# Structure-based Vision-Laser Matching

Abel Gawel*, Titus Cieslewski*†, Renaud Dubé*,
Mike Bosse*, Roland Siegwart* and Juan Nieto*
*Autonomous Systems Lab, ETH Zurich, †Robotics & Perception Group, University of Zurich

***Abstract*— Persistent merging of maps created by different sensor modalities is an insufficiently addressed problem. Current approaches either rely on appearance-based features which may suffer from lighting and viewpoint changes or require pre-registration between all sensor modalities used. This work presents a framework using structural descriptors for matching LIDAR point-cloud maps and sparse vision keypoint maps. The matching algorithm works independently of the sensors' viewpoint and varying lighting and does not require pre-registration between the sensors used. Furthermore, we employ the approach in a novel vision-laser map-merging algorithm. We analyse a range of structural descriptors and present results of the method integrated within a full mapping framework. Despite the fact that we match between the visual and laser domains, we can successfully perform map-merging using structural descriptors. The effectiveness of the presented structure-based vision-laser matching is evaluated on the public KITTI dataset and furthermore demonstrated on a map merging problem in an industrial site.**

## I. Introduction and Related Work

In multi-robot applications, heterogeneous teams of robots can be deployed in the same environment in order to exploit the complementary advantages of different platform characteristics. For instance, lightweight unmanned aerial vehicles (UAVs) equipped with cameras can quickly reconnoiter an unknown terrain, while unmanned ground vehicles (UGVs) can be used to carry heavy payloads. A typical application of such heterogeneous setups is in search and rescue scenarios, which involve both rapid exploration and presence on the ground [1], [2].

An important part of multi-robot applications is relative localization. For example, if UAVs are used for reconaissance to be exploited by UGVs, the UGVs need to be able to localize themselves within the maps created by the UAVs. How this can be achieved depends on the sensors available on the mapping robot $R_M$ and the sensors available on the robot $R_L$ that is localizing against the map. On one hand, if both $R_M$ and $R_L$ are equipped with cameras, visual place recongition approaches such as presented in [3], [4], [5] can be used. On the other hand, if both $R_M$ and $R_L$ are equipped with a LIDAR sensor, place recognition approaches for dense maps such as [6], [7] can be used. Furthermore, if $R_M$ is equipped with both cameras and a LIDAR sensor, multi-modal maps, such as presented in [8], [9], [10] can be created, in which case either approach can be used, depending on the sensors present on $R_L$.

This paper focuses on the remaining case, in which either the mapping robot is equipped with a camera and the localizing robot with a LIDAR sensor, or vice versa. For example, a UAV would be equipped with a camera due to weight constraints, while a UGV would be equipped with a LIDAR sensor which provides a more robust and accurate model of the ground, which is required for its path planning. Previous approaches for achieving localization in such circumstances typically consist of either simulating visual data in laser scans [11] or extracting the dense structure that is characteristic for LIDAR maps from visual sensors [12]. The former approach requires the ability to predict surface reflectivity from camera images. While this has been shown to work in an urban environment, in particular on streets with lane markings, this approach might not generalize to cluttered environments present in search and rescue scenarios. In contrast, the latter approach has shown to be more general. Extracting 3D structure from camera images can for instance be achieved using patch-based multi-view stereo algorithms (PMVS) [13], structure from motion [14] or the commercial Pix4D mapper software. However, these 3D reconstruction techniques are very expensive computationally and often have trouble with untextured surfaces. Furthermore, transmitting dense maps over networks typically requires more bandwidth than the transmission of sparse visual maps.

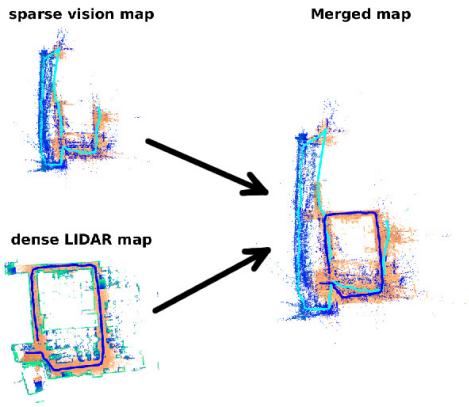


Fig. 1: The figure illustrates two different maps obtained with vision and lidar and the final alignment resulted with our approach.

Therefore one of our priorities is to use sparse visual data.

It was recently shown that that sparse vision maps and dense LIDAR maps can be aligned based on geometry with a good initial guess [15]. In contrast, we show that sparse visual keypoint locations and LIDAR maps contain sufficient

mutual structural information and can be matched using structural descriptors without prior registration.

We propose a framework for structure-based vision-laser matching and evaluate it on a suite of structural descriptors operating solely on 3D structural data, abstracting the neighbourhood around a keypoint location:

- the 3D Gestalt descriptor [7] stemming from the LIDAR mapping community, as a representative of descriptors that performs well on dense LIDAR data,
- the neighbour-binary landmark density (NBLD) descriptor [16] designed for visual feature tracking and performing well on sparse visual keypoint maps and,
- the Boxli descriptor, a generic 3D occupancy descriptor, which essentially downscales the point-cloud resolution.

Furthermore, keypoints are gravity aligned, i.e., IMU measurements are used to estimate the z-axis of the keypoints. This step drastically increases efficiency in searching descriptor matches. We assume limited local errors, i.e., a drift on the open loop solution of up to 5% for odometry measurements and a maximum error for the IMU-based z-axis estimation of 5%.

State-of-the-art methods are employed to efficiently yield high matching results at good computational performance, including descriptor projection and dynamic place segmentation. For efficient feature matching we downscale descriptor dimensionality by employing descriptor projection as presented in [17]. Specifically, this paper presents the following contributions:

- The presentation of an algorithm using structural descriptors to merge sparse vision and dense laser maps.
- A comparison of different structural descriptors w.r.t. vision-laser matching.
- Discussion of the parametrization of the proposed map merging pipeline.
- Evaluation of the approach on the public KITTI dataset [18], and demonstration on an industrial indoor dataset, as depicted in Fig. 1.

## II. Approach

This section describes the steps of our approach starting with data acquisition, filtering and registration, structural description, projection and matching, to place recognition and verification. An overview of the approach is depicted in Fig. 2. The approach requires no prior registration between the used vision and laser data. The structural descriptors however use IMU measurements to estimate gravity alignment of the z-vectors for their orientations' reference frames.

### A. System input and registration

The proposed system has two separate inputs from the LIDAR and the vision pipeline, i.e., depth and IMU measurements on one side, camera image stream and IMU measurements on the other side.

Keypoint maps $\boldsymbol{M}_V$ from vision- and IMU-only data are obtained using an efficient visual inertial odometry algorithm. Our approach is experimentally evaluated using ORB_SLAM2 [5] but is not limited to this choice. This process yields a loop-closed keypoint map, consisting of optimized camera tracks and sparse visual keypoints.

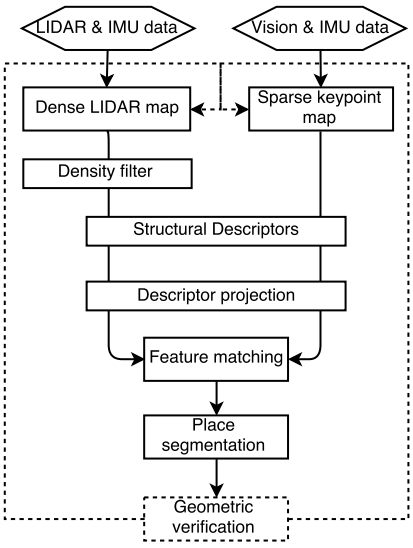


Fig. 2: System diagram of the approach for matching and fusing vision-laser maps using structural descriptors.

Several options are available to build maps from LIDAR data. The registration was tested using scan-registration with ICP and a pose-graph based continuous-time framework which both yield sufficiently accurate maps. Since the vision keypoint maps considered are sparse, the resulting LIDAR map can be density filtered to the average density of the vision base map $\boldsymbol{M}_V$, resulting in the laser base map $\boldsymbol{M}_L$.

Optionally, the quality of the maps can be further improved by performing batch post-processing e.g., loop-closure, which is not a prerequisite for the vision-laser matching of Sec. II-D.

### B. Structural descriptors

A structural descriptor $\boldsymbol{d}_i \in \mathbb{R}^n$ is an abstracted description of the surrounding structure in the neighbourhood $\boldsymbol{\Omega}_i \subset \boldsymbol{M}$ of a keypoint $\boldsymbol{p}_i$.

The $n \times m$ matrix of all descriptors is $\boldsymbol{\Sigma} = \{\boldsymbol{d}_0, \boldsymbol{d}_1, ..., \boldsymbol{d}_{m-1}\}$. We compare three structural descriptor schemes on both $\boldsymbol{M}_L$ and $\boldsymbol{M}_V$, i.e., 3D Gestalt [7], Boxli and Neighbour-binary landmark density (NBLD) descriptors [16]. We select key points $\boldsymbol{P}_L$ and $\boldsymbol{P}_V$ by choosing a random 10% subset of the point clouds $\boldsymbol{M}_L$ and $\boldsymbol{M}_V$, i.e., $\boldsymbol{P}_L \subset \boldsymbol{M}_L$ and $\boldsymbol{P}_V \subset \boldsymbol{M}_V$ respectively. For achieving rotation invariance of the used descriptors, the descriptors are aligned relative to their neighbourhood $\boldsymbol{\Omega}_i$. Therefore the method by [17] is employed to normalize orientations.

The z-axis of each orientation's reference frame is aligned with the gravity vector $\boldsymbol{g}$ which is estimated by an IMU, leaving one rotational degree of freedom in yaw. Let $\boldsymbol{\lambda}_{i,j}$ be the sorted eigenvalues in ascending order of the covariance matrix $\boldsymbol{C_i}$ for each $\boldsymbol{\Omega_i}$. The eigenvector $\boldsymbol{e}_{i,0}$ corresponding to the smallest eigenvalue $\lambda_{i,0}$ is derived, i.e., the surface normal. Subsequently $\boldsymbol{e}_{i,0}$ is projected on the ground plane forming the local x-axis for the descriptor. As this operation

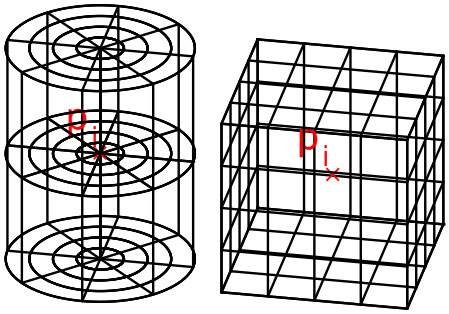


Fig. 3: Left: Cylindrical binning shape used for NBLD and Gestalt descriptors. Right: Cuboid 3D grid used for Boxli descriptor.

has 2 possible solutions, we force the x-component of $\boldsymbol{e}_{i,0}$ to point towards the observer of $\boldsymbol{p}_i$.

In some cases the calculated orientations may be unstable due to small noise drastically changing their values:

1) If the shape of $\boldsymbol{C}_i$ is very cylindrical, i.e., the smallest eigenvalues are of similar size $\frac{\lambda_{i,1}-\lambda_{i,0}}{\sum_{j=0}^{2}\lambda_{i,j}} < 0.9$ the orientation may be any vector within a plane spanned by $\boldsymbol{e}_{i,0}$ and $\boldsymbol{e}_{i,1}$
2) If the surface normal is almost parallel to the z-axis, i.e., $arccos(\boldsymbol{e}_{i,0} \cdot (0,0,1)^T) < 10^\circ$ $\boldsymbol{e}_{i,0}$.

Descriptors with orientations fulfilling any of the two criteria are discarded.

Furthermore, the neighbourhood $\boldsymbol{\Omega}$ has a temporal extent $r_t$, meaning that only points within the neighbourhood which were acquired close in time are considered in the calculation, i.e., we prevent re-observations of places to be accounted for in the same descriptor. In the following we define the three structural descriptor schemes evaluated in this work.

*a) Boxli:* Boxli descriptors define a cuboid 3D grid of expansion $r$ and height $h$ around each keypoint into $n_{dim}^3$ cells as illustrated in Fig. 3:

$$\boldsymbol{\Omega}_i = \boldsymbol{p}_i + \{\boldsymbol{x} \in \mathbb{R}^3 | x_1, x_2 \in [-r, r], x_3 \in [-\frac{h}{2}, \frac{h}{2}]\} \subset \mathbb{R}^3 \quad (1)$$

Each grid cell holds the count $n_{k,l,m}$ of encapsulated points. The Boxli descriptor is then:

$$\boldsymbol{d}_i = (n_{1,1,1}, n_{1,1,2}, n_{1,1,3}, ...) \quad (2)$$

*b) 3D Gestalt:* The 3D Gestalt has proven to be a very successful descriptor for laser-based place recognition [7]. It defines a cylinder of radius $r$ and infinite height around each keypoint $p_i$:

$$\boldsymbol{\Omega}_i = \boldsymbol{p}_i + \{\boldsymbol{x} \in \mathbb{R}^3 | x_3 \in (-\infty, \infty), \|x_1, x_2\| \leq r\} \subset \mathbb{R}^3 \quad (3)$$

Each cylinder is evenly divided into $n_{ra}$ radial and $n_{az}$ azimuthal bins, where each contains average height $\mu_{az,r}$ and variance $\sigma_{az,r}^2$ of the enclosed points, as illustrated in Fig. 3. Additionally, the overall planarity $pl_i$ and cylindricality $cy_i$ values are computed per cylinder from the eigenvalues, yielding the complete descriptor $d_i$.

$$pl_i = \frac{2\lambda_{i,2} - 2\lambda_{i,1}}{\sum_{j=0}^{2}\lambda_{i,j}}, cy_i = \frac{\lambda_{i,1} - \lambda_{i,0}}{\sum_{j=0}^{2}\lambda_{i,j}} \quad (4)$$

$$\boldsymbol{d}_i = (\mu_{1,1}, \sigma_{1,1}^2, \mu_{1,2}, \sigma_{1,2}^2, ..., pl_i, cy_i) \quad (5)$$

The values $pl_i$ and $cy_i$ are a measure for the planarity and cylindricality respectively of a neighbourhood. The values range from 0 to 1.

*c) NBLD:* Neighbour-binary landmark density (NBLD) is a structural descriptor designed for sparse visual keypoints and uses binary comparisons between bins. It has shown to be very successful in place recognition using the 3D locations of sparse visual keypoints, also expressing high performance on data undergoing appearance changes [16]. Like the Gestalt descriptor, the NBLD operates on a virtual cylinder around each keypoint $p_i$. The cylinder is however bounded in z:

$$\boldsymbol{\Omega}_i = \boldsymbol{p}_i + \{\boldsymbol{x} \in \mathbb{R}^3 | \|x_1, x_2\| \leq r, x_3 \in [-\frac{h}{2}, \frac{h}{2}]\} \subset \mathbb{R}^3 \quad (6)$$

Additionally to $n_{az}$ azimuthal and $n_{ra}$ radial bins, NBLD also has a regular binning $n_h$ in z. The descriptors' binary values encode the density differences between bins, i.e., each bin density $\rho_{az,r,z}$ is compared to all densities of its neighbouring bins $\rho_{neighbours}$ and the according feature dimension assigned 1 or 0:

$$bd_{az,r,z} = \begin{cases} 1 & \rho_{az,r,z} > \rho_{neighbours} \\ 0 & \rho_{az,r,z} \leq \rho_{neighbours} \end{cases} \quad (7)$$

$$d_i = (bd_{1,1,1}, bd_{1,1,2}, bd_{1,1,3}, ...) \quad (8)$$

### C. Descriptor projection

In practice, not every dimension of a generic structural descriptor design carries useful information and therefore more dimensions in descriptors do not necessarily increase the separability of matches and non-matches. However, the time required for matching features, which is performed in the next step, increases quadratically with descriptor dimensions. Using only a combination of most expressive descriptor dimensions is desirable. We therefore find a projection $\boldsymbol{A}_b$ of the descriptors to a lower dimensionality $b$ that improves separability, which is a variant of the procedure proposed in [17]:

$$\boldsymbol{\Sigma}_b = \boldsymbol{A}_b \cdot \boldsymbol{\Sigma} \quad (9)$$

with $\boldsymbol{A}_b \in \mathbb{R}^{b \times n} \subset \boldsymbol{A} \in \mathbb{R}^{n \times n}$. For calculating $\boldsymbol{A}$, artificial noise is added to the keypoints while a matching between the keypoint and the noisy keypoint is postulated. Then the distributions of matched and unmatched descriptors are calculated. The best SNR lies in the descriptor dimension with largest eigenvalues for the distribution of differences, i.e., between the covariances $\boldsymbol{C}_M$ and $\boldsymbol{C}_U$ for matched and unmatched descriptors respectively.

$$\boldsymbol{A}^T \boldsymbol{A} = \boldsymbol{C}_M^{-1} - \boldsymbol{C}_U^{-1} \quad (10)$$

This projection training can be performed in 3 different ways:

1) train on laser only,
2) train on vision only,
3) train on both, i.e., postulate matches between laser keypoint and corresponding vision keypoint.

As expected, the third method provides the best results as it models the matching we aim to perform. However,

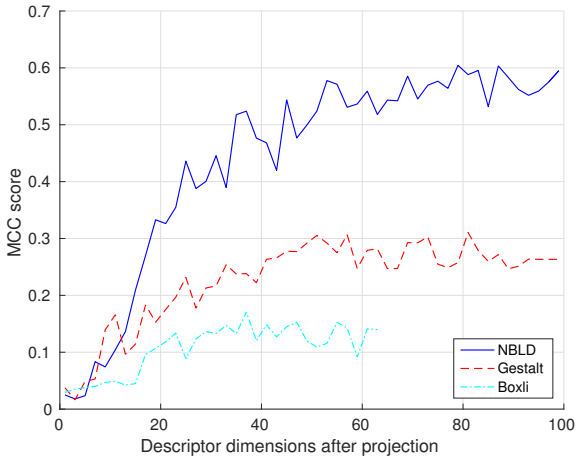


Fig. 4: Comparison of the structural descriptor matching quality over the number of projected descriptor dimensions using Matthews Correlation Coefficient.

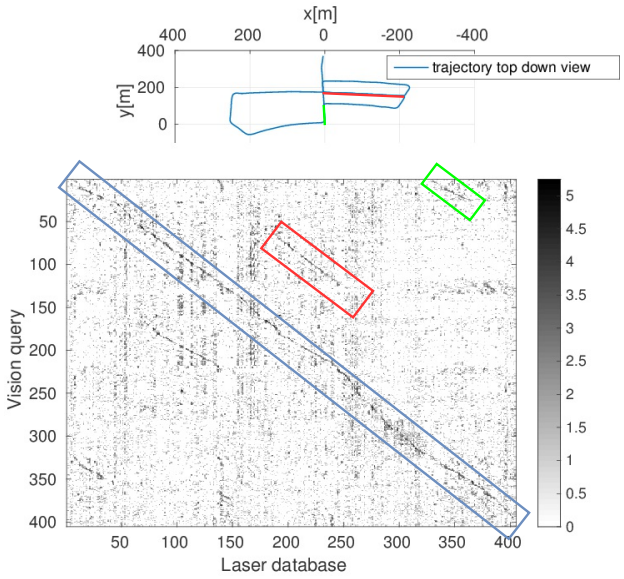


Fig. 5: The vision-laser vote space histogram accumulates all votes from the vision query on the laser database. Areas of high density (black) correspond to the accumulation of many votes. High densities on the main diagonal (highlighted with blue box) correspond to direct place matches, off diagonals (red box and green box) correspond to revisits of places. The corresponding trajectory segments of the evaluation dataset are illustrated on the top.

we want to avoid pre-training to be adaptive to unknown environments and therefore assume no prior knowledge about the registration between LIDAR and vision maps. Our design choice is therefore to train the descriptors on either vision or laser. Intuitively, vision maps in general contain higher noise and training on these may yield the identification of the most robust dimensions. However, our experiments do not show one of the first two methods being superior over the other. In our experiments, we use the vision trained projection matrix $\boldsymbol{A}_V$ for projecting the descriptors $\boldsymbol{\Sigma}$ to a lower dimensionality, choosing the $b$ best dimensions of $\boldsymbol{A}_V$. Fig. 4 shows an exemplary evaluation of different descriptor dimensions projections. Matching quality saturates at much lower descriptor dimensionality than using all descriptors dimensions. Since $\boldsymbol{A}_V$ is trained on real data, it needs to be retrained for different environments. However, as we train $\boldsymbol{A}_V$ uni-modal, i.e. either on vision or on laser data, $\boldsymbol{A}_V$ can adapt to new environments without supervision.

### D. Descriptor matching

A kd-tree of the laser dataset's descriptors $\boldsymbol{\Sigma}_{b,L}$ is built and queried for each keypoints' $k$ nearest neighbours of the vision dataset $\boldsymbol{\Sigma}_{b,V}$. The results are vote scores $\boldsymbol{Z}_{L,V}$ between the datasets. Fig. 5 illustrates a histogram over votes between vision and laser. For creating the image, the same path is followed both by a LIDAR and a camera system, resulting in the expected dominant main diagonal of 1-to-1 correspondences of places and off-diagonals, which correspond to place-revisits, on the vote space. Please note that we chose to evaluate our approach on a dataset where both sensors visit the same area to yield many possible place matches. In the envisioned use-case of map merging, a vote space of distinct vote clusters and not continuous diagonals of clustered votes are expected.

The voting space is further aggregated into places to enable analysis and improvements with place density thresholding.

### E. Place segmentation and geometric verification

Places are commonly defined by spatial regions or keyframes that are queried in a database of descriptors. One common problem in segmenting places is the choice of a feasible place size to aggregate the votes within that region. Popular approaches are fixed-sized grids in the time domain [7] or keyframe queries [19]. We found that dynamic place segmentation, as proposed in [4] yields higher place recognition quality than fixed-size grids or keyframes. In this approach the place segmentation problem is treated as a continuous 2D probability estimate on the matching matrix, i.e., places are segmented along path segments scoring high vote densities. Rotating the vote space, facilitates the segmentation process, by enabling trajectory-aligned vote-space segmentation.

The full rotated vote space is initialized as a single node in this decomposition. The algorithm then recursively splits the space in x- and y-direction, based on density gradients within a node. The decomposition stops if a maximal decomposition depth is reached or if the maximal density gradient within a node is lower than a threshold. Fig. 6 shows the segmentation of the 20 % densest matching segmentations for the evaluation dataset. We threshold on the density for an individual structural descriptor scheme above which segmented areas are considered as place correspondence candidates. To account for varying robot velocities, the time index of keypoints is converted into the travelled distance.

The approach assumes that maps are recorded following robot trajectories, i.e., both the query and the database data

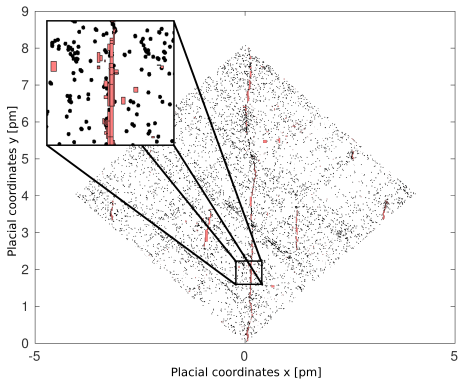


Fig. 6: Place segmentation candidates on rotated vote-space after dynamic place segmentation for KITTI05. The image shows the rotated vote-space of Fig. 5 and the resulting high-density place-matching segments as small boxes overlaid. The 20% densest segments of the evaluation dataset are illustrated as small boxes. Most high density segments accumulate around the (rotated) main diagonal and off diagonals.

were recorded, travelling forward or backward on close continuous trajectories. Since we cannot always assume robots to follow similar trajectories in the context of map-merging, we limit the segmentation size to a low maximal extent.

The place-correspondence candidates which are a collection of descriptor matches $\mathbf{\Sigma}_M$ can finally be checked for a consistent transform from one place to the matching candidate. This is achieved by calculating the individual transforms $\boldsymbol{T}_{i,j}$ of each keypoint match in a segmented region and building a histogram on the transformations. If the likelihood of the dominant transform exceeds a threshold, the place-correspondence is accepted as a match.

Finally, the match constraint is propagated to the vision and laser maps yielding a registration between them.

## III. Experiments

The described procedure has been evaluated on the KITTI dataset [18] using data from stereo cameras and 3D LIDAR data. We use sequence 05 of the dataset consisting of a trajectory of 2.2 km in an urban environment, with several loop closures. The particular sequence was chosen for its richer urban environment, compared to highway sequences and the presence of several loop closures which visually aids understanding the vote histograms.

The vision and LIDAR data were independently converted into maps as described in Sec. II-A. For evaluating the descriptors' performances both vision and laser point-cloud were individually loop-closed and aligned using the GPS ground-truth. The known transformation between stereo camera and 3D LIDAR is only used for the evaluation against ground truth and not in the vision-laser matching system itself.

We furthermore demonstrate the approach in a machine hall, an industrial indoor environment on independently recorded data with partial overlap between vision and LIDAR maps. In a first experiment, a UGV equipped with a 3D laser scanner explored the machine hall. Later, a human carrying a camera explored some parts of the machine hall and the nearby corridors (see Fig. 1).

### A. Evaluation

Our evaluation point is set after place segmentation and before geometric verification, as we regard geometric verification as an additional step, which can be applied to place segmentations independent of the way segmentations were acquired. We use precision, recall and Matthews Correlation Coefficient (MCC) to evaluate the quality of the proposed method and the performance of different structural descriptors on KITTI. A threshold on the vote-density $t_d$ is used to distinguish between positives and negatives. We mark a match as true positive, if the euclidean distance $r_{L,V}$ between the matching keypoints is within the matching radius $r_{gt,near}$, while the density $d_{L,V} > t_d$. False positives are accounted, if $r_{L,V}$ is greater than a threshold $r_{gt,far}$, while $d_{L,V} > t_d$. True negatives fulfill the conditions $r_{L,V} > r_{gt,far}$ and $d_{L,V} < t_d$. Finally false negatives are counted, if $r_{L,V} < r_{gt,far}$ and $d_{L,V} < t_d$.

As we do not have ground truth data for the machine hall dataset, we facilitate this dataset for demonstration purposes of our method. The vision and laser machine hall datasets were manually aligned, and we use the data to identify common true positives and false positives, i.e., in which regions our method performs good or bad.

### B. Parametrization

The method requires several settings:

- the number of point-cloud points to use as keypoints,
- the dimensionality $b$ of the descriptor projection $\boldsymbol{A}_b$,
- the number $k$ of descriptor matches per query descriptor,
- the threshold $t_d$ on the vote density to distinguish positive and negative votes,
- the ground truth radii $r_{gt,near}$ and $r_{gt,far}$ are only needed in the evaluation, to assign the true or false predicate to matching candidates,
- the descriptor settings.

We hand-tuned the descriptors according to the following parametrization choices.

The number of keypoints can be chosen arbitrarily, but should not be too sparse. We used 10% of our point-clouds as keypoints.

As shown in Sec. II-C, the MCC score saturates with a much lower number of descriptor dimensions and therefore most of the descriptor space can be captured in a 50 dimensional subspace.

The number of nearest-neighbours $k$ is set according to the keypoint density within the neighbourhood of a keypoint, i.e. the estimate of self-similar keypoints near an evaluated keypoint, a rule of thumb is to use 5 to 15 matches.

We furthermore set $t_d$ to a value that maximizes the MCC performance of our method, while providing good precision and recall rates. This corresponds to the approximate spacial overlap of the datasets, i.e., the 5% densest regions are selected as positives for full coverage, linearly scaling down with lower coverage.

TABLE I: Parameters used in the evaluation.

| | KITTI | Machine hall |
|---|---|---|
| $\mathbf{\Omega}_d$ | $(r, h, r_t) = (20m, 18m, 50s)$ | $(r, h, r_t) = (6m, 6m, 20s)$ |
| $n_{dim}$ | $(n_{az}, n_{ra}, n_z) = (16, 4, 8)$ $(n_x, n_y, n_z) = (5, 5, 8)$ | $(n_{az}, n_{ra}, n_z) = (16, 4, 6)$ |
| $k$ | 10 | 10 |
| $b$ | 50 | 50 |
| $t_d$ | 5% | 3% |
| $r_{gt}$ | $(near, far) = (20m, 30m)$ | $(near, far) = (6m, 10m)$[1] |

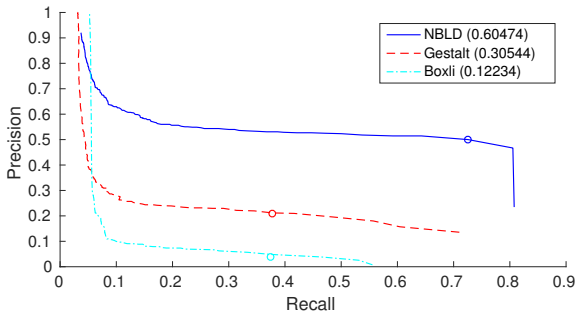


Fig. 7: Precision-recall curves for KITTI05, the parameters were chosen according to Tab. I. The points and values of the maximum MCC scores are illustrated.

The ground truth radii $r_{gt}$ for evaluation should be set equal to or greater than the descriptor radius.

We found the binning of descriptors as depicted in Tab. I to provide good results by trying different settings. The neighbourhood $\mathbf{\Omega}_d$ should be set according to the environment, the sensor characteristic and estimated speed of the robot. The radius $r$ determines the minimal expected overlap between the vision and LIDAR map. However, the choice of too large radii may encode the travelled trajectory rather than the local structure. The height $h$ can be set to values that realistically capture the vertical extent of the measurements and $r_t$ corresponds to the minimal time the robot requires to cover the area around a keypoint.

The parameters for our evaluation are collected in Tab. I. Considering recent advances in robust estimation lower precision rates are commonly acceptable as false positives may be filtered [20], [21].

The overall best choices for feature matching however are the parametrizations scoring highest correlation scores, such as MCC.

### C. Results on KITTI dataset

The precision-recall curves for KITTI05 are illustrated in Fig. 7. These results were obtained by varying the density threshold $t_d$ on the place segmentation.

Both the 3D Gestalt and the Boxli descriptors show poor performance throughout the experiments, whereas NBLD yields good results at moderate precision values for high recall. It turns out that the Gestalt descriptor is susceptible to the two different modalities' distribution characteristics. Gestalt's mean and variance features show to be poorly reproducible between the modalities. Furthermore, does Boxli perform only slightly better than random. Our experiments indicate that the encoding of point counts does not generalize well between modalities. This is mainly due to different sampling characteristics of the sensor modalities' mapping techniques with laser equally sampling surfaces and vision being more densely sampled in well textured areas.

[1]As we do not have external ground truth, these values were used for evaluating the manually aligned dataset.

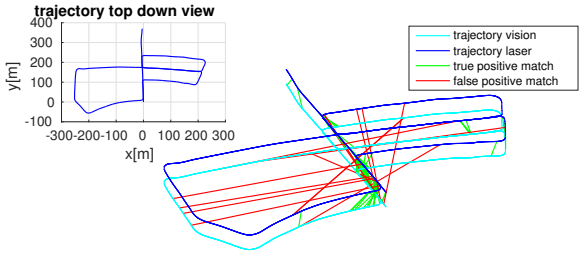


Fig. 8: Overview of the KITTI dataset. The matching is illustrated shown for the 5% densest matches and divided into true positive (green) and false positive (red) matches. For illustration purposes we only plotted 10% of the true positives and false positives respectively.

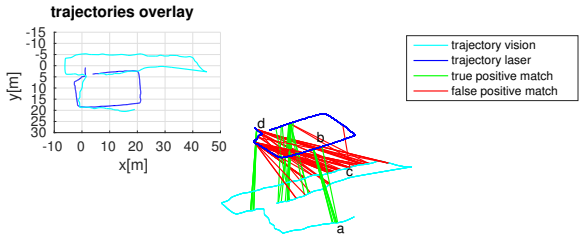


Fig. 9: Overview on the machine hall dataset. The matching is illustrated by the 3% densest matches and divided into true positive (green) and false positive (red) matches. The match a) - b) is an example of a true positive, c) - d) an example of a false positive. a) - d) are further investigated in Fig. 10.

Taking a closer look at the NBLD performance in contrast to the environment, we can identify regions of high and low performance. In Fig. 8 map regions counted as true positives (green) and false positives (red) are highlighted. In KITTI, the matching shows best performance in regions with keypoints covering larger areas, i.e. regions that have points widely distributed over the x-y-plane. Our experiments show that the maximal descriptor radius of NBLD expresses best performance above 15 meters for the KITTI dataset, which is a reasonable choice, meaning that one descriptor can capture the local structure, i.e., street and neighbouring houses on both sides. Since the environment is very homogeneous, we cannot identify a special structural characteristic that has superior performance. The high recall rates indicate that most place-matches are found by the proposed method. Furthermore, consecutive steps can be applied to filter false positives, including geometric verification or techniques of robust estimation.

### D. Results on Machine Hall dataset

We furthermore demonstrate the approach on the machine hall dataset which only has partial overlap between the maps. For this dataset only the NBLD descriptor was used, as it

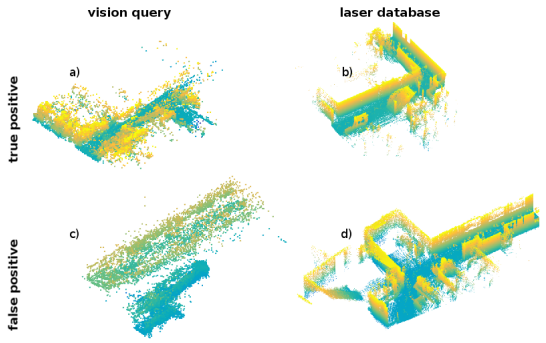


Fig. 10: Example for a true positive match (top row) and a false positive match (bottom row). On the left are the vision queries, on the right side the corresponding laser database matches. The algorithm performs well in distinct structural regions, but fails in self-similar corridors. The indices a) - d) correspond to the matching locations depicted in Fig. 9.

has shown superior performance over the other descriptor candidates. Since the machine hall is an indoor environment, the extent of the descriptors showed good performance on lower extent, i.e., a radius of 6 meters was chosen. In the machine hall dataset, distinct structural corners and transitions between open areas and corridors show best matching performance, whereas the algorithm had difficulties with self-similar corridors. Fig. 9 illustrates the found true (green) and false (red) positives in the machine hall dataset. In Fig. 10 we illustrate a common case for a true positive and a false positive. Furthermore, we discovered a matching trap. Here, the end of a corridor in the laser map received a cluster of votes from several locations of a non overlapping corridor in the vision map, i.e. no segment of this corridor was present in the laser database. This also implies that one location in the laser database was allowed to receive many votes from different locations in the vision map. However, for the presented paper, we do not want to exclude place revisits. We plan to investigate methods within the vote matching and place segmentation step to mitigate this effect in future work.

## IV. Conclusions

Multi-modal matching is a very challenging and still unsolved problem. In this paper, we have presented a variant of vision-laser matching based on structural features and place recognition between those domains. The presented algorithm requires only the point-cloud data of the different sources as inputs, independent from specific visual features or laser intensity values. We have shown an approach that can identify place matching candidates for data without prior registration. These results demonstrate, that a density based descriptor, such as NBLD has great potential for the matching between laser and vision in the application of multi-modal map merging.

For future work we aim to reduce preliminary filtering and extend our evaluation with a variety of datasets. Ultimately, we aim to design adaptive descriptors that express good recognition rates without supervision and therefore adapting to different environments as well as automatic ways for setting parameters.

## V. Acknowledgement

This work was supported by European Union's Seventh Framework Programme for research, technological development and demonstration under the TRADR project No. FP7-ICT-609763.

## References

[1] L. Marconi, C. Melchiorri, M. Beetz, D. Pangercic, R. Siegwart, S. Leutenegger, R. Carloni, S. Stramigioli, H. Bruyninckx, P. Doherty, and Others, "The SHERPA project: Smart collaboration between humans and ground-aerial robots for improving rescuing activities in alpine environments," in *SSRR*, 2012, pp. 1–4.

[2] E. F. Flushing, M. Kudelski, L. M. Gambardella, and G. A. Di Caro, "Connectivity-aware planning of search and rescue missions," in *SSRR*. IEEE, 2013, pp. 1–8.

[3] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.

[4] S. Lynen, M. Bosse, P. Furgale, and R. Siegwart, "Placeless place-recognition," in *3DV*, 2014, pp. 303–310.

[5] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[6] M. Bosse, "Place recognition using regional point descriptors for 3D mapping," *Field and Service Robotics*, pp. 195–204, 2010.

[7] M. Bosse and R. Zlot, "Place recognition using keypoint voting in large 3D lidar datasets," *ICRA*, pp. 2677–2684, 2013.

[8] K. L. Ho and P. Newman, "Loop closure detection in SLAM by combining visual and spatial appearance," *Robotics and Autonomous Systems*, vol. 54, no. 9, pp. 740–749, 2006.

[9] F. Tungadi, W. L. D. Lui, L. Kleeman, and R. Jarvis, "Robust online map merging system using laser scan matching and omnidirectional vision," *IROS*, pp. 7–14, 2010.

[10] G. Pandey, J. R. McBride, S. Savarese, and R. M. Eustice, "Toward mutual information based automatic registration of 3D point clouds," in *IROS*, 2012, pp. 2698–2704.

[11] R. W. Wolcott and R. M. Eustice, "Visual Localization within LIDAR Maps for Automated Urban Driving," in *IROS*, 2014, pp. 176 – 183.

[12] C. Forster, M. Pizzoli, and D. Scaramuzza, "Air-ground localization and map augmentation using monocular dense reconstruction," in *IROS*, 2013, pp. 3971–3978.

[13] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1362–1376, 2010.

[14] C. Tomasi and T. Kanade, "Shape and motion from image streams: a factorization method." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 90, no. 21, pp. 9795–9802, 1993.

[15] T. Caselitz, B. Steder, M. Ruhnke, and W. Burgard, "Matching Geometry for Long-term Monocular Camera Localization," in *ICRA Workshop: AI for long-term Autonomy*, 2016.

[16] T. Cieslewski, E. Stumm, A. Gawel, M. Bosse, S. Lynen, and R. Siegwart, "Point cloud Descriptors for Place Recognition using Sparse Visual Information," in *ICRA*, 2016, pp. 4830–4836.

[17] M. Bosse and R. Zlot, "Keypoint design and evaluation for place recognition in 2D lidar maps," *Robotics and Autonomous Systems*, vol. 57, no. 12, pp. 1211–1224, 2009.

[18] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite," *CVPR*, pp. 3354–3361, 2012.

[19] M. Cummins and P. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *The International Journal of Robotics Research*, vol. 30, no. 9, pp. 1100–1123, 2011.

[20] N. Sünderhauf and P. Protzel, "Towards a Robust Back-End for Pose Graph SLAM," in *ICRA*, 2012, pp. 1254–1261.

[21] Y. Latif, C. Cadena, and J. Neira, "Robust loop closing over time for pose graph SLAM," *The International Journal of Robotics Research*, vol. 32, no. 14, pp. 1611–1626, 2013.