ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Autonomous Systems Lab
Prof. Roland Siegwart

Semester Thesis

Towards Robust Cross-Spectral Optical-Thermal SLAM onboard a fixed-wing UAV

Spring Term 2019

Supervised by:
Timo Hinzmann
Dr. Nicholas Lawrance
Florian Achermann

Author:
Felix Graule

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Towards Robust Cross-Spectral Optical-Thermal SLAM onboard a fixed-wing UAV

Authored by (in block letters):

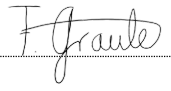
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Graule	Felix

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Zurich, 02.08.2019	

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.

Contents

Abstract

By autonomously soaring in thermal updrafts, fixed-wing glider UAVs are able to extend their flight duration. To achieve this, we propose using a thermal map based on which the UAV can navigate towards predicted locations of such updrafts. Further, thermal maps can be used to fly in poor visual conditions such as fog or at night. In this Semester thesis, we aim to develop approaches to generate such thermal maps in real-time and onboard a UAV. This thermal map needs to be registered to an existing optical map to allow consistent localization of the UAV. To this end, we explore various approaches to align optical and far-infrared images based on a specialized feature descriptors, pose graph optimization and mutual information. Due to the difficulty of this cross-spectral image matching task, those methods are computationally intractable. Hence, we use those methods to generate training data for a deep learning based method to perform cross-spectral matching in real-time. To our knowledge we are the first ones to use a fully end-to-end homography estimation pipeline for a multi-modal image matching problem. We demonstrate that learning cross-spectral descriptors is possible, but more challenging than in the optical to optical setup. We further introduce a large data set consisting of otherwise rare aligned image pairs of optical and thermal images.

Symbols

Acronyms and Abbreviations

ASL	Autonomous Systems Laboratory
CNN	Convolutional Neural Network
DNN	Deep Neural Network
DL	Deep Learning
ETH	Eidgenössische Technische Hochschule
EKF	Extended Kalman Filter
FAST	Features from Accelerated Segment Test
FIR	Far Infrared
GPS	Global Positioning System
GTSAM	Georgia Tech Smoothing and Mapping library
IMU	Inertial Measurement Unit
LMEDS	Least Median of Squares
MI	Mutual Information
NIR	Near Infrared
RANSAC	Random Sample Consensus
SIFT	Scale-Invariant Feature Transform
SLAM	Simultaneous Localization and Mapping
SURF	Speeded Up Robust Features
UAV	Unmanned Aerial Vehicle

Chapter 1

Introduction

1.1 Motivation

In the past years, unmanned aerial vehicles (UAV) have gained great significance in a wide variety of applications such as ground surveillance, search and rescue or maintenance. However, their limited flight duration remains a major drawback of most current solutions since it both restricts the number of possible use cases as well as their economic feasibility since low endurance requires more manual mission support.

Typically, commercial battery-powered multi-copter drones can only stay airborne for up to 30 minutes[1]. Fixed-wing drones on the other hand can fly for as long as 2 hours[2] due to their more efficient, wing-based flight mode. In order to further increase endurance, two major approaches are applied: for one, the UAV's wings can be equipped with solar panels to actively generate energy while flying. The effectiveness of this approach has been shown in projects like AtlantikSolar[3], which achieved a record-breaking 81 hours non-stop flight. A different approach is to use thermal updrafts to gain height by autonomously circling within them. This method is inspired by soaring birds and is at the heart of Microsoft's Project Frigatebird[4], which is tightly connected to this work.

Before a UAV can autonomously soar in thermal upwinds, it needs to find them. There is a manifold of reasons for thermal updrafts to occur, one of them being strong temperature gradients near the ground. Hence, having a detailed understanding of the temperature distribution of the overflown terrain, i.e. a thermal map, is crucial to predict the location of updrafts and act accordingly. The goal of this Semester thesis is to develop algorithms to build such a thermal map of the ground over which the fixed-wing UAV is flying.

1.2 Requirements

We propose that by using a thermal map the UAV could locate thermal updrafts, thus enabling autonomous soaring. Further, the UAV can navigate and localize in bad visual conditions, e.g. at night or in foggy weather. The overall setup and desired behavior of the glider UAV is illustrated in figure 1.1.

[1] http://www.dronesglobe.com/guide/long-flight-time/

[2] https://3dinsider.com/fixed-wing-drones/

[3] https://www.atlantiksolar.ethz.ch/

[4] https://www.microsoft.com/en-us/research/project/project-frigatebird-ai-for-autonomous-soaring/

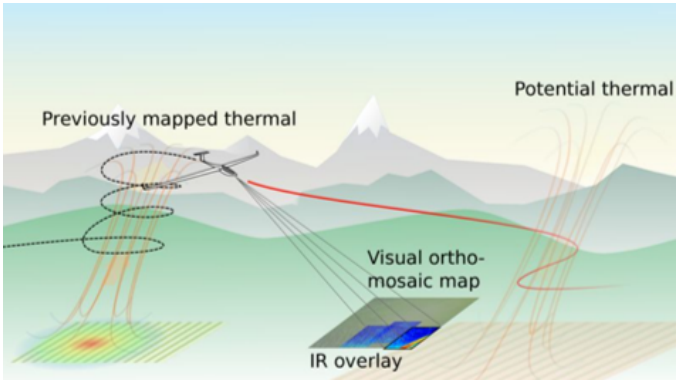


Figure 1.1: Illustration of overall setup (courtesy of Dr. N. Lawrance)

Clearly, the thermal map needs to be globally consistent, should be built on-board the UAV and has to be available in real-time to be used for decision making. Additionally, it should be registered to any existing optical map used for localization and decision making in order to maintain a single set of UAV positions which is used for motion planning and control. With this, we can define the following requirements for the method to be developed:

a) Builds globally consistent thermal map

b) Runs in real-time on-board a fixed-wing UAV

c) Thermal map needs to be registered to existing optical map

From requirement a) we conclude that our method should apply color balancing and loop closure techniques on the thermal map to guarantee global consistency. Requirement c) calls for some sort of cross-spectral image registration method which matches optical to thermal images. Requirement b) adds the need for a computationally tractable method using limited resources. From this, we propose that the cross-spectral matching task can be solved using a Deep Neural Network architecture. This decision is also based on various research results like [1], [2] or [3] showing that DNNs outperform classical image matching methods by a large margin in both matching quality and frame rate, especially when it comes to image pairs with some sort of domain gap between them. Two examples for this are the great performance of LIFT [4] on day-to-night image matching tasks or the high robustness of SuperPoint [5] against strong illumination changes. To train such a DNN architecture, we need perfectly aligned cross-spectral image pairs such that the network can learn the similarity between optical and thermal images. This ground-truth data can be generated offline, hence computational cost is not a first-priority requirement here.

1.3 Key Assumptions

We assume the following about the targeted use case of the method to be developed:

a) The UAV mostly flies over natural scenes such as fields including some, but few human-built structures

b) Images are taken by vertically down-facing optical and thermal cameras

c) The UAV is capable of localizing and acting in reference to the optical map

d) Optical and thermal camera are not tightly time-synchronized, hence giving rise to different optical and thermal poses

1.4 Project Structure

Having defined the requirements and assumptions, we can now define the overall structure of the thesis at hand accordingly. After looking into related work in chapter 2, we describe methods to generate ground-truth in chapter 3 and evaluate them in chapter 4. In chapter 5 we show how to train a DNN for the cross-spectral matching task. Results for the DNN training are shown in chapter 6. Finally, we look into future work and draw conclusions in chapters 7 and 8.

1.5 Previous Work

This project builds upon work done by Bastien Chatton in his Master thesis [6]. Back then, a rigid forward-projection method was used to project thermal images onto the existing optical map using the following transformation:

$$T_{IR}^{G} = T_{opt}^{G} \cdot T_{IR}^{opt} \tag{1.1}$$

Here, T_{opt}^{G} is found by optimizing the pose graph including only the optical poses but disregarding the thermal poses entirely. A subset of the resulting pose graph is illustrated in figure 1.2.

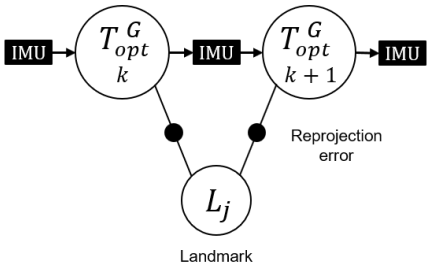


Figure 1.2: Pose Graph optimized in Bastien Chatton's Master thesis [6]

T_{IR}^{opt} is precalibrated and kept constant. As expected, the resulting overlap between the optical and thermal map shows significant inconsistencies as shown in figure 1.3. For example, such an inconsistency can be seen in the blue circle where the border between two different fields does not overlay in the optical and thermal map. One goal of this thesis is to improve the consistency of the overlap between the optical and thermal map by including thermal poses into the optimization problem.

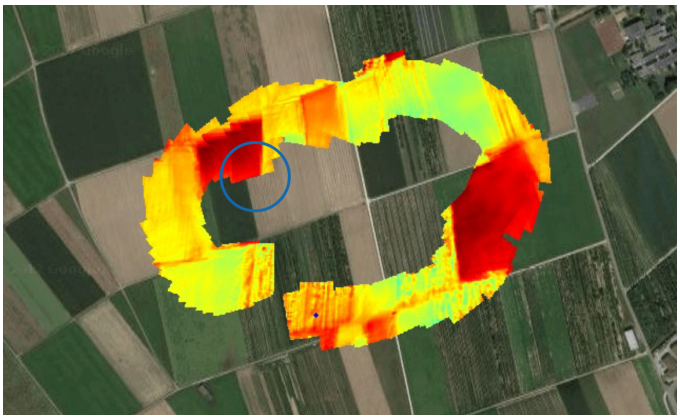


Figure 1.3: Overlap of optical and thermal map in Bastien Chatton's Master thesis

1.6 Cross-Spectral Imaging

The main problem addressed in this project is the matching of optical and thermal (FIR) images. This is a much harder task than conventional optical-to-optical image matching, mainly due to the large frequency gap between the two modalities. While optical images capture radiation of around $0.4 - 0.7\mu m$ [7], the thermal spectrum ranges from $4 - 12\mu m$ which requires special camera devices, e.g. the FLIR Tau 2 used in this work. Thermal imaging is to be separated from near-infrared (NIR) imaging, which happens between $0.7 - 1.0\mu m$ and can be done with most standard CCD sensors [7]. In general, scenes appear much more similar in optical and NIR than they do in optical and thermal.
Due to the large domain gap between optical and thermal, image regions can undergo strong non-linear intensity transformations. A cross-spectral image matching algorithm has to be robust against various transformations and effects which are listed in the following. To our knowledge, there is no existing method to perform cross-spectral matching between optical and FIR in a fast and robust manner.

Gradient Inversion Optical and thermal properties of objects are in many cases completely independent. For example, two black cups filled with cold and hot water respectively, will look perfectly alike in the optical image but differ drastically in thermal. This effect is referred to as Gradient Inversion: a given image gradient between two objects in optical might appear inverted in the thermal image of the same scene. An example is shown in 1.4.

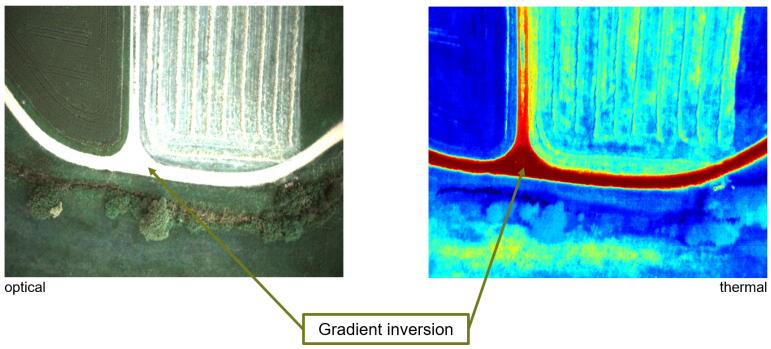


Figure 1.4: Example for gradient inversion

Loss of Sharpness Due to the laws of thermodynamics, objects thermally connected to each other will converge towards a common equilibrium temperature. This causes thermal structures within close proximity to each other to be similar. Hence, thermal images tend to have fewer sharp intensity changes such as corners or edges.

Loss of Texture The same thermodynamic effect combined with the lower resolution and imaging quality of thermal cameras leads to smoothing of image regions with low intensity changes, thus loss of textural information. An example can be seen in figure 1.5.

Angular Dependency of Sun Reflections Depending on the angle we observe a scene from, the amount of sun rays being reflected directly into the thermal camera varies greatly since the angle between the ground and the camera changes. This introduces significant error in the measured thermal distribution, especially in setting where the camera does not face downwards vertically. In this work, such cases appear mostly when flying curved trajectories.

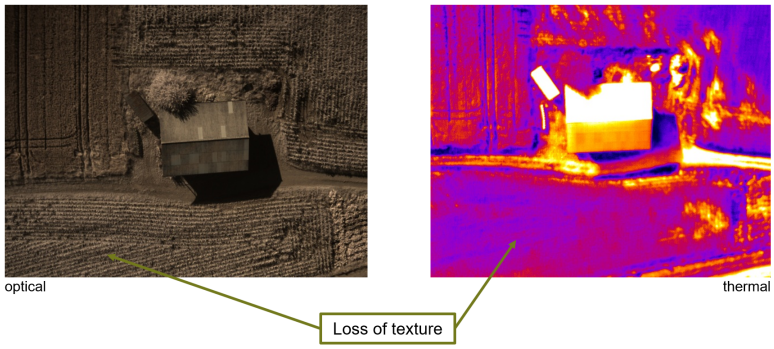


Figure 1.5: Example for loss of texture

Characteristics of Thermal Cameras Two major drawbacks of thermal imaging sensors are their low resolution and the so-called flat-field correction. The lower resolution of thermal cameras leads to a less accurate calibration of the camera intrinsics, thus decreasing the accuracy of all 2D-to-3D operations. Flat-field correction is a periodic recalibration of the thermal detection elements on the sensor during which the shutter is completely closed. This process is necessary to remove artifacts in the thermal images caused by pixel-to-pixel sensitivity variations in the sensor [6]. While flat-field correction improves the overall image quality, it introduces empty frames and discontinuities in the thermal image stream.

Comparison with Optical to NIR Matching To clarify the previously mentioned differences between image matching from optical to FIR and optical to NIR, please refer to the example image pairs shown in figure 1.6. We see that for optical to NIR, no gradient inversion is observed and the loss of texture is less drastic in general. Moreover, sun rays reflected into the camera do not introduce strongly varying disturbances.

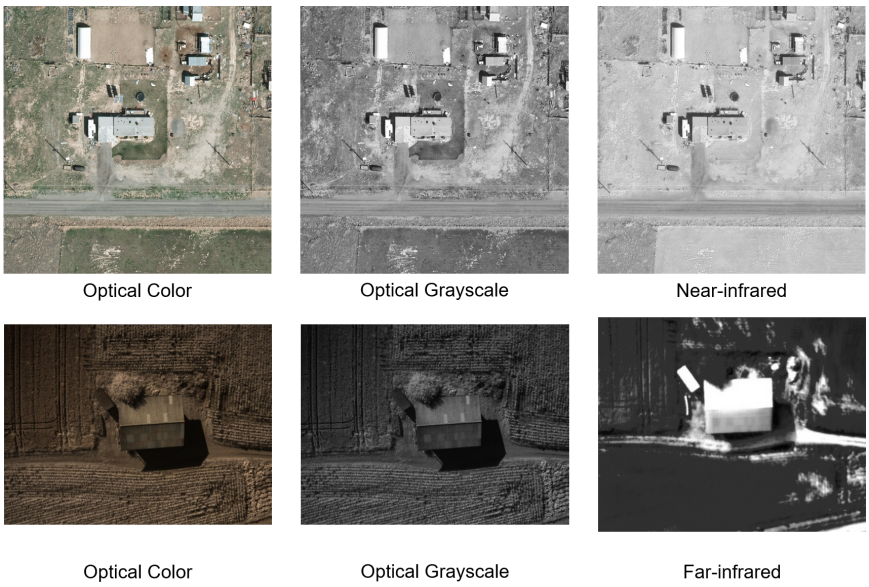


Figure 1.6: Examples for optical to NIR and optical to FIR image pairs

Chapter 2

Related Work

In the following we give an overview of relevant literature connected to our project.

2.1 Soaring and Thermal Imaging for UAVs

The problem of autonomous soaring has been described as a partially-observable Markov decision process (POMDP) by *Guilliard et al.* [8]. They model the thermals as three-dimensional Gaussians and use the thermal's parameters as the state of the POMDP and the altitude gain as the reward. An algorithm called POMDSoar balances exploitation and exploration. However, this approach only solves the soaring task once the UAV is inside a thermal updraft, but does not look at how to predict their location. *Chatton et al.* [6] on the other hand describe the circumstances under which thermal updrafts occur and look into how to predict the 3D shape of the thermal based on the heat map shown in figure 1.3.
To build a thermal map that is fully aligned to an existing optical map, our proposed direct cross-spectral matching is not the only possible approach. One could also continuously fuse two independently generated point-clouds as shown by *Truong et al.* [9]. Alternatively, *Chen at al.* show how one can use pure within-modality matches to update a RGB-T map as part of a SLAM framework [10]. However, such approaches based on within-modality matches do not offer the same amount of robustness against camera model errors and the higher resolution of the optical camera cannot be leveraged to increase the accuracy of the thermal map.
In addition, thermal cameras on-board UAVs can be used for various other tasks than localizing thermal updrafts or night flights. *Kümmerle et al.* effectively use the enhanced visibility of humans in the thermal spectrum to perform detection and tracking in real-time at minimal computational cost [11].

2.2 Traditional Cross-spectral Image Matching

Cross-spectral image matching has been a topic of high interest since decades, not only for optical to FIR but also between other spectra and modalities. Two underlying principles are to be separated: first, some methods are based on detecting salient keypoints in both images and then building descriptors for them to match them. On the other hand, one can use the statistical mutual information measure to globally align the images. Traditional descriptors tend to not work well for cross-spectral matching, as was shown by *Ricaurte et al.* when they applied both gradient-based and intensity-based descriptor methods to match optical and FIR images [12]. To overcome this, a number of specialized cross-spectral keypoint-descriptor methods have been proposed. In order to be robust against the gradient inversion effect

described in section 1.6, *Firmenich et al.* introduced a gradient invariant version of SIFT (GDISIFT) [13] and showed improved performance for optical to NIR matching in comparison to vanilla SIFT [14]. *Aguilera et al.* propose an Edge Oriented Histogram (EOH) [15] descriptor which uses histograms over orientation of edges extracted around the point to be described. This approach was later improved by the same group, resulting in the Log-Gabor Histogram Descriptor (LGHD) [16]. Instead of using edge orientations, here histograms of scales and histograms built on top of high-pass filtered images are used to describe the keypoints. This method is described in detail in section 3.1.1.
Mutual information based image alignment is used extensively in bio-medical imaging, e.g. to align brain scans taken with magnetic resonance imaging (MRI) and computed tomography (CT) as shown by *Pluim et al.* [17]. The same concept has already been shown to work well for robotics, for example as part of a Lucas-Kanade feature tracking system proposed by *Dowson et al.* [18]. The mathematical foundations behind these approaches are explained in chapter 3.

2.3 Deep-Learning Cross-spectral Image Matching

None of the approaches described above were shown to be working in real-time. In contrast, deep learning based approaches allow fast processing of visual data due to massive parallelization on GPUs. Likewise, DL-based approaches were shown to outperform traditional methods in most computer vision task such as stereo vision [1], image classification [2] and patch similarity [3]. In early trials to leverage DL-based methods for cross-spectral tasks, *Aguilera et al.* show three different CNN architectures that can be trained to match from optical to NIR images [19]. They also show that some of the learning generalizes to the task of matching optical to FIR. In more recent work, the same authors proposed Q-net, a quadruple CNN architecture specifically designed for building descriptors for optical to NIR matching [20]. *En et al.* introduced the idea of using both Siamese and Pseudo-Siamese networks to learn characteristic common to both modalities and modality-specific information, respectively. This gives rise to a three-stream architecture, called TS-Net [21]. *Baruch et al.* build on both those works and propose Hybrid CNN, which is another quadruple architecture consisting of two sub-networks, a Siamese CNN and a dual non-weight-sharing CNN [22]. They use a novel auxiliary loss and a hard negative-mining scheme to boost performance. However, all those approaches focus on building descriptors for optical to NIR matching, not optical to FIR. Besides learning cross-spectral descriptors, one also needs learn to detect highly repeatable and distinctive keypoints across both spectra. Such a method is Quad-networks proposed by *Savinov et al.* which uses quadruple patch pairs to train a CNN in an unsupervised keypoint ranking setup [23]. The authors show this to perform well on optical to depth which is a similar task like optical to thermal in terms of loss of texture. Instead of having separate keypoint detector and descriptor networks, the overall homography estimation task which we are ultimately trying to solve, can be trained and performed in an end-to-end fashion. One popular work using this idea is SuperPoint proposed by *DeTone et al.* which consists of a single VGG-encoder [24] and is trained with a self-supervised scheme called Homographic Adaption and optical only data [5] (see section 5.1 for more details). In section 5.2 we propose an approach to alter the original SuperPoint framework to perform the cross-spectral task.

Chapter 3

Ground-truth Generation Methods

In the following chapter we describe four different methods used to generate ground-truth data for the training of a DL-based cross-spectral image matching network.

3.1 Cross-spectral Image Alignment

In a first step, we focus on matching image pairs taken roughly at the same time. As depicted in figure 3.1, only the matches within the same image pair are taken into account. We explore two approaches to perform this image alignment, a keypoint matcher with specialized cross-spectral descriptors explained in section 3.1.1 and a method based on mutual information explained in section 3.1.2.

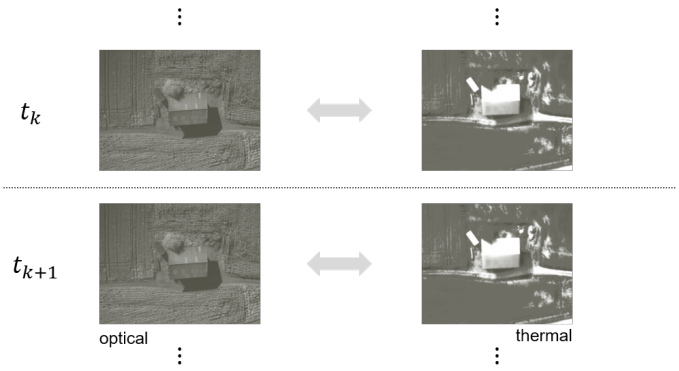


Figure 3.1: Cross-spectral image matching within an image pair

3.1.1 Log-Gabor Histogram Descriptor

As described in chapter 2, we make use of specialized cross-spectral descriptors called LGHD [16] in order to match optical and thermal images. Conceptually, the descriptor is quite similar to SIFT since it also builds on an image-pyramid of scaled and filtered versions of the original image. But, for LGHD the underlying idea is to build a vector representation of an image patch around a keypoint in a high-pass filtered image space. High-frequency components such as edges and corners are supposed to be more invariant to the non-linear intensity changes occurring when going from optical to thermal frequency domain (see section 1.6). In case of

LGHD, the high-pass filtering is done using a bank of Log-Gabor filters with varying scale and orientation. The following equation describes the frequency dependency of these Log-Gabor filters:

$$G(f) = \exp\left(\frac{-(\log(f/f_0))^2}{2(\log(\sigma/f_0))^2}\right) \tag{3.1}$$

Here, f_0 is the center frequency of the filter and σ the bandwidth. Like other wavelets, Log-Gabor filters allow to analyze a signal, the optical or thermal images in our case, in both space and frequency dimension simultaneously (space-frequency signal decomposition). An example of an image pair and one Log-Gabor filtered image is shown in figure 3.2.

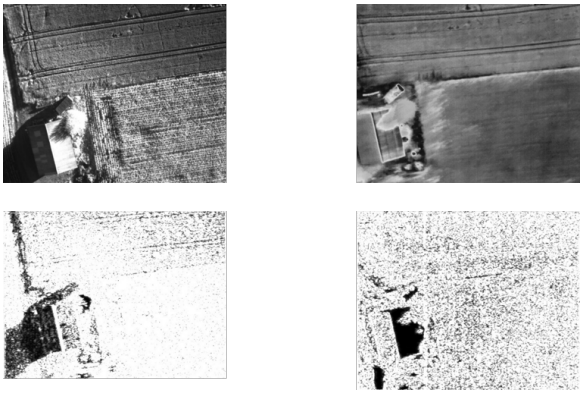


Figure 3.2: Example of cross-spectral image pair and their Log-Gabor filtered images for scale 1 and orientation 1

Those Log-Gabor filter banks can then used to compute the phase congruency of keypoints as done by *Kovesi et al.* in their open-source MATLAB implementation [25]. Phase congruency is generally much more invariant against illumination changes than local image gradients[1]. The authors of LGHD use the work by *Kovesi et al.* directly to implement their descriptor in MATLAB. However, since the existing pipeline for our project is entirely based on C++, we implemented our own version of LGHD. For this, we make use of an open-source C++ phase congruency package[2] by *Pinto et al.*

We build the descriptors over $N_s = 6$ scales and $N_o = 8$ orientations for a patch size of $N_p = 100$ pixels. For our UAV images these parameters proved to work better than the ones proposed in the original paper where 4 scales, 6 orientations and a patch size of 80 pixel patch size are used. The actual descriptor stores the scale and orientation of the maximum filter response for each pixel in the patch. To do so, we extract a patch of size $N_p \times N_p$ for every keypoint and divide it into 4x4 sub-regions. For each sub-region, we iterate over all pixels and build a histogram of the orientations leading to maximum filter response. Hence, each histogram has as many bins as we have orientations. We do this for each scale separately, storing $N_o \times 4 \times 4$ integer values per scale. We then normalize the histograms with respect to all sub-regions. Using our parameters, we therefore end up with $N_o \times 4 \times 4 \times N_s = 8 \times 4 \times 4 \times 6 = 768$ floating point numbers per descriptor.

We further deviate from the original LGHD in the keypoint detection step: instead of using FAST [26] to detect keypoints on both the original optical and thermal

[1] http://cvpr11.cecs.anu.edu.au/files/PhaseCongCVPR.pdf
[2] Code: https://github.com/chvillap/phase-congruency-features

image, we detect the keypoints on images in the joint filtered image space by combining frequency components from all scales and orientation into one image. We found this to give more repeatable keypoints across the frequency gap.

3.1.2 Mutual Information

The general idea for Mutual Information (MI) based image alignment is to find a transformation that maximizes the statistical similarity measure $MI(A, B)$, where A and B are two images which can be represented as probabilistic distributions P_A and P_B. In practice, one can estimate P_A and P_B as the marginal histogram of each image, normalized to 1. Similarly, the joint distribution $P_{A,B}$ can be approximated by normalizing the joint histogram of both images. The mutual information metric is then defined as:

$$MI(A, B) = \sum_{a,b} P_{A,B}(a, b) \log \left(\frac{P_{A,B}(a, b)}{P_A(a) \cdot P_B(b)} \right) \tag{3.2}$$

We see that if P_A and P_B are independent, the sum is zero due to $\log(1) = 0$. We further see that maximizing this terms means maximizing the dependency between both distributions. This is the intuition behind image alignment algorithms based on MI: we aim to find the transformation between A and B that maximizes the term in equation 3.2. In other words, we want to find the transformation for which P_B can be best predicted based on P_A.
Algorithmically, two major solution approaches exist: either optimization methods like gradient ascent or patch-based sliding window search (computationally more expensive but trivial implementation). Due to time-constraints for the project at hand, we only looked into the second approach. The preliminary findings for this are discussed in section 4.1.2.

3.2 Pose Graph Optimization

Instead of only matching in between images taken at the same time, the images can be optimized in a global fashion by solving the pose graph optimization (i.e. bundle adjustment). As illustrated in figure 3.3, cross-spectral matches are combined with within-modality matches for both optical and thermal images in this case.

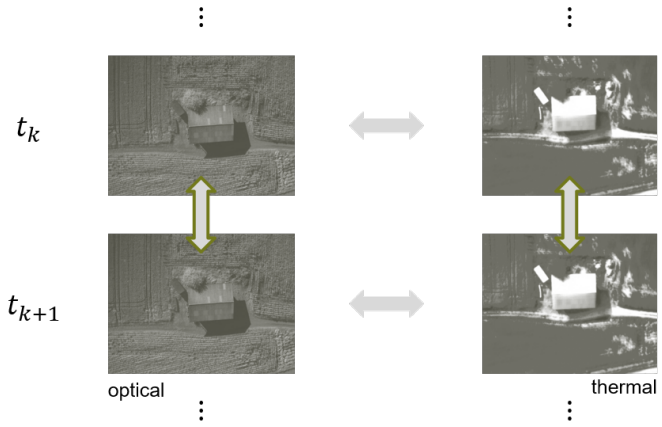


Figure 3.3: Pose graph optimization to align images globally

3.2.1 Pix4Dmapper

The simplest approach to get ground-truth maps from which we can sample patch correspondences, is to use a commercial aerial mapping software like Pix4Dmapper. This tool has shown outstanding performance on optical[3] and thermal[4] images separately. For all mapping tasks, Pix4Dmapper requires the original images, the poses they were taken at (geo-location) as well as the camera intrinsics and distortion parameters. The software then solves the pose graph optimization problem by matching features across images and triangulating the resulting feature tracks to 3D landmarks. Using reprojection error terms, the poses and intrinsics can be optimized in an iterative manner. Finally, we can generate orthomosaics and 3D point clouds of the mapped area.
Following this work-flow, Pix4Dmapper generates good maps when used for optical or thermal separately. In figure 3.4 we show two examples for such maps.

Figure 3.4: Example maps for optical (left) and thermal (right) processed separately

However, generating good maps which are aligned across spectra is a different, apparently more difficult task. For this, the work-flow is as follows: one first creates two sub-projects containing images from only one modality each. Both sub-projects are optimized independently to arrive at an optical and thermal pose graph. Then, both sub-projects are merged into one main project, where they are re-optimized. Finally, the orthomosaics and point clouds can be generated. Here, one can choose the settings as such that Pix4Dmapper only uses the higher resolution optical images to estimate the 3D shape of the environment, but then uses both optical and thermal images to create two different meshes.
Pix4D offers support documents online that explain the entire process for single-modality maps in great detail. Further, a list of requirements for the image sequences to be mapped is given. Those requirements are a lot stricter for thermal imaging than for optical, since in general the matching and triangulation tasks are much harder in thermal. The minimum resolution is 640x480, which is just below what the FLIR Tau 2 used for this project offers, and a very high overlap of 90% in between the images is required, both in flight direction and sideways.
However, when it comes to the merging of two sub-projects from different modalities, less guidance is available online. Nevertheless, there is some directions to follow when running into problems such as:

- Images should not suffer from motion-blur
- Trying out both the standard and alternative calibration methods
- Setting the intrinsics optimization to *All prior* to avoid strong corrections

But, even when following these guidelines and after trying out many different setups, we did not achieve to get globally consistent, cross-spectral maps for the combined project. More details follow in section 4.2.1.

[3] https://support.pix4d.com/hc/en-us/articles/360000235126-Example-projectslabel1
[4] https://support.pix4d.com/hc/en-us/articles/360000235126-Example-projectslabel13

3.2.2 Landmark Triangulation Pipeline

Since we could not solve the pose graph optimization with Pix4Dmapper, we decided to implement our own landmark triangulation pipeline. The big advantage here is that we have full control over the optimization and ways of how to incorporate cross-spectral matches into the process. In contrast to the work done by *Chatton et al.* depicted in figure 1.2, we aim to include both optical and thermal poses directly into the optimization procedure as illustrated in figure 3.3. The idea behind this is to achieve greater robustness against varying delays between the optical and thermal camera shutter times and against errors in the predefined extrinsics of the setup. In the long run, this pipeline also allows us to calibrate the thermal camera during the flight, which was not possible with the previous approach.

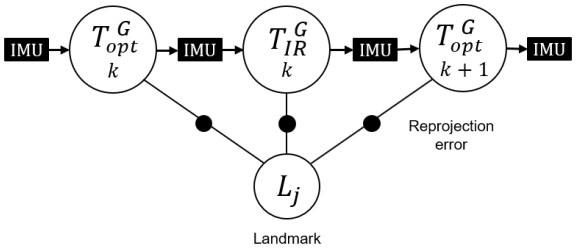


Figure 3.5: Pose Graph including both optical and thermal poses

The algorithm consists of the following steps:

(a) First we find within modality matches by matching sequential images using global SURF feature detection and description and then applying nearest-neighbour matching. We use a two-stage outlier rejection that first applies RANSAC with loose parameters to filter out obvious outliers and then uses LMEDS to refine the found homography between the two images. An example for optical and thermal matches can be seen in figure 3.6 and figure 3.7, respectively. The inliers are then propagated over time to form feature tracks, as can be seen in figure 3.8.

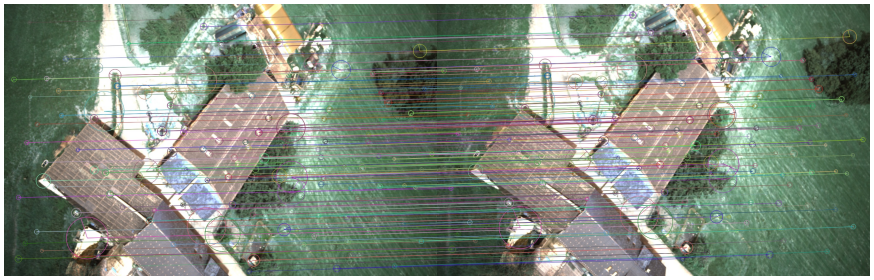


Figure 3.6: With-in modality matches for optical camera

(b) In a first attempt to solve the pose graph, we only use IMU and GPS measurements to get initial pose estimates for the optical and thermal poses. All pose graph optimizations are solved using a Levenberg-Marquardt optimizer built into GTSAM[5]. Based on the optical feature tracks generated in the previous step, we can now refine the optical poses by triangulating 3D landmarks. This works in a much more robust manner for optical images than for thermal, due to the higher resolution and more accurate intrinsics of the optical camera. We add the optical landmarks as reprojection error terms to the pose graph optimization and re-optimize the new graph.

[5]https://github.com/borglab/gtsam

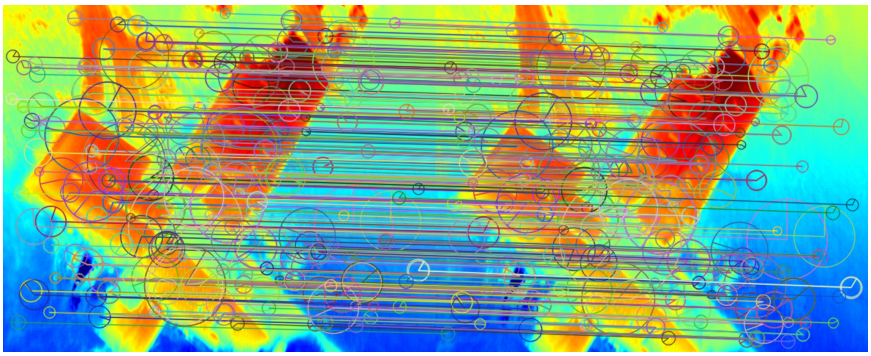


Figure 3.7: With-in modality matches for thermal camera

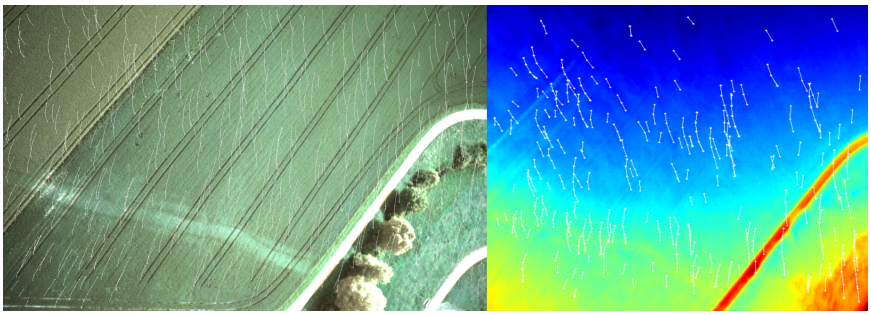


Figure 3.8: Track visualizations for both optical (left) and thermal (right) sequences

(c) Since we now have good estimates of the optical poses, we can apply a simple interpolation scheme to refine the estimates of the thermal poses as well. We use a simple linear velocity interpolation for the position values and Eigen's spherical linear interpolation (Slerp) to interpolate the orientation parameters which are represented as quaternions.

(d) Based on this good estimate of the thermal poses, we can use the thermal feature tracks to triangulate thermal 3D landmarks and add their reprojection errors to the overall optimization problem. The resulting optical and thermal landmarks can be seen in figure 3.9. Reoptimization with GTSAM leads to to a pose graph with global consistency within each modality and fairly good alignment across modalities due to the interpolation scheme. However, we have not yet added an explicitly cross-spectral component to the pose graph.

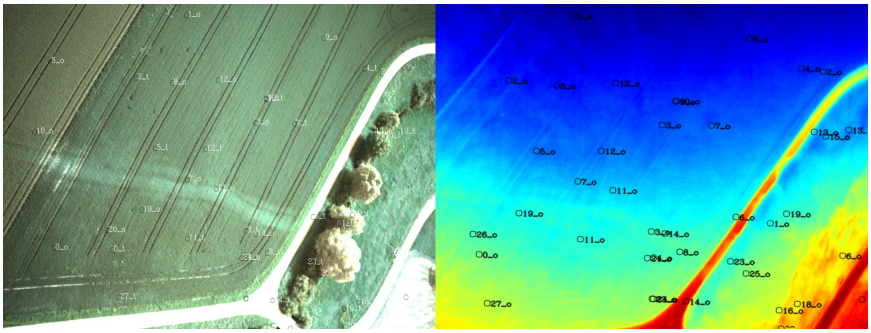


Figure 3.9: Reprojections of optical (left) and thermal (right) landmarks

(e) In a final step, we perform cross-spectral matching using the LGHD method

discussed in section 3.1.1 to correct for inconsistencies across the spectral domain gap. We do this by reprojecting optical and thermal 3D landmarks to the other modality while also matching their original 2D coordinate in a cross-spectral manner. For this, we use the reprojection of the other modality landmark as a position prior to better constrain the matching problem. We then only match points within close proximity of this prior, shown as green squares in figure 3.10. We now use the corrected, corresponding points to triangulate cross-spectral 3D landmarks. With these, we solve the final pose graph optimization as shown in figure 3.5. The resulting maps are shown in chapter 4.

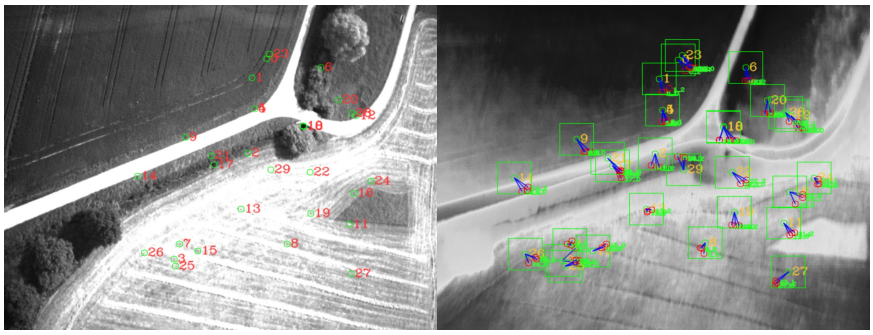


Figure 3.10: Cross-spectral correction in optical (left) and thermal (right) image

Chapter 4

Ground-truth Generation Results

4.1 Cross-spectral Image Alignment

4.1.1 Log-Gabor Histogram Descriptor

To evaluate the performance of LGHD, we compare it to SURF both qualitatively and quantitatively on the following data sets:

ICIP This data set introduced by *Aguilera et al.* [16] consists of ground-based, pre-aligned optical and thermal images of mostly human-built structures.

EOH Similarly to ICIP, this data published by *Aguilera et al.* [15] covers ground-based, pre-aligned cross-spectral scenes of human-built structures.

WARM (ours) This is our own data set generated as part of this project. It consists of aligned, aerial cross-spectral image pairs taken from the UAV.

Looking at the qualitative comparison shown in figures 4.1 to 4.6, we clearly see that LGHD outperforms classical, mono-modality matching methods by a large margin.

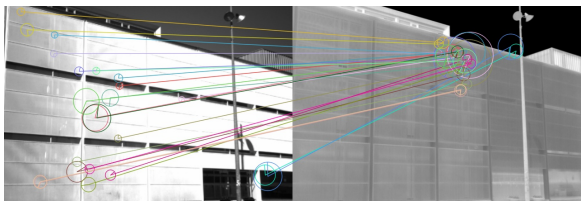


Figure 4.1: SURF matches for ICIP example images

Figure 4.2: LGHD matches for ICIP example images

In figures 4.1 and 4.2, we see that SURF completely fails to register the images taken from ICIP. In contrast, LGHD achieves a dense and well-distributed set of matches which captures the geometric relation between the two images.

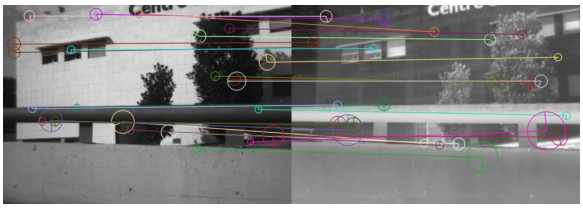


Figure 4.3: SURF matches for EOH example images

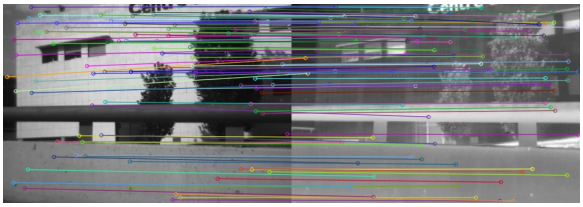


Figure 4.4: LGHD matches for EOH example images

For the example taken from EOH shown in figures 4.3 and 4.4, we see that SURF gets a sparse set of correct matches, but fails to correctly connect large regions of the images. Again, LGHD achieves a dense and spread-out collection of point correspondences.

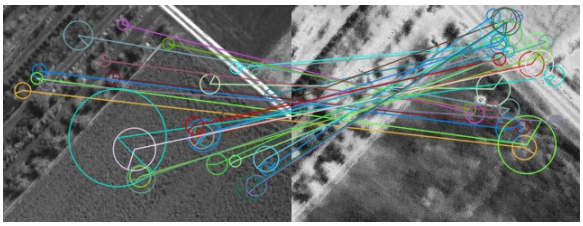


Figure 4.5: SURF matches for WARM example images

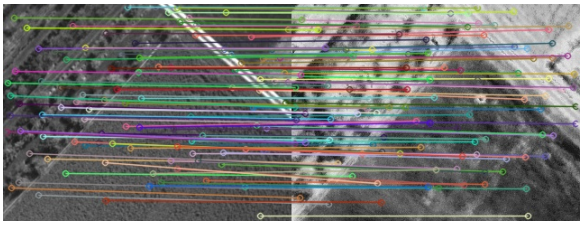


Figure 4.6: LGHD matches for WARM example images

From looking at figures 4.5 and 4.6, we see that SURF fails to register aerial cross-spectral images taken from WARM. LGHD however, achieves a high number of correct matches despite the low amount of texture and good features.
To further underline the large performance gain we achieve by using LGHD, we show the homography error for both LGHD and SURF for all data sets in table 4.1. Here, the homography error is computed as the L2-norm of the difference between the estimated and true homography matrices, hence:

$$E_H = \left|\left|H_{est} - H_{GT}\right|\right|_2 \tag{4.1}$$

Since we know that the images at hand are perfectly aligned, we can set $H_{GT} = I_3$. Both LGHD and SURF matches are fed into a two-stage outlier rejection pipeline

using RANSAC and LMEDS as described in section 3.2.2 where we optimize all parameters for each method independently. The results found using data from the WARM data set are to be interpreted carefully due to a bias towards LGHD, hence the asterisk*. The bias exists since there is no ground-truth alignment given for these images. To overcome this, we ran all frames through the LGHD-based image alignment and then filtered out only the perfectly aligned examples in post-processing. This showed to be very time-consuming, further motivating the implementation of a more robust pose graph optimization as explained in section 3.2.2. Roughly every third frame from our UAV footage would lead to a well-aligned image pair. Naturally, the post-processing introduces a bias since it consists of images where LGHD already proved to work well on. Hence, the results for this data are to be interpreted with caution. It also tends to over-represent feature rich parts of the UAV trajectory versus images of plain fields.

Method	Data set	E_H
SURF	ICIP	211.7
LGHD	ICIP	**8.6**
SURF	EOH	180.0
LGHD	EOH	**21.7**
SURF	WARM	146.9*
LGHD	WARM	**16.0***

Table 4.1: Results for homography estimation with LGHD and SURF

Besides the ability to use LGHD from within the existing code base used by the fixed-wing UAV team of ASL, implementing LGHD in C++ as described in section 3.1.1 also leads to an strong increase in performance. When comparing our C++ implementation to the original algorithm [16] written in MATLAB, we see a decrease in processing time of around 3.5x when applying the method using around 4'000 keypoints.

Language	Timing
MATLAB	70.05s
C++ (ours)	**21.59s**

Table 4.2: Timings for original LGHD implementation and ours in C++ when aligning optical and FIR images with resolution 640x480 pixels

Finally, we can conclude that even though LGHD allows for a lot better and faster cross-spectral image matching, the post-processing step to filter out bad matches remains necessary. This means that the direct cross-spectral image alignment approach does not scale well enough to be used to generate large amounts of training data, which is ultimately why we chose to also solve the overall pose graph optimization problem.
In figures 4.7 and 4.8 we see how a set of point correspondences that looks mostly correct at first, still leads to significant mismatch between the overlapping optical and thermal images. Special attention should be directed to the area in the blue circle which is placed over each image in the same location.

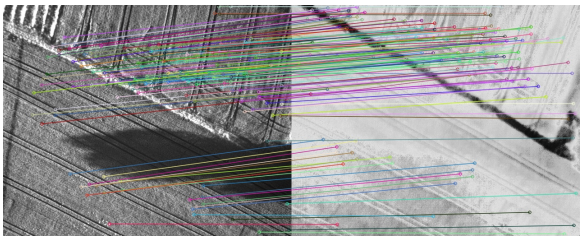


Figure 4.7: Inliers found for sample image pair from WARM

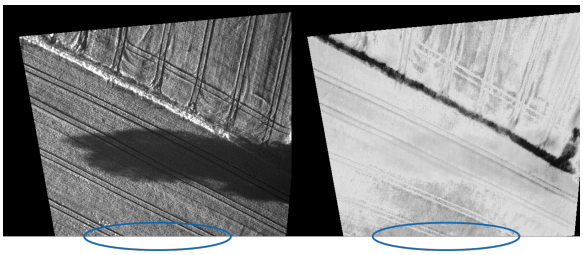


Figure 4.8: Poor alignment between optical and thermal aligned by LGHD

4.1.2 Mutual Information

As described in section 3.1.2, MI-based image alignment often employs some sort of global optimization method regarding the Mutual Information metric. Hence, the surface properties of said metric are crucially important for the convergence and correctness of such methods. Therefore, we investigate the MI surface by using a sliding window approach. For the cross-spectral images shown in figure 4.9, we first use LGHD to pre-align them and then choose a patch from the thermal image. Next, we slide that patch across every pixel in the optical image computing the MI score between the thermal query patch and the optical candidate patch.

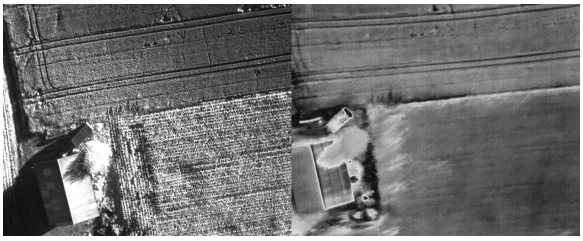


Figure 4.9: Example optical (left) and thermal (right) images

In figure 4.10, we see the first set of images and the query patches. Looking at the MI surface in figure 4.11, we see that the MI is the highest along the two parallel lines in the patch, as would be expected. This forms two sharp, parallel ridges near the right edge of the MI surface which indicates that the images feature strong visual aliasing effects. This is caused by the high similarity of objects as crop rows or field paths. For MI approaches this can be an issue since it leads to many local minima, especially for smaller patch sizes. Furthermore, we see an unexpected, extremely non-linear behavior of the MI score in the lower left corner. Despite these non-linearities, the global maximum MI score roughly coincides with the true location of the patch in the optical image.

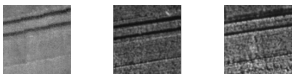


Figure 4.10: Thermal query patch (left), correct optical patch (middle) and optical patch with maximum MI (right)

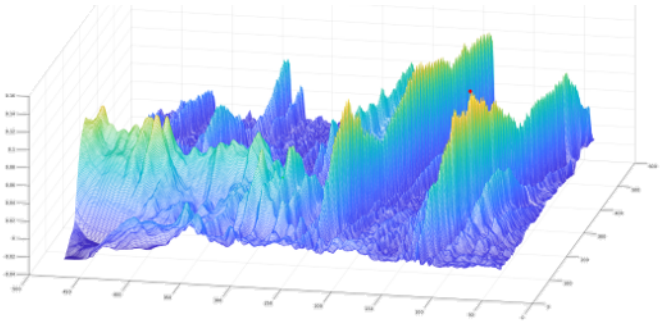


Figure 4.11: MI surface when sliding thermal query patch over optical image

Similarly, for the patches in figure 4.12 and 4.13, the MI surface is extremely non-linear. However, in this case the global MI maximum does not even coincide with the location of the correct patch. This indicates that comparing MI for one small, single patch is not a descriptive enough metric to register images.

Figure 4.12: Thermal query patch (left), correct optical patch (middle) and optical patch with maximum MI (right)

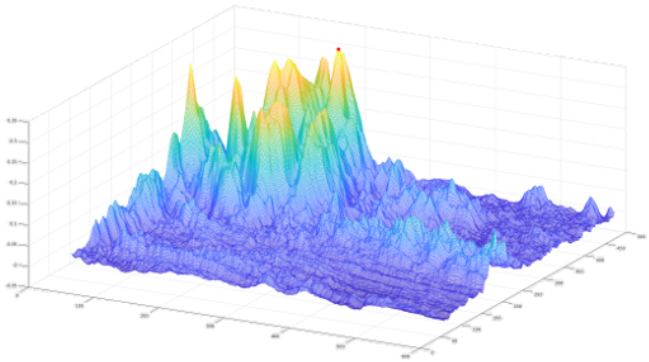


Figure 4.13: MI surface when sliding thermal query patch over optical image

From these experiments we can conclude the following:

(a) The convergence properties of a global optimizer might be very hard to tune due to the highly non-linear surface

(b) Even when converging to the global optimum, we might not find the correct patch correspondence based on a sliding window approach.

Finally, we note that a simple sliding window optimization based on MI is not an appropriate approach for our task. As explained in chapter 7, a constraint MI optimization might lead to much better results.

4.2 Pose Graph Optimization

4.2.1 Pix4Dmapper

As already mentioned in section 3.2.1, we were not able to generate cross-spectral maps with Pix4D. One major problem with the data we used is the low overlap between the images. This was further confirmed to be crucial by the Pix4D support team. Nevertheless, the high amount of open discussions in online support forums hints strongly towards unsolved issues and unsatisfactory performance of Pix4Dmapper when creating cross-spectral maps. Most discussions usually evolve around the easier task of optical to NIR matching, since NIR is widely used in crop observation these days. But even for this use case, it seems to be difficult to get Pix4Dmapper to perform well. In chapter 7, we list ideas to make this work in the future.

4.2.2 Landmark Triangulation Pipeline

We already show intermediate results for the steps of the algorithm in section 3.2.2. In figure 4.14 and 4.15 we compare the resulting orthomosaics from both the optical and thermal modality. These maps are based on both optical and thermal landmarks optimized in a joint manner, just like illustrated in figure 3.5. We can see a high overall consistency of the individual maps as well as the good alignment between both maps. However, we also see an abrupt change of color in the field just above the street. This is due to the absence of color balancing, which should be added in the future (see chapter 7).

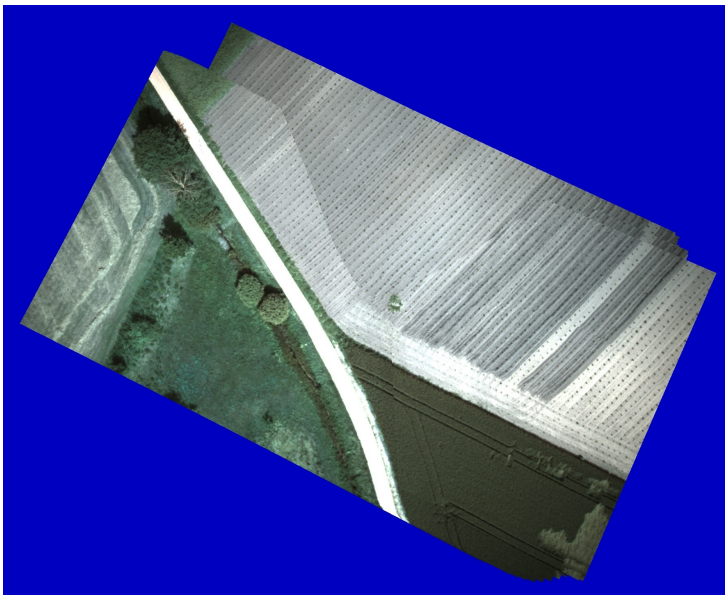


Figure 4.14: Optical orthomosaic based on 30 frames

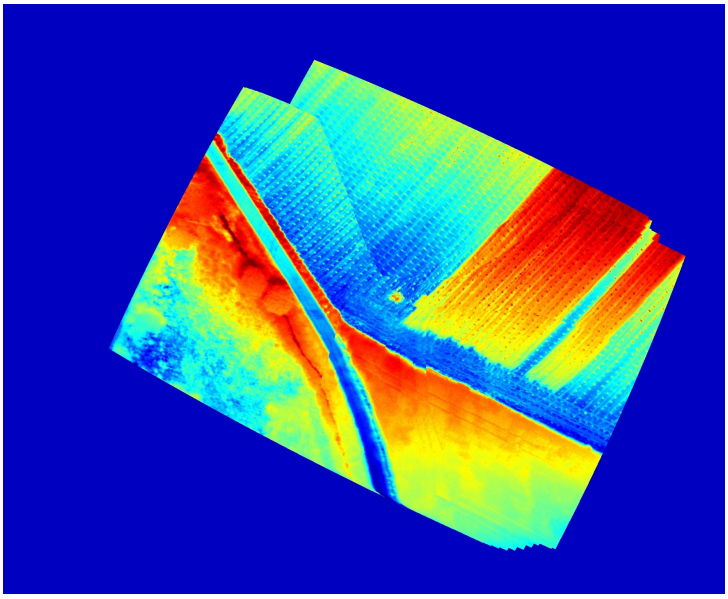


Figure 4.15: Thermal orthomosaic based on 30 frames

4.3 Resulting Training Data

As a result for the first phase of the project, the following data sets were collected or generated for later use as cross-spectral training data:

ICIP & EOH Ground-based, human-built structures (see section 4.1.1).

WARM Aerial UAV images of fields and human-built structures (see section 4.1.1). This data set consists of more than 1'000 image pairs produced by using the LGHD pipeline described in section 3.1.1 and the post-processing step to filter out bad matches as shown in figure 4.7. As normal for manually post-processed image collections, there remains some image pairs with small misalignment between them.

VOT-RGBTIR Security camera footage of mostly urban environments with pedestrians and cars (example shown in figure 4.16). Due to this data being highly repetitive, we only use every of each 10th frame of video clip for training.

Since we showed example image pairs for all data sets except VOT-RGBTIR, we only show one more example cross-spectral pair for said data set here.

Figure 4.16: Optical (left) and thermal (right) image taken from VOT-RGBTIR

In total, we collected or generated over 4'000 cross-spectral image pairs from the optical and FIR spectrum. To our knowledge, this is the cross-spectral image collection with the most diverse set of scenes and the first one to include a large amount of high quality aerial images.

It should be noted, that a lot more data exists for optical to NIR matching, but as shown in figure 1.6 that task is a lot easier than matching optical to FIR. A list of all available data sources that we found is given in the appendix.

Chapter 5

Learned Cross-Spectral Matching Methods

In the following two chapters, we use the generated training data to train a DNN in order to solve the cross-spectral matching task in real-time and with increased performance.

5.1 Vanilla SuperPoint

The original (vanilla) SuperPoint paper [5] describes a fully-convolutional, single encoder deep neural network which can be trained in self-supervised manner using the novel *Homographic Adaption* approach. It solves the homography estimation task in an end-to-end manner, meaning that it does both interest point detection as well as feature description. Figure 5.1 gives a good overview of the architecture.

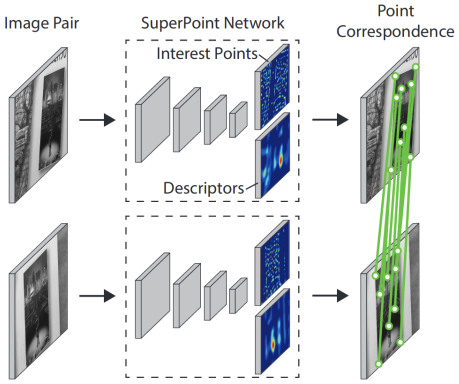


Figure 5.1: Overview of SuperPoint architecture (figure taken from [5])

The network is originally trained on MS-COCO [27], which consists of roughly 80'000 high resolution images of a wide variety of everyday objects and sceneries. The authors report a frame rate of 70 frames per second on a Titan X GPU.
The only learned component of the network is its encoder, based on which a fixed detector and descriptor head generate both the keypoints and descriptor vectors. The encoder has a VGG-like structure [24] consisting of convolutional and pooling layers. The encoding process shrinks the input image of size $W \times H$ to an intermediate representation of size $W/8 \times H/8$. The necessary upsampling is unlearned in order to reduce training and inference time. While the detector head only outputs

a one-dimensional probability indicating the likelihood of each pixel being a corner, the descriptor head outputs a categorical distribution over 64 classes and one dust bin used for pixels that are not interest points. Figure 5.2 offers a detailed view of the architecture with its two different heads.

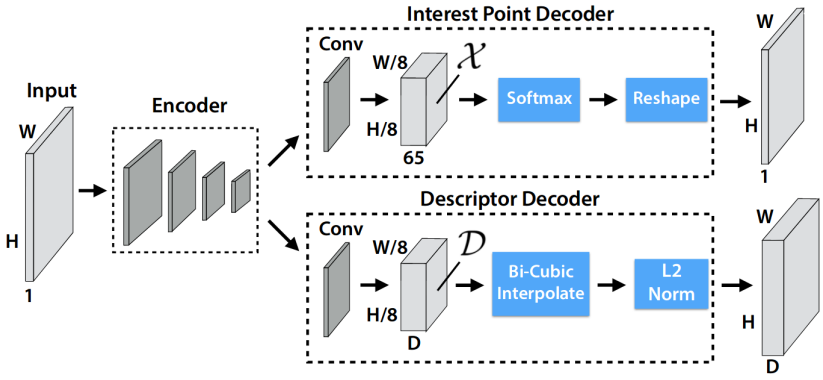


Figure 5.2: Details of the SuperPoint architecture (figure taken from [5])

The overall loss function that is optimized during training is a weighted sum of interest point losses L_p and a descriptor loss L_d. We denote the intermediate representations in the interest point decoder and descriptor decoder as X and D, respectively. We write Y for the 0/1-labels describing if a pixel is an interest point or not. Further, S denotes the set of point correspondences between two images. Using this notation, we can write the total loss for an image I and a warped version of itself I' as:

$$L(X, X', D, D', Y, Y', S) = L_p(X, Y) + L_p(X', Y') + \lambda L_d(D, D', S) \tag{5.1}$$

Here, λ is a tuning parameter to balance the loss types. The individual losses L_p and L_d are defined in section 5.1.2 and 5.1.3.
The training of the network can be separated into three major stages, as seen in figure 5.3. In the following, we describe each of the stages in more detail so that we can later explain how we propose to change each step to use SuperPoint for cross-spectral matching.

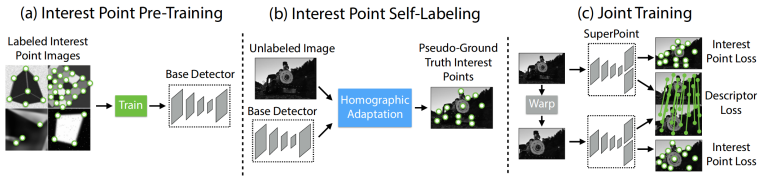


Figure 5.3: Training stages of SuperPoint (figure taken from [5])

5.1.1 Interest Point Pre-Training

The goal of the first training stage (see (a) in figure 5.3) is to initialize the encoder by teaching it what corners are in a fully supervised fashion. For this, we render images from known geometric objects such as cubes or rectangles and show those to the network. Since we know the ground-truth location of the corners, we can improve the network's representation of corner-like points iteratively. The geometric

objects are rendered in various scales and orientation in order to augment the data set we sample from. Further, different kinds of noise and photometric variations such as lighting changes are introduced to generalize the data. This results in the so-called Base Detector, which is used an initialization for the following stages.

5.1.2 Interest Point Self-Labeling

We can now use the Base Detector to generate pseudo ground-truth to start the self-supervised training (see (b) in figure 5.3). For this, we take each input image and run it through the Base Detector, resulting in a set of interest points. We do the same for many warped and otherwise altered versions of each input image. By comparing the locations of detected interest points across all homographic and photometric versions of each individual input image, we can find out which interest points are most repeatable, i.e. reappear under a wide variety of changes applied to the image. Such interest points are more stable, hence we reinforce the network to find keypoints that are found often. This novel procedure introduced by the authors is called *Homographic Adaption* and is depicted in figure 5.4. Stage two of the training process results in a model called MagicPoint.

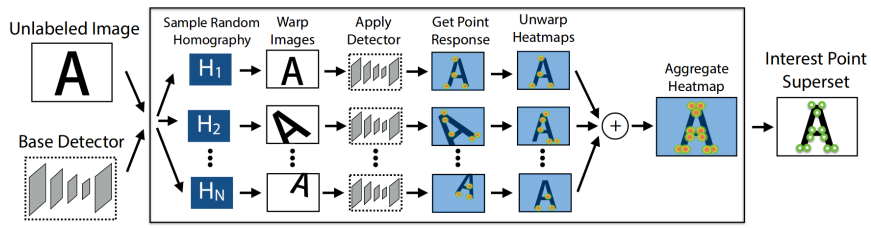


Figure 5.4: Homographic Adaption for self-supervision (figure taken from [5])

The interest point loss function used both for the initial pre-training and self-labeling is a simple cross-entropy loss over the 64 classes and the dust bin. Using $x_{hw} \in X$, the authors define L_p as:

$$L_p(X, Y) = \frac{1}{H_c W_c} \sum_{h=1,w=1}^{H_c,W_c} -\log\left(\frac{\exp(x_{hwy})}{\sum_{k=1}^{65}\exp(x_{hwk})}\right) \tag{5.2}$$

Here, H_c and W_c denote the dimensions of the down-sampled, encoded version of the input image.

5.1.3 Joint Training

So far, only the detector head has contributed to the loss to be optimized during training. In the third and final training stage, we now add descriptor based loss terms by matching between the input image and a warped version of the same image (see (c) in figure 5.3). The process is hence similar to Homographic Adaption, but here we only warp once per loss computation instead of aggregating many warps to one loss term.
Using $d_{hw} \in D$, $s \in S$ and two image I and I', the descriptor loss can be written as:

$$L_d(D, D', S) = \frac{1}{(H_c W_c)^2} \sum_{h=1,w=1}^{H_c,W_c} \sum_{h'=1,w'=1}^{H_c,W_c} l_d(d_{hw}, d'_{h'w'}, s) \tag{5.3}$$

where

$$l_d(d_{hw}, d'_{h'w'}, s) = \lambda_d \cdot s \cdot \max(0, m_p - d^T d') + (1 - s) \cdot \max(0, d^T d' - m_n) \quad (5.4)$$

This term equals a hinge loss or support vector machine (SVM) with positive decision margin m_p and negative decision margin m_n as tuning parameters. For matching points with $s = 1$, we therefore push the scalar product of the descriptors d, d' to be big, i.e. the descriptors to be similar. For non-matching points with $s = 0$, the scalar product is pushed to zero, i.e. the descriptors are encouraged to be different from each other.

5.2 Cross-spectral SuperPoint

Due to the self-supervised training approach of SuperPoint, the adjustments to make it perform cross-spectral matching are quite straightforward. In the following, we explain what changes we propose to apply to each stage of the training procedure shown in figure 5.3.
Since the original paper [5] does not come with open-source code, we chose to base our work on a publicly available TensorFlow implementation by *R. Pautrat* and *P. Sarlin*[1]. Due to the use of a more complex interpolation (bi-cubic instead of linear) and other implementation differences, this network only achieves a frame rate of 2 frames per second. For true real-time performance, one would therefore have to further optimize the network's implementation.

5.2.1 Interest Point Pre-Training

For this stage of the pipeline, we decided to leave everything as is. This is based on the simplifying assumption that corners look more or less the same for both optical and thermal images. As explained in section 1, this assumption is not true in general due to the non-linear intensity changes between optical and thermal frequencies.

5.2.2 Interest Point Self-Labeling

The only thing to be changed in this stage is that when performing Homographic Adaption, instead of warping from an optical input image to a warped version of said image, we warp to the corresponding thermal image of the input. We also do this in the reversed order, hence warping from a thermal input image to its optical correspondence. Both the homographic and photometric augmentation techniques are applied exactly like for vanilla SuperPoint. The decision whether to warp from optical to thermal or from thermal to optical is random, so over the overall training we do both directions equally often.

5.2.3 Joint Training

Just like for the second stage, instead of warping from optical to optical we apply a cross-spectral warping scheme here. Again, we choose by random whether we warp from optical to thermal or from thermal to optical. Everything else is kept exactly as is for the vanilla version. For both steps, we do not allow with-in modality matching as this showed to make the network only learn the much easier non-cross-spectral matching task. The same behavior was observed when training on optical to optical samples, as seen in figures 6.8 and 6.7.

[1] https://github.com/rpautrat/SuperPoint

All adjustments made to the pipeline are illustrated in figure 5.5.

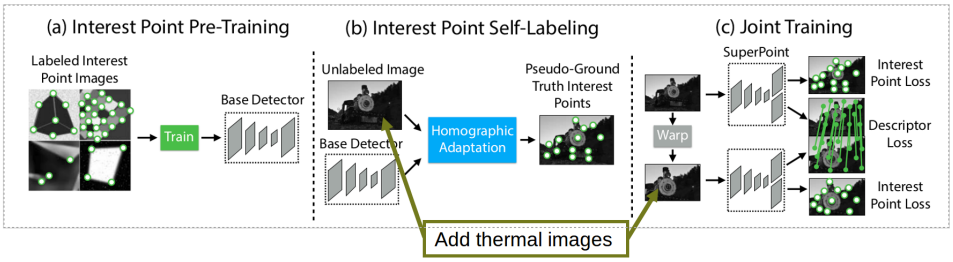


Figure 5.5: Cross-spectral training of SuperPoint (figure adapted from [5])

Throughout the project, we often experienced this stage to account for most of the interest point learning. This means that the detection of keypoints improved the most when training on the overall task in an end-to-end manner.

Chapter 6

Learned Cross-Spectral Matching Results

We conducted different experiments to test the changes made to the pipeline and access the performance of SuperPoint in regards to cross-spectral image matching. Those experiments and their outcome are described in the following sections.

6.1 Training and Testing Procedure

We mostly use the standard training parameters of the TensorFlow implementation[1] by *R. Pautrat* and *P. Sarlin*. The only significant parameters to be changed are the amount of variation in the distribution from which we draw homographies during training (see experiments 5 and 6) and the number of training epochs.

The training pipeline is the same for all following experiments: we train the Base Detector, use it to generate pseudo ground-truth keypoints #1 and based on that train MagicPoint. We then generate pseudo ground-truth keypoints #2 using the MagicPoint model, based on which we then train SuperPoint. As is standard, we choose fully distinct training and test sets for all data set combinations used in the following experiments. All training was run on the GPU cluster by ETH named Leonhard[2]. Based on availability, training took place on single Nvidia Titan X GPUs, multiple such units or Nvidia Tesla V100 cards. We use job chaining[3] to fully automatize the training pipeline and reduce manual work.

For all experiments, we show qualitative examples with the found keypoints marked as red circles and the matches shown in green. For all the following figures, the optical image is shown on the left side while the thermal correspondence is plotted to the right. Further, we show the homography correctness as a quantitative measure in table 6.1. This metric measures the percentage of correct homographies that were found based on the matches outputted by the network. The correctness of a homography is measured upon certain error tolerance, the so-called correctness threshold. For experiments 1a and 1b, we set this threshold to 5 pixels as the authors do. Since the overall matching performance degrades for the later experiments, we have to increase the threshold to 30 pixels to keep all homographies from being rejected. All quantitative results are shown in table 6.1.

[1] https://github.com/rpautrat/SuperPoint
[2] https://scicomp.ethz.ch/wiki/Leonhard
[3] https://scicomp.ethz.ch/wiki/Job_chaining

6.2 Experiment 1: Sanity Check Changes

As a first step, we make sure that the changes described in section 5.2 do not break the pipeline with respect to the original, optical to optical task. This is a simple sanity check to make sure that our code changes do not introduce unexpected, incorrect behavior. For this, we train the adjusted network with optical images only. We pretend to present it optical and thermal images, but the thermal images are really just copies of the optical ones. Doing so, we can test the new pipeline with respect to the original task which we expect to be solved effectively.

6.2.1 1a: MS-COCO

First, we train the network using the roughly 80'000 images from MS-COCO [27] just like in the original paper. The resulting model achieves a homography correctness of 60.4% (see table 6.1). This is close to the performance reported by the authors which is 63.1%. An example of an optical image pair matched with our adapted pipeline is given in figure 6.1. Even for very strong warps between the images, the model often finds good point correspondences. This shows that for the mono-modality case, the network learns scale- and rotation-invariant descriptors.

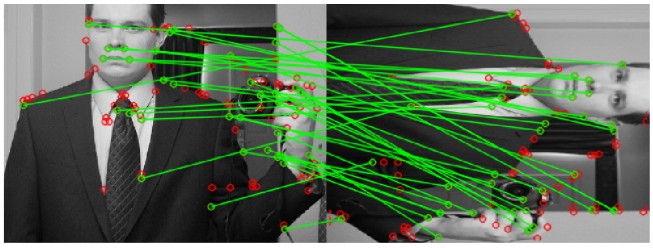


Figure 6.1: Experiment 1 trained and tested on MS-COCO

6.2.2 1b: MS-COCO and Optical UAV Images

Interestingly, when applying the same model trained on MS-COCO to optical images from our UAV, the homography correctness drops drastically to 21.5%. This is caused by the great difference between the two data sets, since in general scenes from MS-COCO are much richer in features than top-view images of rural regions used for agriculture. Figure 6.2 shows a well-working example taken from the UAV data.

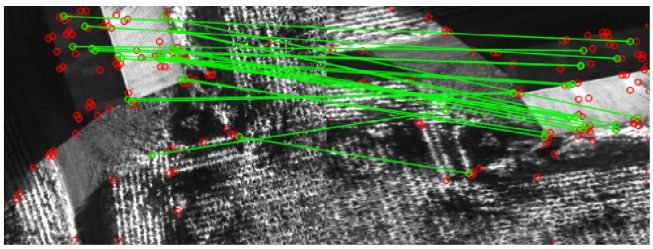


Figure 6.2: Experiment 1 trained on MS-COCO and tested on optical aerial images

These first two experiments indicate that the new pipeline is still capable to learn the task of homography estimation in the optical to optical case, so the changes do not break its functionality. We can further note that learning does transfer to unseen, significantly different data, but only in to a limited very small degree.

6.3 Experiment 2: Cross-Spectral Mix

Naturally, the next experiment is to train and test the network on the cross-spectral data we generated and collected in the fist phase of this thesis. In a first setup, we combine all the cross-spectral data available which adds up to roughly 4'000 images. We call this combination of data set *Cross-Spectral Mix*. Qualitatively, we found that the model captures some similarities across the spectra, but does not converge to a well-generalizing cross-spectral matcher. Some examples are given in figures 6.3 to 6.5. The resulting homography correctness of 16.0% is fairly low, even after increasing the correctness threshold to 30 pixels.

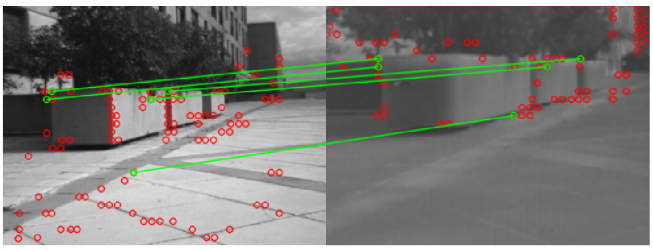


Figure 6.3: Experiment 2 tested on a sample from ICIP

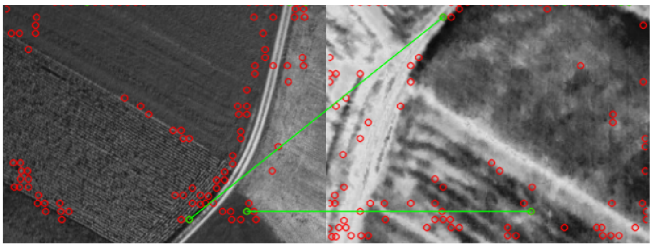


Figure 6.4: Experiment 2 tested on a sample from WARM

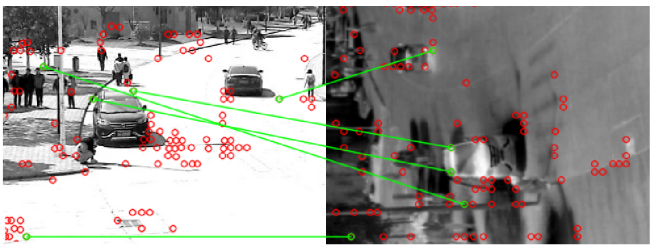


Figure 6.5: Experiment 2 tested on a sample from VOT-RGBTIR

We conclude from this that after training on all the cross-spectral data at once, the network is able find some, but very few correct point correspondences. However, we are not yet able to teach it building well-generalizable descriptors to robustly match points across spectra and under strong translational and rotational warps.

When looking at the loss curves for this experiment shown in figure 6.6, we observe a fast drop in all loss components. However, the detector and therefore the overall loss then stabilize on a relatively high value. From this, we formed the hypothesis that the amount of data we were using was simply too small to get the network to fully converge. More clearly, we suspected that the network is simply reusing already seen data over and over again which does not help it to learn.

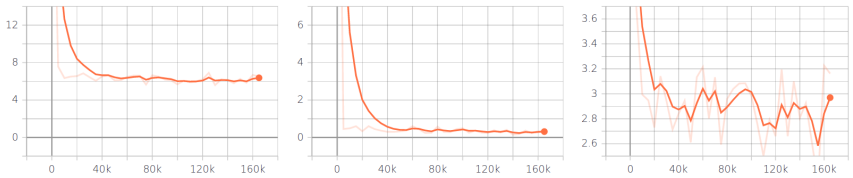


Figure 6.6: Loss curves for average (left), descriptors (middle) and detector (right) for experiment 2

6.4 Experiment 3: Cross-Spectral and MS-COCO

A simple approach to test the hypothesis formed in experiment 2 is to extend the amount of data by adding optical to optical samples from MS-COCO. By adding as many images from MS-COCO as cross-spectral ones, the total number of images is increased to almost 10'000. We experimented both with training first on MS-COCO and then on cross-spectral data as well as training in a fully joint manner, i.e. mixing the cross-spectral and MS-COCO data. For the former, no learning was possible after the first phase due to the gradients diverging. We suspect this to be caused by a large and sudden change in the optimization target. For the second approach, the learning did not diverge but qualitative analysis (see figure 6.7 and 6.8) showed that the network only picks up the far easier optical to optical task. Meanwhile, the learning for cross-spectral matching does not benefit at all from augmenting the data with mono-modality samples. This is also reflected in the very low homography correctness of 2.5% when tested on only the mixed cross-spectral data. The loss behaves almost identical than in experiment 2 which is why we do not show the loss curves again.

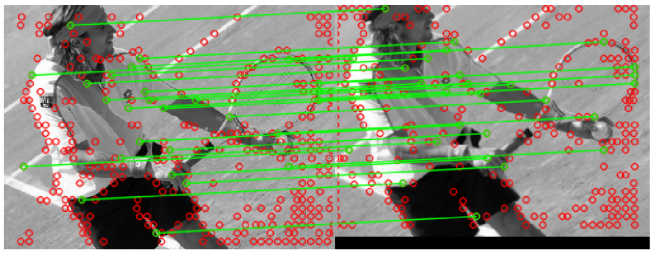


Figure 6.7: Model from experiment 3 tested on MS-COCO sample

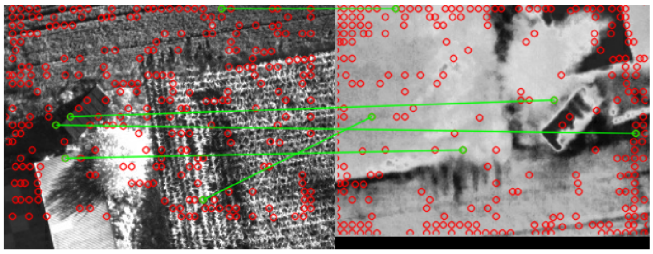


Figure 6.8: Model from experiment 3 tested on WARM sample

From experiment 3, it can be noted that it is a lot easier to teach SuperPoint the optical to optical matching task than the cross-spectral one. We did not achieve to use transfer learning based on the optical to optical task to the cross-spectral matching problem.

6.5 Experiment 4: Network Expressiveness

After seeing these partially unsatisfactory results, our next hypothesis was that having only a single encoder, the network might simply not be expressive enough to capture the non-linear transformations from optical to thermal descriptors. In order to test this, we choose a single cross-spectral image pair from a data set and only train and test on this single image. Basically, we dramatically overfit the network to a single image pair. Since for these experiments we only train and test on one single image, showing a percentage of correct homography estimations does not make sense. Hence, experiment 4 does not appear in table 6.1.

6.5.1 4a: ICIP

For ICIP, we found that the network was not able to match points in the case where there is a significant warp between both images. Figure 6.9 shows almost only outliers which indicates either missing expressiveness of the network or missing rotational invariance of the learned descriptors.

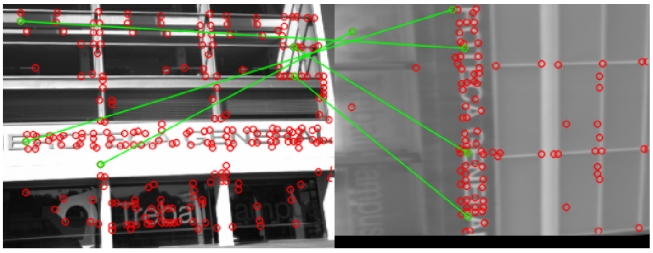


Figure 6.9: Experiment 4 overfitted and tested on ICIP

In contrast, when testing on an image pair with zero warp between them, a rich set of matches is found as illustrated in figure 6.10. During training, the network still saw warped versions of the image, this is hence not a purely non-informative task where the networks simply learns to match to the same coordinate in both images. Rather, we see this as an indicator that the problem is indeed a missing invariance against strong rotations and other transformations. As explained in experiment 5, this could be solved by using less extreme homographies during the last two training stages.

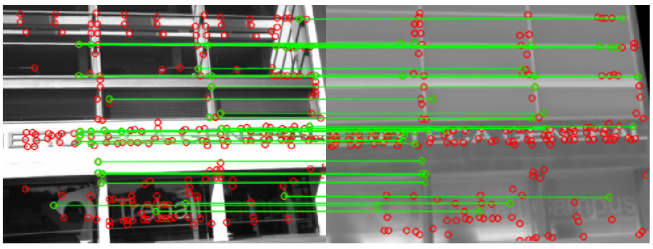


Figure 6.10: Experiment 4 overfitted and tested on ICIP with no warp applied

6.5.2 4b: WARM

A similar behaviour was found when using a sample from our own data set WARM. The test image pair shows very little geometric distortion from the optical to the thermal image, like the second example shown for ICIP. Again, the model is able to match many points correctly to each other as shown in figure 6.11.

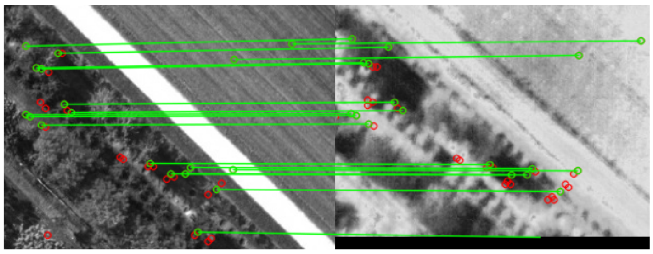


Figure 6.11: Experiment 4 overfitted and tested on WARM

The loss curves for such overfitting training procedures, see figure 6.12, show that the network fully converges and ends up with close to zero loss.

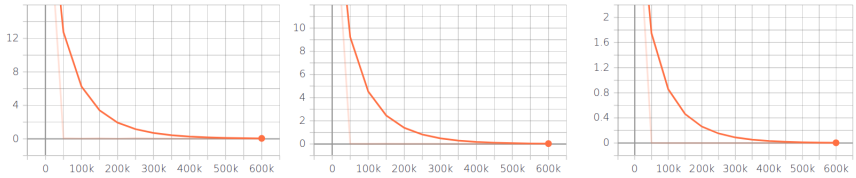


Figure 6.12: Loss curves for average (left), descriptors (middle) and detector (right) for experiment 4

Experiment 4 shows that the single-encoder architecture of SuperPoint is expressive enough to capture cross-spectral matches, at least for some of the data sets. Furthermore, we can observe the same extremely high sensitivity towards rotational warps as seen in experiment 2 when training with highly varying homographies.

6.6 Experiment 5: Only VOT-RGBTIR

Our next hypothesis targeted (a) the missing rotational invariance of the learned descriptors seen in experiment 2 and 4 as well as (b) the high variance within data sets which might keep the model to converge to one single, reasonable optimum performing cross-spectral matching. To cope with problem (a), we decrease the amount of rotation applied to the images when warping during Homographic Adaption and the joint training step. This makes the overall task easier and we expect to see better results when training for the same number of iterations. In order to improve (b), we train our pipeline only on subgroups of the cross-spectral data. In this experiment, we only use the image pairs from the VOT-RGBTIR data set.

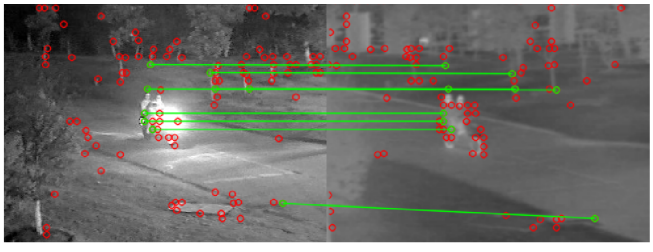


Figure 6.13: Experiment 5 trained and tested only on VOT-RGBTIR

The resulting model shows much better overall matching performance as shown by example in figure 6.13 and 6.14. Accordingly, the homography correctness increases to 22.0% which is significantly better than the corresponding result in experiment

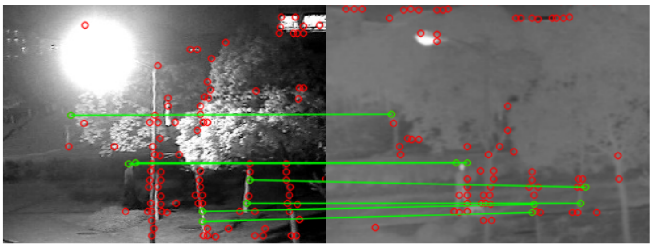


Figure 6.14: Experiment 5 trained and tested only on VOT-RGBTIR

2. However, if we compare this value to the performance achieved by the optical mono-modality model, it is clearly not good enough in order to have to network perform cross-spectral in a robust and generalizable fashion.

6.7 Experiment 6: Only WARM

We then performed exactly the same steps as in experiment 5 on our own data set WARM with cross-spectral aerial image pairs. During training, we again draw homographies from a more constrained distribution to make the task easier and better adjust it to the requirements of the matching to be done while the UAV is flying. This network again shows increased performance compared to the one trained in experiments 2 and 3. Qualitative examples are shown in figure 6.15 and 6.16. The homography correctness of 22.7% is slightly above the one seen in experiment 5. But similar to experiment 5, it is not comparable to to the performance shown for the optical mono-modality task.

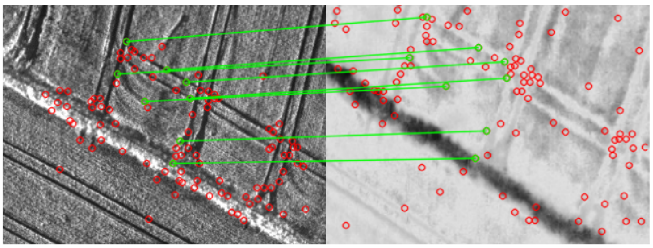


Figure 6.15: Experiment 6 trained and tested only on WARM

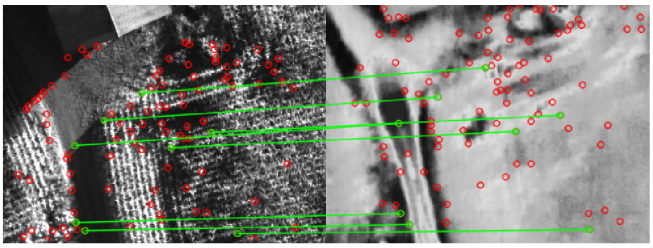


Figure 6.16: Experiment 6 trained and tested only on WARM

In figure 6.17 we plot the loss curves for such runs where we only train on a single data set and with reduced variance in the homographies. We see that the loss still does not quite converge but compared to experiment 2 and figure 6.6 it stabilizes on a much lower level. The reason for this is that since the network can focus on only one type of images, it converges to a point which is closer to the global optimum.

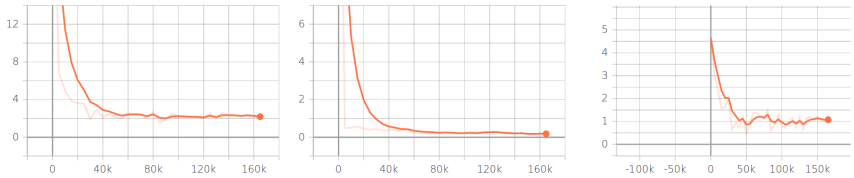


Figure 6.17: Loss curves for average (left), descriptors (middle) and detector (right) for experiment 6

Finally, from experiment 5 and 6 we can conclude that by reducing the range of homographies that we sample from, the matching performance for small homographies in the test data increases significantly. Further, using samples from very different data sets decreases performance significantly, which hints at bad generalization of the learned features.

Exp.	Training Data	Testing Data	Correctness Threshold	Homography Correctness
1a	MS-COCO	MS-COCO	5px	60.4%
1b	MS-COCO	Optical Aerial	5px	21.5%
2	Cross-Spectral Mix	Cross-Spectral Mix	30px	16.0%
3	Cross-Spectral Mix MS-COCO Subset	Cross-Spectral Mix	30px	2.5%
5	VOT-RGBTIR	VOT-RGBTIR	30px	22.0%
6	WARM (ours)	WARM (ours)	30px	22.7%

Table 6.1: Results for all experiments using cross-spectral SuperPoint

Chapter 7

Future Work

Due to time constraints, some approaches could only be explored partially. In future projects, we suggest looking into the following key areas of interest.

7.1 Ground-truth Data Generation

As mentioned in the very beginning of this work (see section 1.2), the global consistency of the thermal map could be improved by using color balancing techniques and by detecting loop closures. Especially for loop closure, having the thermal map registered to an optical map is helpful since optical features change much slower over time than thermal landmarks do.

The generated maps described in section 4.2.2 could be used as input to a rendering based data augmentation pipeline. In fact, one could define various trajectories for a virtual two-camera system to render images from to increase the amount of scale-, rotation- and perspective-invariance given in the ground-truth training data.

For future offline methods aligning images across spectra, it would interesting to look deeper into mutual information based approaches. Especially using MI in a more constraint optimization with multiple patches used simultaneously could lead to MI surfaces that are easier to optimize than the ones described in section 3.1.2.

As reported, we were not able to use Pix4Dmapper to generate cross-spectral maps. In the future this might be possible if the data capture would include optical and thermal Ground Control Points, hence markers with known location and temperature. An additional possiblity would be to collect data using quad-copter drones, which allow for much more precise trajectory planning and better control over the resulting forward and lateral overlap between the images.

7.2 Learned Cross-Spectral Matching

7.2.1 Existing Cross-Spectral SuperPoint

For the proposed cross-spectral variant of SuperPoint, it would be interesting to look deeper into the poor rotational invariance observed in our trials. As shown already, a more well-defined set of warps chosen at training time leads to less sensitivity of the matching capability of cross-spectral points. In fact, sampling homographies from a less spread-out distribution, hence having less extreme warps, is also inline with the intended use case of the method since we do not expect very large homographies across the cross-spectral images, since the images are taken from two rigid camera that are more or less aligned. Further optimizing this set of

homographies to sample from by taking a more in-depth look at the data captured by the UAV would certainly be interesting.

It would also make sense to train the adapted pipeline only on optical mono-modality images taken by the UAV. As mentioned, image from MS-COCO are in general much richer in features than the aerial ones. Hence, such a model would be a much better baseline to compare the cross-spectral, UAV only networks towards.

As seen in experiment 2, transfer learning does not work for optical only images. However, matching optical to NIR images is more similar to matching optical to NIR than compared to mono-modal optical matching. Hence, transfer learning based on optical to NIR images could help the network to solve the cross-spectral task.

Finally, we saw in experiment 5 and 6 that learning based on very different cross-spectral data sets decreases the performance of the network. Hence, instead of using other data set it would be better to generate even more aerial image pairs. For this, the pose graph based offline data generation method could be used. We expect it to produce much more robust matches and thus require less manual work to build even bigger data sets than the introduced WARM.

7.2.2 Pipeline Adaptions

A so far unused opportunity to introduce cross-spectrality into the pipeline is the training of the Base Detector. Instead of using the exact same set of rendered images as in the vanilla SuperPoint, one could introduce data augmentation techniques that specifically model the non-linear intensity transformations described in 1. To this end, one could look into combinations of photographic negatives and gamma / logarithmic transformations[1]. Intuitively, this would allow to model the frequency domain gap and improve the initialization of the subsequent training steps.

As described in chapter 2, most other networks used for cross-spectral matching tasks choose to use some sort of quadruple CNN architecture to separate the learning for each of the modalities. In future projects, some of those other architectures should be tested to evaluate their performance on our aerial optical to FIR data set WARM. Furthermore, the idea of quadruple CNNs could be incorporated into the SuperPoint pipeline by using two encoders instead of only one. One encoder would then be used to learn how to encode optical images, the other one would specialize to thermal images. The detector and descriptor heads could remain the same for such a setup and still be unlearned. This idea is illustrated in figure 7.1.

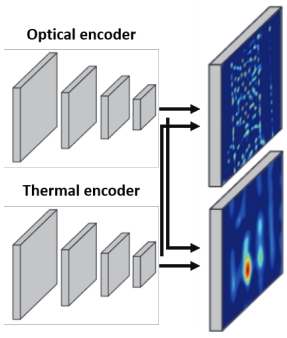


Figure 7.1: SuperPoint architecture with two encoders, one for each modality

Since the current implementation of SuperPoint only runs at around 2Hz, future efforts would also target to speed up the network. This could include using smaller encoder networks which in turn would raise the question of expressiveness of those.

[1] http://www.cs.uregina.ca/Links/class-info/425/Lab3/

Chapter 8

Conclusions

In this Semester thesis, we aim to solve the problem of autonomous soaring by predicting the location of thermal updrafts. We propose to build a thermal map by aligning thermal images to an existing optical map through cross-spectral matching. To better understand the challenges posed by this task, we summarize the differences between mono-modality matching and matching from optical to NIR and FIR. To solve the optical to FIR matching task, we use a two stage approach where offline algorithms provide training data for a much faster DL-based approach.

In the first stage, we implement and test four approaches to match optical and thermal images offline. We show that LGHD, a feature descriptor based on phase congruency and high-pass filtering, offers drastic performance improvements compared to classical methods like SURF. By implementing it in C++ we speed it up compared to the original version and enable it to be used in the ASL software stack. Using LGHD and manual post-processing, we generate over 1'000 aerial optical to thermal image pairs which form our data set WARM.
By adding geometric constraints from 3D landmarks as part of a pose graph optimization, we show how to generate fully registered cross-spectral orthomosaics. These can be used to generate bigger and more generalizable data sets in the future. We further explore the use of mutual information as an image alignment metric and show that a simple sliding window approach does not lead to well-formed optimization surfaces. Furthermore, we try to use Pix4Dmapper to generate cross-spectral maps and find it to have unsatisfactory performance.
Through the efforts listed above and an extensive search for cross-spectral data sets, we generate and collect more than 4'000 samples of otherwise rare training data for optical to thermal matching.

In the second stage of the thesis, we propose adaptions to make an existing DL end-to-end homography estimation pipeline perform cross-spectral matching and test its performance. To our knowledge, we are the first to use a SuperPoint adaption for multi-modal matching. We show that cross-spectral feature descriptors can be learned, but it is much harder than in the mono-modality case. We further conclude that transfer learning from the optical to optical task to the cross-spectral one does not work well. Additionally, we show that a single encoder is in fact expressive enough to capture both optical and thermal encoding at once. Lastly, we find that learning does not generalize well between different cross-spectral data sets and that we can improve the pipeline's performance by tuning the amount of homographic adaption during training to better fit the intended use case.

Finally, by providing a number of directions for future work, we lay out the path towards robust cross-spectral optical-thermal SLAM onboard a fixed-wing UAV.

Bibliography

[1] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *CoRR*, vol. abs/1510.05970, 2015.

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014.

[3] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," *CoRR*, vol. abs/1504.03641, 2015.

[4] K. M. Yi, E. Trulls Fortuny, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," *Computer Vision - Eccv 2016, Pt Vi*, vol. 9910, pp. 17. 467–483, 2016.

[5] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," *CoRR*, vol. abs/1712.07629, 2017.

[6] B. Chatton, "Thermal updraft prediction for a fixed-wing uav," Master's thesis, ETH Zurich, Switzerland, 2017.

[7] N. J. W. Morris, S. Avidan, W. Matusik, and H. Pfister, "Statistics of infrared images," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, June 2007, pp. 1–7.

[8] I. Guilliard, R. Rogahn, J. Piavis, and A. Kolobov, "Autonomous thermalling as a partially observable markov decision process (extended version)," *CoRR*, vol. abs/1805.09875, 2018.

[9] T. P. Truong, M. Yamaguchi, S. Mori, V. Nozick, and H. Saito, "Registration of rgb and thermal point clouds generated by structure from motion," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Oct 2017, pp. 419–427.

[10] L. Chen, L. Sun, T. Yang, L. Fan, K. Huang, and Z. Xuanyuan, "Rgb-t slam: A flexible slam framework by combining appearance and thermal information," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 5682–5687.

[11] J. Kümmerle, T. Hinzmann, A. S. Vempati, and R. Siegwart, "Real-time detection and tracking of multiple humans from high bird's-eye views in the visual and infrared spectrum," vol. 10072, 12 2016, pp. 545–556.

[12] P. Ricaurte, C. Chilán, C. Aguilera, B. Vintimilla, and A. Sappa, "Feature point descriptors: Infrared and visible spectra," *Sensors (Basel, Switzerland)*, vol. 14, pp. 3690–701, 02 2014.

[13] D. Firmenichy, M. Brown, and S. Süsstrunk, "Multispectral interest points for rgb-nir image registration," in *2011 18th IEEE International Conference on Image Processing*, Sep. 2011, pp. 181–184.

[14] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov 2004.

[15] C. Aguilera, F. Barrera, F. Lumbreras, A. D. Sappa, and R. Toledo, "Multispectral image feature points," *Sensors*, vol. 12, no. 9, pp. 12 661–12 672, 2012.

[16] C. A. Aguilera, A. D. Sappa, and R. Toledo, "Lghd: A feature descriptor for matching across non-linear intensity variations," in *2015 IEEE International Conference on Image Processing (ICIP)*, Sep. 2015, pp. 178–181.

[17] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey," *IEEE Transactions on Medical Imaging*, vol. 22, no. 8, pp. 986–1004, Aug 2003.

[18] N. Dowson and R. Bowden, "Mutual information for lucas-kanade tracking (milk): An inverse compositional formulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 1, pp. 180–185, Jan 2008.

[19] C. A. Aguilera-Carrasco, F. J. Aguilera, A. D. Sappa, C. Aguilera, and R. Toledo, "Learning cross-spectral similarity measures with deep convolutional neural networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 267–275, 2016.

[20] C. A. Aguilera-Carrasco, A. D. Sappa, C. Aguilera, and R. Toledo, "Cross-spectral local descriptors via quadruplet network," in *Sensors*, 2017.

[21] S. En, A. Lechervy, and F. Jurie, "Ts-net: Combining modality specific and common features for multimodal patch matching," *CoRR*, vol. abs/1806.01550, 2018.

[22] E. B. Baruch and Y. Keller, "Multimodal matching using a hybrid convolutional neural network," *CoRR*, vol. abs/1810.12941, 2018.

[23] N. Savinov, A. Seki, L. Ladicky, T. Sattler, and M. Pollefeys, "Quad-networks: unsupervised learning to rank for interest point detection," *CoRR*, vol. abs/1611.07571, 2016.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[25] P. D. Kovesi, "MATLAB and Octave functions for computer vision and image processing," available from: <http://www.peterkovesi.com/matlabfns/>.

[26] M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua, "Fast keypoint recognition using random ferns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 448–461, March 2010.

[27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.

Appendix A

List of Data Sets

A.1 Aerial Cross-Spectral Images

Name: WARM
Access: Open source
Notes: TBD
URL: TBD

Name: Solair Rosbags ASL
Access: Closed source, only ASL internal (ask Timo Hinzmann)
Notes: Flight with downward-looking IR camera, low altitude and little overlap
URL: https://github.com/ethz-asl/fw_human_detection_rega/wiki/Datasets

Name: Landmap: Thermal imagery of England
Access: Closed source, CEDA: denied, Geoinformation Group: no reply
Notes: Aerial imagery (aligned orthomosaic in optical and thermal)
URL: https://catalogue.ceda.ac.uk/uuid/bbe9b540281d29efb8e4fa3b82c2d801

Name: ICARUS Marche-en-Famenne
Access: Open source
Notes: Only point-clouds open source, images not to be found anymore
URL: https://projects.asl.ethz.ch/datasets/doku.php?id=jfricarus

Name: Semantic Drone Dataset
Access: Open source, link sent through email after registration
Notes: Provides thermal images too but in very bad quality
URL: https://www.tugraz.at/index.php?id=22387

Name: Rega Rosbags ASL
Access: Closed source, only ASL internal
Notes: Forward-looking IR camera, not appropriate to map ground
URL: https://github.com/ethz-asl/fw_human_detection_rega/wiki/Datasets

A.2 Non-Aerial Cross-Spectral Images

Name: ICIP 2015
Access: Open source
Notes: Aligned optical and thermal images of buildings, but not aerial
URL: https://owncloud.cvc.uab.es/owncloud/index.php/s/1Wx715yUh6kDAO7

Name: VOT-RGB TIR 2019 dataset preview
Access: Open source
Notes: Aligned optical and thermal images of pedestrians on streets, but not aerial
URL: http://www.votchallenge.net/vot2019/dataset.html

Name: LITIV-VAP dataset
Access: Open source
Notes: Aligned optical and thermal images of people, not perfectly aligned
URL: https://www.polymtl.ca/litiv/en/codes-and-datasets

Name: Morris IRDATA
Access: Open source
Notes: Not perfectly aligned optical and thermal images
URL: http://www.dgp.toronto.edu/~nmorris/data/IRData/

A.3 Non-Aligned Images

Name: Interspectral image registration dataset (RFAE)
Access: Open source
Notes: Has FIR images, but they are not aligned
URL: https://dronehub.tk/datasets-96fc4f9a92e5

A.4 NIR Cross-Spectral Images

Name: Vehicle Detection in Aerial Imagery (VEDAI)
Access: Open source
Notes: High resolution satellite images with aligned optical and NIR images (aerial)
URL: https://downloads.greyc.fr/vedai/

Name: RGB-NIR Scene Dataset
Access: Open source
Notes: Aligned optical and NIR images, but not thermal nor aerial
URL: https://ivrl.epfl.ch/research-2/research-downloads/supplementary_material-cvpr11-index-html/

Name: Wheat Field
Access: Open source
Notes: no thermal (FIR) images
URL: https://www.sensefly.com/education/datasets/?dataset=5538&sensors%5B%5D=25

Name: Mixed-Use Fields
Access: Open source
Notes: No thermal (FIR) images
URL: https://www.sensefly.com/education/datasets/?dataset=5632&sensors%5B%5D=25

Name: Crop Field (multi-spectral)
Access: Open source
Notes: No thermal (FIR) images
URL: https://www.sensefly.com/education/datasets/?dataset=5632&sensors%5B%5D=25

A.5 Thermal Images Only

Name: Pix4D Thermal imagery example
Access: Open source
Notes: Thermal images only, no optical overlay available
URL: https://support.pix4d.com/hc/en-us/articles/360000235126-Example-projects#label4

Name: Solar Panel Installation
Access: Open source
Notes: Thermal images only, no optical overlay available
URL: https://www.sensefly.com/education/datasets/?dataset=1416&sensors%5B%5D=26

Name: Thermal infrared dataset ASL
Access: Open source
Notes: Thermal infrared images, no corresponding optical images
URL: https://projects.asl.ethz.ch/datasets/doku.php?id=ir:iricra2014