–687.
- Evrard P, Gribovskaya E, Calinon S, Billard A and Kheddar A (2009) Teaching physical collaborative tasks: Object-lifting case study with a humanoid. In: *IEEE-RAS International Conference on Humanoid Robots*.
- Fernaues Y, Håkansson M, Jacobsson M and Ljungblad S (2010) How do you play with a robotic toy animal?: A long-term study of PLEO. In: *Proceedings of International Conference on Interaction Design and Children*.
- Floyd MW, Drinkwater M and Aha DW (2015) Trust-guided behavior adaptation using case-based reasoning. Technical report, Naval Research Laboratory, Washington, USA.
- Gribovskaya E, Kheddar A and Billard A (2011) Motion learning and adaptive impedance for robot control during physical interaction with humans. In: *IEEE International Conference on Robotics and Automation*.
- Hagras H, Callaghan V and Colley M (2004) Learning and adaptation of an intelligent mobile robot navigator operating in unstructured environment based on a novel online fuzzy-genetic system. *Fuzzy Sets and Systems* 141(1): 107–160.
- Han F, Yang X, Deng Y, Rentschler M, Yang D and Zhang H (2016) Life-long place recognition by shared representative appearance learning. In: *RSS 2016 Workshop on Visual Place Recognition: What is it Good For?*
- Han F, Yang X, Deng Y, Rentschler M, Yang D and Zhang H (2017a) SRAL: Shared representative appearance learning for long-term visual place recognition. *IEEE Robotics and Automation Letters* 2(2): 1172–1179.
- Han F, Yang X, Zhang Y and Zhang H (2017b) Sequence-based multimodal apprenticeship learning for robot perception and decision making. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 2584–2591.
- Jevtić A, Colomé A, Alenya G and Torras C (2018) Robot motion adaptation through user intervention and reinforcement learning. *Pattern Recognition Letters* 105(1): 67–75.
- Kanda T, Shiomi M, Miyashita Z, Ishiguro H and Hagita N (2010) A communication robot in a shopping mall. *IEEE Transactions on Robotics* 26(5): 897–913.
- Keller CG, Enzweiler M, Rohrbach M, Llorca DF, Schnorr C and Gavrila DM (2011) The benefits of dense stereo for pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems* 12(4): 1096–1106.
- Kleiner A, Dietl M and Nebel B (2002) Towards a life-long learning soccer agent. In: *Robot Soccer World Cup*.
- Kosuge K and Kazamura N (1997) Control of a robot handling an object in cooperation with a human. In: *IEEE International Workshop on Robot and Human Communication*.
- Kruijff-Korbayová I, Colas F, Gianni M, et al. (2015) TRADR project: Long-term human-robot teaming for robot assisted disaster response. *KI-Künstliche Intelligenz* 29(2): 193–201.
- Leite I, Martinho C and Paiva A (2013) Social robots for long-term interaction: A survey. *International Journal of Social Robotics* 5(2): 291–308.
- Li Y, Tee KP, Chan WL, Yan R, Chua Y and Limbu DK (2015) Continuous role adaptation for human-robot shared control. *IEEE Transactions on Robotics* 31(3): 672–681.
- Liu C, Conn K, Sarkar N and Stone W (2008) Online affect detection and robot behavior adaptation for intervention of children with autism. *IEEE Transactions on Robotics* 24(4): 883–896.
- Lowry S, Sünderhauf N, Newman P, et al. (2016) Visual place recognition: A survey. *IEEE Transactions on Robotics* 32(1): 1–19.
- Lowry SM, Milford MJ and Wyeth GF (2014) Transforming morning to afternoon using linear regression techniques. In: *IEEE International Conference on Robotics and Automation*.
- Medina JR, Shelley M, Lee D, Takano W and Hirche S (2012) Towards interactive physical robotic assistance: Parameterizing motion primitives through natural language. In: *IEEE International Conference on Robot and Human Interactive Communication*.
- Mikolajczyk K and Schmid C (2001) Indexing based on scale invariant interest points. In: *Proceedings Eighth IEEE International Conference on Computer Vision, 2001 (ICCV 2001), Vol. 1*. IEEE, pp. 525–531.
- Mitsunaga N, Smith C, Kanda T, Ishiguro H and Hagita N (2006) Robot behavior adaptation for human-robot interaction based on policy gradient reinforcement learning. *Journal of the Robotics Society of Japan* 24(7): 820–829.
- Nam H and Han B (2016) Learning multi-domain convolutional neural networks for visual tracking. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4293–4302.
- Neubert P, Sünderhauf N and Protzel P (2013) Appearance change prediction for long-term navigation across seasons. In: *European Conference on Mobile Robots*.
- Nikolaidis S, Nath S, Procaccia AD and Srinivasa S (2017a) Game-theoretic modeling of human adaptation in human-robot collaboration. In: *Proceedings of ACM/IEEE International Conference on Human-Robot Interaction*.
- Nikolaidis S, Zhu YX, Hsu D and Srinivasa S (2017b) Human-robot mutual adaptation in shared autonomy. In: *ACM/IEEE International Conference on Human-Robot Interaction*.
- Oliva A and Torralba A (2006) Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research* 155: 23–36.
- Parker LE (2000) Lifelong adaptation in heterogeneous multi-robot teams: Response to continual variation in individual robot performance. *Autonomous Robots* 8(3): 239–267.
- Redmon J, Divvala S, Girshick R and Farhadi A (2016) You only look once: Unified, real-time object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788.
- Redmon J and Farhadi A (2018) YoloV3: An incremental improvement. arXiv preprint arXiv:1804.02767.
- Ren S, He K, Girshick R and Sun J (2015) Faster r-CNN: Towards real-time object detection with region proposal networks. In: *Advances in Neural Information Processing Systems*, pp. 91–99.
- Rhodes G, Watson TL, Jeffery L and Clifford CW (2010) Perceptual adaptation helps us identify faces. *Vision Research* 50(10): 963–968.
- Ritschel H and André E (2017) Real-time robot personality adaptation based on reinforcement learning and social signals. In: *ACM/IEEE International Conference on Human-Robot Interaction*.
- Rosen DM, Mason J and Leonard JJ (2016) Towards lifelong feature-based mapping in semi-static environments. In: *IEEE International Conference on Robotics and Automation*.
- Rublee E, Rabaud V, Konolige K and Bradski G (2011) ORB: An efficient alternative to SIFT or SURF. In: *2011 IEEE*

- International Conference on Computer Vision (ICCV)*. IEEE, pp. 2564–2571.
- Shibata T and Tanie K (2000) Influence of a priori knowledge in subjective interpretation and evaluation by short-term interaction with mental commit robot. In: *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*.
- Siva S and Zhang H (2018a) Omnidirectional multisensory perception fusion for long-term place recognition. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 1–9.
- Siva S and Zhang H (2018b) Robot adaptation to environment changes in long-term autonomy. In: *ICRA 2018 Workshop on Long-term Autonomy and Deployment of Intelligent Robots in the Real-world*.
- Sminchisescu C and Triggs B (2003) Kinematic jump processes for monocular 3D human tracking. In: *Proceedings 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1*. IEEE.
- Sünderhauf N, Neubert P and Protzel P (2014) Predicting the change—a step towards life-long operation in everyday environments. *IEEE Workshop on Robotics Challenges and Vision*.
- Sunderhauf N, Shirazi S, Jacobson A, et al. (2015) Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free. In: *Robotics: Science and Systems*.
- Tapus A, Tapus C and Mataric M (2010) Long term learning and online robot behavior adaptation for individuals with physical and cognitive impairments. In: *International Conference on Field and Service Robotics*.
- Thrun S, Bennewitz M, Burgard W, et al. (1999) Minerva: A second-generation museum tour-guide robot. In: *IEEE International Conference on Robotics and Automation*.
- Thrun S and Mitchell TM (1995) Lifelong robot learning. In: *The Biology and Technology of Intelligent Autonomous Agents*. Berlin: Springer, pp. 165–196.
- Tung C, Kelleher MR, Schlueter RJ, et al. (2019) Large-scale object detection of images from network cameras in variable ambient lighting conditions. In: *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, pp. 393–398.
- Vinas M, Sawides L, De Gracia P and Marcos S (2012) Perceptual adaptation to the correction of natural astigmatism. *PLoS One* 7(9): e46361.
- Wang X, Han TX and Yan S (2009) An HOG-LBP human detector with partial occlusion handling. In: *2009 IEEE 12th International Conference on Computer Vision*. IEEE, pp. 32–39.
- Watson I and Marir F (1994) Case-based reasoning: A review. *The knowledge Engineering Review* 9(4): 327–354.
- Webster MA (2011) Adaptation and visual coding. *Journal of Vision* 11(5): 3–3.
- White A, Modayil J and Sutton RS (2012) Scaling life-long off-policy learning. In: *IEEE International Conference on Development and Learning and Epigenetic Robotics*.
- Wu J and Rehg JM (2011) Centrist: A visual descriptor for scene categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(8): 1489–1501.
- Xia L, Chen CC and Aggarwal JK (2011) Human detection using depth information by Kinect. In: *2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, pp. 15–22.
- Yuan L, Chan KC and Lee CG (2011) Robust semantic place recognition with vocabulary tree and landmark detection. In: *IROS Workshop on Active Semantic Perception and Object Search in the Real World*.
- Zhang H, Han F and Wang H (2016) Robust multimodal sequence-based loop closure detection via structured sparsity. In: *Robotics: Science and Systems*.
- Zhang Q, Qian H and Zhu M (2005) Parameter adaptation by case-based mission-planning of outdoor autonomous mobile robot. In: *Proceedings of Intelligent Vehicles Symposium*.