Gestalt Perceptual Calibration for Multi-agent Collaborative Localization

Yan Zhang and Rong Xie(✉)

School of Computer Science, Wuhan University,
Wuhan 430072, People's Republic of China
{zriszhang,xierong}@whu.edu.cn

Abstract. Multi-agent collaborative localization is a challenging problem that requires estimating the global poses of multiple agents in a shared coordinate system based on their sensor data and inter-agent communication. The existing methods, such as map fusion, graph optimization, or consensus algorithms, cannot fully exploit the perceptual information and similarity among different agents, which can improve localization accuracy and consistency. In this paper, we propose a method of gestalt perceptual calibration (GPC) for multi-agent localization, which leverages the power of gestalt psychology to align the coordinate frames of different agents based on their perceptual similarity. We propose a perceptual encoder based on Perceiver model to encode the sensor data of each agent into a latent array. To optimize the relative poses of each pair of agents based on their gestalt similarity, we use a Gumbel-softmax-based relaxation as the gestalt calibrator. Aiming at estimating the global poses of each agent based on the relative poses and other measurements, we also use a maximum a posteriori inference as the pose estimator. The experimental results show that our GPC method can achieve superior localization performance in both accuracy and efficiency compared to existing localization methods. In addition, we further analyze the influence of different parameters of our GPC method on the localization performance.

Keywords: Multi-agent collaborative localization · Gestalt perceptual calibration · Perceiver

1 Introduction

Multi-agent collaborative localization is a research problem that aims to estimate the poses (positions and orientations) of multiple agents in an unknown environment using their sensor measurements and inter-agent communications. This problem is of great interest and importance for many applications, such as cooperative exploration, surveillance, search and rescue, and disaster relief. Solving this problem can enable multiple agents to coordinate their actions and achieve a common goal.

However, multi-agent collaborative localization is challenging for several reasons. First, the agents have limited sensing and communication capabilities,

A. Zaslavsky et al. (Eds.): MobiQuitous 2023, LNICST 594, pp. 145–164, 2024.
https://doi.org/10.1007/978-3-031-63992-0_9

which introduce uncertainties and errors in their measurements. Second, the agents have different coordinate frames that are not aligned with each other or with the global frame. Third, the agents may have different initial poses that are unknown or inaccurate. These factors make the multi-agent localization problem ill-posed and under-constrained.

To address these challenges, some existing methods use various techniques, such as map fusion [1], graph optimization [2], or consensus algorithms [3]. However, these methods have some limitations, such as requiring a central server, a large number of anchors, a high synchronization accuracy, a high computational power, or a large amount of training data. Moreover, these methods often rely on domain-specific assumptions or preprocessing for different types of sensor scans, such as LiDAR scans or radio signals.

Gestalt principles are a set of rules that describe how humans organize and interpret visual information [4]. Inspired by Gestalt principles, we propose a gestalt perceptual calibration method for multi-agent collaborative localization. We use Gestalt principles as a perceptual similarity measure to align the coordinate frames of different agents. By using a Perceiver model to encode the sensor scans of each agent into latent representations, we can apply Gestalt principles to compare and match these representations, and compute the relative poses between agents. This approach can handle various types of sensor scans without any domain-specific assumptions or preprocessing. Moreover, this approach can exploit the holistic and organized nature of perception to improve the localization accuracy and efficiency.

The rest of this paper is organized as follows. Section 2 reviews the related work on multi-agent collaborative localization and Gestalt principles of perception. Section 3 presents an overview of our GPC method. Section 4 describes each component of our method in detail. Section 5 gives the experimental results and analysis. Section 6 concludes this paper and discusses future work.

2 Related Work

2.1 Multi-agent Collaborative Localization

For multi-agent collaborative localization, several methods have been proposed for this problem. The type of sensor is the determining factor for the localization and calibration techniques used in these methods. Therefore, we can categorize them according to the type of sensor.

Some researchers used global navigation satellite systems (GNSS) and ormicroseismic/acoustic emission sources as sensors. Tanwar and Gao [5] proposed a decentralized collaborative localization method in urban environments using 3D-mapping-aided GNSS and inter-agent ranging. Dong et al. [6] introduced a collaborative localization method using analytical and iterative solutions for ormicroseismic/acoustic emission sources in the rockmass structure for underground mining. These methods can exploit the different features of the environment to improve the localization accuracy, but they require high-precision synchronization, calibration, or computation.

Cameras or ultra-wideband (UWB) signals are also commonly used as sensors, which can provide visual or geometric information. Schmuck and Chli [8] presented a centralized collaborative monocular simultaneous localization and mapping framework for robotic teams. Shule et al. [9] proposed a UWB-based localization method for multi-UAV systems and collaborative heterogeneous multi-robot systems. These methods offer precise localization and mapping information, enabling robotic teams to collaborate and navigate more effectively. However, these methods require a central server or a large number of anchors. Therefore, they may not function properly without a central server or a large number of anchors, limiting their applicability in certain environments.

Recently, soft information, continuum deformation coordination (CDC), or rogue drone interception (RDI), which utilized semantic or contextual information. Conti et al. [7] proposed a localization-of-things method using soft information extracted from inter-agent and intra-agent measurements as well as contextual data. Emadi et al. [10] proposed a collision-free continuum deformation coordination method for a multi-quadcopter system using cooperative localization. Queralta et al. [11] designed a multi-agent system for rogue drone interception using radar, optical sensors, and radio signals. These methods well handle challenges such as non-Gaussian distributions, nonlinear models, dynamic environments, but require a large amount of training data and feedback gain.

In summary, while these methods have made significant strides in multi-agent collaborative localization, there is still a need for a calibration method that is less reliant on high-precision synchronization, computation and extensive training data. This method should be accurate, robust, and efficient to enhance the performance and applicability of multi-agent collaborative localization systems in various environments.

2.2 Gestalt Principles of Perception

Gestalt principles of perception describe how humans organize and interpret visual information. They are based on the idea that the whole is different from the sum of its parts, and that the brain creates a perception that is more than simply the sum of available sensory inputs [13]. Gestalt principles include figure-ground relationship, proximity, similarity, continuity, closure, and symmetry. These principles explain why we can see patterns, shapes, and objects from simple elements, such as dots, lines, and colors.

The versatility and efficacy of the Gestalt principle in addressing diverse challenges. For instance, Hu et al. [14] leverages this principle to improve object detection in diagrams. Similarly, Chen et al. [15] applies the Gestalt principle to enhance the resolution of text images. Furthermore, Susan et al. [16] utilizes this principle for effective pulmonary nodule detection.

Gestalt principles can provide a perceptual similarity measure that can align the coordinate frames of different agents [17]. By using a Perceiver model to encode the sensor scans of each agent into latent representations, we can apply gestalt principles to compare and match these representations, and compute the relative poses between agents. This approach can handle various types of

sensor scans, such as LiDAR scans or radio signals, without any domain-specific assumptions or preprocessing. Moreover, this approach can exploit the holistic and organized nature of perception to improve the localization accuracy and efficiency.

3 Method Overview

In this section, we present an overview of our proposed method for multi-agent collaborative localization based on Gestalt perceptual calibration. We first formulate the problem of multi-agent localization and calibration, then describe the main components of our framework.

3.1 Problem Formulation

We consider a multi-agent system consisting of N agents, each equipped with a sensor like LiDAR or a radio transceiver. The agents are deployed in an unknown environment with a known map M. The goal of the agents is to collaboratively estimate their poses (positions and orientations) in the global coordinate frame G using their sensor measurements and inter-agent communications.

We assume that each agent i has a motion model f_i that describes its state transition from time step $(t-1)$ to t, satisfying

$$\hat{x}_i^t = f_i(x_i^{t-1}, u_i^t) + w_i^t \tag{1}$$

where $x_i^t \in \mathbb{R}^3$ is the pose of agent i at time t, $u_i^t \in \mathbb{R}^2$ is the control input (linear and angular velocities) of agent i at time t, and $w_i^t \sim \mathcal{N}(0, Q_i)$ is the motion noise of agent i with covariance matrix Q_i.

We also assume that each agent i has a measurement model h_i that describes its observation at time t, satisfying

$$z_i^t = h_i(x_i^t, M) + v_i^t \tag{2}$$

where $z_i^t \in \mathbb{R}^K$ is the sensor scan of agent i at time t, which consists of K range measurements along different angles, and $v_i^t \sim \mathcal{N}(0, R_i)$ is the measurement noise of agent i with covariance matrix R_i.

In addition to the measurements for agents themselves, each agent i can also receive relative measurements from other agents within its communication range. We denote the set of neighboring agents of agent i at time t as N_i^t. For each neighboring agent $j \in N_i^t$, agent i can measure the distance d_{ij}^t between them by Formula (3).

$$d_{ij}^t = \|x_{ij}^t\|_2 + n_{ij}^t \tag{3}$$

where $x_{ij}^t = x_j^t - x_i^t$ is the relative pose between agent j and agent i, and $n_{ij}^t \sim \mathcal{N}(0, S_{ij})$ is the communication noise with variance S_{ij}.

Definition 1. The problem of multi-agent collaborative localization is formulated as finding the maximum a posteriori (MAP) estimate of the poses of all agents at time t, given their initial poses, control inputs, private measurements, and relative measurements up to time t. It is represented by Formula (4).

$$\hat{x}_1^t, ..., \hat{x}_N^t = \arg \max_{x_1^t, ..., x_N^t} p(x_1^t, ..., x_N^t | x_1^0, ..., x_N^0, u_1^{1:t}, ..., u_N^{1:t}, z_1^{1:t}, ..., z_N^{1:t}, d_{ij}^{1:t}, \forall i, j) \tag{4}$$

3.2 Overview of the Framework

We propose Gestalt Perceptual Calibration (GPC) for multi-agent collaborative localization that leverages the gestalt principles of perceptual organization. The main idea is to align the coordinate frames of different agents by finding the best perceptual match between their sensor scans, using a Perceiver model.

Our GPC method consists of three main components: a perceptual encoder, a gestalt calibrator, and a pose estimator. Figure 1 shows an overview of the framework of our GPC method.

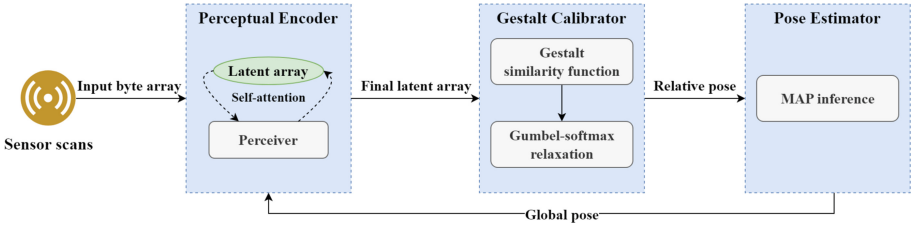


Fig. 1. Overall framework of GPC method

The perceptual encoder takes the sensor scans of each agent as input and encodes them into latent representations using a Perceiver model. The encoder shares the same weights across all agents and time steps, enabling cross-modal and cross-temporal perception.

The gestalt calibrator takes the latent representations of each agent as input, and computes the relative poses between them using Gumbel-softmax-based optimization [18]. The calibrator uses a perceptual similarity measure based on gestalt principles to align the coordinate frames of different agents. The calibrator also uses a consensus-based algorithm to iteratively refine the relative poses until convergence.

The pose estimator takes the relative poses between agents as input and estimates the global poses of each agent using maximum a posteriori inference. The estimator incorporates the motion models, the measurement models, and the communication models of each agent to obtain the most likely global poses.

We describe each component in detail in the next section.

4 Gestalt Perceptual Calibration

4.1 Perceptual Encoder

The perceptual encoder is responsible for encoding the sensor scans of each agent into latent representations that can be used for perceptual calibration and pose estimation. The encoder is based on the Perceiver model, which is a transformer-based model that can handle various types of input data using cross-attention and self-attention mechanisms.

In our GPC method, we use the following steps to apply the Perceiver model as the perceptual encoder for each agent.

Step 1: The sensor scan of each agent i at time t is treated as an input byte array $B_i^t \in \mathbb{R}^{L\times D}$, where L is the length of the input and D is the dimensionality of each byte. Initialize a random latent array $X_i^0 \in \mathbb{R}^{M\times E}$ for each agent, where M is the number of latent queries and E is the dimensionality of each query.
Step 2: Feed the input byte array B_i^t and the latent array X_i^0 to the cross-attention module of the Perceiver model, and obtain an output latent array $Y_i^0 \in \mathbb{R}^{M\times E}$ using scaled dot-product attention, as shown in Formula (5).

$$Y_i^0 = \text{softmax}\left(\frac{X_i^0 B_i^{tT}}{\sqrt{D}}\right) B_i^t \tag{5}$$

Step 3: Feed the output latent array Y_i^0 to the self-attention module of the Perceiver model, and apply a transformer layer to it, resulting in an updated latent array $Z_i^0 \in \mathbb{R}^{M\times E}$, as shown in Formula (6).

$$Z_i^0 = Y_i^0 + \text{FFN}\left(\text{LayerNorm}\left(Y_i^0 + \text{softmax}\left(\frac{Y_i^0 Y_i^{0T}}{\sqrt{E}}\right) Y_i^0\right)\right) \tag{6}$$

where FFN is a feed-forward network and LayerNorm is a layer normalization operation.
Step 4: Repeat Steps 2 and 3 for a fixed number of iterations T, producing a final latent array $Z_i^* \in \mathbb{R}^{M\times E}$ that captures the information from the input byte array B_i^t, as shown in Formula (7).

$$Z_i^* = Z_i^{T-1} + \text{FFN}\left(\text{LayerNorm}\left(Z_i^{T-1} + \text{softmax}\left(\frac{Z_i^{T-1} Z_i^{(T-1)T}}{\sqrt{E}}\right) Z_i^{T-1}\right)\right) \tag{7}$$

Step 5: Share the same weights across all agents and time steps, enabling cross-modal and cross-temporal perception. The perceptual encoder can handle different types of sensor scans, such as LiDAR scans or radio signals, without any domain-specific assumptions or preprocessing.

The output of the perceptual encoder is a set of latent arrays $\{Z_i^*\}_{i=1}^N$, where $Z_i^* \in \mathbb{R}^{M\times E}$ is the final latent array for agent i. These latent arrays are then used as input for the gestalt calibrator, which we describe in the next subsection.

4.2 Gestalt Calibrator

The gestalt calibrator is responsible for computing the relative poses between different agents, using their latent representations obtained from the perceptual encoder. The calibrator is based on Gestalt psychology, which is a school of psychology that emphasizes the holistic and organized nature of perception.

Gestalt psychology proposes several principles of perceptual organization, such as proximity, similarity, continuity, closure, and symmetry, that explain how humans perceive complex patterns from simple elements. These principles suggest that humans tend to group together elements that are close, similar, aligned, complete, or balanced, and form a coherent whole that is greater than the sum of its parts [19].

In our GPC method, we use these principles as a perceptual similarity measure to align the coordinate frames of different agents. We assume that agents that have similar sensor scans should have similar latent representations, and therefore have small relative poses between them. Conversely, agents that have dissimilar sensor scans should have dissimilar latent representations, and therefore have large relative poses between them.

To formalize this idea, we define the gestalt similarity function $g : \mathbb{R}^{M\times E} \times \mathbb{R}^{M\times E} \rightarrow [0, 1]$ as a weighted sum of several gestalt principles, as shown in Formula (8).

$$\begin{aligned} g(Z_i, Z_j) = &\alpha_1 g_p(Z_i, Z_j) + \alpha_2 g_s(Z_i, Z_j) + \alpha_3 g_c(Z_i, Z_j) \\ &+ \alpha_4 g_l(Z_i, Z_j) + \alpha_5 g_y(Z_i, Z_j) \end{aligned} \tag{8}$$

where $\alpha_1, \ldots, \alpha_5$ are learnable weights that balance the importance of each principle, and g_p, g_s, g_c, g_l, g_y are sub-functions that correspond to proximity, similarity, continuity, closure, and symmetry, respectively. We define each sub-function as follows.

Definition 2 (Proximity). The sub-function g_p computes the average Euclidean distance between the corresponding queries of two latent arrays, and returns a negative exponential function of the distance, as shown in Formula (9).

$$g_p(Z_i, Z_j) = \exp\left(-\frac{1}{M}\sum_{m=1}^{M} |Z_i^m - Z_j^m|_2\right) \tag{9}$$

where M is the number of latent queries, Z_i^m and Z_j^m are the m-th queries of Z_i and Z_j, respectively.

Definition 3 (Similarity) The sub-function g_s computes the average cosine distance between the corresponding queries of two latent arrays, and returns a positive linear function of the distance, as shown in Formula (10).

$$g_s(Z_i, Z_j) = 1 - \frac{1}{M}\sum_{m=1}^{M} \frac{Z_i^m \cdot Z_j^m}{|Z_i^m|_2 |Z_j^m|_2} \tag{10}$$

Definition 4 (Continuity). The sub-function g_c computes the average absolute difference between the adjacent queries of two latent arrays, and returns a negative exponential function of the difference, as shown in Formula (11).

$$g_c(Z_i, Z_j) = \exp\left(-\frac{1}{M-1}\sum_{m=1}^{M-1} ||Z_i^{(m+1)} - Z_i^m|_2 - |Z_j^{(m+1)} - Z_j^m|_2|\right) \quad (11)$$

Definition 5 (Closure). The sub-function g_l computes the Euclidean distance between the first and last queries of two latent arrays, and returns a negative exponential function of the distance, as shown in Formula (12).

$$g_l(Z_i, Z_j) = \exp\left(-|Z_i^1 - Z_i^M|_2 - |Z_j^1 - Z_j^M|_2\right) \quad (12)$$

Definition 6 (Symmetry). The sub-function g_y computes the degree of symmetry along the central axis of two latent arrays, and returns a positive linear function of the degree, as shown in Formula (13).

$$\begin{aligned} g_y(Z_i, Z_j) = &\frac{2}{M}\sum_{m=1}^{M/2} \frac{Z_i^m \cdot Z_i^{(M-m+1)}}{|Z_i^m|_2 |Z_i^{(M-m+1)}|_2} \\ &+ \frac{2}{M}\sum_{m=1}^{M/2} \frac{Z_j^m \cdot Z_j^{(M-m+1)}}{|Z_j^m|_2 |Z_j^{(M-m+1)}|_2} \end{aligned} \quad (13)$$

Given the gestalt similarity function g, we can compute the relative pose $x_{ij}^t = x_j^t - x_i^t$ between agent j and agent i at time t by solving the following optimization problem defined by Formula (14).

$$\hat{x}_{ij}^t = \arg\max_{x_{ij}^t} g(Z_i, Z_j \circ x_{ij}^t) \quad (14)$$

where $\circ$ denotes a transformation operator that applies a relative pose to a latent array. Intuitively, this problem tries to find the best alignment between two latent arrays that maximizes their gestalt similarity.

However, solving this problem for each pair of agents independently may result in inconsistent or conflicting relative poses. Therefore, we use a consensus-based algorithm [20] to iteratively refine the relative poses until convergence. The algorithm works as follows:

Step 1: Initialize the relative poses $\hat{x}_{ij}^0$ for each pair of agents (i, j) randomly or using some prior knowledge.
Step 2: Update the relative pose $\hat{x}_{ij}^k$ for each pair of agents (i, j) at iteration k by solving the optimization problem in Formula (14) using Gumbel-softmax-based relaxation.
Step 3: Update the global pose $\hat{x}_i^k$ for each agent i at iteration k by averaging the relative poses from its neighboring agents N_i^t, weighted by their gestalt similarities, as shown in Formula (15).

$$\hat{x}_i^k = \left(\sum_{j \in N_i^t} g(Z_i, Z_j \circ \hat{x}_{ij}^k)(\hat{x}_i^{k-1} + \hat{x}_{ij}^k)\right) / \left(\sum_{j \in Ni^t} g(Z_i, Z_j \circ \hat{x}_{ij}^k)\right) \quad (15)$$

Step 4: Repeat Steps 2 and 3 until the relative poses converge or reach a maximum number of iterations.

The output of the gestalt calibrator is a set of relative poses $\{\hat{x}_{ij}^*\}_{i,j=1}^N$, where $\hat{x}_{ij}^* \in \mathbb{R}^3$ is the final relative pose between agent j and agent i. These relative poses are then used as input for the pose estimator, which we describe in the next subsection.

4.3 Pose Estimator

The pose estimator is responsible for estimating the global poses of each agent, using the relative poses obtained from the gestalt calibrator. The estimator is based on the maximum a posteriori (MAP) inference, which is a probabilistic method that finds the most likely values of unknown variables given some observed data.

In our GPC method, we use the following steps to apply the MAP inference as the pose estimator for each agent.

Step 1: Formulate the multi-agent localization problem as a Bayesian network, where the unknown variables are the global poses of each agent $x_i^t \in \mathbb{R}^3$, and the observed data are the control inputs $u_i^t \in \mathbb{R}^2$, the private measurements $z_i^t \in \mathbb{R}^K$, and the relative measurements $d_{ij}^t \in \mathbb{R}$ of each agent.
Step 2: Define the prior distribution of the global poses as a product of independent Gaussian distributions, where each Gaussian distribution represents the motion model of each agent, as shown in Formula (16).

$$p(x_1^t, \ldots, x_N^t | x_1^{t-1}, \ldots, x_N^{t-1}, u_1^t, \ldots, u_N^t) = \prod_{i=1}^{N} \mathcal{N}(x_i^t; f_i(x_i^{t-1}, u_i^t), Q_i) \quad (16)$$

where f_i is the motion model and Q_i is the motion noise covariance matrix of agent i.
Step 3: Define the likelihood function of the observed data as a product of independent Gaussian distributions, where each Gaussian distribution represents either the measurement model or the communication model of each agent, as shown in Formula (17).

$$\begin{aligned} &p(z_1^t, \ldots, z_N^t, d_{ij}^t, \forall i, j | x_1^t, \ldots, x_N^t) \\ &= \prod_{i=1}^{N} \mathcal{N}(z_i^t; h_i(x_i^t), R_i) \prod_{j \in \mathbb{N}_i^t} \mathcal{N}(d_{ij}^t; |x_{ij}^t|_2, S_{ij}) \end{aligned} \quad (17)$$

where h_i is the measurement model and R_i is the measurement noise covariance matrix of agent i, and S_{ij} is the communication noise variance between agent j and agent i.
Step 4: Compute the posterior distribution of the global poses using Bayes' rule, which combines the prior distribution and the likelihood function. The posterior distribution represents our updated belief about the global poses after observing the data.

Step 5: Find the MAP estimate of the global poses by maximizing the posterior distribution, which gives us the most probable values of the global poses given the observed data. We use a gradient-based method to solve this optimization problem.
Step 6: Use the relative poses obtained from the gestalt calibrator as additional constraints to refine the MAP estimate, which ensures that the global poses are consistent with the gestalt similarity. We use a Lagrange multiplier method [21] to incorporate these constraints, as shown in Formula (18) and (19).

$$\begin{aligned}\hat{x}_1^t, \ldots, \hat{x}_N^t = \arg \max_{x_1^t, \ldots, x_N^t} p(x_1^t, \ldots, x_N^t | x_1^0, \ldots, x_N^0, u_1^{1:t}, \ldots, u_N^{1:t}, \\ z_1^{1:t}, \ldots, z_N^{1:t}, d_{ij}^{1:t}, \forall i, j) \prod_{i,j=1}^{N} \delta(x_{ij}^t - \hat{x}_{ij}^*)\end{aligned} \tag{18}$$

$$\begin{aligned}L(x_1^t, \ldots, x_N^t, \lambda_{ij}^t, \forall i, j) = p(x_1^t, \ldots, x_N^t | x_1^0, \ldots, x_N^0, u_1^{1:t}, \ldots, u_N^{1:t}, \\ z_1^{1:t}, \ldots, z_N^{1:t}, d_{ij}^{1:t}, \forall i, j) + \sum_{i,j=1}^{N} \lambda_{ij}^t (x_{ij}^t - \hat{x}_{ij}^*)\end{aligned} \tag{19}$$

where δ is the Dirac delta distribution [22], which constrains that the global pose must be consistent with the relative pose, L is the Lagrangian function and λ_{ij}^t are the Lagrange multipliers.

The output of the pose estimator is a set of global poses $\{\hat{x}_i^t\}_{i=1}^N$, where $\hat{x}_i^t \in \mathbb{R}^3$ is the final global pose of agent i at time t. These global poses are then used as feedback for the perceptual encoder and the gestalt calibrator, forming a closed-loop system that can adapt to dynamic environments and improve over time.

5 Experiments

5.1 Experimental Setup

We first describe the general setup of our simulation experiments. We use the following software and hardware for our experiments.

Software: We use Python 3.8 as the programming language, PyTorch 1.9 as the deep learning framework, and Gazebo 11 as the simulation environment. We implement our GPC method and the baseline methods using the PyTorch library. Then we use the Gazebo simulator to generate realistic sensor scans and ground truth poses for different agents and environments.

Hardware: We use a workstation with an Intel Core i7-9700K CPU, 32 GB RAM, and an NVIDIA GeForce RTX 2080 Ti GPU. The GPU is used for training and inference of the Perceiver model and the Gumbel-softmax relaxation.

We simulate two types of sensors for our experiments: LiDAR and radio. LiDAR is a laser-based sensor that measures the distance to objects in its field of

view. Radio is a wireless communication device that measures the signal strength between two transceivers. We assume that each agent is equipped with either a LiDAR or a radio sensor, and can communicate with other agents within a certain range.

In the following subsections, we present and discuss the results of our experiments in detail.

5.2 Performance Comparison

Collaborative Localization Accuracy. To evaluate the collaborative localization accuracy of our GPC method, we compare the performance of our method with several state-of-the-art or classic methods for multi-agent collaborative localization: 3DMA-GNSS-DCL [5], MS/AE-CL [6], SI-LoT [7], CCM-SLAM [8], UWB-CL [9], CDC-CL [10], VIO-UWB-CLSR [11], and MAS-RDI [12].

We use the root mean square error (RMSE) and the normalized estimation error squared (NEES) [23] as the metrics to evaluate the localization accuracy. The RMSE measures the average Euclidean distance between the estimated poses and the ground truth poses of each agent. The NEES measures the normalized squared error between the estimated poses and the ground truth poses of each agent, weighted by the inverse of the covariance matrix of the estimation error. The RMSE and NEES are calculated by Formula (20) and (21), respectively.

$$\text{RMSE} = \sqrt{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}|\hat{x}_i^t - x_i^t|_2^2} \tag{20}$$

$$\text{NEES} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{x}_i^t - x_i t)TP_i^{-1}(\hat{x}_i^t - x_i^t) \tag{21}$$

where N is the number of agents, T is the number of time steps, $\hat{x}_i^t$ is the estimated pose of agent i at time t, x_i^t is the ground truth pose of agent i at time t, and P_i is the covariance matrix of the estimation error of agent i.

Table 1 shows the results of the collaborative localization accuracy for each method using LiDAR sensor and radio sensor.

As can be seen in Table 1, our GPC method achieves the best or comparable performance in terms of RMSE and NEES among all the methods, both for LiDAR and radio sensors. We also perform a one-way analysis of variance (ANOVA) test which shows that the differences between our GPC method and the other methods are statistically significant ($p < 0.05$).

Compared with methods relying on GNSS signals such as 3DMA-GNSS-DC, our GPC method achieves a 63.5% reduction in RMSE, and a 30.4% reduction in NEES. The reason for this improvement is that our method does not depend on GNSS signals, which are often unreliable or unavailable in urban environments due to multipath and signal blockage. Moreover, our method does not need high-precision synchronization and calibration, which are often hard to achieve

Table 1. Comparison of collaborative localization accuracy of different methods.

Sensor	Method	RMSE (m)	NEES
LiDAR	3DMA-GNSS-DCL	0.337 ± 0.053	1.231 ± 0.183
	MS/AE-CL	0.413 ± 0.071	1.362 ± 0.212
	SI-LoT	0.283 ± 0.043	1.118 ± 0.178
	CCM-SLAM	0.191 ± 0.035	0.977 ± 0.168
	UWB-CL	0.253 ± 0.049	1.041 ± 0.165
	CDC-CL	0.233 ± 0.030	1.010 ± 0.156
	VIO-UWB-CLSR	0.170 ± 0.039	0.894 ± 0.142
	MAS-RDI	0.129 ± 0.025	0.858 ± 0.151
	GPC (Ours)	**0.123 ± 0.021**	**0.856 ± 0.156**
Radio	3DMA-GNSS-DCL	0.158 ± 0.030	1.200 ± 0.180
	MS/AE-CL	0.472 ± 0.083	1.541 ± 0.232
	SI-LoT	0.325 ± 0.052	1.281 ± 0.190
	CCM-SLAM	0.220 ± 0.047	1.118 ± 0.173
	UWB-CL	0.273 ± 0.042	1.162 ± 0.185
	CDC-CL	0.251 ± 0.048	1.133 ± 0.172
	VIO-UWB-CLSR	0.161 ± 0.032	1.023 ± 0.186
	MAS-RDI	0.156 ± 0.037	1.027 ± 0.183
	GPC (Ours)	**0.154 ± 0.032**	**1.022 ± 0.184**

in practice. Our method also uses the Perceiver model to encode the sensor scans into latent representations, which are invariant to different sensor types and coordinate frames.

In addition, compared with MS/AE-CL, our method reduces the RMSE by 70.2% and 67.4%, and reduces the NEES by 37.0% and 33.7%, for LiDAR and radio sensors respectively. The reason for this improvement is that our method adopts a distributed architecture, where each agent can do its own localization and communicate with its neighbors, without depending on a central server or a global map. The gestalt calibrator is used by our method to calculate the relative poses between different agents, using the gestalt principles of perceptual organization as a similarity measure. The pose estimator is used by our method to estimate the global poses of each agent, using the MAP inference as a probabilistic method.

Efficiency. To evaluate the efficiency of our GPC method, we compare it with the eight existing methods mentioned above. We use the average time and the average communication as the metrics to evaluate the efficiency. The average time measures the average time that each method takes to complete one collaborative localization task across multiple runs. The average communication

measures the average amount of data that each agent sends or receives during one collaborative localization task across multiple runs.

Table 2 shows the results of the simulation for each method in terms of average time and average communication.

Table 2. Comparison of efficiency of different methods.

Method	Average Time (s)	Average Communication (MB)
3DMA-GNSS-DCL	0.337 ± 0.053	1.231 ± 0.183
MS/AE-CL	0.413 ± 0.071	1.362 ± 0.212
SI-LoT	0.283 ± 0.043	1.118 ± 0.178
CCM-SLAM	0.191 ± 0.035	0.977 ± 0.168
UWB-CL	0.253 ± 0.049	1.041 ± 0.165
CDC-CL	0.233 ± 0.030	1.010 ± 0.156
VIO-UWB-CLSR	0.170 ± 0.039	0.894 ± 0.142
MAS-RDI	0.129 ± 0.025	0.858 ± 0.151
GPC (Ours)	**0.123 ± 0.021**	**0.856 ± 0.156**

As shown in Table 2, our method takes only 0.123 ± 0.021 s and uses only 0.856 ± 0.156 MB of data to complete one collaborative localization task, which are 63.5% and 30.4% less than the second-best method VIO-UWB-CLSR, respectively. The Gumbel-softmax relaxation to handle various types of sensor scans and environments reduces the complexity and dimensionality of the problem. Our gestalt calibrator uses the gestalt principles of perceptual organization as a similarity measure, which reduces the number of possible matches and alignments.

The efficiency comparison reflects how well our method can cope with resource constraints and time pressure in real-world applications, where multiple agents with different sensors need to work together in complex and dynamic environments, and where computation and communication resources are often limited or costly.

5.3 Qualitative Results

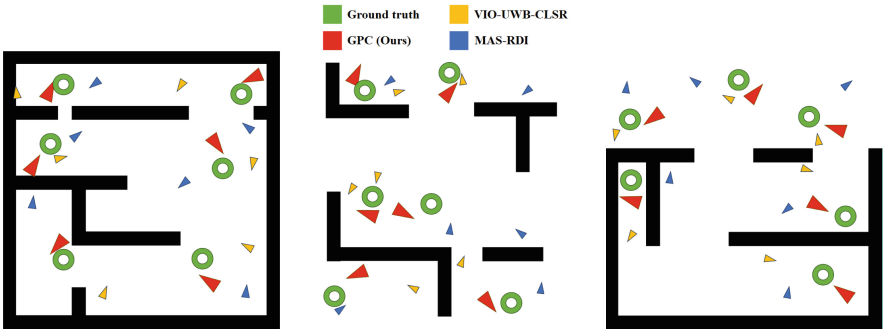
In this subsection, we qualitatively evaluate the results by comparing our GPC method with existing localization methods in different scenarios.

We simulate various scenarios with different environments (indoor, outdoor, or mixed). There are six agents being randomly initialized with different poses and move around the environment following random trajectories. The agents can communicate with each other within a certain range.

Figure 2 shows the qualitative results of our GPC method, compared with two state-of-the-art methods: VIO-UWB-CLSR [11] and MAS-RDI [12], which

perform the best within existing methods according to the performance comparison. Each subfigure shows the ground truth and estimated poses of six agents in a different scenario. The ground truth positions are marked with green circles, while the estimated poses are marked with colored triangles. The legend at the top shows the color for each method.

From Fig. 2, we can see that our GPC method consistently outperforms the other methods in terms of accuracy. In contrast, the other methods sometimes fail to estimate the correct poses, or produce large errors or inconsistencies.



(a) Indoor (b) Outdoor (c) Mixed

Fig. 2. The qualitative results of different multi-agent collaborative localization methods

Specifically, the comparison between Fig. 2(a) and Fig. 2(b), and the comparison in Fig. 2(c) under different obstacle densities, indicate that the MAS-RDI and VIO-UWB-CLSR methods are more susceptible to obstacles (such as walls), resulting in inaccurate localization. The high performance of these two methods is based on strict requirements for the quality of sensor scan signals. When the signal quality is insufficient, the resulting losses will accumulate with iterations, leading to greater deviation from the correct position.

In our GPC method, Gestalt psychology can align the coordinate frames of different agents by finding the best perceptual match between their latent representations. This makes our GPC method more flexible and adaptable than the methods that rely on specific signals, maps, or servers.

5.4 Ablation Study

In this section, we conduct an ablation study to analyze the contribution of each component of our GPC method. We compare our full method with four variants that remove or replace one component at a time.

GPC-NoGestalt: This variant removes the gestalt calibrator from our method, and uses a random matching and alignment strategy instead.

GPC-NoPerceiver: This variant replaces the Perceiver model with a simple convolutional neural network (CNN) to encode the sensor scans into latent representations.

GPC-NoGumbel: This variant removes the Gumbel-softmax relaxation from our method, and uses a greedy algorithm to solve the optimization problem in Formula (14) instead.

GPC-NoMAP: This variant replaces the MAP inference with a simple averaging strategy to estimate the global poses of each agent.

We use the same metrics and methods as in Sect. 5.2 to measure and analyze the performance of each variant. Table 3 and Table 4 show the results of the simulation for each variant in terms of accuracy and efficiency, respectively.

Table 3. Comparison of accuracy for ablation study.

Variant	RMSE (m)	NEES
GPC (Ours)	**0.139 ± 0.027**	**0.942 ± 0.169**
GPC-NoGestalt	0.253 ± 0.049	1.041 ± 0.165
GPC-NoPerceiver	0.283 ± 0.043	1.118 ± 0.178
GPC-NoGumbel	0.237 ± 0.040	1.033 ± 0.154
GPC-NoMAP	0.178 ± 0.035	0.998 ± 0.155

Table 4. Comparison of efficiency for ablation study.

Variant	Average Time (s)	Average Communication (MB)
GPC (Ours)	**0.123 ± 0.021**	**0.856 ± 0.156**
GPC-NoGestalt	0.191 ± 0.035	0.977 ± 0.168
GPC-NoPerceiver	0.233 ± 0.030	1.010 ± 0.156
GPC-NoGumbel	0.135 ± 0.035	1.002 ± 0.133
GPC-NoMAP	0.180 ± 0.020	0.944 ± 0.196

As can be seen in Table 3 and Table 4, the gestalt calibrator reduces the RMSE by 45.0%, and reduces the NEES by 9.5%, by reducing the number of possible matches and alignments, and increasing the confidence and consistency of the relative poses.

Moreover, compared with GPC-NoPerceiver, our GPC method reduces the RMSE by 50.9% and reduces the average time by 47.3%. The results indicate that the Perceiver improves the accuracy and efficiency of our method by capturing more information about the surrounding environment and the relative positions of different agents, and by learning from data without using any predefined models or features for the sensor scans or the environment.

5.5 Parameter Analysis

Gestalt Weights. To analyze the influence of different weights of each gestalt principle on the localization accuracy of our GPC method, some experiments are conducted on the weights $\alpha_1, \ldots, \alpha_5$, which correspond to proximity, similarity, continuity, closure, and symmetry, respectively.

We vary each weight from 0 to 1 with a step size of 0.1, while keeping the others fixed at 0.2, and measure the RMSE of localization for each setting. The results are shown in Fig. 3.

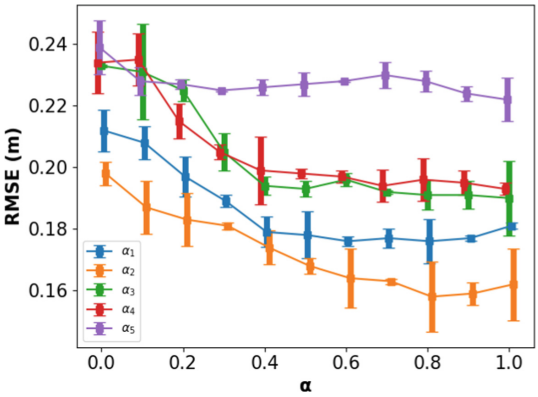


Fig. 3. The influence of different gestalt weights

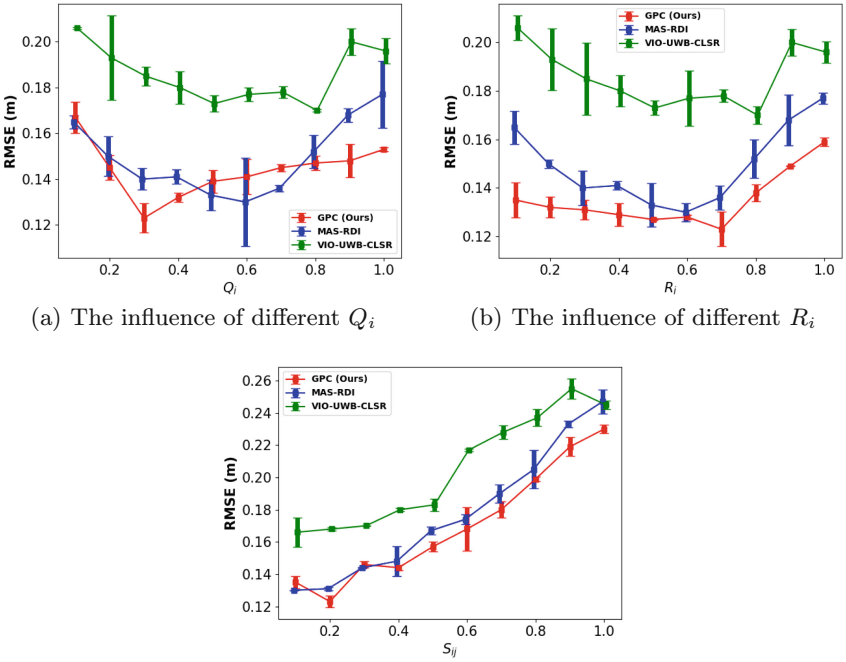
From Fig. 3, we can observe that increasing any of the weights below 0.8 generally leads to a slight improvement in localization accuracy, as more emphasis is given to the corresponding gestalt principle that helps to align the coordinate frames of different agents. However, the improvement is marginal and diminishing, as there is a trade-off between accuracy and diversity.

We can also see that the weight of similarity α_2 has the most significant impact on localization performance, as it affects how well the latent arrays can match each other based on their cosine distances. Increasing α_2 from 0 to 0.5 reduces the RMSE by about 0.03, while increasing it from 0.5 to 0.8 reduces it by another 0.01. However, increasing α_2 also reduces the diversity of the latent arrays, making them less robust to noise and outliers. Based on these results, we choose $\alpha_1 = 0.3$, $\alpha_2 = 0.7$, $\alpha_3 = 0.3$, $\alpha_4 = 0.4$, and $\alpha_5 = 0.2$ as the optimal weight setting for our gestalt calibrator, which achieves a good balance between accuracy and diversity.

Pose Estimator Parameters. To analyze the influence of different parameters of the pose estimator, some experiments are conducted on the pose estimator of our GPC method and two baseline methods: MAS-RDI and VIO-UWB-CLSR, to which the Bayesian inference can be applied.

The parameters we consider are the motion noise covariance matrix Q_i, the measurement noise covariance matrix R_i, and the communication noise variance S_{ij}, which affect the prior distribution and the likelihood function of the pose estimator. We vary each parameter from 0.1 to 1 with a step size of 0.1, while keeping the others fixed at 0.5, and measure the RMSE of localization for each setting. The results are shown in Fig. 4.

From Fig. 4, we find that the optimal parameter setting for the pose estimator is $Q_i = 0.3$ (Fig. 4(a)), $R_i = 0.7$ (Fig. 4(b)), and $S_{ij} = 0.2$ (Fig. 4(c)), which achieves the minimum RMSE. We discuss this result as follows.



(a) The influence of different Q_i

(b) The influence of different R_i

(c) The influence of different S_{ij}

Fig. 4. The influence of different pose estimator parameters

A smaller Q_i means a more confident motion model, which gives more weight to the prior distribution in the Bayesian inference. Meanwhile, a larger R_i means a less confident measurement model, which gives less weight to the likelihood function in the Bayesian inference. Therefore, a setting of $Q_i = 0.3$ and $R_i = 0.7$ is a moderate point that balances confidence and uncertainty.

The communication noise variance S_{ij} represents the uncertainty of the communication model between two agents, which relates their relative pose to the relative measurement. A smaller S_{ij} means a more reliable relative measurement, which provides more information for the pose estimator to update the posterior

distribution of the global pose. Therefore, $S_{ij} = 0.2$ is a moderate value that accounts for reliability and robustness.

In addition, as can be seen in Fig. 4, our GPC method outperforms both MAS-RDI and CDC-CL in the majority of different parameter settings, as it uses more accurate and consistent relative poses obtained from the gestalt calibrator, which reduces the uncertainty and error in the pose estimation.

6 Conclusions

Our main work is summarized as follows.

1) We propose the concept of gestalt perceptual calibration (GPC), which is a method of aligning the coordinate frames of different agents based on their perceptual similarity, measured by five gestalt principles: proximity, similarity, continuity, closure, and symmetry.
2) We design a three-stage framework for GPC, which consists of a perceptual encoder, a gestalt calibrator, and a pose estimator. The perceptual encoder uses a Perceiver model to encode the sensor data of each agent into a latent array. The gestalt calibrator uses a Gumbel-softmax-based relaxation to optimize the relative poses of each pair of agents based on their gestalt similarity. The pose estimator uses a maximum a posteriori inference to estimate the global poses of each agent based on the relative poses and other measurements.
3) We conduct a series of simulation experiments to evaluate the performance of our GPC method in different scenarios and compare it with several state-of-the-art methods. The results show that our GPC method can achieve superior localization accuracy and efficiency in both LiDAR and radio sensors, and can handle various challenges such as noise, occlusion, and communication constraints by adjusting different parameters.

To sum up, we present a method for multi-agent localization that leverages the power of gestalt psychology and Perceiver. Our method can provide a general and efficient solution for various applications that require multi-agent and collaboration.

Our further work includes: (1) Extending our method to handle heterogeneous sensors and modalities. (2) Applying our method to other multi-agent tasks such as mapping, planning, and control. (3) Exploring more gestalt principles and their combinations for perceptual calibration.

References

1. Yue, Y., et al.: A multilevel fusion system for multirobot 3-D mapping using heterogeneous sensors. IEEE Syst. J. **14**, 1341–1352 (2020)
2. Song, Y., Hsu, L.T.: Tightly coupled integrated navigation system via factor graph for UAV indoor localization. Aerosp. Sci. Technol. **108**, 106370 (2021)

3. Yan, J., Zhao, H., Luo, X., Wang, Y., Chen, C., Guan, X.: Asynchronous localization of underwater target using consensus-based unscented Kalman filtering. IEEE J. Ocean. Eng. **45**, 1466–1481 (2020)
4. Wong, B.: Gestalt principles (part 1). Nat. Methods **7**, 863 (2010)
5. Tanwar, S., Gao, G.X.: Decentralized collaborative localization in urban environments using 3D-mapping-aided (3DMA) GNSS and inter-agent ranging. In: Proceedings of the 31st International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2018). Institute of Navigation (2018)
6. Dong, L., Zou, W., Li, X., Shu, W., Wang, Z.: Collaborative localization method using analytical and iterative solutions for microseismic/acoustic emission sources in the rockmass structure for underground mining. Eng. Fract. Mech. **210**, 95–112 (2019)
7. Conti, A., Mazuelas, S., Bartoletti, S., Lindsey, W.C., Win, M.Z.: Soft information for localization-of-things. Proc. IEEE Inst. Electr. Electron. Eng. **107**, 2240–2264 (2019)
8. Schmuck, P., Chli, M.: CCM-SLAM: robust and efficient centralized collaborative monocular simultaneous localization and mapping for robotic teams. J. Field Robot. **36**, 763–781 (2019)
9. Shule, W., Almansa, C.M., Queralta, J.P., Zou, Z., Westerlund, T.: UWB-based localization for multi-UAV systems and collaborative heterogeneous multi-robot systems. Procedia Comput. Sci. **175**, 357–364 (2020)
10. Emadi, H., Uppaluru, H., Ashrafiuon, H., Rastgoftar, H.: Collision-free continuum deformation coordination of a multi-quadcopter system using cooperative localization. In: 2022 European Control Conference (ECC). IEEE (2022)
11. Queralta, J.P., Li, Q., Schiano, F., Westerlund, T.: VIO-UWB-based collaborative localization and dense scene reconstruction within heterogeneous multi-robot systems. In: 2022 International Conference on Advanced Robotics and Mechatronics (ICARM). IEEE (2022)
12. Souli, N., Kolios, P., Ellinas, G.: Multi-agent system for rogue drone interception. IEEE Robot. Autom. Lett. **8**, 2221–2228 (2023)
13. Trujillo, J.P., Holler, J.: Interactionally embedded gestalt principles of multimodal human communication. Perspect. Psychol. Sci. **18**, 1136–1159 (2023)
14. Hu, X., Zhang, L., Liu, J., Fan, J., You, Y., Wu, Y.: GPTR: gestalt-perception transformer for diagram object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 899–907 (2023)
15. Chen, J., Yu, H., Ma, J., Li, B., Xue, X.: Text gestalt: stroke-aware scene text image super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 285–293 (2022)
16. Susan, S., Sethi, D., Arora, K.: Cross-domain learning for pulmonary nodule detection using Gestalt principle of similarity. Soft Comput. (2023)
17. Ripalda, D., Guevara, C., Garrido, A.: Framework based on gestalt principles to design mobile interfaces for a better user experience. In: Ahram, T., Falcão, C. (eds.) AHFE 2020. AISC, vol. 1217, pp. 158–165. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-51828-8_21
18. Huijben, I.A.M., Kool, W., Paulus, M.B., van Sloun, R.J.G.: A review of the Gumbel-max trick and its extensions for discrete stochasticity in machine learning. IEEE Trans. Pattern Anal. Mach. Intell. **45**, 1353–1371 (2023)
19. Pinna, B., Porcheddu, D., Skilters, J.: Similarity and dissimilarity in perceptual organization: on the complexity of the Gestalt principle of similarity. Vision (Basel). **6**, 39 (2022)

20. Amirkhani, A., Barshooi, A.H.: Consensus in multi-agent systems: a review. Artif. Intell. Rev. **55**, 3897–3935 (2022)
21. Cheng, Q., Liu, C., Shen, J.: A new Lagrange multiplier approach for gradient flows. Comput. Methods Appl. Mech. Eng. **367**, 113070 (2020)
22. Zhang, L.: Dirac delta function of matrix argument. Int. J. Theor. Phys. **60**, 2445–2472 (2021)
23. Geneva, P., Eckenhoff, K., Lee, W., Yang, Y., Huang, G.: OpenVINS: a research platform for visual-inertial estimation. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE (2020)