Pose Estimation of Thermal Cameras and Lidar Scan in a Cave Environment

Julian Jandeleit

October 27, 2021

Abstract

Pose Estimation is a classical problem in computer vision. This paper focuses on registering calibrated cameras with a lidar scan of the scene. The challenge of this task lies in the fact subterranean environment is overall similar in temperature which results in low structure captured by the thermal cameras.
This paper proposes an approach based on correspondences between a thermal image and a depth map generated from the lidar scan. This allows for pose estimation from only few known feature points, which can be annotated manually. The algorithm is applied to the Ushichka dataset and shows good results of the estimated pose. The result is then used as ground truth data to compare it with SIFT and Phase Congruency based approaches. We find a mean reprojection error of 8 pixels. The other two methods do not result in a estimated pose at all, which shows the superiority of the proposed algorithm.

1 Introduction

Pose estimation is one of the central problems computer vision deals with. The problem comes in different shapes. The common part is that the position of one object, often a camera that observes the scene, should be estimated in some frame of reference. It is an essential step for the structure-from-motion pipeline, which simultaneously estimate camera position and reconstructed scene from image feature correspondences [6]. Another classical use case is the area of robot navigation and autonomous driving, where the position of the car inside the environment or the position of an object relative to the car is estimated.
Different kinds of sensors have different characteristics. In this paper, we focus on two types of sensors. Lidar sensors have the advantage to provide very accurate depth information [5]. This can be used to create a detailed model of the observed scene. Here, we work with lidar data that is compiled into a mesh, a set of 3D points, so called point cloud, and a set of faces, which each connects 3d points to represent a surface. Thermal cameras capture the temperature of the scene, but don't provide the depth information. In order to combine the sensors to make use of both advantages, the data needs to be fused together. This is done by estimating the camera position and rotation inside the frame of reference of the lidar mesh and will be investigated in this paper.
We work on the *Ushichka* dataset [19]. It captures echolocating bats in their natural cave environment. Among others, it consists of 3 thermal cameras and a lidar scan

of the cave. The thermal cameras are calibrated into a common coordinate system, which we will call the *calibrated* coordinates. They capture the cave background which is comparatively uniform and high temperature bats flying around. The lidar scan was done beforehand, with the aim to visualize the bats in their natural environment. The coordinate system of the lidar points will be called *world-* or *lidar space*. The aim of this work is to find a transformation from calibration to lidar space that solves the pose of the cameras. By extension, this prepares the ground to view the bats flight path in world coordinates.
In the following, we propose a semi-automatic algorithm which computes such a transformation and compare it with approaches that rely on feature matching in the thermal images, by applying them to the Ushichka data.

2 Related Work

There are papers that deal with several aspects of the task. We will introduce them in this section.
Truong et al. [12] register point clouds generated from thermal and rgb images using structure-from-motion. They achieve their result by manufacturing a calibration chessboard pattern that is recognizable by both thermal and rgb cameras. Similarly, Kim and Park [15] use a chessboard for camera to lidar calibration. Kang and Doh [18] describe a method to calibrate a camera to a lidar scan by aligning edges in the point cloud with edges in the image. However, they focus on a 360 °rgb camera which produces a spherical panoramic image.
Hajebi and Zelek [8] describe how to generate Structure from infrared stereo images. They use log-gabor filters and the related phase congruency [4] to detect and describe features. It is evaluated in a office scene and compared to classical feature matching approaches.
There are several approaches that immerse themselves into the topic of phase congruency descriptors. HOP [10] and LGHD [11] use the information at different orientations created during phase congruency computation. HOMPC [13] merges the phase with the magnitude information. All approaches are unite by the use of histograms for feature representation.
To the best of my knowledge, there is no published approach that focuses on thermal-lidar alignment for subterranean environments, without using a calibration object.

3 Proposed Algorithm

In this section, we will discuss the proposed algorithm to find the pose of the cameras. We will refer to the algorithm as DMCP in short for *depth-map-correspondences*. It requires a depth map that is calibrated to the lidar scene, a (thermal) camera in its calibrated coordinate system and correspondences between both in pixel coordinates. Note, that the depth map can be computed by backprojecting the mesh into an arbitrary virtual camera, that observes the scene.
From this input, it will compute a transformation A that registers points in calibration space with points in lidar space. Figure 1 shows the general flow of information as an outline of the algorithm. It can roughly be divided into two parts:

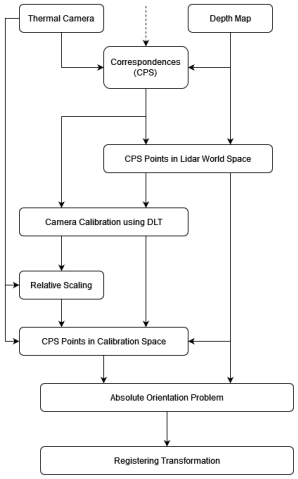


Figure 1: DMCP Information Flow

1. Pose calibration for one camera
2. Computing the registering transformation

Naming Conventions Let $P = K \cdot E$ be the projection matrix in the sense of the pinhole projection model [6]. Then K is the intrinsic camera matrix and E is the extrinsic camera matrix, which transforms world to camera coordinates. We will call $C = E^{-1} = [R\,T]$ the extrinsic matrix which describes the transformation from camera to world coordinates. R describes the rotation and T the position of the camera in world space.
We will distinguish different cameras using subscript and coordinate systems using superscript. Further, we will generally assume points and matrices to be given in their homogenous form.

Pose calibration Let P_{th}^{calib} be the projection matrix of a selected *thermal* camera, given in *calibration* coordinates. Let P_{dm}^{lidar} be the projection matrix for the camera that observed the *depth map*, given in *lidar* coordinates. Let K_{th} and K_{dm} be their respective intrinsic camera matrices. Let CPS_{th} be the ordered set of corresponding points in the thermal image and CPS_{dm} the respective points in the depth map.

Let $I_{dm}(x, y)$ denote the depth value in the depth map at position (x, y). We can now compute the 3D locations of the corresponding points CPS_{lidar} in the lidar space using the intrinsic camera matrix K_{dm} and the cameras extrinsic matrix C_{dm}.

$$CPS_{lidar} \coloneqq \{\, C_{dm} \cdot I_{dm}(x, y) \cdot K_{dm}^{-1} \cdot [x, y, 1]^T \mid [x, y, 1]^T \in CPS_{dm} \,\} \tag{1}$$

The next step is to estimate the pose of the thermal camera in the lidar space. The corresponding point sets CPS_{th} and CPS_{lidar} represent the problem of camera calibration, which can be solved using direct linear triangulation (DLT) [6].

$$P_{th}^{lidar} \coloneqq DLT(CPS_{th}, CPS_{lidar}) \tag{2}$$

P_{th}^{lidar} represents the estimated pose for the selected thermal, camera th, in the form of a pinhole projection matrix. Absolute pose and orientation can be derived using the relationships stated above in paragraph *Naming Conventions*.

Registering Transformation To transform any points in calibration space to lidar space, a transformation that registers those points needs to be computed.
Both coordinate systems may have a different scale, which needs to be taken account for. It can be extracted from comparing the estimated projection matrix P_{th}^{lidar} with the calibrated P_{th}^{calib}, more specifically their respective extrinsic camera matrices, which have the form $E = [r\ t]$. The matrix r is a rotation matrix and the norm of each column vector describes the scale in the respective axis. Let sv_{calib} and sv_{lidar} be the vector that describes the respective scales, then

$$scale \coloneqq \frac{\|sv_{calib}\|}{\|sv_{lidar}\|} \tag{3}$$

describes the relative scale between both coordinate systems. This factor takes account for the different scales. Now we have enough information to transform CPS_{lidar} first into thermal camera space and from there into calibrated space.

$$CPS_{calib} \coloneqq \{\, C_{th}^{calib} \cdot scale \cdot E_{th}^{lidar} \cdot [x, y, 1]^T \mid [x, y, 1]^T \in CPS_{lidar} \,\} \tag{4}$$

This works, because P_{th}^{lidar} and P_{th}^{calib} describe the same camera, only in different frames of reference. We now have corresponding 3D points in lidar and calibration space. This allows us to calculate the registering transformation A by solving what is known as the *absolute orientation problem* using Horns method [1].

$$A \coloneqq [R\ T] = solveAbsoluteOrientation(CPS_{calib}, CPS_{lidar}) \tag{5}$$

It computes R, T such that $CPS_{lidar} = R \cdot CPS_{calib} + T$, as an affine matrix $CPS_{lidar} = [R\ T] \cdot CPS_{calib} = A \cdot CPS_{calib}$. The projection matrices P_i^{calib} of each camera i in the calibration space can then be transformed with:

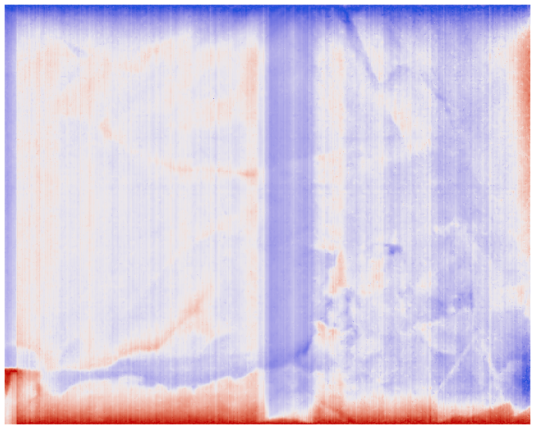
$$P_i^{lidar} \coloneqq P_i^{calib} \cdot A^{-1} \tag{6}$$

This solves the pose estimation problem for all cameras in calibration space.

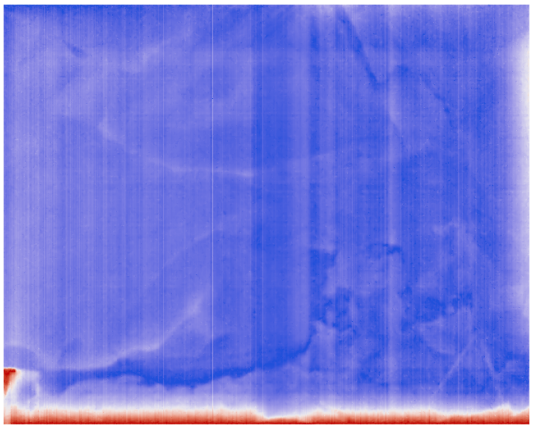
4 Experimental Evaluation

Experiments are carried out on the *Ushichka* dataset [19]. It contains a lidar scan of the *Orlova Chuka* cave in Bulgaria. The scan is represented in form of a mesh. Additionally, there are three thermal cameras, which are calibrated relative to each other in a *calibration* coordinate system. Figure 3b shows the relevant part of the mesh. Representative images for each thermal cameras are obtained by taking the median at each pixel along the time axis. A video of 100 frames has been chosen for this experiment. Figure 2a shows the normalized representative image for camera 1.
The programming languages MATLAB [20] was used to implement the algorithm and to run the experiments. The depth-map generation from backprojecting the mesh was done using the PyVista library [16], which implements this step for OpenGL-like camera models.

Preprocessing The walls and structures all have a similar temperature which, in addition to general infrared imaging characteristics [8], result in comparatively low gradients at structure boundaries.



(a) Thermal image from camera 1, normalized to $[0, 1]$



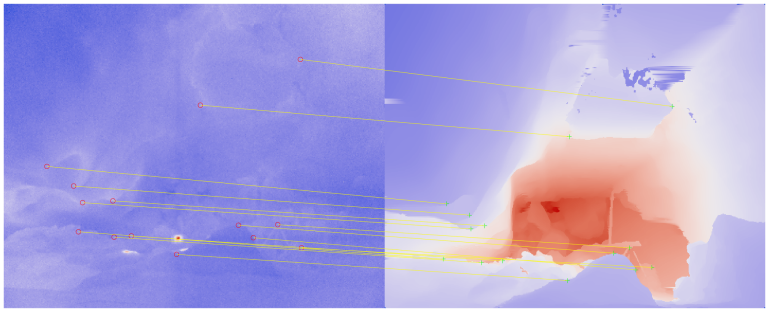
(b) Image from camera 1 without FPN, normalized to $[0, 1]$

Figure 2: Fixed Pattern Noise

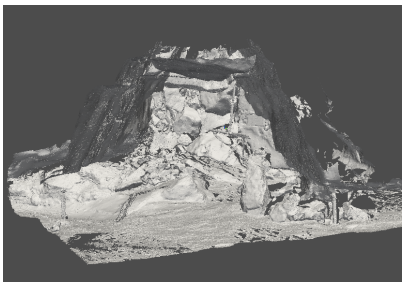
In this case, there also is a lot of fixed pattern noise (FPN) present. Wang et al. proposes a method for removing this kind of noise from infrared images [17]. Here, we will extract the FPN by selecting all horizontal and vertical components of the Fourier-Transformation of the representative image and then subtracting it from the original. Figure 2 shows the result on the representative image for camera 1.

Experiments We execute the proposed depth-map-correspondence (DMCP) algorithm on the dataset using the manually annotated correspondences, as shown in Figure 3a. The estimated pose for each camera is then used to backproject the lidar mesh into a depth map. Note that, aside from the sparse thermal to depth map correspondences, we do not have to specify any hyper-parameters. The minimum number of input correspondences is defined as the maximum of the minimum points needed for camera calibration (DLT) and the absolute orientation problem. As Horns method needs a minimum of 5 points and DLT a minimum of 6, a total number of 6 correspondences should be sufficient. Here, we use a total of 14 points to minimize the error in a least-squares sense.
Further, we will compare DMCP to approaches that estimate a point cloud from correspondences between the thermal images only. The idea behind these methods is, that the point cloud should then be registered to the point cloud obtained from the lidar scan using iterative-closest-point algorithm (ICP) [2]. The following two methods will be investigated:

- SIFT based feature detection and matching
- Phase congruency based feature detection and matching



(a) Annotated correspondences between thermal image (left) and depth map (right)



(b) Primary section of lidar mesh

Figure 3: DMCP Experiment

SIFT Tareen and Saleem [14] recommend SIFT [7] to be the descriptor that in general, for images taken with a normal camera, is the most accurate algorithm. We use it to detect features in the thermal images and then match them using Lowes method to reject ambiguous matches.
We use the VLFeat libaray [9] for SIFT computation. The parameter *PeakThresh* is set to 0.5. We consider pixels to be fundamental inliers if the matched point is closer that 5 pixels to the epipolar line. The epipolar line is calculated using the fundamental matrix obtained from the already calibrated cameras.

Phase Congruency Phase congruency is a quantity which provides information invariant to contrast and can be used to detect edges and corners [4]. Hajebi and Zelek [8] suggests it to be used on thermal image data. The LGHD descriptor [10] is one of many proposed descriptors that describe local features using phase congruency. We will apply the LGHD descriptors to the phase congruency feature points and match the descriptors using cosine similarity as suggested by Hajebi and Zelek [8].
Phase congruency is computed using the library from Peter Kovesi [21]. We use the parameters $sigma = 1.5$ and $n\_pts = 500$. The LGHD descriptor is computed using the implementation by the authors. Similar to above, we consider pixels to be fundamental inliers if the matched point is closer that 5 pixels to the epipolar line.

5 Results

In this section we will describe the results of the experiments.

DMCP When executing the proposed DMCP algorithm on the correspondences shown in Figure 3a, we obtain the estimated camera positions shown in Figure 4. The mesh and, in particular the annotated points in 3D space, can then be backprojected into the cameras to obtain an estimated depth map. The registration error can be computed by comparing the depth map with the original image. On a perfectly registered camera, points that correspond to the same 3D point, should lie on the same pixel coordinates. We compute the distance in pixels between the annotated points in the thermal and estimated depth image. The differences are illustrated in Figure 5a. The mean distance is 7.9 pixels. You can find the barchart of all distances in Figure 5b.

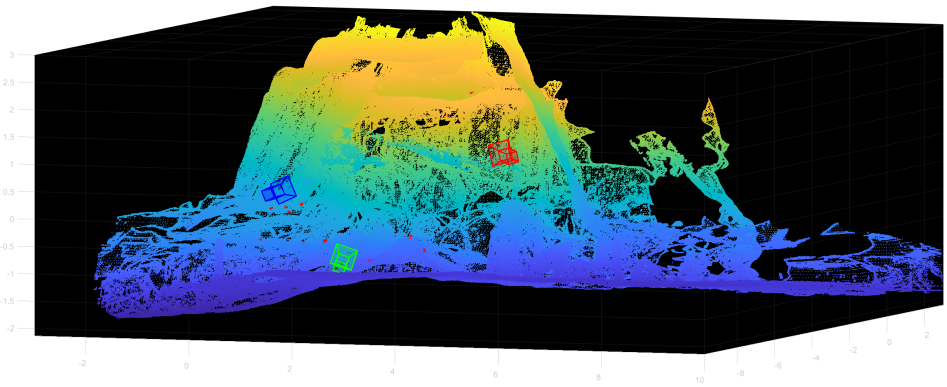
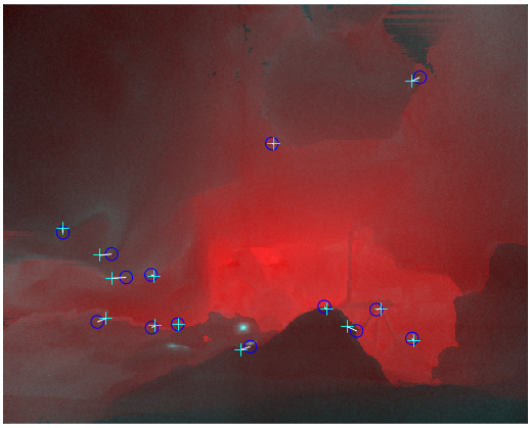


Figure 4: Estimated camera positions, transformed using transformation obtained from DMCP. Camera 1 is shown in red, camera 2 in green and camera 3 in blue. Also shows visible annotated correspondences in 3D as small red points.

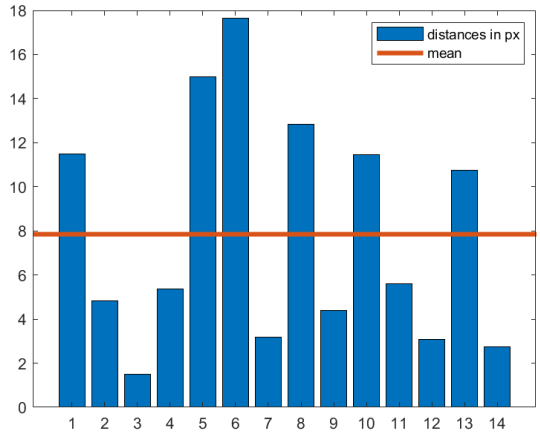
Also note, that we estimated a scaling factor of about 1.65 between calibrated and lidar space.

SIFT Figures 6a and 6b show the points that are detected using SIFT. Figure 6c shows the matching inliers, that satisfy the epipolar constraint.
You can see that SIFT detects points roughly at structure boundaries. However there are also many points distributed between those. There is only one match that satisfies the epipolar constraint. When visually looking at the match, we find that the matched point indeed does correspond.

LGHD Figures 7a and 7b show the points that are detected using LGHD. Figure 7c shows the matching inliers, that satisfy the epipolar constraint as described in SIFT.
We can see that the detected points lie on most structure boundaries that can be identified visually by humans. There are several matches that satisfy the epipolar constraint. Nevertheless, a few of those matches can be identified as false positives, as they do not describe the same 3D point.

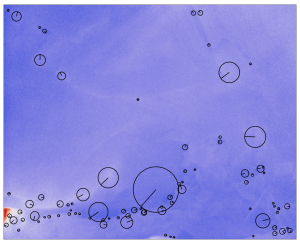


(a) Overlayed thermal image (camera 2) and its estimated depth map. Points in thermal image are marked as cyan crosses, points in the corresponding depth map are marked as blue circles.

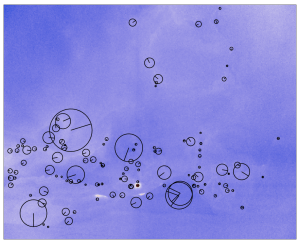


(b) Distances between points in thermal image and estimated depth map.

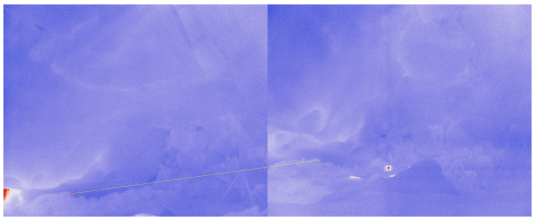
Figure 5: Evaluation DMPC



(a) Detected points in camera 1.

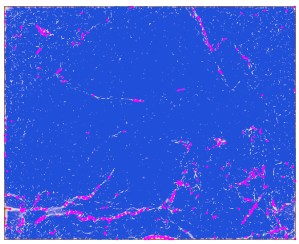


(b) Detected points in camera 2.

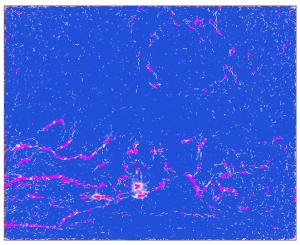


(c) SIFT fundamental inliers

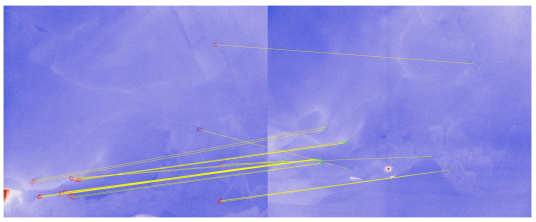
Figure 6: SIFT Experiment



(a) Detected points in camera 1.



(b) Detected points in camera 2.



(c) LGHD fundamental inliers

Figure 7: LGHD Experiment

6 Discussion

The result of the DMCP Experiment shows, that the proposed algorithm succeeds in estimating the pose of the camera. There still is some error in the registration, which can be explained by the fact that the initial correspondences were selected manually. Errors made in this step lead to errors in the DLT algorithm and the estimation of the transform. There could also be errors made by the DLT algorithm itself if the input represents a degenerate configuration. This results in stark contrast to the other two approaches.

Although SIFT, when used with the right threshold, does find many points, they are so unstable that only one match satisfies the epipolar constraint. This makes it extremely difficult to apply ICP-registration as it needs several well distributed points to work. ICP also relies on a initial transform that roughly aligns two clouds. 3D point cloud descriptors rely on a point cloud that is dense enough to describe the local shape. This is why the initial transformation can also not be computed.
The phase congruency approach does find more matches than the SIFT approach. However some matches, while satisfying the epipolar constraint, are not valid. This reduces the overall quality of the estimated point cloud which also still is very sparse. Thus, this approach suffers from similar problems as described above. Advantages however are, that the detected points are very closely aligned to the structure boundaries which is a sign for more stable features. However, on large baselines, these structure boundaries may look different. Another advantage of the phase congruency approach is, that there are more valid matches. This shows that the phase congruency approach in general is preferable to the classical SIFT method.

The proposed DMCP algorithm is the only one that achieved to estimate the pose of the cameras. Its main drawback is that it relies on thermal to depth map point correspondences which are currently annotated manually. The ICP-methods would allow for fully automatic pose estimation, but we showed that it is difficult to compute an adequate feature matches for this dataset. The phase congruency approach looks promising, but it seems that there is still work do be done to describe and correctly match the detected points. Kovesi [3] states, that one of the main advantages of phase congruency is, that it is able to detect features along wide classes of images without having to make an explicit assumption about the luminance profile of the feature. This indicates that, when developing an appropriate descriptor, that takes into account large viewpoint changes, phase congruency might also be used to identify the correspondences between thermal and depth map, which are needed for a fully automated DMCP approach.
For future projects similar to the ushichka dataset, the use of a calibration target that is detectable from both, lidar and thermal sensors, might provide more stable results that are achievable using simpler computations. We covered possible approaches in Section 2.

7 Conclusion

In this paper, we approached the task of pose estimation for thermal cameras in a cave environment. We present the DMCP algorithm, which estimates the transformation

from camera calibration space to the world space of the lidar mesh by using sparse correspondences to a backprojected depth map. With an average error of about 8 pixels, it is a feasible approach to the task.
We compared DMCP with methods that estimate feature matches to be used for later ICP-registration. They still present open challenges as both, the SIFT and the phase congruency experiment, did not result in a point cloud, dense enough for ICP.
When refining the phase congruency approach to improve the feature description and matching, ICP registration might be successful. Progress in this area could also improve the biggest drawback of the DMCP algorithm, which is the manual selection of corresponding points in a thermal image and a depth map.
To conclude, the semi-automatic DMCP solves the pose estimation task, while leaving the challenge of feature matching for subterranean thermal images open for future work.

References

[1] Berthold K. P. Horn. "Closed-form solution of absolute orientation using unit quaternions". In: 4.4 (Apr. 1987), p. 629.

[2] Paul J Besl and Neil D McKay. "Method for registration of 3-D shapes". In: *Sensor fusion IV: control paradigms and data structures*. Vol. 1611. International Society for Optics and Photonics. 1992, pp. 586–606.

[3] Peter Kovesi. *Image Features from Phase Congruency*. 1999.

[4] Peter Kovesi. "Phase Congruency Detects Corners and Edges". In: *DICTA*. 2003.

[5] Stephen E Reutebuch et al. "Accuracy of a high-resolution lidar terrain model under a conifer forest canopy". In: *Canadian Journal of Remote Sensing* 29.5 (Oct. 2003), pp. 527–535.

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

[7] David G. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints". In: 60.2 (Nov. 2004), pp. 91–110.

[8] Kiana Hajebi and John S. Zelek. "Structure from Infrared Stereo Images". In: *2008 Canadian Conference on Computer and Robot Vision*. 2008, pp. 105–112.

[9] A. Vedaldi and B. Fulkerson. *VLFeat: An Open and Portable Library of Computer Vision Algorithms*. `http://www.vlfeat.org/`. 2008.

[10] Cristhian A. Aguilera, Angel D. Sappa, and Ricardo Toledo. "LGHD: A feature descriptor for matching across non-linear intensity variations". In: *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, Sept. 2015.

[11] Hussin K. Ragb and Vijayan K. Asari. "Histogram of oriented phase (HOP): a new descriptor based on phase congruency". In: ed. by Sos S. Agaian and Sabah A. Jassim. SPIE, May 2016.

[12] Trong Phuc Truong et al. "Registration of RGB and Thermal Point Clouds Generated by Structure From Motion". In: IEEE, Oct. 2017.

[13] Zhitao Fu et al. "HOMPC: A Local Feature Descriptor Based on the Combination of Magnitude and Phase Congruency Information for Multi-Sensor Remote Sensing Images". In: 10.8 (Aug. 2018), p. 1234.

[14] Shaharyar Ahmed Khan Tareen and Zahra Saleem. "A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK". In: *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*. Mar. 2018, pp. 1–10.

[15] Eung-su Kim and Soon-Yong Park. "Extrinsic Calibration between Camera and LiDAR Sensors by Matching Multiple 3D Planes". In: *Sensors* 20.1 (Dec. 2019), p. 52. ISSN: 1424-8220.

[16] C. Bane Sullivan and Alexander Kaszynski. "PyVista: 3D plotting and mesh analysis through a streamlined interface for the Visualization Toolkit (VTK)". In: *Journal of Open Source Software* 4.37 (May 2019), p. 1450.

[17] Ende Wang et al. "Infrared stripe correction algorithm based on wavelet decomposition and total variation-guided filtering". In: 16.1 (Dec. 2019).

[18] Jaehyeon Kang and Nakju L. Doh. "Automatic targetless camera–LIDAR calibration by aligning edge with Gaussian mixture model". In: *Journal of Field Robotics* 37.1 (2020), pp. 158–179. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1002/rob.21893`.

[19] Thejasvi Beleyur. "Theoretical and empirical investigations of echolocation in bat groups". PhD thesis. Konstanz: Universität Konstanz, 2021.

[20] *MATLAB version 9.10.0.1649659 (R2021a) Update 1*. The Mathworks, Inc. Natick, Massachusetts, 2021.

[21] P. D. Kovesi. *MATLAB and Octave Functions for Computer Vision and Image Processing*. Available from: <https://www.peterkovesi.com/matlabfns/>.