Map Merging of Heterogeneous Topometric Maps Using Attributed Node Embedding

Maximiliaan van Schendel
student #4384644
m.vanschendel@tudelft.nl

1st supervisor: Edward Verbree
2nd supervisor: Pirouz Nourian
external supervisor: Robert Voûte

24/01/2022

Contents

1 Introduction

Collaborative mapping allows multiple agents to work together to create a single, global map of their environment. By working together large areas can be mapped in a short amount of time. Most existing research has focused on when mapping agents are homogeneous, meaning they sense their environment and behave similarly (**?**). However, there are situations where mapping with heterogeneous agents can be advantageous (**?**). For example, a human carrying lightweight, low-end sensors collaborating with a robot carrying heavy, high-end sensors to map an environment where some areas are not accessible by humans. By compensating for the weaknesses of one agent with the strengths of another more environments and situations can be handled.

Existing collaborative mapping approaches are not well suited for mapping with heterogeneous agents as they often rely on agents being able to communicate with eachother. This is especially the case in indoor environments where external positioning signals are highly attenuated. In this case, a global map can only be created by merging the partial maps of the environment created by each agent individually based on their overlapping areas. This is called map merging (see figure **??**). When partial maps are created by heterogeneous agents they might represent different aspects of the environment at different a different scale, resolution or accuracy, which further complicates map merging as overlapping areas may not appear the same between partial maps.

In this thesis we propose to use both the hierarchical topological relationships of indoor environments, meaning the connectivity between distinctive places and their nesting relationships, and their metric characteristics, the geometry of the environment, to solve the heterogeneous map merging problem. We do this by extracting 3D hierarchical topological-metric maps from heterogeneous partial maps. We then use both the topological and metric characteristics of the partial maps to robustly identify overlapping areas that might appear differently due to being captured by heterogeneous agents. Our hypothesis is that the connectivity and hierarchy of places within the environment are identifiable between heterogeneous partial maps. We further hypothesize that using the metric characteristics of the environment in conjunction with its topological characteristics will improve identification of overlapping areas over a purely topological approach, despite geometrical differences between heterogeneous partial maps. The most important contributions of our work are: 1) applying 3D hierarchical topological-metric map merging to indoor environments. 2) extracting 3D hierarchical topological-metric maps from heterogeneous partial maps. In the rest of this section we will give more detailed definitions for the concepts mentioned above and others that are commonly used in this report.

1.1 Definitions

In this section we will give a number of definitions for the regularly used concepts in this thesis.

./figures/raster/map_merging.png

Figure 1: Partial maps (1) captured by three different agents and resultant global map (2) after map merging.

Map A symbolic representation of an environment which contains information about its characteristics.

Map Representation The choice of characteristics of the environment that a map shows. Hybrid map representations are representations that shows a combination of multiple characteristics of the environment, e.g. a hybrid topological-metric map containing both the environment's large-scale structure and small-scale geometry.

Agent According to **?** "An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators.". In the context of this thesis an agent specifically refers to any human or robot that is capable of perceiving its environment by means of a monocular camera, stereo camera or lidar and is capable of exploring their environment. We further define heterogeneous agents as having different sensing capabilities and being unable to communicate or position themselves relative to eachother.

Partial Maps A collection of maps without a common coordinate frame that each represent a part of the environment. We denote the set of all partials maps as $I = \{m_i\}_{i=1}^{n}$, where n is the number of partial maps. We define heterogeneous partial maps as partial maps captured by heterogeneous agents, thus having different scale, resolution, accuracy or precision.

Global Map A single, more complete map constructed by merging multiple partial maps. To do so, a coordinate transformation T must be applied to the partial maps to bring them into a common coordinate frame. We denote the global map constructed by merging a subset of partial maps $J \subset I$ using coordinate transformations $\{T_i\}_{i=1}^{|J|}$ as G_J, such that $G_J = \{T_i(m_i)\}_{i=1}^{|J|}$.

Map Merging The map merging problem can be stated as follows: given two partial maps $m_1 \in I, m_2 \in I$, find the coordinate transformation T that minimizes a dissimilarity function $\psi(m_1, T(m_2))$ (**?**). The goal of this is to "maximize the overlap of regions that appear in two or more partial maps" (**?**). Depending on the approach, the coordinate transformation, dissimilarity function or optimization method differ.

Metric Map Metric maps represent the geometry of an environment. They are usually derived from sensor range measurements, either directly or by using structure from motion or simultaneous localization and mapping algorithms. A common metric map representation is the occupancy grid (**?**). Another common metric map representation is the point cloud, which represents the surface of the environment as a collection of points. Both map representations are shown in figure **??**.

Topological Map Topological maps are a qualitative graph representation of an environment's structure, where vertices represent locally distinctive places, often rooms, and edges represent traversable paths between them (see figure **??**) (**??**). Topological maps are inspired by the fact that humans are capable of spatial learning despite limited sensory and processing capability and only having partial knowledge of the environment. This is based on observations that cognitive maps, the mental maps used by humans to navigate within an environment, consist of multiple layers with a topological description of the environment being a fundamental component (**??**). We denote the topology of an environment as a graph G, where vertices V represent n distinctive places v_i and edges E represent the presence of m traversable paths between neighbouring pairs of places $\{v_j, v_k\}$, such that $G = (V, E)$, $V = \{v_i\}_{i=1}^{n}, E = \{\{v_j, v_k\}_i\}_{i=1}^{m}, v_j \in V, v_k \in V$.

Figure 2: Diagram showing the topological map of a building with four rooms connected by doors or stairs.

In the context of indoor mapping the graph is often embedded in 2D or 3D euclidean space as a spatial graph. Given the embedding $f : G \rightarrow R^n$, $\widetilde{G} := f(G)$, we denote $\widetilde{G}$ as the spatial graph of G (**?**). Each vertex in $\widetilde{G}$ has an associated 3D coordinate describing its position in the coordinate frame of the map. If the specific paths between vertices are known they can be associated with the edges in $\widetilde{G}$, otherwise an edge only represents a possible path between two points in space.

Point Cloud An unordered collection of points representing the geometry of an object or environment in 3D euclidean space, defined as $\mathcal{P} = \{p_i\}_{i=1}^{n}, p_i \in \mathbb{R}^3$, where n denotes the number of points (**?**). See figure **??** for an example of a point cloud.

Occupancy Grid Also known as a voxel grid, an occupancy grid is a "multi-dimensional (typically 2D or 3D) tesselation of space into cells, where each cell stores a probabilistic estimate of its state." (**?**). A 3D occupancy grid with a number of cells along each dimension, $O_{dim} \in \mathbb{Z}^3_{\geq 0}$ can be represented by a collection of positive integer cell indices that have a non-zero chance of being occupied $O = \{c_i\}_{i=1}^{n}$, $n \in \mathbb{R}$, $n \in [1, \prod_{i=1}^{3} O_{dim,i}], c \in \mathbb{Z}^3_{\geq 0}$. The probability of a cell being occupied is given by a cell's associated state variable $s(c) \in \mathbb{R}$,

$s(c) \in [0,1]$. Cells that are within the extent of the grid that are not stored have zero chance of being occupied, such that $\forall c \in \{c \notin O | c_i \in [0, O_{dim,i}]\} \Rightarrow s(c) = 0$. In the case of a binary occupancy grid a cell can either be occupied or not, such that $s(c) \in \{0,1\}$. In this case, explicitly storing a state variable is not necessary as only the cells in O are occupied, as given by $\forall c \in O \Rightarrow s(c) = 1$. See figure **??** for an example of an occupancy grid.

./figures/raster/voxelization.png

Figure 3: Example of an occupancy grid (right) derived from a point cloud (left).

Topological-Metric Map A hybrid map representation combining both the topological and metric characteristics of the environment. This map representation allows the end-user to use either topological or metric information depending on the needs of the situation, e.g. the topological layer can be used for large-scale navigation and abstract reasoning while the metric layer can be used for landmark detection or obstacle avoidance. In the context of this thesis a topological-metric map refers to a 3D representation of an environment containing both a metric occupancy grid map $\mathbb{M}_M$ representing its geometry and a spatial graph $\widetilde{\mathbb{M}_T} = (V, E)$ representing its topology. Each vertex $v \in V$ has an associated variable $m(v)$ representing a subset of the full metric map describing that place, such that $m(v) \subset \mathbb{M}_M$. See figure **??** for an example of a topological-metric map.

2 Map Extraction

In the map extraction module we extract topometric maps from each individual partial point cloud map. In this section we will discuss the algorithms and data structures used for this purpose.

2.1 Overview

2.2 Voxel representation

A voxel is the 3D equivalent of a pixel. A voxel represents a single cell in a bounded 3D volume divided into a regular grid, a voxel grid, and its associated properties. For example, a voxel may contain information about whether it is occupied and what its color is. We consider a voxel as a three-dimensional vector representing its coordinates along the x, y and z axes of the voxel grid.

To generate a voxel grid we divide a 3D axis-aligned volume V, defined by minimum and maximum bounds $V_{min}, V_{max} \in \mathbb{R}^3$ into a grid of cubic cells with edges of length $e_l \in \mathbb{R}^+$. A voxel $\boldsymbol{v} = (x, y, z) \in \mathbb{Z}^{\nvDash+}$ represents a subvolume of V bounded by a single cell. The minimum and maximum bounds of this subvolume are given by $\boldsymbol{v_{min}} = V_{min} + \boldsymbol{v} * e_l$, and $\boldsymbol{v_{max}} = V_{min} + (\boldsymbol{v} + 1) * e_l$. The voxel's centroid is given by $\boldsymbol{v_c} = (\boldsymbol{v_{min}} + \boldsymbol{v_{max}}) * 0.5$. Given a point $\boldsymbol{p}$ within the bounds of V, the corresponding voxel is given by $\boldsymbol{v_p} = (\boldsymbol{p} - V_{min})//e_l$, where $//$ denotes integer division. A property of a voxel is given by the function $\mathcal{V}_{property} : \mathbb{Z}^{3+} \mapsto \mathbb{R}^{m*n}$.

Due to the regularly spaced nature of the voxel grid a voxel coordinate only consists of integer values. Voxel $\boldsymbol{v_{V_{min}}} = (0, 0, 0)$ represents the first cell along each of the voxel grid's axes and the minimum of the volume's bounds, voxel represents $(0, 1, 1)$ the first cell along the x and the second along the y and z axes, etc. We also restrict voxel coordinates to only be positive as negative coordinates would fall outside of the bounds of the volume. For the same reason, a voxel's coordinates can not be larger than that of the voxel representing the volume's maximum bounds $\boldsymbol{v_{V_{max}}} = (V_{max} - V_{min})//e_l$. We define a voxel grid as a set of occupied voxels $\mathcal{V} = \{v_i\}_{i=1}^n,\ n \in [1, \prod \boldsymbol{v_{V_{max}}}]$ with an associated bounded volume $V_{\mathcal{V}}$ and edge length e_l.

2.2.1 Sparse Voxel Octree

Several operations on voxel grids benefit from using a spatial index, including range searching, radius searching and level of detail generation. We use a data structure called a sparse voxel octree (SVO) to achieve this. A normal octree recursively subdivides a volume into 8 cells, called octants. This operation results in a tree data structure, with nodes representing octants at a certain level of subsidivision. The root node of the tree structure represents the entire volume while the leaf nodes represent batches of 1 or more data points. In the case of a voxel octree, the leaf nodes represent individual voxels. In a sparse voxel octree only the octants which are occupied are represented in the tree.

To generate the SVO we first create a Morton Order for the voxel grid. A Morton order maps the three-dimensional coordinates of the voxels to one dimension while preserving locality. It does by interleaving the binary coordinates into a single binary number, called a Morton code. The ordered vector of Morton codes gives the Morton order. We define the Morton order of a voxel grid as $M_{\mathcal{V}} = \{m_i\}_{i=1}^{|\mathcal{V}|},\ m_i < m_{i+1},\ m_i \in \mathbb{Z}^{+}$. We then divide the Morton order into buckets with width 8, such that each bucket contains at most 8 Morton codes, with a maximum difference of 8, in Morton order. Each non-empty bucket represents a parent node of at most 8 child nodes in the octree. By recursively performing this step until only one bucket remains, the root node, a sparse voxel octree is constructed.

We denote the function that returns all n voxels within range r of a voxel as $radius : \mathbb{Z}^{3+},\ \mathbb{R} \mapsto \mathbb{Z}^{n \times 3}$.

2.3 Navigable volume

To extract the topology from the voxel grid map we first extract a navigable volume. The navigable volume tells us which voxels a theoretical agent in the environment would use to navigate through that environment. In practice, this means the areas of the floor and stairs that are at a sufficient distance from a wall and a ceiling. We compute the navigable volume using a three step algorithm.

2.3.1 Voxel convolution

Voxel convolution involves moving a sliding window, or kernel, over each voxel in the grid to retrieve its neighbourhood and then computing a new value for the voxel based on computing a function $\mathcal{K}_f : \mathcal{K}_w \mapsto \mathbb{R}^{m*n}$ over its neighbours. The neighbourhood can be a radius around the voxel, its Von Neumann neighbourhood, its Moore neighbourhood, or any other arbitrary shape. We can define a kernel $\mathcal{K}$ as a voxel grid, with an associated weight for each voxel $weight : \mathbb{Z}^{3+} \mapsto \mathbb{R},\ v \in \mathcal{K}$ and an origin voxel $\boldsymbol{o_{\mathcal{K}}} \in \mathbb{Z}^{\nvDash}$. To apply a kernel to a voxel we first translate the kernel so that its origin lies on the voxel, such that $\mathcal{K}_{\boldsymbol{v}} = \{\boldsymbol{v_{\mathcal{K}}} + (\boldsymbol{v} - \boldsymbol{o_{\mathcal{K}}}) \mid \boldsymbol{v_{\mathcal{K}}} \in \mathcal{K}\}$. We then get the property which we wish to convolve of each neighbour and multiply it by the neighbour's weight, such that $\mathcal{K}_w = \{weight(\boldsymbol{v}) * \mathcal{V}_{property}(\boldsymbol{v}) \mid \boldsymbol{v} \in \mathcal{K}_v \cap \mathcal{V}\}$. The property after convolution is then given by $\mathcal{K}_f(\mathcal{K}_w)$. We denote the convolution of a property of every voxel in $\mathcal{V}$ with $\mathcal{K}$ as $\mathcal{V}_{property,\ \mathcal{K}} = c(\mathcal{V}_{property}, \mathcal{K})$. If the property is left out it is implied to be occupancy.

The first step in the navigable volume extraction uses voxel convolution with a stick-shaped kernel $\mathcal{K}_{stick}$ based on the research by (GORTE REFERENCE). Each voxel in the kernel has a weight of 1, except the origin voxel which has weight 0. The associated function is summation. Convolving the voxel grid's occupancy property with the stick kernel results in voxels with an obstructed value of 0 when no other voxels are in the stick kernel. This indicates that these voxels have enough space around and above them to possibly be navigable. We denote this convolution as $\mathcal{V}_{\mathcal{K}_{stick}} = c(\mathcal{V}, \mathcal{K}_{stick})$. We then filter out all non-zero voxels, such that $\mathcal{V}_{unobstructed} = \{\boldsymbol{v} \in \mathcal{V}_{\mathcal{K}_{stick}} \mid \mathcal{V}_{obstructed}(\boldsymbol{v}) = 0\}$.

2.3.2 Dilation

The next step of the algorithm is to dilate the unobstructed voxels upwards by 20-25cm. This connects the unoccupied voxels separated by a small height differences into a connected volume. The result is a new voxel grid $\mathcal{V}_{dilated}$.

2.3.3 Connected components

The final step of the algorithm is to split $\mathcal{V}_{dilated}$ into one or more connected components. A connected component $\mathcal{V}_c$ of a voxel grid is a subset of $\mathcal{V}$ where there exists a path between every voxel in $\mathcal{V}_c$.

The neighbourhood graph $\mathcal{G}_{\mathcal{V}} = (V,\ E)$ of $\mathcal{V}$ represents the voxels in $\mathcal{V}$ as nodes. Each node has incident edges towards all other voxels in its neighbourhood, which is defined by a kernel $\mathcal{K}$, such that $V = \mathcal{V},\ E_V = \{(\boldsymbol{v}, \boldsymbol{v_{nb}}) \mid \boldsymbol{v} \in \mathcal{V},\ \boldsymbol{v_{nb}} \in \mathcal{K}_{\boldsymbol{v}}\},\ |E_V| \leq |\mathcal{K}| * |\mathcal{V}|$.

There exists a path between two voxels $\boldsymbol{v_a},\ \boldsymbol{v_b}$ in $\mathcal{V}$ if there exists a path between their corresponding nodes $V_a,\ V_b$ in $\mathcal{G}_{\mathcal{V}}$. We denote the set of all possible paths between two nodes as $paths(V_a,\ V_b)$. A connected component is then defined as $\mathcal{V}_c = \{\boldsymbol{v_a} \mid \boldsymbol{v_a} \in \mathcal{V},\ \boldsymbol{v_b} \in \mathcal{V},\ |paths(V_a, V_b)| \neq 0\}$. We define the set of all connected components as $\mathcal{V}_{cc} = \{\mathcal{V}_{c,\ i}\}_{i=1}^{n}$. Finally, we extract the navigable volume $\mathcal{V}_{nav}$ by finding the connected component with the most voxels.

2.4 Maximum visibility estimation

To compute the isovists necessary for room segmentation it is first necessary to estimate hypothetical scanning positions that maximize the view of the map. We compute these by finding the local maxima of the horizontal distance field of the navigable volume. The steps to achieve this are as follows.

2.4.1 Horizontal distance field

For each voxel in $\mathcal{V}$ we compute the horizontal Manhattan distance to the nearest boundary voxel. A boundary voxel is a voxel for which not every voxel in its Von Neumann neighbourhood is occupied. To compute this value we iteratively convolve the voxel grid with a circle-shaped kernel on the X-Z plane, where the radius of the circle is expanded by 1 voxel with each iteration, starting with a radius of 1. When the number of voxel neighbours within the kernel is less than the number of voxels in the kernel a boundary voxel has been reached. Thus, the number of radius expansions tells us the Manhattan distance to the boundary of a particular voxel. We denote the horizontal distance of a voxel to its boundary as $dist : \mathbb{Z}^{3+} \mapsto \mathbb{Z}$. Computing the horizontal distance for every voxel in $\mathcal{V}$ gives us the horizontal distance field (HDF), such that $HDF = \{dist(\boldsymbol{v}) \mid \boldsymbol{v} \in \mathcal{V}\}$. We then find the maxima of the horizontal distance field within a given radius $r \in \mathbb{R}$. The local maxima of the horizontal distance field are all voxels that have a larger or equal horizontal distance than all voxels within r, such that $HDF_{max} = \{\boldsymbol{v} \mid dist(\boldsymbol{v}) \geq max\{dist(\boldsymbol{v_r} \mid \boldsymbol{v_r} \in radius(\boldsymbol{v},\ r))\}\}$. Increasing the value of r reduces the number of local maxima and vice versa. All voxels in HDF_{max} lie within the geometry of the environment, which means the view of the environment is blocked by the surrounding voxels. To solve this, we take the centroids of the voxels in HDF_{max} and translate them upwards to a reasonable scanning height h, to estimate the positions with the optimal view of the map. We denote these positions as $views = \{\boldsymbol{v_c} + (0, h, 0) \mid \boldsymbol{v} \in HDF_{max}\}$

2.5 Visibility

The next step in the room segmentation algorithm is to compute the visibility from each position in $views$. We denote the all voxels that are visible from a given position as $visibility : \mathbb{R},\ \mathbb{Z}^{n\times 3} \mapsto \mathbb{Z}^{m\times 3},\ m \in \mathbb{R},\ n \geq m$. A target voxel is visible from a position if a ray cast from the position towards the centroid of the voxel does not intersect with any other voxel. To compute this we use the digital differential analyzer (DDA) algorithm to rasterize the ray onto the voxel grid in 3D. We then check if any of the voxels that the ray enters- except the target voxel- is occupied. If none are, the target voxel is visible from the point. We perform this raycasting operation from every position in $views$ towards every voxel within a radius $r_{visibility}$ of that position. Only taking into account voxels within a radius speeds up the visibility computation, and is justifiable based on the fact that real-world 3D scanners have limited range. We denote the set of visibilities from each point in views as $visibility_{views} = \{visibility(\boldsymbol{x}) \mid \boldsymbol{x} \in views\}$.

2.6 Room segmentation

After computing the set of visibilities from the estimated optimal views we apply clustering to group the visibilities by similarity. This is based on the definition of a room as a region of similar visibility. Remember that each visibility is a subset of the voxel grid map. To compute the similarity of two sets we use the Jaccard index, which is given by $J(A, B) = \frac{|A\cap B|}{|A\cup B|}$. Computing the Jaccard index for every combination of visibilities gives us a similarity matrix $S^{n\times n} \in [0, 1]$. The similarity matrix is symmetric because $J(A, B) = J(B, A)$. Its diagonals are 1, as $J(A, A) = 1$. We can also consider $S^{n\times n}$ as an undirected weighted graph $\mathcal{G}_S$, where every node represents a visibility and the edges the Jaccard index of two visibilities. This means we can treat visibility clustering as a weighted graph clustering problem. To solve this problem we used the Markov Cluster (MCL) algorithm (CITE MCL), which has been shown by previous research to be state-of-the-art for visibility clustering.

The main parameter of the MCL algorithm is inflation. By varying this parameter between an approximate range of $[1.2, 2.5]$ we get different clustering results.

We denote the clustering of $visibility_{views}$ that results from the MCL algorithm as $\mathbf{C}_{visibility} = \{c_0, c_1, \ldots c_{n-1}, c_n\},\ n = |views|,\ c \in \mathbb{Z},\ n \geq c$, where nth element of $\mathbf{C}_{visibility}$ is the cluster that the nth element of $visibility_{views}$ belongs to, such that for a given value of c, the elements in $visibility_{views}$ for which the corresponding c in $\mathbf{C}_{visibility}$ has the same value belong to the same cluster. As each visibility is a subset of the map, each cluster of visibilities is also a subset of the map. We denote the union of the visibilities belonging to each cluster as $\mathcal{V}_c$.

It is possible for visibility clusters in $\mathcal{V}_c$ to have overlapping voxels. This means that each voxel in the partial map may have multiple associated visibility clusters. However, the goal is to assign a single room to each voxel in the map. To solve this we assign to each voxel the cluster in which the most visibilities contain that voxel. The result is a mapping from voxels to visibility clusters (which we will from now on refer to as rooms), which we denote as $room : \boldsymbol{v} \mapsto \mathbb{Z}$, such that $room(\boldsymbol{v}) = c,\ c \in \mathbf{C}_{visibility},\ \boldsymbol{v} \in \mathcal{V}$. This often results in noisy results, with small, disconnected islands of rooms surrounded by other rooms. Intuitively, this does not correspond to a reasonable room segmentation. To solve this, we apply a label propagation algorithm. This means that for every voxel we find the voxels within a neighbourhood as defined by a convolution kernel. We then assign to the voxel the most common label, in this case the room, of its neighbourhood, if that label is more common than the current label. We iteratively apply this step until the assigned labels stop changing. Depending on the size of the convolution kernel the results are smoothed and small islands are absorbed by the surrounding rooms.

2.7 Topometric map extraction

The above steps segment the map into multiple non-overlapping rooms based on visibility clustering. In the next step we transform the map into a topometric representation $\mathcal{T} = (, \mathcal{V})$, which consists of a topological graph $= (V, E)$ and a voxel grid map $\mathcal{V}$. Each node in represents a room, and also has an associated voxel grid which is a subset of $\mathcal{V}$ and represents the geometry of that room. Edges in represent navigability between rooms, meaning that there is a path between them on the navigable volume that does not pass through any other rooms. This means that for two rooms to have a navigable relationship they need to have adjacent voxels that are both in the navigable volume. To construct the topometric map we thus add a node for every room in the segmented map with its associated voxels, we then add edges between every pair of nodes that satisfy the above navigability requirement.

3 Map Matching

The process of identifying overlapping areas between partial maps is called map matching. In the case of topometric map matching, this refers to identifying which nodes represent the same rooms between two partial maps. We denote our two partial topometric maps as $\mathcal{T}_A = (\mathcal{G}_A,\ \mathcal{V}_A)$ and $\mathcal{T}_B = (\mathcal{G}_B,\ \mathcal{V}_B)$. The goal of map matching is to find a mapping $match : v_A \mapsto v_B,\ v_A \in \mathcal{G}_A,\ v_B \in \mathcal{G}_B$ which corresponds to the real world and is robust to differences in coordinate system, resolution and quality between partial maps. To identify matches between nodes we need to be able to compute the similarity between them. To do so, we must first transform each node into a feature vector which represents both the node itself and its relationship to its neighbourhood. The feature vectors of two nodes with similar geometry and a similar neighbourhood should be close to eachother, meaning their Minkowski distance is small. Conversely, the feature vectors of two dissimilar nodes should be far away from eachother. The first step of this process, encoding the node's geometry into a feature vector, is called geometrical feature embedding. The second step, encoding both the geometrical feature embedding of the node itself and of its neighbourhood into a new feature vector is called attributed node embedding. We hypothesize that the attributed node embedding will have better performance for map matching, especially when differences between partial maps are large, because it involves not just the node itself but also its neighbourhood in the similarity measure. This can be compared to human place recognition, where places are identified not just by their appearance but also by their relationship to their context. In this section we will discuss multiple algorithms used for geometrical feature embedding and attributed node embedding. We will also discuss how we identify matches between nodes based on their feature vectors.

3.1 Geometrical Feature Embedding

Geometrical feature embedding means transforming a geometric object into a feature vector $f \in \mathbb{R}^m$, where m is the dimensionality of the vector. We denote the function that embeds a set of voxels into a feature vector as $embed_{geometry} : \mathbb{Z}^{n \times 3} \mapsto \mathbb{R}^m$. We implement this function using three different approaches, which we discuss below.

3.1.1 Engineered Features

The first approach uses a number of manually engineered features to construct the feature vector from a room's geometry. These features include, for example, the height of the room and its volume. A full list of features and their explanation is given below. More features were tried, but only the ones that were found to contribute to the accuracy of the clustering by trial and error are included here. The features are computed using a point cloud derived from centroids of the occupied voxels of the voxel grid, which we will denote here as $\mathbf{P}$.

Volume The axis aligned bounding box (aabb) of $\mathbf{P}$ is given by the minimum and maximum value along each axis. This results in two 3-dimensional vectors $aabb_{min}$ and $aabb_{max}$. With these vectors we compute the length of each axis of the aabb by computing $\mathbf{l} = aabb_{max} - aabb_{min}$. We then find the volume of the aabb by finding the product of each element of $\mathbf{l}$.

Height Using the same approach as described above we compute the length of each axis of the aabb, $\mathbf{l}$. The height of the point cloud is simply the y-value of $\mathbf{l}$.

Horizontal Area Once again we first compute $\mathbf{l}$. To find the horizontal area of $\mathbf{P}$ we multiply the x- and y-values of $\mathbf{l}$.

Mean distance to centroid To compute the centroid $\mathbf{c}$ of $\mathbf{P}$ we compute the mean value of each axis of all points in $\mathbf{P}$. We then compute the Euclidean distance of each point in $\mathbf{P}$ to $\mathbf{c}$ and compute the mean distance. This metric is closely correlated with an object's volume.

Number of points To compute this value we simply count the number of points in the point cloud. Larger objects will generally contain more points.

Quotient of eigenvalues By using principal component analysis we can determine the 3 eigenvectors and eigenvalues of $\mathbf{P}$, which indicate the directions of maximal variance in the point cloud and the amount of variance along those directions. If all eigenvalues are approximately equal then no direction dominates. We compute if this is the case by finding the quotient of eigenvalues (the first eigenvalue divided by the second and third). If the quotient is close to 1 then no direction dominates.

Ratio of smallest eigenvalue to sum of two largest eigenvalues We take the two largest eigenvalues and find their sum, then we divide the smallest eigenvalue by it. Objects for which the largest two eigenvalues are much larger than the smallest eigenvector have a single direction that is non-dominant.

Verticality of largest eigenvector We take the largest eigenvector, normalize it, and then find the dot product of the eigenvector and the unit vector in the z-direction. This gives us the degree to which the point cloud is vertically aligned. If the largest eigenvalue is non-vertical then the object is mostly horizontal, which is the case for fences, buildings and cars. If the largest eigenvalue is vertical then the object is vertical, which is the case for poles and trees.

Roughness For each point we find their n nearest neighbours. We then fit a plane through the point's neighbourhood, we do this by finding the eigenvectors of the neighbourhood. The smallest eigenvector gives us the normal vector of the neighbourhood, which along with the neighbourhood's centroid gives us the best fit plane. We then determine the sum distance of each point in the neighbourhood in the plane. The roughness of the point cloud is then given by the mean of the sum distance of each point's neighbourhood to its best fit plane.

3.1.2 ShapeDNA

3.1.3 Deep Learning

Another approach to geometrical feature embedding uses deep learning. This works by using a pretrained model, used for segmentation of objects in indoor environments for example, and using the output of the last hidden layer as the feature vector. We use two different network architectures and models to achieve this, PointNet and DGCNN, which we will describe below.

PointNet

DGCNN

3.2 Attributed Node Embedding

Attributed node embedding aims to find a feature embedding for each node in a graph that uses both an attribute of the node, in our case a geometrical feature embedding, and the node's relationship to the rest of the graph. We denote the function that embeds a node's attribute f_{attr} and its graph into a feature vector as $embed_{node} : \mathbb{R}^m,\ \mathcal{G} \mapsto \mathbb{R}^m$, such that $embed_{node}(f_{attr},\ \mathcal{G}_{attr}) = f_{node}$. Finding the attributed node embedding of a node n in the topometric map $\mathcal{T}$ with topological graph $\mathcal{G}_T$ is then equal to computing $f_{node} = embed_{node}(embed_{geometry}(n),\ \mathcal{G}_T)$.

We use a number of different algorithms to solve this problem in our research, which we will describe below.

SINE

MUSAE

FEATHER-N

3.3 Feature Matching

The above steps are applied to both partial maps. This gives us two sets of feature vectors $\mathcal{E}_A,\ \mathcal{E}_B$ that represent the partial maps' attributed node embedding. Our goal is to find a injection between the elements of both sets. To do so, we first find the Cartesian product $\mathcal{E}_{AB} = \mathcal{E}_A \times \mathcal{E}_B = \{(a, b) \mid a \in \mathcal{E}_A,\ b \in \mathcal{E}_B\}$. We then compute the Euclidean distance between every element in $\mathcal{E}_{AB}$, such that $\mathcal{D}_{AB} = \{\sqrt{(a-b)^2} \mid (a, b) \in \mathcal{E}_{AB}\}$. $\mathcal{D}_{AB}$ describes the pairwise distance between each combination of nodes in the partial maps. To extract an injection between the two sets of nodes from this we use the following algorithm: