Aligning thermal cameras with LiDAR scenes in subterranean environments

Julian Jandeleit
Universität Konstanz
Konstanz, Germany
julian.jandeleit@uni-konstanz.de

Thejasvi Beleyur
Universität Konstanz
Max-Planck Institute of Animal Behavior
Max-Planck Institute for Biological Intelligence
Konstanz, Radolfzell and Seewiesen, Germany
thejasvib@gmail.com

Bastian Goldlücke
Universität Konstanz
Konstanz, Germany
bastian.goldluecke@uni-konstanz.de

Abstract

Modern animal behaviourists are able to collect vast amounts of data from multiple sensors including cameras. Sensor fusion however still remains a challenge given the unique sensor types biologists use, and especially the unusual settings in which these sensors are deployed. Here, we study thermal camera-LiDAR alignment for thermally uniform, feature-deficient cave scenes. To approach these conditions, we introduce the depth-map correspondence (DMCP) algorithm for user-assisted alignment without explicit calibration objects. We quantify the accuracy of DMCP's alignment by evaluating a set of points known to lie on cave walls and compare it with state of the art methods. DMCP's alignment shows a median of 9 cm error, while other methods show at least order of magnitude higher median errors. DMCP thus sets an important baseline result for this challenging sensor-fusion task.

1. Introduction

Animal behaviorists today are able to collect types of data from multiple sensors such as thermal video and even LiDAR scans of the natural environment. Contextualising the animal's behaviour however requires bringing the multi-sensor data to a common coordinate-system. Animals occupy a wide set of physical environments (aquatic, terrestrial, subterranean), each of which constrain sensor performance. However, state-of-the art methods to handle and align sensor data are typically centred around rgb cameras or human and industry environments (e.g. artificial and urban settings)[8] for thermal cameras. State-of-the-art methods may thus fail when used with unconventional data from natural environments, necessitating the development of new algorithms.

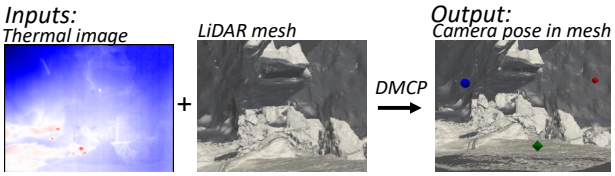


Figure 1. Schematic showing the required raw data inputs and resulting output from the depth map correspondence (DMCP) algorithm. While designed for thermal camera pose estimation, DMCP can work for any type of image-pointcloud or image-mesh sensor data.

1.1. In situ research requires new fusion methods

Many echolocating bat species live in caves, and fly around in groups before emerging to forage. The study of their echolocation and flight patterns is of central interest to the field of active-sensing and collective behaviour [1, 2]. The *Ushichka* dataset [1] is a multi-sensor dataset with multi-channel audio, multi-camera thermal video and LiDAR scans of the Orlova Chuka cave system in Bulgaria. Each of the sensors in *Ushichka* captures an important aspect of the animals' behaviour, but a contextual understanding is achieved only by aligning the three sensors into a common coordinate system.

To understand a bat's flight decisions, its flight path with respect to the cave's walls needs to be known. Aligning

the LiDAR and the camera coordinate systems allows this contextual understanding. Aligning 3D meshes and thermal cameras is typically done with calibration objects in the scene [15, 21]. The *Ushichka* scene however does not contain calibration objects visible to both LiDAR and thermal sensors. We thus need to make use of the naturally available features in the data.

Methods using environmental features developed for visible light camera and LiDAR alignment without calibration objects exist [17, 20]. However, alignment attempts based on feature and photo-consistency based methods fail to provide enough scene points necessary for alignment [11]. A main reason for the failure of these methods on the *Ushichka* dataset is likely the unique nature of the subterranean thermal scenes, which exhibit extremely low contrast across the scene. The cave system maintains a fairly stable temperature of around 10°C all year round, and shows little spatial variation in temperature from one part to another (*pers. obs.*). Also, the thermal scene is very self-similar. Rocks in the scene are difficult to uniquely identify and smooth walls sometimes dominate sizeable portions of the camera view, resulting in features with low descriptive power.

1.2. Overcoming featureless scenes

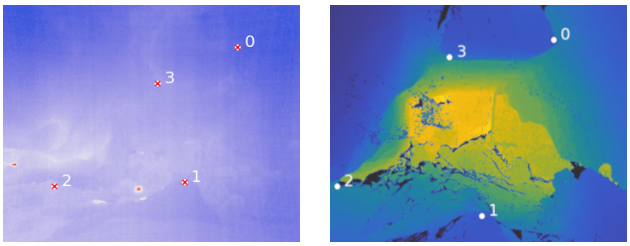
We present the semi-automated Depth Map Correspondence (*DMCP*) algorithm, developed to handle the image-mesh alignment of feature deficient scenes such as those in *Ushichka*'s thermal data (Fig. 1). Additionally, we compare it with the results from state-of-the-art feature detectors provided by the software package *hloc* [23, 24]. *DMCP* consists of the following steps:

1. The user chooses a view of the mesh interactively to approximately match the thermal image. A depth map of the mesh is then generated from the viewpoint. A depth map is image-like and thus eases comparison with the thermal image.
2. The user annotates a minimum of 4 point-to-point correspondences between the thermal image and the depth map (Fig. 2). Each 2D point in the depth map corresponds to a 3D point in the LiDAR scene.
3. The extrinsic pose of the camera in the LiDAR scene is estimated from the user-annotated correspondences.
4. The transformation matrix that converts 3D points from thermal camera world space to LiDAR space is calculated by passing points through camera space and solving the absolute orientation problem.

2. Methodology

2.1. The DMCP algorithm

Terminology Terminology and mathematical background is generally used as defined in [10]. The extrinsic matrix



(a) Thermal image from second camera for session on *2018-08-17*. Crosses and numbers refer to the user annotated points.

(b) Depth map generated for annotation. White points and numbers refer to user annotated points.

Figure 2. Example image and depth map from *Ushichka* experiment. White points indicate annotated pixels. A red *x* indicates the reprojected depth map correspondence.

E of each camera has a corresponding pose matrix C as its inverse. Objects in homogenous coordinates are denoted with a hat (e.g $\hat{E}$).

We call the coordinates the LiDAR mesh is expressed in *lidar world space*. The pivot camera to be aligned is given in its own associated space defined by its extrinsic matrix, we call *thermal space*. The system referred to is denoted by subscript.

Given a thermal image $I_{thermal}$ captured by a camera $P_{thermal} = K_{thermal}E_{thermal}$ in thermal space, the task is now to find a projective transformation M from thermal to lidar space such that the coordinates of a 3D point A in the two systems are related by the affine transformation $\hat{A}_{world} = M\hat{A}_{thermal}$.

2.1.1 Point Annotation

A user first chooses a view point to approximately match the captured thermal image $I_{thermal}$. From the selected view point, we render a depth map I_{dm} of the mesh with [19] according to the chosen projection matrix P_{dm}. The pose $C_{dm} = \begin{bmatrix} R_{dm} & T_{dm} \end{bmatrix}$ of the virtual camera that captured the depth map is saved. The user now annotates at least 4 corresponding points on I_{dm} and $I_{thermal}$. We denote the two corresponding sets of points by cps_{dm} and $cps_{thermal}$, respectively.

Let $a \in cps_{dm}$ be the pixel coordinates of one of the corresponding points on I_{dm}, then $I_{dm}(a)$ is the corresponding depth. The respective point A_{camera} in the camera space of the depth camera can then be computed by $A_{camera} = I_{dm}(a) \cdot K_{dm}^{-1}\hat{a}$, and further be transformed to lidar space with

$$A_{world} = C_{dm}\hat{A}_{camera} = C_{dm}I_{dm}(a) \cdot K_{dm}^{-1}\hat{a}. \quad (1)$$

The transformed set of correspondences in lidar space is denoted by cps_{lidar}. Thus, $cps_{thermal}$ and cps_{lidar} now represent correspondences from thermal camera pixel coordinates to lidar space coordinates. Because correspondences are annotated and originate from a feature deficient scene,

we assume their number to be close to the minimum number of 4 points necessary for the *Sparse Camera Alignment* step. Figure 2 shows example annotations.

2.1.2 Sparse Camera Alignment

Suppose we have a thermal camera defined by its intrinsic camera matrix $K_{thermal}$ and pinhole projection matrix $P_{thermal} = K_{thermal}E_{thermal}$. Additionally, we assume at least 4 corresponding points $cps_{thermal}$ and cps_{lidar} generated in the annotation step. Sparse Camera Alignment is defined in two steps: 1) camera pose estimation and 2) transformation estimation.

Camera Pose Estimation The position of the camera in lidar space is estimated using image to lidar space correspondences. The resulting task is known as the Perspective-n-Point problem [7]. The lowest number of points required to estimate a camera pose is 4 points. To work on only 4 points reliably, we build on an existing solution [14] of the Perspective-Three-Point Problem (P3P) implemented in [5]. P3P requires exactly three correspondences and gives up to 4 distinct valid solutions in the form of extrinsic matrices E_i. The solutions are collected for all combinations of three points in the correspondence set. The fourth point is used to determine the correct solution among the E_i. A point $A_{world} \in cps_{lidar}$ can be reprojected into the image via $a' = K_{thermal}E_i\hat{A}_{world}$. As every solution E_i is generated from 3 points, there is at least one point in cps_{lidar}, that is not used to generate a solution. Thus, the solution that generalizes best can be selected as the solution where the total reprojection error $e_r(i) = \sum_{A_{world} \in cps_{lidar}} ||K_{thermal}E_i\hat{A}_{world} - a||$ is minimized. Here, $a \in cps_{thermal}$ is the point which corresponds to the projection a' of the current point A_{world} for which the term is evaluated.

The final estimated extrinsic matrix and the solution of the camera pose estimation step is therefore

$$E_{lidar} = E_i \text{ with } i = \arg\min_i e_r(i), \tag{2}$$

representing the estimated pose with respect to lidar space.

Transformation Estimation In this step, the transformation that transforms points in thermal space to lidar space is computed from the thermal camera $P_{thermal} = K_{thermal}E_{thermal}$, the estimated lidar space camera $P_{lidar} = K_{thermal}E_{lidar}$ (note that $K_{thermal} = K_{lidar}$ here) and the annotated lidar space points cps_{lidar}.
First, each lidar space point $A_{world} \in cps_{lidar}$ is transformed and moved to thermal space

$$A_{thermal} = C_{thermal}E_{lidar}\hat{A}_{world}. \tag{3}$$

This way, the coordinates of each lidar space point in thermal space are known. The registering transformation can now be defined by solving the absolute orientation problem. Here, Umeyama's method [27] as implemented in [5] is used. The result is the estimated transformation M, which computes

$$A_{world} = MA_{thermal} \text{ and} \tag{4}$$
$$P_{world} = P_{thermal}M^{-1}. \tag{5}$$

Thus, M solves the pose estimation problem for all cameras and points in thermal space. Alternatively, M could also be computed directly using the inversion $M = C_{world}E_{thermal}$. We use Umeyama's method in our implementation since we found it leads to slightly more robust results.

2.2. Ushichka computational experiments

The *Ushichka* dataset [1] consists of multiple nights of multi-sensor data, capturing the flight behaviour of echolocating bats in the same recording volume. On each night, three uncooled thermal cameras were placed in similar locations in the cave across all nights. The DLT coefficients of each camera were used to obtain the intrinsic and extrinsic camera parameters in thermal space. A high-resolution LiDAR scan of the cave was performed on one night ($\leq$ 6 mm resolution) [13]. The generated point cloud was downsampled to a centimetre level mesh in this paper for ease of handling.

2.2.1 Feature matches alignment with ICP

To compare *DMCP* with state of the art, robust transformation estimates are obtained for each night using *DMCP*. Poses from 5 repeated rounds of *DMCP* are averaged. *Ushichka* thermal data was captured as video frames. Images for annotation on each night were generated by using the median value for each pixel along a time axis of 100 frames. The result was further processed by removing horizontal and vertical fixed pattern noise using the fourier transform. A result can be seen in Figure 2a.

We compare alignment from *DMCP* with alignment results from 5 other modern feature descriptors: hloc[23, 24], Disk[26], R2D2[22], SIFT[18], SOSNET[25] and Superpoint[6]. Detected features get matched and matched points projected into 3D using the calibrated camera matrices from the *Ushichka* dataset. Using the DMCP results as initial transformation, the reconstructed structure gets aligned with the LIDAR mesh using ICP[4]. The results are compared to the baseline provided by *DMCP*. That this is a valid strategy will be confirmed in Section 3.1. While small differences in the estimated transformation indicate good performance or even slight improvement over *DMCP*, large deviations indicate wrong estimations.

Method	Median translation (m)	95 %ile interval (m)
DMCP (this paper)	0 (reference)	0 (reference)
Disk [26]	8.63	2.55 - 16.22
R2D2 [22]	10.59	4.43 - 20.49
SIFT [18]	16.04	1.30 - 52.40
SOSNET [25]	6.13	0.64 - 25.84
Superpoint [6]	6.70	0.57 - 1413.56

Table 1. Summary of spatial discrepancy in inferred camera pose between DMCP and other methods using ICP. Here we only report median translation of each ICP transformation.

2.2.2 Microphone and cave point alignment

To verify *DMCP*'s 3D alignment accuracy, we compare the positions of microphones and other cave points after conversion to the LiDAR coordinate system. Many microphones in *Ushichka* were placed directly on the cave surface, and visible on the thermal camera images. Thus, a perfect alignment of microphone and LiDAR mesh means the microphone/cave points will lie perfectly on the mesh surface. Alignment error is quantified by computing the distance to the nearest-mesh point for each transformed microphone and wall point. For comparison, we repeat the evaluation for estimated transforms from Section 2.2.1.

3. Results

3.1. Microphone and cave point alignment

Ideal camera-LiDAR alignment should result in 0 distance between camera-triangulated points and nearest mesh points. We observe a range of alignment distances for DMCP (Figure 3) (median: $0.09m$, 95%ile: $0.007 - 2.5m$, $N = 77$ points). Other methods show much higher alignment errors, proving DMCP's better baseline performance. It is important to caution that a lower nearest-point distance points to a good fit of camera triangulated points with the mesh, and is only a proxy for reliable alignment.

3.2. Feature matches alignment with ICP

Table 1 shows the results of ICP alignment of reconstructed feature matches. Large median translations of 6m and upwards indicate significantly worse performance than *DMCP*, especially as the estimated *DMCP* poses show a median alignment error of $0.09m$ in Section 3.1 and were seen to be correct from visual inspection.

In general, the experiment shows that the *DMCP* framework empowers a user to register calibrated cameras to a mesh, even in difficult scenes. Especially difficult images that do not result in optimal transformations can be estimated by trying again with more effort to find good correspondences.

4. Discussion

Animal behaviourists often collect multi-sensor datasets in unconventional natural settings. To aid in subterranean thermal camera-LiDAR alignment, we developed the DMCP algorithm. DMCP estimates camera pose using user-assisted correspondences between thermal images and LiDAR depth maps. We highlight the success of the method despite the limited thermal camera resolution (640x512 pixels), challenging image conditions (2a), along with the low number (4 points) of correspondences required from the user's side. Having aligned the thermal-camera and LiDAR sensors, we

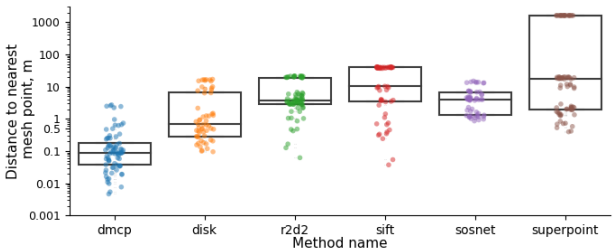


Figure 3. Quantifying alignment error: distance between transformed microphone and cave points and the nearest LiDAR mesh point. Y-axis in log_{10} scale. Box-plot represents 25,50,75 percentiles. DMCP currently outperforms other methods by an order of magnitude.

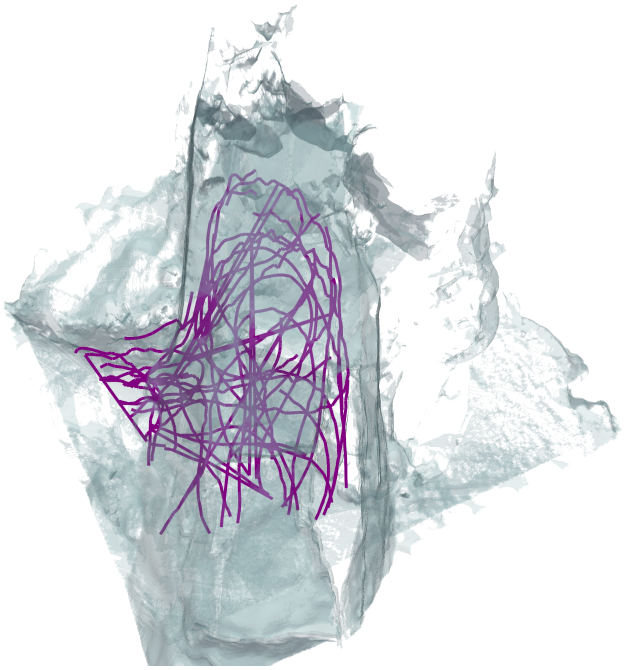


Figure 4. Multiple overlaid bat flight trajectories aligned to the cave LiDAR scan using DMCP (2018-08-17 session). The visualisation shows the view from above, and suggests bats follow walls when flying in caves.

see the power of DMCP with the bat flight trajectories in the cave setting in Figure 4. Preliminary observations already show evidence for wall-following, which would not have been apparent with only flight trajectories in the camera coordinate system.

DMCP is a user-assisted algorithm, poor input correspondences will result in a poor pose estimation. In particular, point annotations made close to edges in the depth map can tip the balance between a good and bad pose estimation. The intrinsic and extrinsic parameters of the experimental cameras are assumed to be known, by calibrating them into a common coordinate system. Errors in camera parameter estimation (depending on the calibration workflow used) will therefore influence DMCP's accuracy as well.

Even though DMCP is developed to solve thermal camera-LiDAR alignment, we stress that it is generalisable to any kind of camera-pointcloud or camera-mesh data. Other potential uses of DMCP are in pose estimation in meshes/point-clouds generated from 3D scanning devices or structure-from-motion type workflows. Thermal-LiDAR data and code used in this paper are available here [3, 12].

5. Conclusion

The current formulation of DMCP estimates transforms from single cameras independently. Future work could generalize the algorithm to multiple cameras. This could be achieved by estimating a consensus transform using point correspondences from all cameras together. The combined approach transforms may result in a more robust transform estimate by removing outliers that do not satisfy the relative camera pose constraint. Automation of two portions of the workflow require further development: 1) generation of corresponding points, and 2) generation of depth maps. To automatically find corresponding points, future work could look into feature detectors optimised from thermal images, such as phase-congruency [9, 11, 16], or machine learning approaches for lidar-depth-map feature descriptors. Instead of a single depth map used to obtain corresponding points, an automated method could sample multiple views of the mesh with varying numbers of corresponding points from each depth map annotated by a thermal-to-depth feature descriptor. The corresponding points across depth maps could then be pooled to obtain a robust pose estimate. We share the code of *DMCP* and relevant raw data to encourage further research in these directions.

We found it surprising that there does not seem to be an automatic solution to successfully align a set of calibrated thermal cameras to a mesh. Finding correspondences between thermal images to reliably obtain 3D points on the mesh is extremely challenging in scenes with low temperature variation. DMCP sets an important baseline for future work on a fully automated sensor-fusion pipeline in feature-deficient scenes.

References

[1] Thejasvi Beleyur. *Theoretical and empirical investigations of echolocation in bat groups*. PhD thesis, Universität Konstanz, Konstanz, 2021. 1, 3

[2] Thejasvi Beleyur and Holger R Goerlitz. Modeling active sensing reveals echo detection even in large groups of bats. *Proceedings of the National Academy of Sciences*, 116(52): 26662–26668, 2019. 1

[3] Thejasvi Beleyur and Julian Jandeleit. Ushichka pose estimation dataset, 2022. Available at https://doi.org/10.5281/zenodo.6620671. 5

[4] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, pages 586–606. International Society for Optics and Photonics, 1992. 3

[5] G. Bradski. The OpenCV Library (v4.5.5). *Dr. Dobb's Journal of Software Tools*, 2000. 3

[6] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 3, 4

[7] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981. 3

[8] Man Lok Fung, Michael Z. Q. Chen, and Yong Hua Chen. Sensor fusion: A review of methods and applications. In *2017 29th Chinese Control And Decision Conference (CCDC)*, pages 3853–3860, 2017. 1

[9] Kiana Hajebi and John S. Zelek. Structure from infrared stereo images. In *2008 Canadian Conference on Computer and Robot Vision*, pages 105–112, 2008. 5

[10] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2004. 2

[11] Julian Jandeleit. *Lidar assisted Depth Estimation for Thermal Cameras*. Bachelor thesis, Universität Konstanz, Konstanz, 2022. 2, 5

[12] Julian Jandeleit. DMCP Code, 2024. Available at https://doi.org/10.5281/zenodo.6621387. 5

[13] Asparuh Kamburov, Holger R Goerlitz, and Thejasvi Beleyur. Geospatial modelling inside the "Orlova Chuka" cave in bulgaria. *XXVIII International symposium on modern technologies, education and professional practice in geodesy and related fields*, 2018. 3

[14] Tong Ke and Stergios I. Roumeliotis. An efficient algebraic solution to the perspective-three-point problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[15] Eung-su Kim and Soon-Yong Park. Extrinsic calibration between camera and lidar sensors by matching multiple 3d planes. *Sensors*, 20(1):52, 2019. 2

[16] Peter Kovesi. Phase congruency detects corners and edges. In *DICTA*, 2003. 5

[17] Jesse Levinson and Sebastian Thrun. Automatic online calibration of cameras and lasers. In *Robotics: Science and Systems*, page 7, 2013. 2

[18] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 3, 4

[19] Matthew Matl. Pyrender 0.1.45, 2021. 2

[20] Miguel Ángel Muñoz-Bañón, Francisco A. Candelas, and Fernando Torres. Targetless camera-lidar calibration in unstructured environments. *IEEE Access*, 8:143692–143705, 2020. 2

[21] Trong Phuc Truong, Masahiro Yamaguchi, Shohei Mori, Vincent Nozick, and Hideo Saito. Registration of rgb and thermal point clouds generated by structure from motion. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 419–427, 2017. 2

[22] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems*, 32, 2019. 3, 4

[23] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 2, 3

[24] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 2, 3

[25] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. Sosnet: Second order similarity regularization for local descriptor learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11016–11025, 2019. 3, 4

[26] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33, 2020. 3, 4

[27] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991. 3