JOINT BUNDLE ADJUSTMENT OF THERMAL INFRA-RED AND OPTICAL IMAGES BASED ON MULTIMODAL MATCHING

A.Sledz*, C. Heipke

Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Germany
sledz@ipi.uni-hannover.de

COMMISSION I, WG I/6

KEY WORDS: Thermal infrared and optical image registration, multimodal feature matching, phase congruency, bundle adjustment

ABSTRACT

Despite the fact that Thermal Infrared (TIR) cameras have been in use for decades, processing of TIR images poses a variety of challenges when compared to optical images, which are captured in the visible part of the electromagnetic spectrum. The estimation of the exterior orientation of TIR cameras by bundle adjustment is a difficult task due to the limited geometric resolution of a TIR camera and the low image quality in terms of contrast and texture compared to optical images. Optical images have a potential to increase TIR external orientation accuracy by incorporating them into a joint bundle adjustment. However, because the modality gap between those two image types is large, classical point matching algorithms typically fail to find matches, making processing both image types in the joint bundle challenging. In order to locate matching points in both modalities, this study suggests using the Edge Histogram Descriptor (EHD) in the frequency domain representation of the images based on phase congruency. To properly allocate edges from the phase congruency, which are then employed in EHD, non-maximum suppression and hysteresis thresholding are used. Considering that both sensors are fixed rigidly to a single platform, the search region for the matching point candidate of the TIR image is determined based on stereo calibration of a thermal/optical stereo setup combined with geometric constraints. The final matching is based on the cosine distance, while RANdom SAmple Consensus (RANSAC) is used in order to eliminate outliers. The findings of this study show that using a joint bundle adjustment with optical images versus a bundle adjustment only with TIR images improves TIR image orientation, which is supported by the increased accuracy of the adjusted Ground Control Point (GCP) coordinates.

1. INTRODUCTION

Thermal imaging systems were originally developed for military purposes, but they have since migrated into other fields and have found various applications in areas such as surveillance, pedestrian detection, driving assistance, as a backup tool for firefighters, in a variety of inspection and assessment tasks and in many more.

Thermal Infrared (TIR) photogrammetric processing has been thoroughly studied, yet there are still certain challenges to overcome, and there is always potential for further development. Hoegner and Stilla (2015) present a method for automatically texturing building facades using terrestrial TIR image sequences. Pech et al. (2013) investigate the topic of capturing multi-temporal thermal images and creating thermal orthophotos with a TIR camera mounted on an Unmanned Aerial Vehicle (UAV). Khodaei et al. (2015) show that DSMs derived from aerial thermal data may be as accurate as DSMs produced from visual images. The influence of interior camera orientation, tie point matching, and ground control points on the resulting accuracy of the bundle adjustment and dense point cloud generation with a commonly used photogrammetric workflow for UAV based thermal imagery in natural environments is investigated by Boesch et al. (2017). Hoegner et al. (2014) combine TIR and time-of-flight depth images to produce an accurate 3D point cloud for scene segmentation and people recognition.

Maset et al. (2017) demonstrate that commercial photogrammetric computer vision software may be used to autonomously orient sequences of TIR images obtained from a UAV and to generate 3D point clouds without the need for GNSS or Inertial Navigation System (INS) data regarding the images' location and attitude. Furthermore, co-registration of images captured with both, optical and TIR cameras can improve the accuracy and density of the 3D point cloud to be produced (Hoegner and Stilla 2016; Hoegner et al. 2016).

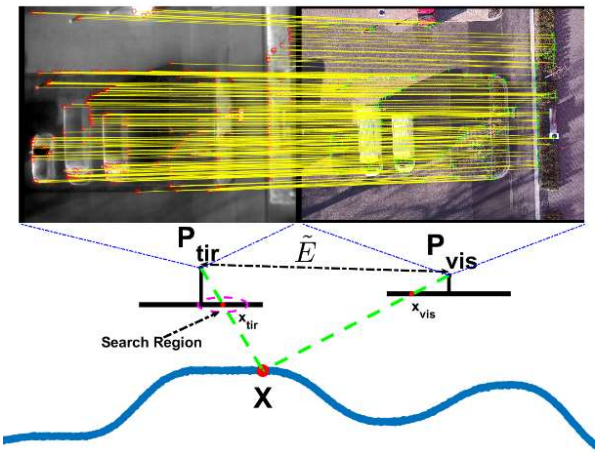

Figure 1. Diagram of multimodal image matching with epipolar constraint (see text for details)

This study focuses on multimodal matching between TIR and optical images, while the final goal is to evaluate joint bundle adjustment results of both image types. *Optical images* in this study hereinafter are referred to as *visible images*., as the latter term is more common in the TIR community. Multimodal matching is

* corresponding author

carried out in the frequency domain using phase congruency (PC). The PC transformation concept of an image was introduced by Kovesi (1999). PC is based on the local energy model and depicts the frequency domain behaviour of an image. PC is invariant to changes in light and contrast and may be used to identify and locate edges and corners (Kovesi, 2003). To obtain equivalent contents of TIR and visible images, consistent edge and corner structures are first detected from the PC. In the following stage, characteristics of the Points Of Interest (POI) of thermal infrared and visible images are retrieved based on consistent edge structures using the Edge Histogram Descriptor (EHD). A confined search region in the TIR image is determined, in which matching candidates for each POI x_{vis} from the visible image are situated, as illustrated in Figure 1. The search region is defined by connecting x_{vis} to the correspondent point X in object-space, defined by locating the intersection of the ray through x_{vis} and the 3D model derived from the photogrammetric processing of the visible images alone. Following that, X is projected into the image plane of the TIR image in order to calculate x_{tir} and to define the search region for possible matching candidates. The exterior orientation of the TIR camera is approximated from stereo calibration between the visible and the TIR camera, which was performed offline. Feature correspondences are determined using the cosine distance between feature descriptions, outliers are eliminated using the RANdom SAmple Consensus (RANSAC) algorithm (Fischler and Bolles, 1981). Finally, the identified tie points between each multimodal stereo pair and all tie points generated from either thermal or visible images are used as input in a joint bundle adjustment. The advantage of the joint bundle adjustment is twofold: improved accuracy in terms of the adjusted GCP coordinates, as demonstrated in the experimental section of this paper; and the ability to combine TIR and visible images into a single block, allowing for further synergistic analysis of the observed scene.

Our paper is structured as follows: Chapter 2 provides an overview of multimodal image registration and point matching. The theoretical aspects of PC and multimodal point matching are discussed in Chapter 3. The experimental section shows the results of the suggested technique as well as key justifications for how a joint bundle adjustment of visible and TIR images can increase the total TIR bundle accuracy in terms of Ground Control Point (GCP) coordinates. Conclusions and open questions for future research are presented in the last chapter.

2. RELATED WORK

TIR and visible images capture distinct wavelength intervals of the electromagnetic spectrum, and thus depict different information in their Field of View (FOV). Visible images, in general, collect light in the visible part of the electromagnetic spectrum ($0.3 - 0.7\mu m$). When taken under good light conditions, they frequently capture a detailed depiction of an object, including its texture. TIR images, on the other hand, capture thermal infrared radiation, which is emitted and/or reflected from a different part of the electromagnetic spectrum ($0.7 - 14\ \mu m$). As a result of those differences, TIR images generally depict an object in a considerably less detailed manner, while textures are essentially nonexistent. In addition, differences in thermal capacity, emissivity and material transitions frequently produce a different appearance in the thermal image. While visual and thermal radiation have similar properties in terms of reflection, refraction, and transmission, they depend not only on the reflected or emitted radiation, but also on the scene geometry from which light is reflected or emitted. Thus, multimodal analysis comes into place to relate the two types of images to each other. A pre-requisite for multimodal analysis is to close the gap between modalities such that each one is expressed by a common representation.

Multimodal image matching is a fundamental and critical problem in a wide range of applications, including medical imaging, remote sensing, photogrammetry and computer vision. It involves identifying and then matching the same content from two or more images with significant modality or nonlinear appearance differences. Over the last few decades, a growing number of different techniques have been suggested with the goal of reducing the modality differences inherent in multimodal imaging.

Jiang et al. (2021) present an in-depth review of multimodal matching. Each application field, for example medical imaging (James et al. 2014, Mani et al., 2013) or remote sensing (Ghassemian 2016), receives a separate overview of techniques and methodologies. The registration of visible and TIR images is the more relevant part of this review related to this study. The feature-based technique is one of the most typical ways in multimodal matching, although it is not the only one. Typically, a feature-based pipeline goes through the steps of feature detection, description and matching. This pipeline is more extensively utilized in the image matching community because sparse features may be thought of as a basic representation of an image, making it more flexible and resistant to geometric and illumination changes and to noise. The detected features often reflect certain structures in an image or in the real world and may be categorised as corner, blob, line or edge, and area features. The term "feature description" refers to the process of translating the local grey value surrounding of a feature point into a stable and discriminative form (often as a vector), which allows for quick and easy matching of the detected features. The generated descriptors of two matched features should be as close as possible in the descriptor space and the descriptors of two non-matching features should be as far away as feasible. At the same time, feature description should be resilient to geometrical transforms, image appearance changes, and different image quality. Feature matching seeks to create proper feature correspondences between extracted feature sets. If modality differences are sufficiently suppressed in the feature detection and description phases, generic approaches may perform well in the matching stage.

A study by Hrkać et al. (2007) investigated how to register infrared and visible image pairs that were obtained from slightly different viewpoints of the same buildings. The researchers rely on the assumption, that the corners are most stable features points in both images type. They located Harris corners from these two images, and then applied a simple similarity transformation to register infrared and visible images. The partial Hausdorff distance was chosen as the measure of similarity between the two descriptors. A frequency-based corner and edge detector, in conjunction with an EHD approach, was proposed by Mouats and Aouf (2013) in order to compute correspondences between images of the visible and infrared spectrum. The convolution of the Fourier transform of the image with a bank of Log-Gabor filters at different orientations and frequencies yields the desired features, which are highly localised and invariant to changes in image contrast and illumination. The descriptor is based on a combination of frequency information at the position of the key point and the spatial distribution of the contours in a window surrounding the feature points that have been extracted. Taking into account the large variation in terms of resolution and appearance produced by different image sensors, Du et al. (2018) suggested a scale-invariant Partial Intensity Invariant Feature Descriptor (PIIFD) for corner feature description and matching that is both, fast and accurate. Additionally, a locality preservation requirement was paired with a false match removal approach in order to improve the estimation of the employed affine matrix in a Bayesian framework. Cui and Zhong (2018) suggested a technique, which involves extracting corners from phase congruency images using their extremal moments and then describing them with

Log-Gabor filters. Once the matching process was completed, the descriptors were compared to identify credible point correspondences, and the RANSAC approach was utilised to verify these correspondences. Zeng et al. (2019) extracted edges using the morphological gradient approach, and then employed the C_SIFT detector, an adaptation of SIFT (Lowe 2004), on edge maps for distinct point search and BRIEF (Calonder et al. 2010) for description, resulting in scale- and orientation-invariant matching and a morphological gradient method for description.

In this study, we use a PC with a mixture of edges and corners as POIs. PC was already proven to be a useful tool for dealing with low-light images, which could be characterized by weak edges and corners (Mouats et al, 2015 and Mehltretter et al, 2018). The PC approach by Kovesi (2003) is used with some modification to extract edges and corners. The proposed modification of allocating edges and corners in this study relies on non-maximum suppression and hysteresis thresholding in order eliminate weak edge. Furthermore, while the majority of the previous studies employed visible and TIR images of similar image resolution, this study used a high-resolution visible image and a low resolution TIR image. This difference in image resolution leads to additional complexity in finding corresponding points. To solve the multi-scale matching issue this study proposes that distinct filter banks be employed for different image types.

3. METHODS

The focus of this chapter is on multimodal matching of TIR and visible images. Providing the reader with relevant theoretical background information as well as a full description of the proposed approach are the primary objectives, and adaptations of the employed technique to the problem at hand are discussed.

3.1. Phase congruency

The spatial frequency transform, resulting in magnitude and phase information, is one of the most essential and commonly used tools for image representation and analysis. Because of the relevance of phase information, it has been implemented in a variety of tasks such as edge and corner detection, image segmentation, and similar steps. Due of phase's strong invariance to noise and low contrast, it is a desirable technique for image processing in general, as well as for TIR image processing in particular.

Kovesi (1999) introduced the Phase Congruency (PC) transformation of an image. It is based on the local energy model (Morrone et al., 1986) postulating that a feature can be best perceived once the frequency components of the signal are all in phase. The transformation is invariant to changes in light and contrast and may be used to identify and locate edges and corners (Kovesi, 2003). It is necessary to reconstruct a signal over a large frequency range in order to compute PC. One goal is to keep as many frequency components in the signal as feasible. To accomplish this goal, a filter bank is generated in which each filter's transfer function overlaps with its neighbours in such a way that the total of all transfer functions produces a relatively uniform spectrum coverage. The reconstructed signal is a band pass version of the original signal, amplified according to the scaling and overlap of the transfer functions. Log-Gabor wavelets are employed in the filter bank, which is a common technique.

The transformation may be calculated at different scales *n* and orientations *o*, as shown in equation (1). In this equation, $A_{no}(x)$ is the amplitude of the transformed signal at a given scale and orientation. $A_{no}(x)$ is calculated as shown in equation (2). $I(x)$ denotes the input signal and M_{no}^{e} and M_{no}^{o} denote the even and odd Gabor wavelets at a scale n and orientation o. T_o is a noise threshold that is calculated independently for each orientation, and W_o is a weight function that seeks to compensate an uneven distributions of the filter response. The ε in equation (1) is a small value, which is used to prevent division by zero. The optimal value of ε is determined by the accuracy with which convolutions and other operations on the signal can be performed; it is not determined by the signal itself (Kovesi 1999) and is set to 0.01 in this study. The impact of noise can be reduced by using large values of ε.

$$PC(x) = \frac{\sum_o \left(W_o(x) \left(\sum_n \sqrt{(I(x) * M_{no}^{e})^2 + (I(x) * M_{no}^{o})^2} - T_o \right) \right)}{\sum_o \sum_n A_{no}(x) + \varepsilon} \quad (1)$$

$$A_{no}(x) = \sqrt{(I(x) * M_{no}^{e})^2 + (I(x) * M_{no}^{o})^2} \quad (2)$$

A difficulty inherent in PC is its response to noise (Kovesi 1999). In the vicinity of a step (edge), PC is only high at the point of the step. Away from the step, however, noise-induced variations are large in comparison to the surrounding signal. This will happen regardless of how uniform the environment is. The upper constraint on the influence of noise on the total of the wavelet response amplitudes is provided by the sum of the estimated noise responses over all wavelet scales, parameter T_o in the equation (1).

Applying a sigmoid function to the filter response, as given in equation (3), is used to create PC's weight function. As mentioned, the primary purpose of this weight function is to penalise uneven filter response distributions, particularly in the area of steps corresponding to line features. The parameter *c* is the cut-off value of the filter response spread, below which phase congruency values are penalised, and the parameter *g* is the gain factor that regulates the sharpness of the cut-off.

$$W(x) = \frac{1}{1 + e^{g \cdot (c - s(x))}} \quad (3)$$

$$s(x) = \frac{1}{N} \left(\frac{\sum_n A_n(x)}{\varepsilon + A_{max}(x)} \right) \quad (4)$$

Many applications, such as stereo matching, motion tracking and image registration, need an accurate detection of so-called "corners" throughout image sequences. This is where the Harris corner detector (Harris and Stephens, 1988) can be applied. The response of the Harris operator, as well as that of other corner operators, varies largely depending on image contrast, which makes determining adequate thresholds for extended image sequences difficult, if not impossible. Following his first phase congruency suggestion, Kovesi (2003) enhanced his method and offered a new corner and edge detector. This new operator employs the primary moments of the phase congruency information. The resultant corner and edge operator has a much better localisation accuracy and exhibits image contrast insensitive responses.

Kovesi (2003) proposes the following to incorporate information on how phase congruency varies with orientation: using equation (1), calculate phase congruency in each orientation separately, then compute moments of phase congruency and examine the variation of the moments with orientation. The primary axis, which corresponds to the axis around which the moment is minimised, indicates the feature orientation. The magnitude of the maximum moment, which corresponds to the moment around an axis perpendicular to the major axis, indicates the feature significance. If the minimum moment is likewise big, it means the feature point has a strong 2D component and should be categorised as a corner.

For each point in the image, three different components are calculated as shown in the equations (5), (6) and (7).

$$a = \sum_{\theta} \left(PC(\theta)cos(\theta)\right)^2 \quad (5)$$

$$b = 2\sum_{\theta} PC^2(\theta)cos(\theta)sin(\theta) \quad (6)$$

$$c = \sum_{\theta} \left(PC(\theta)sin(\theta)\right)^2 \quad (7)$$

where $PC(\theta)$ refers to the phase congruency value determined at orientation θ, and the sum is performed over the discrete set of orientations used (typically six with the step of 30°). The angle of the principal axis Φ is given by

$$\Phi = \frac{1}{2}atan2\left(\frac{b}{\sqrt{b^2+(a-c)^2}}, \frac{a-c}{\sqrt{b^2+(a-c)^2}}\right) \quad (8)$$

The maximum (eq. (9)) and minimum (eq. (10)) moments, M and m respectively, are given by

$$M = \frac{1}{2}\left(c+a+\sqrt{b^2+(a-c)^2}\right) \quad (9)$$

$$m = \frac{1}{2}\left(c+a-\sqrt{b^2+(a-c)^2}\right) \quad (10)$$

The phase congruency edge and corner strength images are little impacted by image contrast, as shown in Figure 2, and may be easily thresholded (in this case with a value of 0.5 for the maximum value of the minimum moments) to provide a clear set of features.

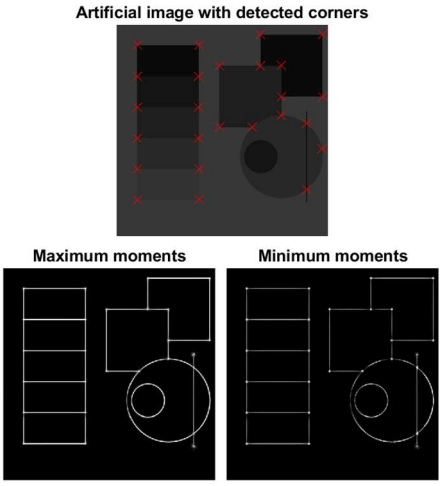

Figure 2. Maximum and minimum momments of an artificial image with low contrast

3.2. Multimodal matching

In his work, Kovesi (2003) compares his approach to the Harris corner detector. Because the latter depends on the image gradient covariance matrix, it is extremely sensitive to image contrast fluctuations, making threshold setting very challenging. Unlike image intensity gradient values, phase congruency values are normalised quantities with no units. When the moments are normalised for the number of orientations evaluated, the phase congruency moment values vary from zero to one. As a result, the maximum and minimum phase congruency moments may be utilised to determine whether there is a substantial edge and/or corner point.

The approach, proposed by Kovesi (2003), assigns corners and edge to features by using a fixed threshold. However, while such a threshold approach may be used intuitively, some noise will be classified as an edge and weak edges will not be found. The approach suggested to detect edges in this study is inspired by the edge detection technique developed by Canny (1986). This study recommends using maximum moments M (eq. 9) and the angle of the principal axis Φ (eq.8) instead of the image gradient amplitude and gradient angles, as in Canny's edge detector. The following is what has been suggested:

- In non-maximum suppression, the edge strength from maximum moments M (eq. 9) of the current pixel is compared to the edge strength of the pixel in the positive and negative direction Φ (eq.8). The value will be kept if the current pixel's edge strength is the highest compared to its neighbours. The value will be suppressed otherwise.
- The double threshold step identifies three types of pixels: strong, weak, and irrelevant pixels. Strong pixels are those that have a larger edge strength than the high threshold and undoubtedly contribute to the final edge. Weak pixels are pixels with an edge strength value that is not high enough to be considered strong, but not low enough to be regarded irrelevant for edge detection. Weak edges are identified if their edge strength falls in between the high and low thresholds. All other pixels are ignored when calculating the edge.
- Hysteresis edge tracking is based on the thresholding findings; hysteresis transforms weak pixels into strong pixels in an iterative way, if and only if at least one of the pixels around the one being processed is a strong one.

The EHD's fundamental concept, as proposed by Mouats et al. (2013) and used by others (Wang et al, 2020, Xu et al., 2020 and Mehltretter et al., 2019), works as follows: from the edge map, an area of NxN pixels centred on a POI is extracted. Local spatial maxima from the maximum and minimum moments of PC, see equations (9) and (10), are considered as POIs. Local edge histograms are generated for each image patch, which is partitioned into 4x4=16 sub-regions. Horizontal, vertical, 45° diagonal, 135° diagonal, and isotropic (edge without orientation) are the five types of edges investigated. Accordingly, there are four direction histogram bins and one non-direction bin. The last bin corresponds to places where there are no edges. To detect the aforementioned edge orientations, five filters are utilised, as demonstrated in Figure 3. Every pixel in each sub-region contributes to the histogram, and the filter with the highest response is picked to vote for the appropriate bin. Subsequently, the histogram vector is normalised; it has 80 bins (4x4x5). The descriptor depicts the spatial distribution of the region's edges.

Mouats et al. (2013) also proposed adding a second part to the descriptor derived from the phase congruency result. For each POI in the image, 24 Log-Gabor coefficients corresponding to six orientations at four frequencies (scales) are produced during PC calculation. These coefficients are utilised as the descriptor's second part. Finally, for the i^{th} POI a descriptor D_i is composed of 104 elements, with 80 of them coming from the histogram and 24 from the Log-Gabor coefficients.

In the current study, the feature vectors are retrieved on POIs using the EHD approach. All local maxima of the minimum moments m (eq. 10) of PC that are associated with corners and all local maxima of the maximum moments M (eq. 9) of PC that are related to edges are assigned a POI. In contrast to the original EHD extraction technique, only the section of the EHD feature

vector that corresponds to the estimated histogram is used in this study, because of the varying number of scales used in the PC computation for the different image modalities (TIR vs. visible). The justification for the varying number of scales for each modality type will be given in the section 4.2.

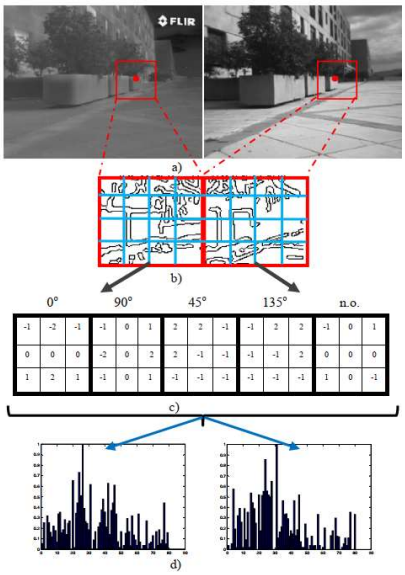

Figure 3. EHD extraction flow diagram (Mouats et al., 2013)

The cosine distance *dist* (eq. 11), which is commonly used in comparing wavelet-based descriptors, is used as similarity measure in the matching process. Each descriptor in the visible image (D^{VIS}) at a given location (x_i^{VIS}, y_i^{VIS}) is compared to all descriptors (D^{TIR}) in the TIR image within a circular region with radius c_r centred at that location (x_i^{TIR}, y_i^{TIR}).

$$\mathrm{dist}\left(D_k^{TIR}, D_n^{VIS}\right) = 1 - \frac{\sum_j d_j^{TIR} d_j^{VIS}}{\sqrt{\sum_j \left(d_j^{TIR}\right)^2 \sum_j \left(d_j^{VIS}\right)^2}} \qquad (\ 11\ )$$

where $\left(D_k^{TIR}, D_n^{opt}\right)$ are the descriptors of the POIs to be compared, while indices *k* and *n* denote the corresponding POIs. The feature in the TIR image that minimizes the distance for a particular feature in the visible image is chosen as potential match. Then, a threshold is used to keep only the strongest matches. The suggested technique, like any feature detection-description-matching scheme, is still prone to false matches. After initial matching, an outlier removal step is performed to reduce the number of blunders. This process is executed by fitting an essential matrix approximating the epipolar geometry to the matched locations using the RANSAC technique.

As illustrated in Figure 1, the coordinates of point $\boldsymbol{x}_i^{TIR}$ are computed by projecting point $\boldsymbol{X_i}$ from a reconstructed 3D model (eq. 12). A thorough photogrammetric processing of a block including only visible images is used to generate this 3D model. Despite the fact that both image sets observe the same scene in space and time, visible images with far higher resolution and including significantly more information are selected as a more reliable source for 3D model reconstruction. As we use a rigid dual-sensor configuration for data capture, the interior orientation ($\boldsymbol{K^{TIR}}$ in eq. 12) and the exterior orientation ($\boldsymbol{R^{TIR}}$ *and* $\boldsymbol{T^{TIR}}$ in eq. 12) of the TIR camera can be recovered offline from the image orientation of the visible sensor and stereo calibration of the ridged sensor set up.

$$\begin{bmatrix} \boldsymbol{x}_i^{TIR} \\ 1 \end{bmatrix} = \boldsymbol{K^{TIR}}[\boldsymbol{R^{TIR}}|\boldsymbol{T^{TIR}}]\begin{bmatrix} \boldsymbol{X_i} \\ 1 \end{bmatrix} \qquad (\ 12\ )$$

Due to a variety of uncertainties, including inaccuracies in the interior and exterior orientation of both cameras, as well as the uncertainty of the reconstructed 3D model, a search radius c_r for related point candidates must be employed to find matches.

4. EXPERIMENTS AND RESULTS

The purpose of this chapter is to provide a more in-depth discussion of the approaches outlined in the previous chapter and to present the outcomes of this study. A quick introduction to hardware, software, and the surveyed scene will be provided first. The paramertization of specific methods, as well as their rationale, will be explained next. Finally, the experiments and results will be presented and discussed in relation to the use of the joint bundle adjustment of TIR and visible images.

To begin with, we define the experimental objectives as well as the assessment criteria. As previously stated, the primary goal of this research is to improve the accuracy of the TIR camera exterior orientation using supplementary information derived from visible images. Despite the importance of the multimodal matching element itself from a scientific standpoint, this step is considered a secondary aim in this study. Due to the lack of ground truth, the multimodal matching between visible and TIR images is evaluated based on average re-projection errors of the tie points and the GCP and the deviations of the GCP coordinates in object space. An improvement of the accuracy of those measures is the success criteria in this study.

4.1. Hardware and Software

A DJI M200[2] UAV, a VTOL quadcopter equipped with a GNSS receiver, an inertial measurement unit (IMU), a barometer, and the stereoscopic camera system DJI Zenmuse XT2[3], was used to acquire the data. The Zemuse XT2 is a gimbal-stabilized system in a dual-sensor configuration that firmly combines a FLIR radiometric thermal imager with 512x640 pixels with 17 μm pixel pitch and a CMOS optical camera in 12 MP resolution with pixel pitch of 1.85 μm.

Stereo calibration was carried out according to standard protocol. An aluminium board with black squares was created to obtain a strong thermal contrast in the TIR images. The board remained stationary facing the sky during the calibration operation in order to capture the cold reflection of the skies on the aluminium portion of the board. As a consequence of the employment of black paint, which increases the board's emissivity, a good thermal contrast was produced, as illustrated in Figure 4.

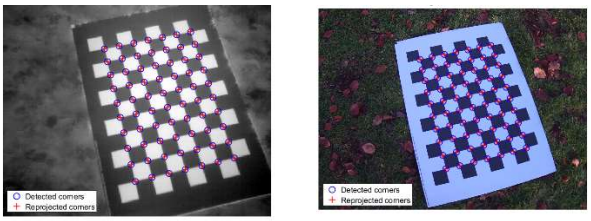

Figure 4. DJI Zenmuse XT2 images of calibration test field. Left: TIR image, right: visible image

[2] www.dji.com/de/matrice-200-series

[3] www.dji.com/de/zenmuse-xt2

A test flight was conducted in the German city of Hannover. Because of the larger Ground Sampling Distance (GSD), flight planning was reliant on the TIR camera parameters. The flying height was set at 30 metres above ground, based on a pixel size of 17 μm and a focal length of 13 mm, resulting in a GSD of 3.9 cm for the TIR images (the corresponding GSD of the optical camera was 0.71 cm). A total of 152 image pairs of visible and TIR images were captured. Figure 4 shows the placement of six GPCs (white boards with black circle) in the study region. RTK GNSS equipment was used to carry out the GCP measurements.

Agisoft Metashape[4] was used to do rigorous photogrammetric processing, which included GNSS measurements for the locations of the image projection centres as initial values.

Figure 5. Visualization of the surveyed area with GCPs marked with red circles

4.2. Multimodal matching

Following the detailed description of PC construction in chapter 3.1, this sub-chapter concentrates on the demonstration of the PC application to multimodal matching. Kovesi's original implementation recommended employing four frequencies (scales) and six orientations for the filter bank, based on Log-Gabor wavelets. The noise threshold, T_o parameter, see equation (1), is calculated using the median value of the sum of the amplitude responses of all scales. The weight function cut-off, parameter c, see equation (3), is set at 0.5.

Figure 6 shows test images in the visible and the IR spectrum of the same scene in the 1st column; the PC of both images is shown in the 2nd column, while the 3rd column shows enlarged regions, which are marked by red rectangles in the 1st column. For the TIR image, the PC has kept practically every significant edge and has removed the majority of the noise from uniform areas. For the visible image, the situation is similar in terms of object edges; however, PC also includes a lot of details, especially texture representation; those details are associated with high frequencies in the image. The outcome of using of the default PC parameters is that the representation difference between two PCs is still substantial, partly due to the different GSD of the two sensors.

Figure 7 shows the experimental results for particular filter bank settings for the PC calculation, where the GSD difference was partly compensated: Default parameters for the PC of the TIR image were used, i.e. four scales with a starting frequency of three pixels. For the PC of the visible image, five scales were utilised, with the initial frequency of 10 pixels. As can be observed in the third column of Figure 7, the PC of the visible image has far less high frequency components and resembles the PC of the TIR image considerably more. The result is that the modality representation gap may be reduced by optimizing the selectin of PC calculation settings for each image type.

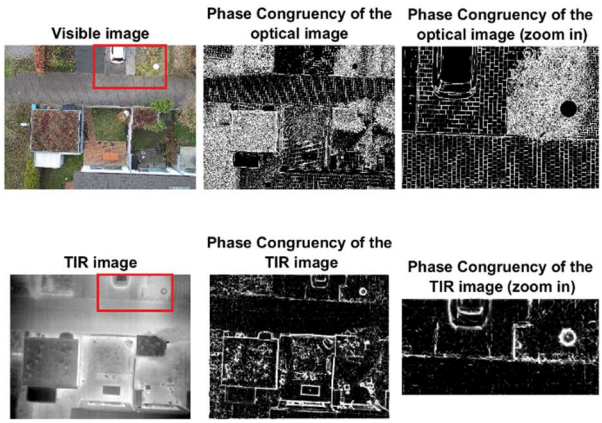

Figure 6. Phase Congruency of TIR and visible images with default parameters

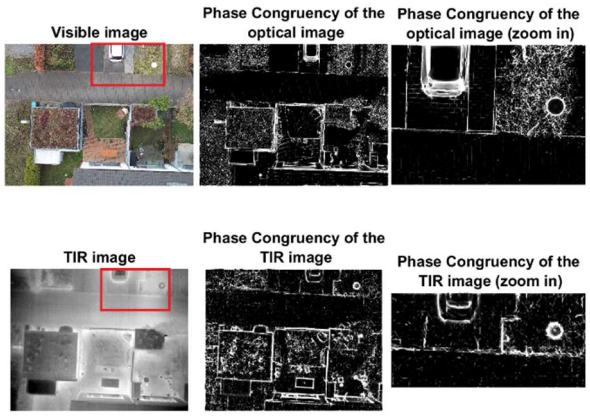

Figure 7. Phase Congruency of TIR and visible images with optimized parameters

The multimodal matching is based on EHD (see chapter 3.2) and the cosine distance (eq. 11) followed by RANSAC. Unfortunately, it is not possible to integrate additional phase information in the EHD construction, as recommended by Mouats et al. (2013): The EHD description of the TIR image would include 24 elements for each point on the image, while the EHD descriptor of the visible image would contain 30 entries, resulting in descriptor lengths that are conflicting. Therefore, we only use the part of the EHD vector for the histogram elements. The final results of multimodal matching are shown in Figure 8: every corresponding point was allocated on an edge or a corner, which is an inherent characteristic of the method.

Because the visible and TIR cameras are set up in a rigid-body dual sensor configuration, there is no requirement for orientation compensation from the descriptor side. However, the GSD of the cameras varies (because of differences in focal length and pixel size). Thus, from the standpoint of well-known descriptors, there is a requirement for scale invariant matching. Based on the sensor's prior knowledge, scale invariance is accomplished by using varying window sizes for the EHD extraction. A 48x48 pixel window is used for the TIR images. The EHD of the visible images, on the other hand, is extracted on a 240x240 pixel window.

[4] https://www.agisoft.com/

This ratio of EHD window sizes set to five based on the GSDs ratio, being approximately 5.4.

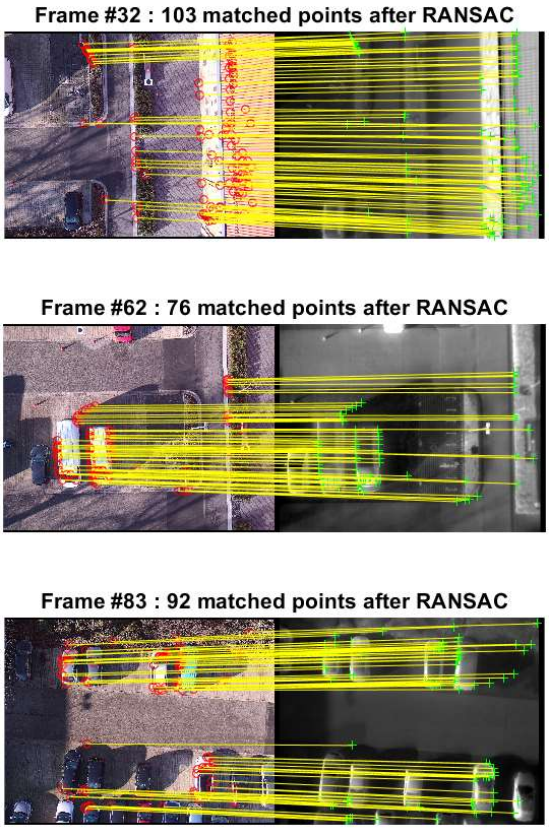

Figure 8. Examples of multimodal matching

4.3. Joint bundle adjustment of visible and TIR images

The primary finding of this study is presented in this section: the effect on the bundle adjustment of TIR images with the support of supplementary information in form of visible images. This aid is provided in the form of tie points, which connect each pair of visible and TIR images. A variety of tests were carried out to evaluate the suggested approach. In each case, observations for the bundle adjustment were the image coordinates of all related tie points, the image coordinates of the six GCPs, which were measured automatically, and 3D coordinates for the GCPs. The standard deviations for those observations were set to 0.5 pixel for all image coordinate measurements and 1 cm for the 3D GCP coordinates.

In the 152 visible images, approximately $410 \cdot 10^3$ tie points in image space were derived, whereas in the 152 thermal images a little more than 10% of tie points could be found. This ratio reflects the fact that the texture is much poorer in the thermal images as mentioned before. However, also the thermal images have more than enough tie points to reconstruct a stable photogrammetric block.

Each bundle adjustment is assessed using two key criteria: the error in the adjusted GCP coordinates, stated in cm, and their mean; the Root Mean Square (RMS) values of the tie point residuals in image space and the RMS values of GCPs reprojection errors in image space, expressed in pixels and in micrometres (μm) and their average.

The following experiments were conducted:

- Experiment 1 provides a baseline for comparing the effects of bundle adjustment. Each image set was processed independently and no multimodal matching was carried out.
- Experiment 2. In this experiment, the bundle adjustment was carried out using only the visible images. The TIR images were added to the block after the bundle adjustment was completed, with their exterior orientation computed based on the orientation parameters from the stereo calibration. Again, no multimodal matching was carried out.
- Experiment 3 provides an initial idea of the performance of the joint bundle adjustment. Both image types were processed in the joint bundle adjustment, with image coordinates of the GCPs acting as tie points between two modalities. In the visible images, each GCP was seen 36.3 times on average, whereas in the TIR images, it was seen only 22.2 times. The gap in the average number of GCP observations is due to different footprints of the two cameras. In total, $(36.3+22.2) \cdot 6$ or approximately 350 multimodal tie points in image space were contained in this adjustment, these were added to the tie points from experiment 1.
- Experiment 4 is the joint adjustment of both image types with conjugate points from multimodal matching and image coordinates of GCPs as tie points. Multimodal matching yielded approximately $49 \cdot 10^3$ tie points in image space and these were added to the tie points from experiment 3.

As demonstrated in Table 1, in the baseline experiment a subpixel accuracy for matching was reached. It is interesting to note that a slightly better result was achieved for the tie points of the thermal images (0.7 vs. 0.9 pixel). The RMS value of the GCPs in object space reflects this result as well. Not surprisingly, the visible image accuracy outperforms the TIR image accuracy in all criteria: the GCPs coordinates (1.4 cm vs 2.9 cm), the RMS value of the tie points (1.7 μm vs. 12.1 μm) and when considering the RMS value of the GCP reprojection errors (0.7 μm vs 9.4 μm). The reason for this difference is twofold, which relates to the better texture representation of the observed scene and the smaller GSD of the visible images.

Table 1 Results of experiment #1 – separate bundle adjustment of the visible and the TIR images

	Visible images			*TIR images*		
	3D Error (cm)	RMSE		3D Error (cm)	RMSE	
		Pix	μm		*Pix*	μm
TPs	**-**	**0.9**	**1.7**	**-**	**0.7**	**12.1**
GCP 1	1.2	0.3	0.6	1.2	0.6	10.7
GCP 2	1.7	0.4	0.6	1.4	0.7	12.2
GCP 3	1.2	0.4	0.8	4.2	0.3	5.8
GCP 4	1.0	0.5	0.9	2.6	0.4	7.1
GCP 5	1.2	0.4	0.7	5.6	0.6	10.7
GCP 6	1.9	0.3	0.6	2.6	0.6	9.7
GCP mean	**1.4**	**0.4**	**0.7**	**2.9**	**0.6**	**9.4**

Experiment 2 yielded an unexpected result (see Table 2). For the visible images the results are the same as for experiment 1, of course. However, the GCP reprojection errors for the TIR images increased dramatically. This is a sign that either the prior stereo calibration is not accurate enough, the sensor mounting is not

rigid or the images were not captured at the same epoch, pointing to a time synchronisation problem.

Table 2 Results of experiment #2 - bundle adjustment of the visible images, while the TIR exterior orientation was calculated based on the stereo calibration

	3D Error (cm)	*Visible images* RMSE		*TIR images* RMSE	
		Pix	μm	Pix	μm
TPs	**-**	**0.9**	**1.7**	**-**	**-**
GCP 1	1.2	0.3	0.6	11.4	193.0
GCP 2	1.7	0.4	0.6	12.6	213.5
GCP 3	1.2	0.4	0.8	12.0	203.8
GCP 4	1.0	0.5	0.9	12.1	206.0
GCP 5	1.2	0.4	0.7	13.3	226.6
GCP 6	1.9	0.3	0.6	13.6	230.5
GCP Mean	**1.4**	**0.4**	**0.7**	**12.5**	**212.2**

When the image coordinates of the GCPs are the only multimodal tie points as in experiment 3 (see in Table 3), the accuracy of the visible images again remains constant in comparison to the baseline. The main reason is the much larger number of tie points from the visible images. An important, yet not surprising result is that the GCP accuracy in object space also remains constant. This means that for the thermal images, the standard deviation of the unknowns of the adjustment is now better by a factor of 2 (1.4 cm vs. 2.9 cm) as a consequence of the joint adjustment. On the other hand, the RMS values for the thermal tie points are slightly worse (13.4 μm vs. 12.1 μm) and the RMS values of the GCP reprojection errors of the TIR images are increased by about a factor of 3.5. Although these results are much better than of those of the experiment 2, the usage the GCP image coordinates as tie points apparently deforms the thermal part of the block.

Table 3 Results of experiment #3 - jont adjustment of the visible and the TIR images with only GCPs used as tie points

	3D Error (cm)	*Visible images* RMSE		*TIR images* RMSE	
		Pix	μm	Pix	μm
TPs	**-**	***0.9***	**1.7**	***0.8***	**13.4**
GCP 1	*1.2*	*0.6*	1.1	*1.2*	19.9
GCP 2	*1.4*	0.7	1.3	2.0	33.8
GCP 3	*0.8*	0.9	1.6	1.7	29.4
GCP 4	*1.2*	0.8	1.5	1.5	25.3
GCP 5	*1.9*	0.8	1.5	3.4	58.0
GCP 6	*1.7*	0.7	1.3	2.6	43.7
GCP Mean	***1.4***	**0.7**	**1.4**	**2.1**	**35.0**

Experiment 4 (see Table 4), in which the correspondences between the two image sets are established by the result of multimodal matching (and, in addition, by the image coordinates of the GCPs), follows the findings that were obtained in the third experiment. The GCP accuracy in object space is again dominated by the visible images, thus significantly improving the results of the thermal images of the baseline. While the RMS values of the tie points are practically unchanged compared to experiment 3, the RMS values of the GCP reprojection errors in the TIR images are reduced by 20% (28.9 vs 35.0 μm). They are, however, still significantly higher than those from the baseline, pointing to remaining systematic errors in the thermal part of the common block. The fact that multimodal correspondences are created only across stereo image pairs, rather than throughout the whole block with both image modalities, might explain the large disparity.

Table 4 Results of experiment #4 - jont adjustment of the visible and the TIR images with GCPs and result of mulimodal matching used as tie points

	3D Error (cm)	*Visible images* RMSE		*TIR images* RMSE	
		Pix	μm	Pix	μm
TPs	**-**	***0.9***	**1.6**	***0.8***	**13.2**
GCP 1	*1.2*	*0.6*	1.1	*1.1*	18.9
GCP 2	*1.4*	0.7	1.2	1.6	26.7
GCP 3	*0.8*	0.9	1.6	1.9	33.0
GCP 4	*1.2*	0.8	1.4	1.0	17.5
GCP 5	*1.9*	0.7	1.4	1.9	33.0
GCP 6	*1.6*	0.7	1.3	2.6	44.2
GCP Mean	***1.4***	**0.7**	**1.3**	**1.7**	**28.9**

5. CONCLUSION AND DISCUSSION

This research contributes to multimodal point matching of visible and TIR images. It is demonstrated that point matching between visible and TIR images is feasible after pre-processing the images using the concepts of phase congruency (PC) and edge histogram descriptors (EHD). Although PC and EHD are not scale invariant, scale differences can be compensated by including prior knowledge of the ground sampling distance of the images.

Based on the multimodal matching results a joint bundle adjustment was carried out. It could be demonstrated that the accuracy of the unknowns of the adjustment, namely the RMS values of the employed GCPs in object space, are improved by a factor of 2 in the joint adjustment, compared to only using the thermal images. As a result, the proposed approach can be considered a valuable tool for enhancing the accuracy of photogrammetric processing of thermal imagery. For instance, a thermal orthophoto of higher accuracy or, alternatively, an orthophoto with co-registered visible and thermal channels can be generated, allowing for a more detailed examination of the surveyed region.

At the same time, we found certain problems which need further investigations. First, our supposingly rigid and synchronous stereo set up showed errors. Second, the joint bundle introduced some systematic errors into thermal part of the block, resulting in larger residuals in image space. These issues will be further analysed in future work.

Several additional questions remain unresolved. It remains unclear how to identify correspondences between all involved images from the two modalities. One problem is to determine how to account for rotation between two image types during the descriptor construction step. Given our dual-sensor setup, the rotation of a pair of visible and TIR images, which are not a specific

stereo pair, might be estimated by using stereo calibration between visible and thermal images and, in addition, the orientation of visible images computed from bundle adjustment. Another option is to utilise the PC's phase information (Φ in eq. 8) to determine the principal axis of the given POI. From a scientific standpoint, it is also necessary to investigate the impact of establishing tie point connections between all of the images of both modalities on the bundle adjustment accuracy.

ACKNOWLEDGMENT

The work is supported by the Arbeitsgemeinschaft industrieller Forschungsvereinigungen (AiF) under IGF-grant no. 19768 N. This support is gratefully acknowledged. The authors would like to thank our partners Fernwärme-Forschungsinstitut (FFI) GmbH, Hemmingen (Germany) and Enercity AG, Hannover (Germany), and in particular Volker Herbst (FFI) and Werner Manthey (Enercity) for their strong support.

REFERENCES

Boesch, R., 2017: Thermal Remote Sensing With Uav-Based Workflows, Int ArchPhRS, Vol. XLII-2/W6, 41-46.

Calonder, M., Lepetit, V., Strecha, C., Fua, P., 2010: BRIEF: Binary Robust Independent Elementary Features. ECCV 6314, 778-792. 10.1007/978-3-642-15561-1_56.

Canny, J., 1986: A Computational Approach To Edge Detection, IEEE TAMI, 8(6), 679–698

Cui, S., Zhong, Y., 2018: Multi-modal remote sensing image registration based on multiscale phase congruency, in: Proceedings IEEE IAPR WS on Pattern Recognition in Remote Sensing, 1–5

Du, Q., Fan, A., Ma, Y., Fan, F., Huang, J., Mei, X., 2018: Infrared and visible image registration based on scale-invariant PIIFD feature and locality preserving matching, IEEE Access 6, 64107–64121

Fischler M. A., Bolles R. C., 1981: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Comm. ACM, vol. 24 (6) 381–395

Ghassemian, H., 2016: A review of remote sensing image fusion methods, Inf. Fusion 32, 75–89

Harris, C., Stephens, M., 1988: A Combined Corner and Edge Detector, In Proc. of the 4th Alvey Vision Conference, 147-151

Hoegner, L., Hanel, A., Weinmann, M., Jutzi, B., Hinz, S., Stilla, U., 2014: Towards people detection from fused time-offlight and thermal infrared images. IntArchPhRS XL(3), 121-126.

Hoegner, L., Stilla, U., 2015. Building facade object detection from terrestrial thermal infrared image sequences combining different views. ISPRS Annals II-3/W4, 55–62.

Hoegner, L., Tuttas, S., Stilla, U., 2016. 3D building reconstruction and construction site monitoring from RGB and TIR image sets, 12th IEEE International Symposium on Electronics and Telecommunications (ISETC), Timisoara, 305-308.

Hoegner, L., Stilla, U., 2016. Automatic 3D reconstruction and texture extraction for 3D building models from thermal infrared image sequences. Proc. of QIRT'16.

Hrkać, T., Kalafatić, Z., Krapac, J., 2007, Infrared-visual image registration based on corners and hausdorff distance, in: Scandinavian Conference on Image Analysis, Springer, 383–392.

James, A.P., Dasarathy, B.V., 2014, Medical image fusion: A survey of the state of the art, Inf. Fusion 19, 4–19

Jiang, X., Ma, J., Xiao, G., Shao, Z., Guo, X., 2021. A review of multimodal image matching: Methods and applications. Information Fusion. 73. 10.1016/j.inffus.2021.02.012.

Khodaei, B., Samadzadegan, F., Javan, F. D., Hasani, H., 2015. 3D surface generation from aerial thermal imagery. IntArchPhRS XL(1), 401-405.

Kovesi, P., 1999: Image Features from Phase Congruency. Videre: Journal of Computer Vision Research 1(3), 1–26

Kovesi, P., 2003: Phase Congruency detects Corners and Edges. In: Sun C., Talbot H., Ourselin S. Adriaansen T. (Eds.): DICTA. Proc. VIIth Digital Image Computing: Techniques and Applications. The Australian Pattern Recognition Society Conference, 309–318

Lowe, D.G., 2004: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60, 91–110

Mani, V.R.S., Arivazhagan, S., 2013: Survey of medical image registration, J. Biomed. Eng. Technol. 1 (2), 8–25.

Maset, E., Fusiello, A., Crosilla, F., Toldo, R., Zorzetto, D., 2017: Photogrammetric 3D Building Reconstruction From Thermal Images. ISPRS Annals IV-2/w3, 25-32.

Mehltretter, M., Heipke, C., 2018: Illumination Invariant Dense Image Matching based on Sparse Features. 38. Wissenschaftlich-Technische Jahrestagung der DGPF und PFGK18 Tagung in München, Band 27, 584-596

Mehltretter, M., Kleinschmidt, S., Wagner, B., Heipke, C., 2019: Multimodal Dense Stereo Matching: 40th German Conference, GCPR 2018, Stuttgart, Germany, October 9-12, 2018, Proceedings. 10.1007/978-3-030-12939-2_28

Morrone M.C., Ross J.R., Burr D.C., Owens R.A., 1986. Mach bands are phase dependent. Nature, 324(6094):250–253.

Mouats, T., Aouf, N., 2013: Multimodal Stereo Correspondence based on Phase Congruency and Edge Histogram Descriptor. In: Proceedings of the 16th International Conference on Information Fusion. 1981–1987. IEEE

Mouats, T., Aouf, N. & Richardson, M.A., 2015: A novel image representation via local frequency analysis for illumination invariant stereo matching. IEEE Transactions on Image Processing, 24(9), 2685-2700.

Pech, K., Stelling, N., Karrasch, P., Maas, H.-G., 2013: Generation of multitemporal thermal orthophotos from uav data. ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 1(2), 305–310.

Wang, Q., Gao, X., Wang, F., Ji, Z., Hu, X., 2020: Feature Point Matching Method Based on Consistent Edge Structures for Infrared and Visible Images, Applied Sciences. 10. 2302. 10.3390/app10072302

Xu, C., Li, Q., Ma, X., Ma, Y., Zhou, Y., 2020: EOLGD: an edge feature descriptor-based method for long-wave infrared and visible image registration, J. Electron. Imag. 29(4) 043017, doi.org/10.1117/1.JEI.29.4.043017

Zeng, Q., Adu, J., Liu, J., Yang, J., Xu, Y., Gong, M., 2019: Real-time adaptive visible and infrared image registration based on morphological gradient and C_SIFT, J. Real-Time Image Process. 1–13.