Multi-Spectral Imaging Based GPS Denied Localisation Solution for Autonomous Platforms

Amar Ali N. Khan
Supervisor: Prof. Nabil Aouf

School of Science & Technology
Department of Electrical & Electronic Engineering
City University of London

A thesis presented for the degree of
Doctor of Philosophy
October 2024

Table Of Contents

List of Figures

List of Tables

Dedication

I dedicate this thesis and the work done during it to my supervisor Nabil Aouf and my sponsors. I would like to thank all these people for the support, guidance and opportunities that they have given me during these years.

1 Introduction

Visual Odometry (VO) has its roots in the work of David Nister [1] as an alternative to the problem encountered by the lunar rover. During NASA's exploration of the moon using the lunar rover a method to track the trajectory of the agent was necessary to determine the location of the observations on the surface of the moon. In order to determine the location of the observations the rover was equipped with an encoder on the axis of its wheels, this would track the number of turns of the wheels which could then be integrated to determine the trajectory of the rover. Once combined with the initial conditions it was possible to estimate the position of the robot on the surface of the moon. In practice, this became problematic due to the nature of the moon's surface, which led to the rover suffering from the phenomenon of wheel slip. This would lead to vast inaccuracy in the trajectory estimate over time.

Nister's introduction of visual odometry relied upon the existence of key points in images, such as Harris corners [2], which could be traced between images and so projected into the three-dimensional world and projected back into the image plane given a stereo configuration. Which allowed for the estimation of the rotation and translation undergone by the agent during this time.

With the progression of time, it became known that the visual odometry problem is a system of eight polynomial equations [3] which can be reduced to four using the help of techniques such as Grobner basis. Various other facts also came to light with the progression of time such as the effects of flow decoupling [4] and its use in calibrations [5].

The field also expanded to encompass many new sensors such as thermal cameras [6] and Inertial Measurement Units (IMU) [7]. This led to the the development of new frontiers in the field such as data fusion and robustness analysis.

The introduction of the various sensor types has led to the adoption of various data fusion methods in the field. This includes many types of filters [8] but also more sophisticated types of data fusion such as those employing Artificial Intelligence (AI) [9]. The introduction of these techniques has led to the development of many frameworks and packages for programming such solutions [10].

Whilst it is possible to utilise graph-based fusion techniques to combine all possible sensor solutions [11] the approach does not facilitate the ability to use the unique properties of a sensor to compensate for the downfalls of another.

The advancement of AI has led to many unique innovations in the field of visual odometry. This is often in the form of a deep learning model with a different domain [12] or the removal of a bias in the parameter estimation of the six degrees of freedom regression problem.

It is due to these changes that the current state of the field is a continuous expansion of the sensor configurations and data fusion methods employed. This is currently working hand in hand with the propagation of the field towards end-to-end deep learning solutions. A move away from the traditional approach to the field. The two aforementioned trends seem to have forgotten in large part the reconsolidation of the various sensors it has developed to encompass.

1.1 Overview of the Thesis

It is a principle objective of this section of the thesis to introduce the reader at a very high level to the contents of the thesis reminiscent of a bird's eye view. It provides a detailed understanding of the motivations and objectives of the thesis, in such a way as to enable the reader to find the actual scope of this thesis and the relationships between its chapters. This is known as the essential narrative of this thesis or the representation of its novel contributions to the field and why it is of significance. This section is rather concise as it does not require the same level of technical depth or expertise that is mandated by later sections.

1.1.1 Motivation

It is the principle objective of this section of the thesis to convey to the audience the sole motivating factor behind the study conducted during the PhD program, and how it resulted in the final context of the study.

The original motivation for this thesis was the ability to improve the robustness of vision-based navigation through the use of auxiliary sensors. With particular emphasis being placed upon the use of various modalities such as the thermal or infrared portion of the electromagnetic spectrum, however, the initial scope of the study was only to understand the nature of the trade-offs between improved robustness and the lowered accuracy of the final position estimation. The nature of this relationship was initially viewed as a tradeoff, due to the fact that feature matching between modalities is harder then within a single modality. This was later developed into a study of how various sensing modalities may result in larger than expected errors when combined into a single unified system. This then presented the novel opportunity to attempt to develop some method or framework by which the various sensors or modalities could interact with each other in a single unified solution to the position estimation problem without increasing the error beyond what would be expected from a single modality-based solution.

Whilst almost identical to the initial scope, the unique distinction between the two scopes of the thesis is articulated best when considering the nature of the estimation error accumulation within the position estimation systems. As the estimation error increases superlinearly in the time domain and the majority of its accumulated error results from errors in the estimation of the rotation matrix and not necessarily that of the translation vector, it quite clearly becomes apparent that any minute change is in the future matching subsection of the visual odometry pipeline would result in outside final errors in the estimation trajectory of the agent in question. This primarily results from the inconsistency in feature detection and feature mapping from the 3-dimensional space into the 2D image space between different camera sensors, arising in principle from the various parts of the electromagnetic spectrum employed by the cameras to act as the detection phenomenon.

Upon realisation that this startling fact had occurred, it became apparent that the whole of the electromagnetic spectrum could be viewed as a sensing apparatus of the real world by which solutions could select a series of trade-offs in this sampling schema, the final result of which was different samples of a three-dimensional scene known concretely as images or pictures. The fact that some phenomena would only be present in a subset of the electromagnetic spectrum and not in its entirety meant that no matter what solution was employed, there was never going to be any guarantee that a single multi-modality-based solution would always be able to match every single point between modalities - as sometimes the point would simply not exist in at least one of the modalities. This then further reduces the scope of the PhD thesis to only accurately match the feature points across modalities that were known to uniquely exist in each modality that was employed across the solution. The natural extension to this was to attempt to generalize the methods of the thesis is to be able to allow an arbitrary selection of sensing modalities in any solution to work.

Ultimately, this resulted in a novel method or taxonomy of the field of visual odometry which became readily apparent as a lens by which to view the error of accumulation of any and all solutions which opted to use senses from various parts of the electromagnetic spectrum as a basis by which to sample the three-dimensional world into a two-dimensional image and then employ a series or pipeline by which feature matching would be exploited to generalize a three-dimensional image position estimation of the agent.

This notion of generalizability across any arbitrary selection of camera sensors quickly led down the path of artificial intelligence due to the kernel optimization process and its innate generalizable nature. This, however, proved to be extremely difficult as the neural networks were not easy to optimize after training for a different set of modalities. However, the backpropagation algorithm enabled the construction of a method to produce a generalizable framework by which it was possible to do so. This was itself problematic as it did not converge to either of the preset kernels but rather would result in a new middle-ground type of kernel optimization.

This inability to produce a consistent set of neural network kernels that could apply equally well across the different portions of the electromagnetic spectrum led to the notion of the construction of a new image space in which a single kernel optimization will prove equally usable for any and all sensor configurations.

This was highly motivating and captivating as a field of study; however, the lack of diversities in the sensors which were available and the unset of the COVID-19 pandemic prevented the final set of explorations required to fully verify the extensibility of the new system (through the use of artificial intelligence) into all components or aspects of the electromagnetic spectrum when viewed as a sensing phenomenon or sampling schema, thus resulting in a final unanswered question of the nature of the latent space representation of the final components of the pipeline.

It further prevented the use of cluster analysis on the latent space representation of the fused images, which may have proved substantially beneficial in detecting outliers in the future points between modalities thereby preventing false matches and thus further improving the pose estimation. It should be noted that in this context, outliers do not refer to the typical notion of an outlier feature error in the future matching pipeline, but rather refer to a feature which only exists in a subset of modalities, for example, a queue that was observed in a visual fashion but not in a thermal sensory environment. A non-classical but easily identifiable example of this is an individual blinking. The act of blinking would result in the pixels covered by the eyelids changing in the intensity and colour values as depicted in the visible image; however, no such change would be observable in the thermal image.

The desire to fuse these images into a single image led to the introduction of various data fusion techniques into the PhD study. Due to this, various methods such as inertial navigation systems were also studied particularly for their ability to reduce drift in other systems, such as visual navigation systems through combination with each other into a single unified system known as a visual inertial navigation system. This is achieved in this particular thesis through the use of deep learning architectures and the extended kalman filter; however, it is also possible to achieve this in other ways such as the use of the particle filter.

1.1.2 Objective

Due to the relatively stagnant nature of the field - relative to its adoption of electromagnetic spectrum imagery other than that of the visible light band of the spectrum, there exists ample space for the development of such systems in addition to the reconciliation of the various alternatives to visual odometry to the existing body of work. This is the principal objective of the thesis.

It should be noted that this does not require every paper to be directed at solely solving this problem; however, each of the papers must attempt to solve at least one of the main problems underlying the reconciliation.

The three major problems concerning the reconciliation of the field and the ability to freely implement solutions ranging from a single modality to any possible combination of modalities are:

- The ability to match feature points accurately between modalities;
- The portability of traditional solutions to different modalities and combinations of modalities;
- The development of a truly multi-spectral loop closure solution.

If the PhD thesis can solve the above problems during its course, then the taxonomy of the field would shift from that in Figure 1 to that in Figure 2. This thesis was not long enough to tackle loop closure, however, the ability to port solutions into other modalties is still useful to all such loop closure developments.

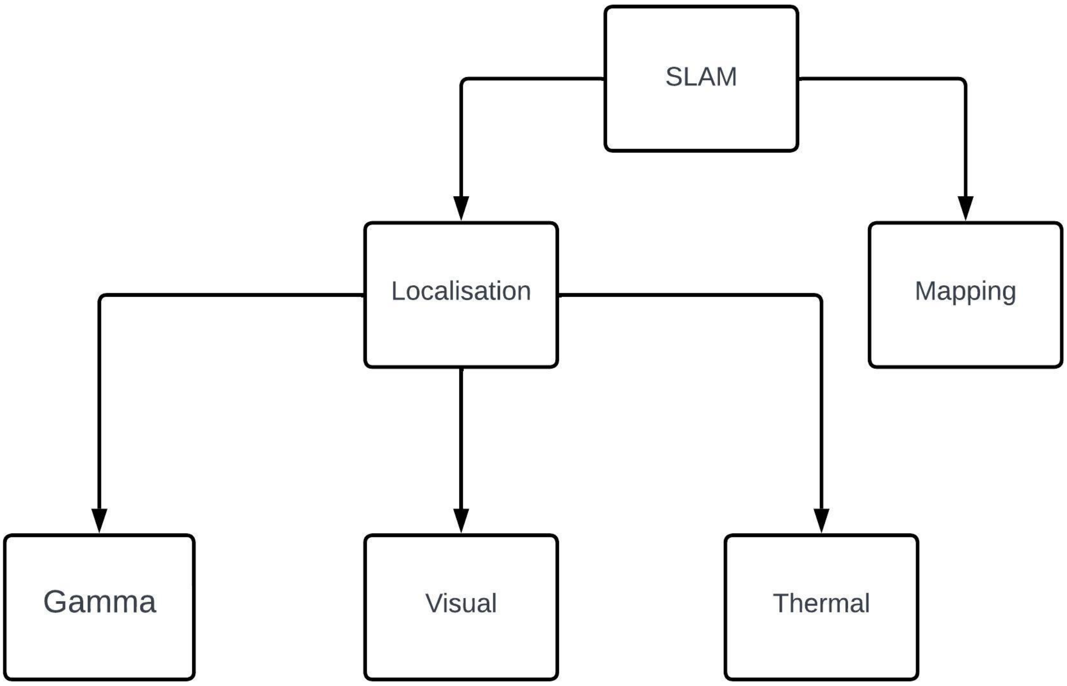

Figure 1: A graphical depiction of the current taxonomy of the field.

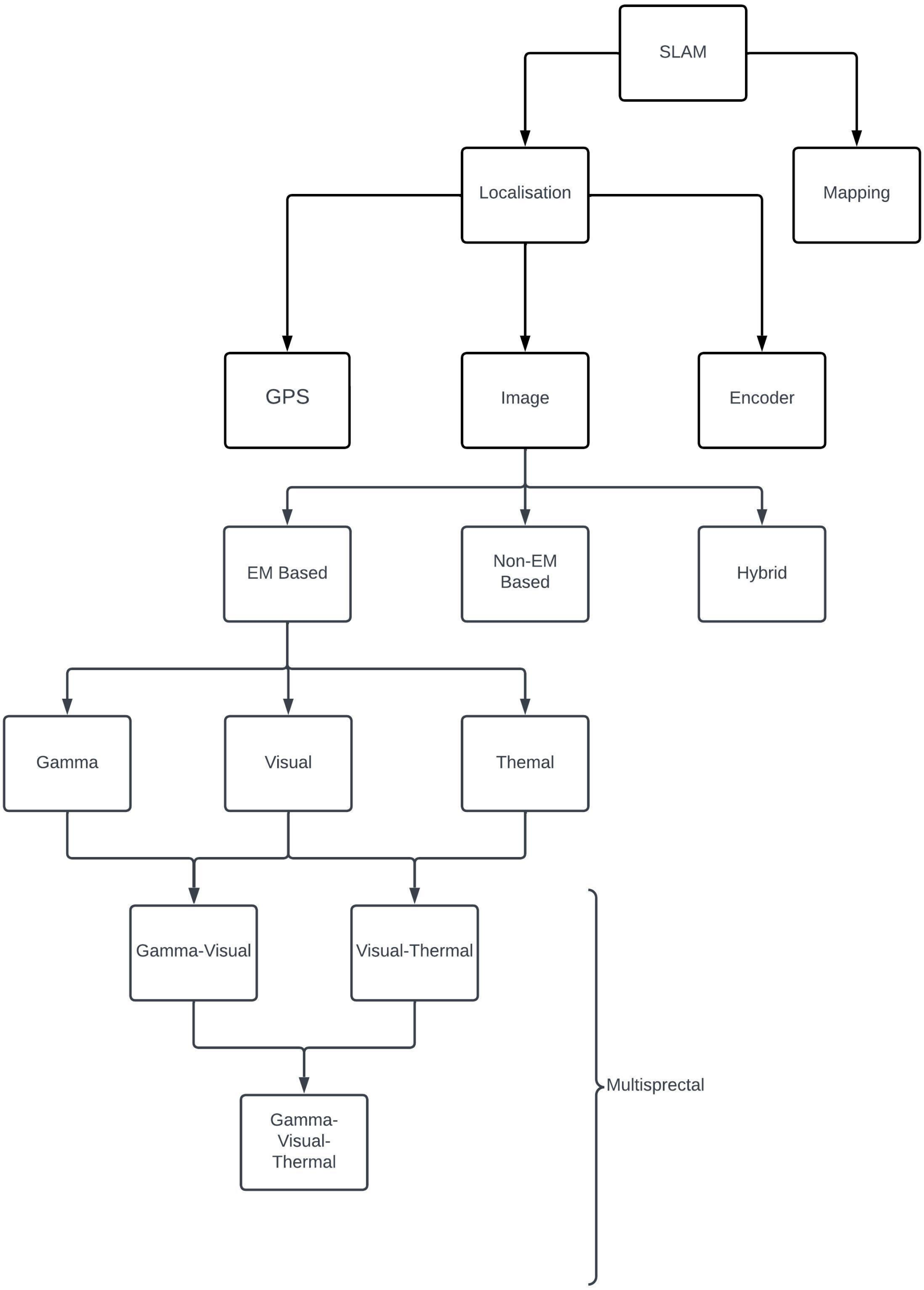

Figure 2: A graphical depiction of the theorized taxonomy of the field.

1.2 Scope of the Thesis

As is demonstrated by Figure 2, the localisation problem can be segregated into systems/solutions based upon the type of input they receive. This is equivalent to the segregation of the field based on sensor types. Due to the limited time span of this thesis, the scope of this work is limited to the image plane, i.e. camera sensor-based solutions.

The camera sensor was chosen as the funding body of this thesis, desired a solution based on passive sensors for GPS-denied environments. The problems with GPS include both the fact that the absence of a reference signal would reduce its precision to tens of meters which is often not sufficient and that the

GPS is not available in all locations on Earth. This is further complicated in space, where localisation is required, but the GPS infrastructure is likely not assessable.

The process of converting some subset of the information in a 3D environment into a 2D image using a camera sensor can be done in many unique and interesting ways, each corresponding to a different camera model. This body of work is limited to the imagery generated by the pinhole camera sensor and so the pinhole camera model. This is one of the most widely used and studied camera types and a fitting place to begin. Future work can expand the novel innovations developed as part of this thesis to other camera models.

It is often the case that employing data fusion can improve the results of a solution; for example, combining the data received by inertial sensors such as an IMU with a pinhole camera is superior to using either independently. For this reason, this work does consider employing the use of data fusion; however, to maintain the passive sensoring, it does not explore GPS or Lidar-based fusion and instead focuses on inertial fusion.

There currently exists no known suitable data set for the purposes of this PhD thesis; as such, this thesis has resulted in a novel data set for the work conducted here. However, it should be noted that many forces such as the outbreak of the COVID-19 virus and the effects of Brexit conspired to delay the production of the data set. This led to some work being done on the KITTI data set in the early stages of the program.

The development of the novel data set was limited by the hardware available, as such practical limitations resulting in only the thermal and visible portions of the EM spectrum being explored during the thesis. The work developed during this thesis was conducted in a manner that enables it to operate in all multi-spectral combinations, whilst being practically tested to work as intended on the stereo-visual/thermal and the visual/thermal multi-spectral domain.

As can be observed by the novel taxonomy in Figure 2, the new expansion of the field will lead to the discovery of many new avenues of study; however, due to time limitations, most of the newly discovered avenues are not within the scope of this thesis. The scope of this thesis is restricted to the development of the fundamental building blocks required to address the foundational issues with the employment of multi-spectral odometry.

The issues that are required to be solved before multi-spectral odometry can advance as a subfield of odometry are:

- The inability to port existing visual odometry solutions to other modalities and combinations of modalities;
- The inability to swap modalities in pre-trained AI-based solutions;
- The inability to exploit all the relevant information available in a system during loop closure;
- The inability to accurately match feature points between varying modalities.

The benefit to developing the mechanism by which portability can be achieved is a vast saving of time. Due to the historical developments of the field of visual odometry, most of the work done has been restricted to the visible light portion of the EM spectrum. Redeveloping this vast body of work for each modality would be extremely expensive in terms of time. The ability to port existing systems into other modalities and to optimize the results for that modality can radically reduce this time requirement.

The purpose of solving the interchangeability of the modalities in a solution is the robustness of the solution in its practical deployment. If we consider that the number of modalities in the solution has some effect on the complexity of the solution, then it follows that for some applications a particular degree of complexity is optimal. This would then limit the number of modalities in the solution; however, each modality is an encoding of different information in a scene, and as such the desired degree of complexity limits the availability of information in a practical solution. Depending on the chosen setting of the deployment of the solution, the optimal results may require the use of different combinations of modalities. The end-user of such systems (i.e. the funding body of this research) may then benefit from the interchangeability of the modalities as it would be far more cost-efficient than developing an entirely new solution for each environment. One example of this would be the employment of a solution in an urban environment (requiring a fusion of visual and thermal modalities), which is then re-deployed underwater (requiring a fusion of visual and ultrasonic modalities) in such a setting that the only practical cost to the end-user would be to replace a single sensor.

The benefit of enabling loop closure to exploit all the information in a multi-spectral system is an improvement in the accuracy of the solutions to all multi-spectral solutions. In the existing visible light-based solutions, loop closure can improve the estimated trajectory by exploiting revisited areas to reduce the drift in the system. However, if the same loop closure solution is applied to a multi-spectral solution, it would only be able to take into consideration a single modality, this would lead to sub-optimal results as the loop closure would not consider the information encoded by other modalities that are not available in the visible light wavelength. This may be addressed by having a separate loop closure instance for each modality; however, this would increase the time complexity of the solution, forcing the hardware cost to increase to maintain the real-time performance of some systems. This would also not take into consideration any information gained from the combination of modalities - which is the chief benefit of multi-spectral solutions. Thus, a novel loop closure method that enables the use of all information available in the system would both prevent some multi-spectral solutions from having increased manufacturing costs and enable the optimal estimate of the trajectory.

Because each modality represents an encoding of different information in a scene, the distribution of feature points in the image plane are encoded differently, which makes it difficult to accurately match feature points between modalities. This significantly hinders the performance of the solution. By solving this problem, it would become possible to address the fundamental problem of accurately using the traditional pipeline in the new multi-spectral domain.

Given the motivation by the funding body to study image based odometry and the intent to study the new multi-spectral domain, three questions naturally arise:

- Why follow the EM spectrum?
- Why not study the non-EM or hybrid solutions?
- Why not study the optimal weighting of fusion different modalities?

Given that the study of EM-based image odometry is relatively new outside of the study of the visible wavelength, there are foundational questions (which act as prerequisites) that must be asked and studied before the secondary questions. This in combination with the time limit of the PhD program has conspired to prevent me from studying the secondary question of optimal fusion. The question of optimal fusion is rather time intensive to study as it must be done in a manner considering the intrinsic characteristics of each modality and the combination thereof - this question can easily accommodate several PhDs worth of effort.

The fact that the optimal combination of the EM-based modalities has not been studied in detail and that the non-EM based methods have only a small body of sporadic work, it is currently not feasible to study sporadic methods. The study of the non-EM based division of the novel taxonomy would not conform to the desired output of the PhD. The funding body of this research had at the offset specified the requirement of producing an odometry solution combining both visual odometry and thermal odometry. To comply with this requirement, this thesis explores only the EM-based solutions.

1.2.1 Outline and Contributions

It is the principle objective of this section of the thesis to describe the novel contributions produced within the thesis, not in terms of a single empty list but in terms of an outline and contribution merit list which can be used as a descriptive measure of the contents of the thesis and therefore used to evaluate the reading process.

The first major contribution of this thesis to the field is a novel taxonomy by which to express the relationship between different aspects of the electromagnetic spectrum and sensing schema which do not arrive from the electromagnetic spectrum in relation to the visual navigation problem. This schema, in conjunction with the work done in the literature review, is sufficient to produce a novel taxonomy which can act as not just a new lens by which to view the field, but also a significant contribution in the form of one to two novel ideas and surveys which may enable the audience of the field we dissect the context of the field through the new taxonomy with appropriate referencing to further enable the development of novel multi-spectral drift reduction techniques. It is likely that the combined body of work is far too large to publish as a single journal article and therefore must be displayed into multiple papers. This could easily be done in the proceedings of a conference and or general resulting in two possible publications.

The second great contribution of this thesis towards the field is a novel application of the extended Kalman filter to fuse the inertial and visual data streams. Whilst it is common knowledge that such applications have been demonstrated previously and so are not novel, this work represents the first known publication of a stereo visual inertial navigation system employing the extended Kalman filter in the thermal modality. It is further expanded by the introduction of a novel data set and the ability to contrast the performance of the filter in various ways. For example, the use of the filter on the well-known KITTI data set allows a contrast between the impacts of using a stereo visual input stream with an inertial navigation system. Further to this, the versions of the fusion that enables stereo visual inertial odometry in the visible portion of the electromagnetic spectrum and the multi-spectral solution allow for further and far richer comparison than any which is known to exist in literature. This provides substantial insight into multi-spectral solutions and how the drift varies between them.

The next significant development of a novel program within this thesis is represented in the ability to faithfully encode or adapt the EKF based visual inertial odometry solution into an artificial intelligence based equivalent, exploiting the relationship between deep learning and back propagation. This was done in a way that exploited optical flow which is not typically the method employed in such end-to-end deep learning based or navigation solutions. This is represented in the novel architecture and findings of the study.

Following this in the thesis, one finds that there is a possible method or a generalizable framework which has been developed, in order to enable optimization of multi-spectral deep learning-based solutions. The ability to apply this to any end-to-end multiple deep learning base navigation solution is quite impressive. This was achieved through the exportation of the omnipresent back propagation method, which is the principle method by which deep learning based solutions are optimized and therefore can be assumed to exist in any and all deep learning-based solutions without loss of generality.

This thesis then concludes with the ability to construct latent space representations of multiple input image streams, which enables the fusion of multiple modalities into a single image, thereby enabling a single optimization of the kernels to produce a far more generalizable schema. This is achieved through the use of autoencoders which in themselves are a form of deep learning-based compression.

1.2.2 Published and Submitted Manuscripts

It is the principle objective of this part of the thesis to list the various publications of manuscripts which have been accepted and have been submitted during the course of the study and encompassing the PhD program.

Each result has a brief summary of why it belongs in the thesis followed by the details of the publication in an itemized list. Such details include the publication to which the manuscript was submitted to and accepted for publication in. The details also include the name of the manuscript and its listed authors where possible a unique identifier or paper ID is produced; however, it should be noted that this would have varied largely between the various publications and some publications may not even present such a unique identifier.

Finally, it should be noted that each and every entry in such a list is a reproduction of a section from the receipt of publication presented by the publication, upon acceptance of the manuscript post the peer review process. It should further be noted that only the highest level of peer review is acceptable for the manuscripts published in this PhD - this is namely a double blind review process conducted by seasoned professionals in the field.

The first manuscript that has been submitted for publication corresponds to the transition from traditional feature-based data fusion navigation systems into end-to-end base deep learning solutions. It is the final product of the work conducted in the section entitled: Deep Multispectral Inertial Odometry. The details of this publication are presented below in an itemized manner in accordance with the format described in the introduction to this section.

- Publication Title: 2021 21st International Conference on Control, Automation and Systems (ICCAS 2021)
- Title: Encoding A Mathematically Faithful DeepVIO Solution
- Author(s): Khan, Amar A. N.*; Aouf, Nabil
- Author E-mail: Amar.Khan@city.ac.uk
- Paper ID: P00563

The second manuscript that has been submitted for publication corresponds to the elimination of multispectral drift from deep end-to-end solutions through the exploitation of the back propagation algorithm. It is the final product of the work conducted in the section entitled: Multispectral Error Elimination. The details of this publication are presented below in an itemized manner in accordance with the format described in the introduction to this section.

- Publication Title: 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)
- Article Title: Backpropagation Based Deep Multispectral VO Drift Elimination
- Author(s): Khan, Amar A. N.*; Aouf, Nabil
- Author E-mail: Amar.Khan@city.ac.uk
- eCF Paper Id: 407073

The next publication of this doctoral work is concerned with the ability to fuse multispectral images and an inertial datastream into a format that can be used to predict the pose of the agent. The novel innovation here is the use of a stereo thermal EKF that has filled a long-standing hole in the literature.

- Publication Title: Conference: 2023 IEEE International Conference on Robotics and Biomimetics.
- Article Title: Towards Robust Modality Agnostic Pose Estimation via Unbiased Data Fusion
- Author(s): Khan, Amar A. N.*; Aouf, Nabil
- Author E-mail: Amar.Khan@city.ac.uk
- Submission number: 122.

An additional two papers have been yet to be accepted for publication. The first in a conference and the second in a peer-reviewed journal. The submission details are listed below:

- Publication Title: Conference: 2024 IEEE International Conference on Robotics and Biomimetics.
- Article Title: Convolutional Autoencoder Based Deep Monocular Multisprectal Visual Odometry
- Author(s): Amar Ali N. Khan and Nabil Aouf
- Author E-mail: Amar.Khan@city.ac.uk
- Submission Passcode: 8X-D6J4E5B3E9.

The paper entitled Meaningful Multisprectal Latentspace Odometry via Variational Autoencoders is to be submitted to the journal Advanced Intelligent Systems.

2 Theoretical Background and Tools

In this chapter, the fundamental concepts that serve as the basis for the contributions made throughout this thesis are introduced. It begins with a review of the techniques of traditional stereo visual odometry and continues by reviewing the process of image formation employed by modern cameras, in both the visible and thermal infrared modalities. Paying considerable focus to the underlying physical processes and mathematical models adopted by the field. Subsequently, the chapter makes an attempt to develop the foundational concepts relating to the reference frames required to align the different types of data used in this thesis. This is followed by a thorough review of the various data fusion methods employed in the field. The chapter then presents the basic concepts of artificial intelligence that are required for the innovations that are presented in subsequent chapters. Finally, the chapter concludes with an in-depth review of the various datasets employed within this thesis.

2.1 On Stereo Visual Odometry

As one of the two primary cases of visual odometry, stereo visual odometry has been the subject of considerable attention during the entire history of visual odometry and as such has been developed greatly. Stereo visual odometry is the penultimate form of visual odometry and as such allows the accurate computation of depth, but adds complexity to the system. It is through the lens of stereo visual odometry that two of the five fundamental problems of visual odometry are derived.

Stereo visual odometry attempts to take two sets of input features and deduces from them the precise rotation and translation of the camera. In doing so, it allows the computation of the trajectory of an agent, given that the initial conditions are known. The computation of the rotation and translation applied to the agent between time steps inside a stereo system defines two of the five fundamental problems of visual odometry.

In order to do this, stereo visual odometry captures a static image of the scene, from which it derives a set of feature points (by an arbitrary algorithm, for example, Harris corners). It then proceeds to have the agent propagate through the vector space (in some unknown fashion). Once the agent has reached a halt, a second static image of the new scene is captured and the same algorithm is used to extract a set of features from it. The two sets of the feature points are then compared to find a set of correspondence between the feature points. Once found, these points can be used to calculate the rotation and translation that uniquely define the propagation of the agent. Given that all static images are taken at equidistant time intervals, a full trajectory of the path can be computed.

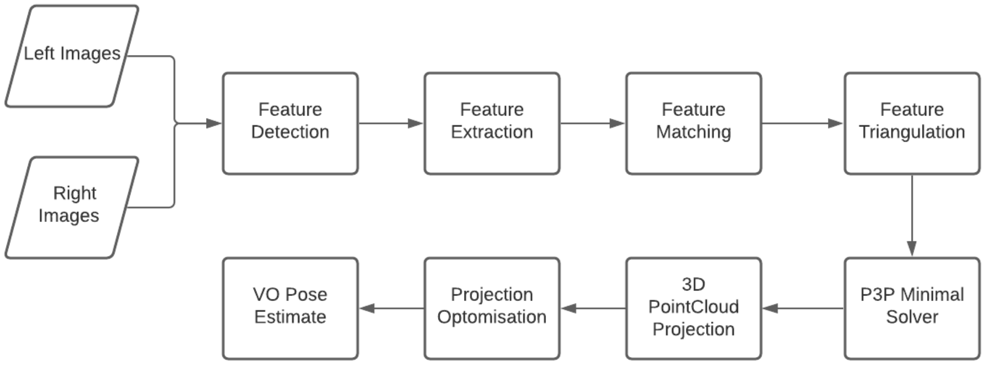

Figure 3: A graphical portrayal of the traditional stereo VO pipeline.

2.1.1 A Review of Stereo Visual Odometry

Visual odometry (VO) is traditionally viewed as the process of estimating the egomotion (position and orientation) of an agent [13] using a sequence of visual inputs collected from one or two cameras. The term VO is derived from the seminal work of Nister in 2004 [14]. The VO procedure rests on the fundamental assumption that the trajectory of the agent is piecewise deconstructable; however, this is only possible given the uninterrupted propagation of the agent.

Visual odometry provides an alternative to other forms of odometry such as wheel odometry, which has the adverse property of being subject to wheel slips; thus VO is often preferred in many rough terrain environments. The inherent desirability of VO is increased by the fact that recent algorithms can bound the relative position error from 0.1 to 2 percent [13]. This desirability is further increased in GPS-denied environments.

The study and development of visual odometry were first developed by the work of Moravec [15] and published in his PhD thesis, which was sponsored by NASA. The early adoption of VO was primarily employed by NASA on the Mars Planetary Rover [15], [16], [17] [18]. The problem of VO itself is a subset of the structure-from-motion problem (SFM), which can be traced back to [19]. SFM primarily attempts to tackle the problem of reconstructing a three-dimensional scene, given a series of two-dimensional inputs. Both SFM and VO commonly employ the use of bundle adjustment to minimise errors [20].

The early work on VO focused almost exclusively on feature-based methods and so led to the development of several feature detectors. The first feature detector was published by Hannah [21], which is now known as the Moravec corner detector [22], which was further improved upon by the work of Forstner [23] and then by Harris and Stephens [19], [2].

The initial work tested by Moravec was employed on a Mars rover with a single sliding camera, which allowed for the removal of outliers by checking for depth inconsistencies in the eight images taken at each time interval.

Further advancements occurred when the employment of the error covariance matrix was employed in the motion estimation step. Notable work includes [16], [17], which managed to achieve a 2% relative error on a 5.5-m path. [17], [24] further advanced the field by employing the inclusion of an absolute orientation sensor such as a compass or omnidirectional camera. [24] also showed that the rate of growth of the error is super linear in the distance transversed when based solely upon cameras, which eventually led to a relative position error of 1.2% on a 20-m path.

The next stage of improvement came from the relation that the employment of dense stereo methods, which was superior in selecting key points [25], [26] as there is a strong correlation between feature depth and the shape of the correlation curve [13]. This fact was later combined with the Harris corner detector and RANSAC methods by [27], [28] to improve upon the results delivered by [26].

A notably different approach was employed by [25], which encompassed the use of Shi-Tomasi corners and the iterative closest-point algorithm [29]. The work of Nister [14] significantly differed from the aforementioned work as it detected features in all frames, as opposed to tracking them, and is the first real-time long-run implementation of a VO algorithm that had a robust outlier-rejection scheme.

[30] pioneered approaches based upon local windowed bundle adjustment to recover the motion of an agent and 3D maps. This approach also relied upon the use of five-point RANSAC [31] to remove outliers. The next major breakthrough in the field came from the work of [32] which decoupled the rotation and translation allowing the solution of each by independent constraints.

The decoupling of rotation and translation in the stereo case by works such as [33] has led to improvements in many different areas of computer vision, which directly affects all pre-existent visual odometry systems. This fact is easily conveyed with [5], which employs decoupling to better approximate the intrinsic camera parameters associated with all visual cameras; this has led to better trajectory estimates in visual odometry solutions by the reduction of radial distortion effects. Works such as [4] attempt to employ decoupling to minimise the re-projection error by employing the constraints on rotation independently from translation.

[33] builds upon the decoupling of rotation and translation by identifying that the optical flow [34] observed upon a unit sphere centred at the centre of projection undergoes distinct transformations when translated or rotated. By realising that the effects of the two transformations did not result in some sort of an interference pattern upon the unit sphere, [33] was able to determine the rotation matrix by exploiting the use of a new epipolar constraint; however, it was unable to determine the translation undergone by the agent or to triangulate the features.

The introduction of the sliding window bundle adjustment has led to it seeing widespread adoption [35], [36], [37], [38]. One of the most significant results parting to the employment of sliding window bundle adjustment originates in [37] and demonstrates that the sliding window bundle adjustment can decrease the final position error by a scale factor of up to 5.

[39] further improved the accuracy of the motion estimation by incorporating the quadrifocal tensor into their methods as this facilitated the direct estimation of motion from 2D-to-2D images as appose to the 3D-to-3D and 3D-to-2D methods that predated it.

In a stereo visual odometry system, [33] has shown that it is possible to achieve the exact rotation—in a noiseless environment—by the use of decoupling methods; however, [33] fails to do this in a monocular system which can arise from most stereo systems. This implies that should a camera fail or the baseline become too small, it may become impossible to retrieve the depth and accurate motion given by stereo methods. This then gives birth to the question of how small a baseline can be before a stereo system essentially degrades into a monocular system. This question has been subject to great interest, most notably in [40], which develops the ability to triangulate distances from a baseline as small as 8 mm; however, its maximum range is only 400 mm. [40] further builds upon the works of [41], [42], [43] by providing an alternative to the focal stack method, which is robust over small and large data sets.

The quest to optimise baselines in order to derive depth, by triangulation, from as little or as large a disparity between camera positions, leads to the question of disparity optimisation. To this end, [44] has developed a construct to preserve the stereo motion in a series of images that where[!!!] later subject to disparity manipulation, allowing the effective implementation of works like [45]. [44] concludes by presenting a method by which to create stereo image pairs conforming to the optimised disparity map, providing a framework for works such as [46], [47].

The visual odometry literature is rich with computational examples of the drift found within an algorithm; however, there exist relatively few mathematical models of drift from which inferences may be deduced. [48] found that the drift observed in visual odometry systems can be fitted accurately with a variant of Allan variance [49] based upon hidden Markov models, proving drift to be super linear in distance travelled.

2.1.2 The Techniques of Visual Odometry

Stereo visual odometry consists of four fundamental algorithms upon which much of the field resides:

- Triangulation: calculating the precise depth of a feature in the three-dimensional space, allowing the formation of a depth map of the scene;
- Essential Matrix: the essential matrix allows for the calculation of the propagation of the agent by 2D-to-2D correspondence of features.
- Point Clouds: the use of point clouds allows the calculation of the propagation of the agent by feature correspondence in the three-dimensional space.
- Bundle Adjustment: bundle adjustment allows for the calculation of the propagation of the agent by corresponding features between 3D and 2D.

Triangulation is often the goal of many visual odometry applications; however, it does not act as the basis upon which many other algorithms are derived. The other three algorithms act as representatives of the three main classes under which all stereo visual odometry algorithms must fall.

2.1.2.1 Triangulation

Given two cameras in a three-dimensional vector space, it is possible to find their relative positions, from which a baseline can be constructed. If the focal lengths of both cameras are known, then it becomes possible to depict the basis of a rectangular region relating the two centres of projection. Noting that under the pinhole camera model, each point in the three-dimensional space (inside the field of view of the camera) is represented on the two-dimensional image plane from the data inferred from a single beam of light and that both images have the same features, it becomes possible to draw a ray from each centre of projection through the corresponding point on their respective image planes and onto the three-dimensional feature point. The perpendicular bisector of the baseline and the three-dimensional feature point is equivalent to the sum of the depth and the focal length.

As the perpendicular bisector is equal to the depth and the focal length F, by subtracting the focal length from the perpendicular bisector, the actual depth of the feature in the three-dimensional space becomes known. Geometrically, this is equivalent to the length of the perpendicular bisector originating from the three-dimensional feature point and the line parallel to the baseline B which is closer to the feature point than the baseline by a distance equal to the magnitude of the focal length. The same point in both images is taken to form the base of one of the two triangles, this point is called x the subscript l and r determine if its in the left or right image respectively.

Then, by similar triangles, the depth Z is given by:

$$Z = \frac{FB}{x_l - x_r} \tag{1}$$

2.1.2.2 The Essential Matrix

Given two cameras of equal dimensions in the three-dimensional space, the transformation between them is no longer dependent upon a scaling factor; as such it can be expressed as follows:

$$C' = RC + T \tag{2}$$

where the second camera C' is located at an arbitrary position in the three-dimensional space, which can be related to the first camera C by some rotation R and translation T. From this relation, it is possible to derive a single matrix (the essential matrix E) which relates the camera positions and therefore encodes the rotation and translation between the two camera positions.

$$TxC' = Tx(RC + T) = TxRC + TxT = TxRC \tag{3}$$

$$C'.(TxC') = C'.(TxRC) = 0 \tag{4}$$

$$C'EC = 0|E = TxR \tag{5}$$

The derivation of the essential matrix proves useful in practice as it can be fitted to the features detected by each camera allowing the construction of a rigid transform between the two camera positions. If this is done between camera positions at different times, it becomes possible to decompose the essential matrix into the rotation and translation undergone by the camera during the elapsed time. It should be noted that Equation 4 is zero as the angle between the camera and itself is zero forcing the product to be zero.

Visual odometry relies upon the use of the essential matrix to retrieve the rotation and translation undergone by the agent between two input images at a time step. The continuity of this process to all time steps allows the computation of a relative path that, when combined with the initial pose of the agent, forms a trajectory.

2.1.2.3 The Fundamental Matrix

The fundamental matrix F is the extension of the essential matrix E. It applies the relationship the pinhole camera model forced upon the three-dimensional feature points and their two-dimensional correspondence to the relationship described by the essential matrix. The fundamental matrix relies upon the fact that under the pinhole camera model, a point in the three-dimensional space can be mapped onto the image plane by the intrinsic matrix. This is done through the exploitation of the intrinsic matrix K which describes the parameters of the camera and its lense.

$$P' = Kp', P = Kp \tag{6}$$

where the primed feature point P' refers to the point P after the time step.

$$P'EP = 0 \Rightarrow (Kp')^T E(Kp) = 0 \tag{7}$$

$$(Kp')^T E(Kp) = (p')^T K^T EKp = 0 \tag{8}$$

$$p'^T Fp = 0 | K^T EK = F \tag{9}$$

Notice that whilst the essential matrix employs the use of the world coordinate frame of reference, the fundamental matrix depicts a relationship between the images coordinates. This results in both matrices being heavily used in practice, but for different tasks. Some open-source libraries such as OpenCV often require the derivation of the fundamental matrix to derive the essential matrix, which is then deconstructed into the rotation and translation undergone by the agent at the corresponding time step; however, software such as Matlab allows the derivation of the essential matrix directly from the point correspondence. Equation 7 is equal to zero due to the constrains of epipolar geometry.

2.1.2.4 Point Clouds

Point clouds are significantly more than an alternative to the essential matrix. While the essential matrix allows the derivation of the rotation and translation undergone by an agent via the use of two-dimensional correspondence, point clouds allow the derivation of an agent's propagation via the use of three-dimensional correspondence, thereby opening up a subfield of visual odometry.

Due to the central application of point clouds being in the three-dimensional vector space, they are often only utilised for a certain subset of sensors including LIDAR, which generates the point cloud (a collection of points in a three-dimensional vector space) directly; however, alternative uses of point clouds rely on the mapping of two-dimensional features to three dimensions.

The point-cloud algorithm takes as input two sets of corresponding features in the three-dimensional vector space and computes the inverse transformation from the second set to the first, the inverse of which corresponds to the rigid transformation between the two sets of the propagation of the agent. The series of rigid transformations can then be concatenated into a single matrix representing the change from the agent's initial pose to its final pose or used to express the trajectory of the agent.

2.1.2.5 Bundle Adjustment

Bundle adjustment is the fundamental algorithm of visual odometry using feature correspondence between the two-dimensional and three-dimensional cases. The basic case of bundle adjustment focuses on minimising the error formed by projecting the three-dimensional features onto the two-dimensional image plane subject to some specified loss function—normally the sum of square errors.

Unlike the essential matrix and the point cloud methods, bundle adjustment does not attempt to directly infer a transformation or inverse matrix, but instead attempts to employ a similar concept to that of integration (by use of the limit operation) to derive a globally optimal linear transformation, which is then expressible as a matrix.

Given the position of the feature point in the three-dimensional space and its corresponding point in the image plane, the projection of the three-dimensional feature point can be used to uniquely define a plane. Within that plane, three points correspond to a triangle where the base of the triangle is given by the error between the projection and correspondence of the three-dimensional feature point.

The main component in the bundle adjustment algorithm attempts to find the projection transformation that would minimise the distance between the projected point and the corresponding point, or put more succinctly, attempts to minimise the area of the triangle which in the limit becomes zero once the projection ray converges to the feature ray.

In practice, this is often impossible to achieve due to the interference caused by the existence of noise. In an attempt to minimise the error caused by the noise, most visual odometry systems attempt to repeat this with as many features as possible (subject to time and hardware constraints); however, as all the features attempt to determine the same transformation, they are often used as a single system of equations to minimise some objective loss function. This amounts to a minimisation problem which has been well studied in operational research.

2.1.2.6 RANSAC

RANSAC attempts to remove any outliers from the data set; however, it does so based on probabilities and is therefore prone to failure. However, it has achieved good results in practice. RANSAC begins by selecting a subset S of the data set and fits to it the predefined model, which allows the construction of non-overlapping partitions of S into the data records generated by the model (inliers) and those not generated by the model (outliers), where the outliers are defined as records with a distance from the estimate of the model greater than some threshold d. The threshold is normally set by the employment of a predefined distance model, implying that the distance model generates the error model. The set of inliers is taken to be the consensus set generated from S.

The algorithm is then set to terminate when the size of the consensus set exceeds some predefined threshold T or is allowed to run N iterations and returns the largest value of the consensus set. In practice, the most useful understanding of the RANSAC algorithm arises from the relationship of the number of samples that need to be drawn to achieve a certain probability of success.

In order to derive this, let $1 - e$ be the probability that a point is generated by the model. Then $(1 - e)s$ becomes the probability that all the points in a sample are inliers. This means $1-(1-e)s$ is the probability that every member of a sample is an outlier and as such $[1-(1-e)s]N$ is the probability that every member of every set sampled is an outlier. By setting some arbitrary probability $1 - \epsilon$ that bounds the probability that all points sampled are outliers, the following relationship is derived:

$$[1 - (1 - e)s]^N = 1 - \epsilon \tag{10}$$

$$\ln([1 - (1 - e)s]^N) = \ln(1 - \epsilon) \tag{11}$$

$$N * \ln([1 - (1 - e)s]) = \ln(1 - \epsilon) \tag{12}$$

$$\therefore N = \frac{\ln(1 - \epsilon)}{\ln([1 - (1 - e)s])} \tag{13}$$

Noting that the subtraction and division laws or logs [50] ensure:

$$\frac{\ln(1 - \epsilon)}{\ln([1 - (1 - e)s])} = \ln(1 - \epsilon + (1 - e)^s - 1) = \ln((1 - e)^s - \epsilon) \tag{14}$$

The effectiveness of RANSAC leads it to be one of the most widely employed algorithms in computer vision, despite its tendency to fail.

2.2 On Image Formation

A camera is a device that samples from a three-dimensional dynamic scene and maps the sample to a 2D representation of the 3D scene. This mapping loses information that is contained in the 3D world. Whilst the mapping fails to capture several types of information in the sample, the main concern is normally the loss of the depth information in the scene. This, however, is not the principal concern of this thesis; rather, the principal concern of this thesis is the employment of multispectral data samples and their effects on the robustness of the trajectory.

The advancement of modern technology has made the employment of visual cameras ubiquitous in society. Whilst this has introduced the general public to the uses of visible cameras (and to some degree thermal cameras), most people are unaware of the wide range of cameras available [51]. The full range of this taxonomy of sampling devices is well outside the scope of this thesis; however, this thesis will provide a treatment of the image formation process undergone by both the visible and infrared cameras.

2.2.1 Visible Wavelength

Today's term "camera" generally refers to solid-state cameras rather than their predecessor vacuum-tube counterparts which produced an analogue voltage output proportional to light hitting photo conductive electrodes [52]. Solid-state camera usages have evolved substantially compared to their vacuum-tube predecessors as evidenced by smaller sizes, higher robustness ratings, and resistance against damages caused by higher illumination intensities.

At present, area image sensors use a dual stage process: the first stage is the conversion of energy from photons into electric charge at each pixel level; secondly, amplifying and converting this charge into electrical signals for the amplifier. These processes enlist silicon semiconductor elements (typically photodiodes, photo capacitors, or photo conductors) which absorb electrons by creating electron-hole pairs; this final stage requires readout elements with most typically being charge coupled devices (CCD) or complementary metal oxide conductors (CMOS).

CCD cameras work by transporting charge packets from image-sensoring elements' surfaces directly to the output, where they are converted into voltage [53]. As part of this process, physical storage areas also come into play to make sure that any extraneous charge levels do not hinder image quality - anti-blooming circuits may even help mitigate them!

CMOS sensors attach transistors to every individual pixel to allow individual amplifying and enable on chip image processing, making CMOS sensors the current standard in digital cameras due to their smaller form factor and increased functionality compared with their counterparts such as CCD sensors and traditional SCM ones [54]. Their reduced form factor also makes them popular choices among mobile phone cameras; some even support artificial intelligence technology which leads to AI enhanced images!

2.2.2 Infrared Wavelengths

An infrared (IR) camera's ability to map thermal emissions from three-dimensional scenes onto two-dimensional image planes is made possible via its parallel mechanism with that of visible cameras. The focal plane array (FPA), located on the thermal camera's IR detection unit, serves as the front of its detection process and comes in either scanning or starting-based formats. Scanning FPA employs linear arrays which scan across the scene across its horizontal field of view (FOV), using rotating mirrors as two-dimensional image creators to produce two-dimensional imagery. Starting-based FPAs employ dedicated pixels that equal CCD or CMOS sensors in terms of visible light operation; all cameras used here fall within either category regardless of what part of the electromagnetic spectrum they operate within.

Infrared detection can be accomplished using either photon or thermal detectors within FPAs that, when activated by measurement of photon or thermal decay, convert their strength into electrical signals for detection purposes. Photon-based FPAs come in two variants: photovoltaic and photoconductive; both require cooling elements for effective results [55]. Intrinsic detectors tend to operate at higher temperatures while dissipating less power compared with extrinsic detectors. Variability in results may be attributable to the materials employed in producing sensors; typically, these include mercury cadmium telluride, indium antimony, and doped silicon. Thermal detectors typically absorb any incidental radiation on scene and record any change in system temperature; consequently, they usually do not need cooling. Resistive thermal detectors utilize resistive materials near their silicon readout to change local resistance values according to temperature changes in their surroundings, while capacitive detectors based on pyroelectric effects use different materials altogether.

2.3 On Data Fusion

2.3.1 Defining Data Fusion

The field of data fusion has seen considerable interest due to the advancement of sensor technology; this has resulted in several generic [56], [57], [58] and spacic [59], [60], [61], [62], [63] reviews of the field. This has resulted in a wide series of definitions of the field ranging from practical to abstract. [64] acts as an example of the litter of definitions data fusion enjoys, as it simply augments prior work to allow for the inclusion of multiple sensors. [65] has conducted a review of a subset of definitions for the field resulting in the view of data fusion being "the study of efficient methods for automatically or semi-automatically transforming information from different sources and data points in time to a representation that provides support to human or automated decision making"; however, as this definition is a rather unfaithful representation of the underlying principals and mechanisms of the field, it is the view of this report that data fusion is simply the employment of data from multiple sources to approximate the abstract parameters of a probability distribution which is assumed to model the system in question.

The employment of data fusion over a single sensor system has several advantageous benefits [66]. For most practical applications, the most advantageous benefits lie in being able to reduce data ambiguity and increasing reliability or the robustness of the solution; however, such benefits are not without cost as the transmission of data between a large numbers of agents or nodes may cause potential collisions and the transmission of redundant data.

This range of advantages and drawbacks coupled with the myriad of definitions of sensor fusion has led to the development of several frameworks which attempt to encompass data fusion. The most renown of which is the JDL model; however, as the JDL model does not extend to the processing of data, Dasarathy's framework [67] acts as both an extension and alternative by taking the view of a software engineer. A further and much more powerful extension comes in the form of [68] which acts as an abstraction facilitating the fusion of decision uncertainties and the decisions themselves. The most general, and therefore all encompassing, framework originates in [69] which employs the use of category theory; however, the realm of mathematics offers a variety of constructs that may act as the basis of all knowledge, not just category theory.

2.3.2 The Plagues of Data Fusion

The employment of data fusion into a real-world solution often results in the addition of a diverse range of issues that would otherwise not limit the performance of the solution to as great an extent. The plagues brought forth by the application of data fusion can be categorized into the following non-exhaustive catalogue:

- Data Imperfections: noise is present in all real-world sensor reading; however, when employing the use of multiple sensors of varying types, the noise in a system may approach heights which make the solution impossible, absent a method to reduce noise through data redundancy;
- Outliers and Spurious Data: ambiguity and inconsistency in the environment such as temperature variation through the day may be confused with errors that arise from sensor ambiguities and noise;
- Conflicting Data: the applications of belief systems may result in unexpected errors;
- Data Modality: by employing a vast array of sensor types, it is possible to incorporate a rich array of data types into the solution; however, this results in the need for a scheme to combine different sensor readings.
- Data Correlation: when sensor nodes are exposed to the same environment that are likely to be exposed to the same noise, without processing this could result in large errors;
- Data Alignment: the fusion of various data entry from the array of nodes requires for each nodes data to be represented into a common frame of reference prior to processing;
- Static vs Dynamic: it is often necessary for data fusion schemes to incorporate a recent history of measurements into the fusion process.

In order to produce results with minimum noise, a series of algorithms exists which take as input the nature of the cross covariance of the data; however, as the sensors measure the same common phenomena, it is possible and indeed often the case that the data is correlated with an unknown cross covariance [70], [71], both highlighting the dangers of double-counting data or, as it is more commonly known, data incest. The most problematic outcome of data incest is the evolution of a convergent algorithm into a divergent one [72]. There exist several methods which attempt to deal with this, such as the Kalman filter; however, they are plagued with quadratic scaling [72]; nevertheless, it is possible to keep track of all the nodes in the fusion pipeline. However, this scales egregiously with the number of nodes [73], and it also limits the range upon which a single solution may span.

Due to the placement of nodes in the real world, it is often the case that data incest arises due to the varying paths the data may propagate upon in order to arrive at its desired location—this problem is far more prevalent in distributed systems [74]. This problem can be dealt with either by the removal of data incest [75] or the reconstruction of measurements [76]. Both of these solution families currently impose topological restrictions on the arrangement and placement of sensors, prohibiting the employment of such solutions in highly irregular geographical topologies or high flexibility tasks. There has been some work [77], [78] attempting to remove data incest from arbitrary topologies; however, the work is far from perfect. Extensions to this work do consider far more complex systems in which other problems are present such as data clutter [79]. Covariance intersection is perhaps the most common method to deal with the problem of data incest [80]; CI, as it is more commonly known, has been shown to be an optimal solution when attempting to find the upper bounds of combined covariance [81]. The most impressive fact about this work is its application to all abstract probability distributions [82]; however, like all work which is applicable in the abstract, the process is computationally demanding. This has led to more restricted variants of CI becoming quite frequent in the literature [83], [84].

Like most popular algorithms, there exists a popular alternative to the CI algorithm, the most appealing of which is the Largest Ellipsoid (LE), which represents a family of alternative approaches based upon it. This family of alternative algorithms cannot at present be employed in conjunction with any framework more robust than the family of approaches based upon the Kalman filter.

Spurious data is also a considerable problem and requires several alterations to the algorithm; for example, sensors may be subject to short spike faults and slowly developing failure. One such possibility is the breakdown of Kalman filters [85]. Most techniques that attempt to deal with this problem rely upon the use of prior data which is not available in most applications [86].

Due to the topologies of the solution and the method of communication between nodes, many systems suffer from data arriving for processing out of sequence (as measured by a time date stamp present upon all sensor readings). The trial solution is commonly employed; however, it leads to an excessive loss of information and therefore limits the performance of the solution. The main issue regarding this is how to incorporate data from previous times into the current reading [8], [87];

This often prohibits the employment of RNNs or relegates them to processing a submodule of the algorithm.

The pioneering work in the field of out-of-sequence data corners only single-lag data [88], [89]; however, the advancement of the field has led to the extension of such work to apply to arbitrary lags [90], [91], [92]. [92] is the most abstract of them, as it allows for the construction of a single unifying framework of out-of-sequence data.

The vast majority of research in the field of out-of-sequence data concerns single target filtering and not the problem of data clutter, which is inherent in all multi-sensor data fusion problems [93]. [93] extends the pre-existing literature by considering likelihood computation and hypothesis management in multi-target systems. This is further improved upon by the work of [94], which considers disordered tracks as opposed to measurements.

Conflicting data always present both a problem and potential innovation. [95] demonstrates that naive applications of Dempsters rule to conflicting data, resulting in counter-intuitive results. Such behaviour has left the rule subject to much criticism [96], resulting in several solutions to the rule [97], [98], [99]. Somehow other scholars have defended the rule [100]. In practice, the restrictions imposed by the aforementioned methods render many solutions infeasible; this has led to the evolution of several simplified methods [98], [99]. One of the best approaches to the fusion of conflicting data is to employ the Bayesian framework [101], [102].

The existing literature on data fusion suggests the existence of several areas of research which will yield the most fruit in the upcoming years; the remainder of this section briefly develops on them.

Opportunistic data fusion is more of a paradigm than a method and attempts to use all the nodes in a network as shared resources [103], allowing the fusion of data in an opportunistic manner, potentially solving both data incest and the out-of-sequence data problems. This is done through the on-the-fly discovery of nodes in the network [103], [104].

The advent of adaptive fusion allows the process of data fusion to occur in dynamically changing environments through the use of on-the-fly re-estimation of existing parameters. This approach has led to the discovery of several new Kalman filters such as the novel adaptive Kalman filter (NAKF), which achieves adaption through the use of mathematical functions based on covariance matching [105]. This also enables the use of several machine-learning methods [106].

Whilst not the most practical of outlooks, automated fusion is feasible and allows the development of distinct models [107]. The examination of the reliability of belief has given rise to the notion of the degree of uncertainty and models which exploit it.

This notion has led to a range of domain-specific methods which rely heavily upon contextual information [108], artificial intelligence [109], possibility theory [110], and human expertise [111]. This in turn has led to fundamental works on uncertainty this work[!!!] into pre-existing frameworks such as Demper-Shafer theory [112], transferable belief model [113], and probability theory [114].

2.4 On Data Sets and The Experimental Setup

2.4.1 Open-Source Data Sets for Motion Estimation

The KITTI data set is an industry standard benchmark data set used for visual odometry and related computer vision tasks such as stereo and depth estimation. Conceived by Karlsruhe Institute of Technology and Toyota Technological Institute in Chicago, this benchmark data set includes multiple forms of car-mounted sensor suite data such as stereo camera images, Velodyne laser scanner measurements, GPS/IMU measurements, and ground truth poses captured from its sensors.

The visual odometry component of this data set features sequences of monocular and stereo camera images captured from moving cars on different routes within Karlsruhe city in Germany, taken at 60 frames per second at resolutions between 93-442×375 pixels with timestamps, camera calibration parameters, and ground truth camera poses included as part of this collection. 22 sequences make up this portion of this data set with each sequence consisting of between 93-8800 images at resolution 1242×375 pixels; these sequences consist of between 93-44224 images at resolution 1242×375 pixels each! There are 22 sequences total in total consisting each including timestamps, as well as timestamps along with camera calibration parameters, as well as ground truth camera poses provided for every camera frame.

KITTI was collected with the use of a car-mounted sensor suite consisting of multiple sensors synced up together in order to capture data simultaneously. This system was mounted onto an altered car equipped with roof racks specifically tailored for holding sensors and driven around various routes in Karlsruhe, Germany.

KITTI data set featured two grayscale stereo cameras to capture images of the road ahead. Mounted on the roof rack and separated by approximately 54 cm, these 12-42×375 resolution cameras were calibrated using standard chessboard patterns in order to correct distortion while also estimating intrinsic and extrinsic camera parameters.

KITTI sensor suite included not only stereo cameras but also a Velodyne laser scanner to capture 3D point cloud data of its environment. Mounted on the roof rack of a car and rotating at a 10-Hz rate to cover the 360-degree view with 64 laser beams with an approximately 100-m range, each displaying a vertical resolution of 0.4 degrees and horizontal resolution of 0.08 degrees, respectively, this scanner was capable of gathering full 360-degree view data of its surroundings.

KITTI sensor suite also included a GPS/IMU system to record measurements of the car's position, velocity, and orientation. A GPS receiver mounted on top of the car connected with IMU installed inside of it provided measurements for latitude, longitude and altitude while IMU provided acceleration and angular velocity readings simultaneously in three dimensions.

All sensors included in the KITTI data set were coordinated using an audio trigger signal generated from stereo cameras to timestamp all sensor data so it could easily be combined and utilized in computer vision applications.

Overall, the KITTI sensor suite was intended to capture data that would serve various computer vision tasks - visual odometry, stereo and depth estimation, and 3D object detection and tracking being among them - with high-quality ground truth data providing valuable resources. Careful calibration and synchronization ensured the sensors combined with quality ground truth make the KITTI data set an indispensable asset to researchers working within various fields of computer vision research.

The KITTI data set scoreboard is an open leaderboard maintained by the KITTI team to track the performance of various computer vision algorithms on this data set. This scoreboard features different evaluation metrics for each task, as well as rankings of the top performing algorithms with links back to their code or publications.

The KITTI scoreboard evaluates visual odometry algorithms using various metrics, including absolute trajectory error (ATE) and relative pose error (RPE). ATE measures differences between estimated camera poses at each time step versus ground truth camera poses, while RPE indicates errors over an extended sequence of frames. It also offers metrics for rotation/translation error separately, as well as combined pose error measurements.

The KITTI scoreboard provides algorithms with an objective evaluation for stereo and depth estimation tasks using metrics such as mean absolute error (MAE), root mean squared error (RMSE), and the percentage of pixels with errors below a certain threshold. Additionally, metrics for endpoint errors and disparity accuracy can also be provided on this scoreboard.

The KITTI scoreboard provides metrics to evaluate 3D object detection and tracking algorithms such as average precision (AP), which measures the accuracy of object detection, multi-object tracking accuracy (MOTA), which assesses accuracy over a longer sequence of frames, as well as metrics specific to object classes like cars, pedestrians, and cyclists.

The KITTI scoreboard is regularly updated to display the results of participating algorithms, providing researchers with a useful way to benchmark them with others in the field and compare algorithms against each other. Open to all researchers, the leaderboard encourages codes and techniques sharing to advance computer vision research.

The problem with the KITTI dataset is the fact that it does not provide any multi-spectral components such as thermal or ultrasonic imagery. This is rectified by several others however these daasets are often taken indoors with little illumination variation and often do not have a large enough scale of the trajectory to enable adequate testing of state-of-the-art algorithms designed for extended flights on on unmanned aerial vehicles or unmanned ground vehicles. This has been a systematic problem within the literature since its inception however recently there has been some movement in this direction for example the MS2 data set enables multi-spectral algorithms to be tested on a suitable trajectory with adequate length under some degree of illumination variation and heating variation with both visual and thermal modalities. However there are still several problems with this data set firstly it is limited to the visual and thermal modalities although it doesn't enable two distinct variations of the thermal mortality, this prevents the algorithms and neural networks designed within this thesis from being tested upon or adapted to trimodaility settings and beyond.

It should also be noted that this particular dataset was taken within the Korean Peninsula whereas the popular KITTI dataset was taken within Europe the vast difference in the road conditions and infrastructure of these two datasets prohibit direct comparison of the trajectories, it would further suggest that training on this dataset limits the model to the geographic location which the dataset was taken and as the conditions of the road and driving environments deviate from those of its original settings the results of the model will get substantially worse. This does mean that It is unlikely that the models and algorithms produced within this thesis are able to suitably operate within remote rural locations. It should be noted that this is not the central problem of this thesis, but may affect the results of each of the models as the scenery within the locations differs. For example, trajectories which have a park setting will be grossly underestimated or represented within the training sets of each and every model and therefore produce an inherent bias towards urbanization in each and every model. Whilst this is worthy of note it does not prohibit the contribution of this thesis which is the reconciliation of the multispectral taxonomy.

The dataset is presented in four components two taken in daytime and two taken at nighttime where one of the day times is done in rainy conditions and the other In clear sky conditions [115]. Both nighttime components are done with clear sky conditions however they represent different components of an urbanised environment those with high levels of illumination and a variety of light sources and those with low levels of illumination due to a lack of light sources. This does to some extent aid in removing the bias from the training set, however, it should be noted in both cases the feature points detected belong to an urban setting such as the sides of a skyscraper which is not likely to be present in rural environments.

The dataset has approximately 10 trajectories which are downloadable with pre-synchronization of the various sensors [115], like the KITTI dataset this may be subject to change, the trajectory is from approximately 8 to 25 gigabytes in size and runs over thousands of images. As demonstrated within the paper of the data set, At the time the data set was produced a consisted of the widest arrangement of sensors for the visual task.

In reference to the dataset collected during the course of this thesis, the MS2 dataset [115] is far more extensive and offers a better configuration of sensors several of which are superior to their counterparts in the dataset generated during these. The main difference between the two datasets is that the type of agent and driving conditions very from smalls robots with low speeds in the inside environment to road-driving cars this results in rapid degradation of the results of any algorithm based on machine learning artificial intelligence but takes inputs from only one of these two datasets and test the results on the other. The advantage of the dataset generated during the course of this thesis is that it enables a wider range of testing and has superior ground truth precision, however, the precision of the ground truth is for the most part irrelevant to the task at hand. It can be said that the Korean dataset is far superior to that which was generated during the course of this thesis, to the extent that if it was available at the beginning of this thesis the may have been no reasonable motivation to develop the in-house dataset.

2.4.2 Dataset Design and Instrumentation

In order to facilitate the analysis that exists in this thesis, a novel data set was produced as there existed no known data set that was constructed to the specification required in this thesis. The specification required in this thesis is a data set that had rigidly combined two thermal and two visual cameras together with an IMU. We achieved this by mounting the sensors onto a ClearPath Jackal robot and employing the ROS waypoint navigation method to navigate the robot to the preset nodes. The coordinates of position are fed to the robotic agent by broadcasting the optitracks stream over the SSH protocol. Figure 4 shows our robotic system with the necessary sensors attached. An example of each acquired image type is presented in Figure 6.

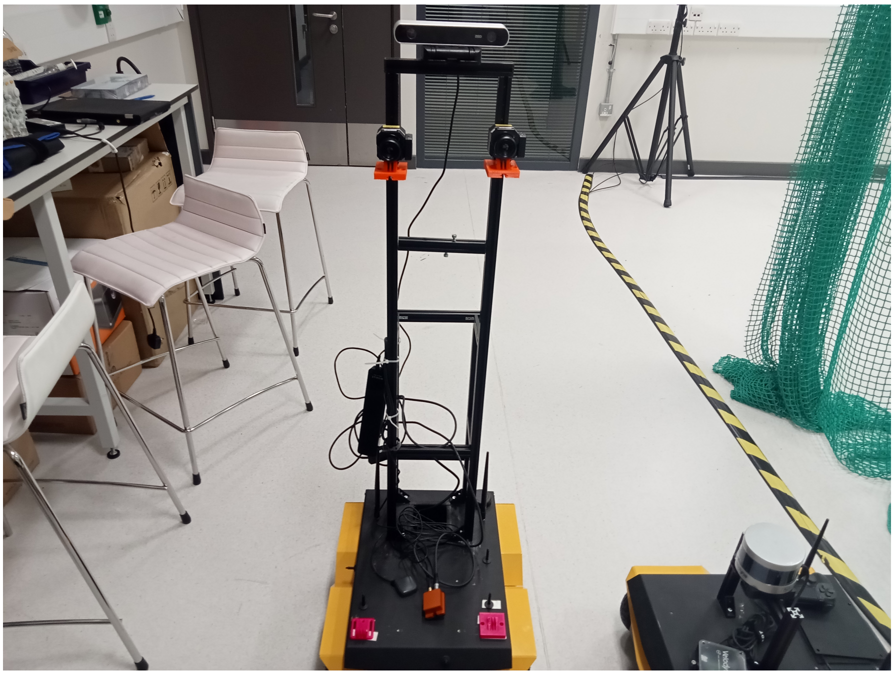

Figure 4: A photographic depiction of the final agent with the necessary sensors attached. This image orginates from the Autonomous Lab at City St. George of London.

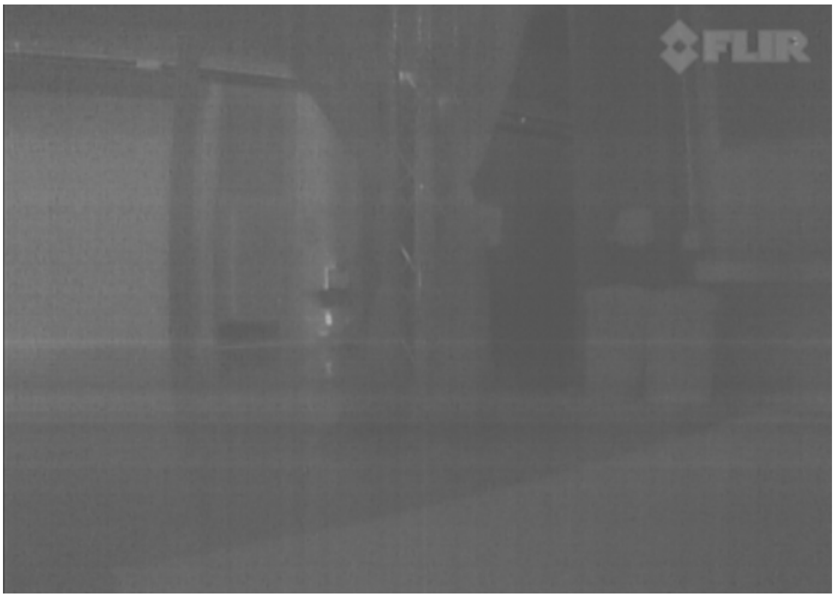

Figure 5: An example of the thermal image used in the project.

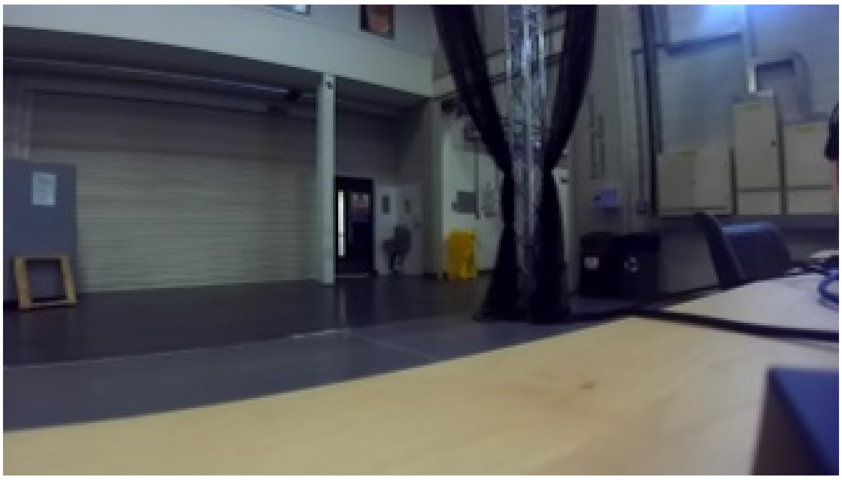

Figure 6: An example of the visual Image used in the project.

In order to test the interference in each of the modalities independently and jointly, we build four unique classes of difficulty scenarios: the first has no interference and so is the easiest modality; two of the classes have interference in one of two modalities (either thermal or visual but not both). These represent a medium level of difficulty; the final class has interference in both modalities and consequently is the most difficult of all the classes. This is summarised in Table 1. In order to ensure the trajectories can have a valid comparison between classes, we opted to preset four trajectories that the agent transverses. As these trajectories are consistent between classes, they can be directly compared across classes, and it, therefore, becomes possible to isolate the effects of the interference in any or all the modalities.

The next design decision was taken to isolate the changes in the results, which are caused by a change in contrast (feature texture) in the scene. This led to the second round of data acquisition with the exact same process as the first round being duplicated; the only distinction is that the second round included the introduction of tables and chairs used to dramatically change the texture in the scene, as illustrated in Figure 7. As there are four trajectories that are repeated once for each of the four classes and then the whole process is duplicated for a second round, the total number of trajectories is $4 \times 4 \times 2 = 36$. The interference was generated using external heat and light generators. That where observed in the images of some trajectories.

Difficulty	Visual Interference	Thermal Interference
Easy	-	-
Medium	X	-
Medium	-	X
Hard	X	X

Table 1: A table showing the presence of interference in the data set and how it relates to the difficulty of the sequence. An X marks the presence of interference in a modality.

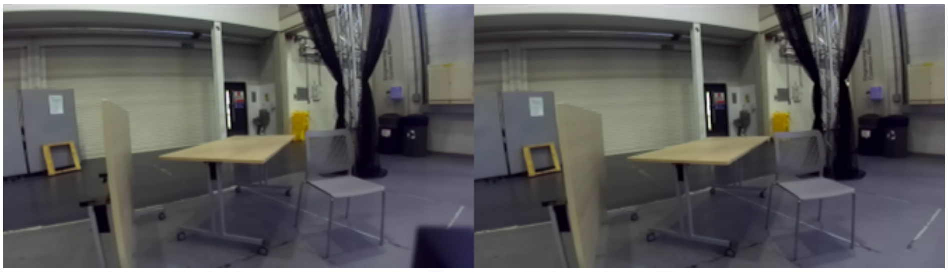

Figure 7: A depiction of the three types of objects used in the data set. The tables employed in the data set are foldable and were used in both positions.

In order to take continuous samples of the three-dimensional world and convert them into two-dimensional images that can be utilized as data input feed or series of input feeds in multiple modalities, a unique sensor configuration was employed.

Firstly, a stereo ZED camera was utilized as a source of RGB or visual images. Due to the nature of the camera, i.e. two smaller cameras were integrated into a single rigid frame, only a single visible sensor unit was required for the configuration. This was attached to the top of the metal frame fixed to the rigid body of the clear-path robot. It was integrated directly into the motherboard of the robot through the use of the external USB connection. Secondly, to sample in the thermal modality and construct a stereo image pair at each sampling point, two indistinguishable thermal camera units were employed. These cameras were the FLAIR Vue Pro series of cameras and were also rigidly fixed to the frame of the robot's metal attachment. These two cameras had a very surreal limitation in that they did not allow for wireless or wired external communication through an SDK on their own; however, it was possible to locate a pair of external modules which allowed the migration of the real-time image feed into the robot storage unit through an HDMI feed. It should be noted that the HDMI cables purchased required HDMI to mini HDMI adaptor on one end to fit into the module. This meant having two converting attachments. A second problem resulting from the use of these sensoring units was the power input; namely, in order to put in power to the cameras via the external module, another short cable was required. This cable was very small and could not directly attach itself to the motherboard of the robot, as such a series of extension cables was utilized.

This led to another problem. This was the tearing of the images as there was not sufficient power coming out of the motherboard to power all the sensors once we accounted for the IMU. By looking carefully at the documentation for the robot, it became apparent that the battery module for the robot had a higher power output and external power breakout board, which could be used as an ulterior source of power for the sensors; however, this board also fed into the main motherboard and so there was a minor problem. This problem was overcome by directly screwing on the entirety of the motherboard and decided modules, and then connecting jumper wires from the side modules through the end sensors bypassing the motherboard and using a second external power port on the board to power the main motherboard. Further to this, the end connections established on the Jackal robot were so terrible that we replaced them with new connectors. In addition to this, the connectors on the battery were rewired due to health and safety issues that arose from these sparks which came anytime somebody wanted to plug the battery into the robot or take it out. This was likely due to some faulty capacitors in place at some point. As the FLAIR cameras had an external HDMI feed, but no external USB feed and the robot IO ports did not allow for this, each camera output was converted to an extra USB feed through the use of MAGEWELL HDMI to USB converters

Perhaps the most interesting part of this sensor configuration was the joint GPS and IMU module purchased from Xsense. This module was of the MTi-G-710 variant. It also boasted the ability to fuse the INS and GPS modules on board through some form of kalman filter; however, this was never utilized in the project. The values for the error propagation of the INS came from the data sheet of the IMU supplied by the manufacturer. This data sheet was not located on a website but was in the packaging in the IMU box once it was delivered. It should also be noted that the IMU used serial to USB connection, which itself required an interesting cable which was not sold with the IMU and was quite hard to locate and bring into the UK due to the ongoing trade ambiguities resulting from Brexit with mainland Europe. The IMU itself was plugged into the motherboard. The use of so many cables on these sensoring robot left the cables dangerously in the way of any unknown traffic, the robot itself, and some of the senses. In order to correct for this, a series of zip ties and Velcro cable ties were used to safely attach the cables to the metal body of the robot and attach the frame in such a fashion that they would not harm any individual or the robot, or obstruct the sensing.

Finally, it was impossible to use the GPS module to assert a ground truth value due to the fact that the experiments were taking place underground in the basement of the university. To especially state the reason why, it was the fact that the concrete ceilings prevented any satellite signal from breaching the actual IMU/GPS unit which prevented it from gaining an accurate location or localizing. In place of this, an optitrack platform was used to identify the location of the robot at any given point in time, as well as orientation, resulting in the full pose of the robot, from which the rotational and angular velocities can be derived. The optitrack platform employs the location of various ball stickers on the robot to identify its orientation and position within the configuration of the volume. In order to ensure accurate orientation at all points, it was impossible to attach the stickers in a symmetric configuration as this would enable the cameras to be able to identify the location within the volume of the robot but not its orientation.

The code used to enable this data capture was written in C++ using the modules provided by the sensor providers in addition to some custom-made code that unified the various SDKs and constructed various loops and functions which enabled accurate asynchronous capture of data.

3 Stereo Visual Inertial Odometry

The applications for visual-odometry-based agents are not restricted to any singular domain [116], [117], [118], [119]. This has in turn led to the development of several visual odometry methods which attempt to operate upon microaerial vehicles [120], [121], [122], [123]. Due to the restricted size of the agent, significant work has been done on employing nature-based solutions to the problems of visual odometry [124]. Other works have been done attempting to exploit data fusion in cooperative systems [125], [126], [127], [128], [129].

The existence of several data fusion architectures for visual odometry leads one to the belief that the most intuitive one is the centralised architecture, which takes all sensor readings as inputs into a single node for processing. This is both optimal and computationally expensive [130]. [131] implemented particle filters for the tracking of UAVs. [132] expanded such systems to be applicable to cooperative networks, which further allows the employment of AI-based methods through the various nodes in the network.

Due to the computational demands of the central processing architecture, parallel architectures have become somewhat popular alternatives [133]. Such systems often employ local estimates which are then transmitted to a master filter, which can then employ the local estimates and their covariance to derive the global estimate [134].

Stereo visual odometry has been successfully applied in many UAVs such as [135]; however, they are not suitable to MAVs [136]. This is due to the ratio of the mass of the camera to the drone. As hardware improves, it is likely that stereo systems will be employed on all but the smallest drones due to the increased information given by the system such as depth. [137] has successfully applied stereo visual odometry to the exploration of caves in order to find the safest mining path and locate viable resources.

[120] has successfully demonstrated the fusion of visual and non-visual sensors in the visual odometry field through the employment of optical-flow-based methods. Showing that both the velocity and the elevation of the agent may be approximated through the employment of optical flow. The most challenging component of these optical-flow-based techniques is the requirement of scaling the sensor readings to be compatible with each other [138].

Range cameras are capable of generating voxel maps for use in visual odometry systems. The two main types of range cameras that enjoy widespread employment are time-of-flight and structured light-based cameras. The main difference between stereo-based voxel imagery and range-camera-based voxel imagery is the application to MAV; however, only voxel maps generated by range cameras can be assumed to complete description of the depth in a scene. [139] successfully showed the ability to control the height of a UAV through the employment of voxels maps via a calibrated Kinect sensor. A detailed review of the employment of voxel maps is found in [140].

IMUs have been one of the most frequently used type of sensor in odometry in general; however, they have a tendency to suffer from drift in their estimates of velocity overtime. This has led to many attempts of applying data fusion by the employment of many other sensors such as GPS to remove or reduce the drift inherently found in IMUs, thereby allowing for the use of classical mechanics to estimate the position of the agent. This approach has been shown to require low energy usage, as well as to display high scalability [141], [142].

Visual odometry combined with an IMU unit offers far better results, in terms of availability, than GPS-augmented IMUs as it is completely self-contained (in the agent) and passive. Furthermore, it is free from signal-masking problems as the system requires no external signal for processing. In addition to this, the on-board camera system can be employed to estimate the elevation of the agent and find the agents vertical trajectory [143].

This approach of combining visual odometry with IMU sensors has led to the computationally efficient construction of geographical terrains, from otherwise computationally infeasible SLAM solutions with minor [144] employed data fusion to develop an outlier removal method which employed the use of IMU data to augment a visual feed. However, it is only possible for this solution to work in a cooperative environment.

[141] employed a visual integrated IMU system to estimate the height of an agent in a three-dimensional vector space. This was made possible through the use of a Kalman filter that allowed the optimal combi-

nation of both types of data.

[145] developed a novel algorithm which is capable of fusing vSLAM and IMU to form a solution that allows a UAV to operate in GPS-denied locations. The work employed a monocular scheme which could detect landmarks in an urban environment and EKF-based SLAM to estimate the position of the agent. By employing a dual-axis accelerometer, the authors were able to detect and measure all movement patterns undergone by the agent. This also provides the benefit of knowing the absolute scale.

The employment of various sensor types in visual odometry solutions allows the extension of its applications far beyond what would be possible through the employment of the visual spectrum [6]. However, it also forces the design of such a system to consider various sensor specific problems, such as the cyclic nature of readings from a thermal sensor [146].

Thermal sensors capture variations in temperature in a scene, allowing the extension of pre-existing visual odometry algorithms to environments with low lighting such as caves and deep water terrains [6]; however, this adds considerable complexity to the system as it forces the necessity of considering low signal-to-noise ratios and non-uniform noise.

It has been shown that the application of thermal visual odometry may detect different features from visual light odometry and as such result in different results [147]. For this reason, [6] presents a comparison of feature detectors employing a similar benchmark to [148] and [149]. The paper then continues to explore thermal 3D reconstruction through the use of uncommon optimisation techniques such as the Levenberg-Marquadt and the Double Dogleg techniques. It also compares the performance of such techniques to that of the more classical solution, the Gauss-Newton method.

[150] demonstrates the fact that a visual odometry pipeline can be subdivided into a series of subtasks, allowing the development of partial solutions such as [151], which employed the use of an infrared camera in a monocular system to estimate the egomotion of the agent. This approach was later adopted by [152] to employ both a thermal and a visual camera. This work, however, employed the use of each camera independently. The first instance of data fusion between a thermal and a visual camera from the purposes of visual odometry was in the work of [153], which later led to a study on the suitability of infrared cameras in night-time visual odometry [154].

Many such surveys have been rendered obsolete due to advancements in sensor technologies. [155] demonstrates the ability of enhancing the estimation of road geometry from the use of thermal data—something that many such surveys have considered impossible to due hardware limitations.

Notable work in the field of thermal stereo visual odometry has been conducted in [156]. This work extended the triangulation abilities of visual stereo odometry to work with a pair of thermal cameras and produces a series of voxel maps as a result [157]. [158] extended such work to the topic of pedestrian detection by attempting to find the optimal sensor configuration of a stereo odometry solution to detect people in static images.

Works such as [159], [160], [161], [162], [163] have exploited the recent trend in thermal stereo cameras to develop a series of tools for the initial calibration of thermal stereo systems, allowing for the development of real-world thermal visual odometry solutions as seen in [147]. This body of work also posed novel solutions to a series of problems which prohibited the wide spread adoption of thermal cameras in visual odometry solutions. These problems include the large radial distortion resulting from the design of the lens of the thermal camera.

The employment of thermal cameras in visual odometry systems has led to piecewise deconstructable models of the dynamics in a scene [164] due to the cyclic nature of temperature variations. It has been shown that for tasks such as object detection with thermal cameras, such models yield superior results.

[165] employs the use of thermal visual odometry to build upon the work of [166] to allow the autonomous navigation of a visual-odometry solution through obscurants like fog and smoke, which was previously not possible.

Due to the size of most UAVs and indeed all MAVs, little work has been done in favour of RADAR-assisted visual odometry, and hence only a few approaches exist, which enable visual odometry in GPS-denied environment and few still in environments that both inhibit GPS and have obscurants present in them [167], [168].

[169] presents an overview of the topic of combining selective thermal and visual sensors in the field of visual odometry; however, it is rather limited in scope and does not tightly integrate the two sensors in the optimal data fusion method for the system, nor does it enable the employment of non-ground based agents. [170] takes upon the task of cooperative visual odometry by employing a range of sensors including the thermal ones. [171] develops a method for combining appearance and thermal information, resulting in a method with the potential of expanding thermal imagery in visual odometry beyond basic sensor fusion. Works such as [172] demonstrate the possibility of using stereo visual odometry with thermal lenses to track the flow field of the heat in a scene and retrieve from it the trajectory of the smoke in the scene. [173] has presented a method by which to calibrate multiple thermal cameras in a single solution; however, to date, there exist no robust visual odometry solutions which employ thermal data fusion in GPS-denied environments with high obscurants.

In order to calibrate thermal cameras to work in a multi-sensor data fusion environment, a series of thermal camera intrinsic model calibrations were formed [173], [174], [175], [176]. This has helped tremendously in incorporating thermal cameras with IMU sensors [177]. It has also led to the open-source work of [7].

3.1 Motivation

The use of kalman filters has become standard in practical applications of visual odometry and SLAM. This is due to the ability of the filter to optimally fuse the pose estimates derived from different localisation systems based upon their reliability.

Most solutions that employ Kalman filters aim to fuse the pose predictions from complementary sensors to offset the reliability issues of the systems. This thesis employs the use of IMU-based INS to offset the drift accumulation from a pure stereo-VO based solution; however, there are monocular solutions that attempt to employ this same fusion to estimate the scale from the INS.

It has been shown that the Kalman filter can be used to fuse an arbitrary number of such systems. This has led to Kalman filters becoming a key component in large scale location solutions for self-driving cars. However, Tesla is attempting to move all components of their self-driving system into a large neural network.

Most practical applications of the kalman filter do employ some variant of the kalman filter as its underlying assumption of linear transformations is not compatible with the requirements of the localisation problem. This adoption of the traditional kalman filter normally takes the shape of the extended kalman filter (EKF) which employs the use of the taylor series expansion to optimally estimate the transformation functions around a given input.

The use of the EKF has several practical limitations such as computational complexity, which limits the ability of the solutions to be employed in a real-time environment. To prevent the computational complexity of the system from growing too large, this thesis employs the error-state adoption of the EKF.

The loosely-coupled EKF proposed in this thesis optimally fuses the results of the INS (with pre-integrated IMU measurements) and the VO solution into a single VINS estimate of the pose of the agent.

This then leads to the construction of two separate navigation solutions that will be integrated into a single unified solution. The first of these two solutions, the INS, is derived from the fundamental equations of kinematics and is developed in the proceeding section, whilst the VO solution is developed in the following section. The final section of this chapter will develop the error-state EKF fusion of the two systems.

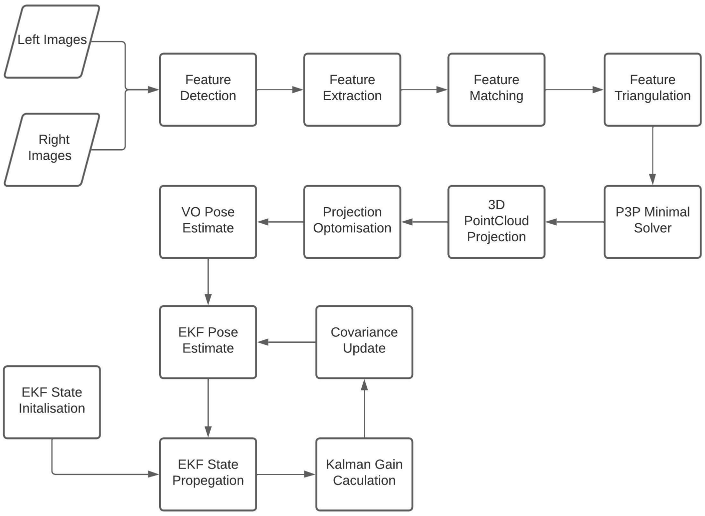

Figure 8: A high-level overview of the EKF.

As mentioned, the main benefit of the practical application of the error state kalman filter is computational complexity. This results from the fact that the orientation error state is minimal. This also prevents over-parameterisation and significantly reduces the possibility of a singularity.

The fact that the error state is always operating close to the origin, far from most over-parameterisation and gimble lock issues, guarantees that the linearisation always holds. As the error-state is always small, the second-order products are negligible and the jacobians are fastly computable.

3.2 Methodology

This section of the thesis proceeds to develop two independent solutions to the localisation problem. The first is a simple visual odometry algorithm exploiting the bundle adjustment pipeline, the second solution is a standard visual inertial odometry solution.

Both solutions are then combined via an optimal data fusion method - the extended kalman filter. This allows for a single pose estimate that encodes all available information in the most optimal manner.

The two central ideas to note in the filter are the ability to exploit the confidence matrix to form an optimal fusion and the fact that the visual odometry solution only produces a six-degree-of-freedom pose estimate (constituting of a three-element position vector and a three-element orientation vector) not the full five-element state required by the kalman filter fusion state. The missing velocity, acceleration, and angular velocity components of the state are derived as secondary observations from the pose estimate. This is done by assuming that the sampling rate of the solution is sufficient to assume that any two sequential pose estimates have a constant velocity, acceleration, and angular velocity throughout that period. This then allows simple integration via multiplication with the change in time.

The process of extracting the covariance matrix from the bundle adjustment (BA) involves computing the hessian matrix, which captures the curvature and the nonlinear relationship between the residuals and the camera and 3D point parameters. The hessian matrix is then inverted to obtain the covariance matrix, which provides information about the variances, covariances, and correlations of the estimated parameters.

The hessian matrix is computed as the sum of the outer product of the jacobian matrix with itself over all observations. The jacobian matrix is a matrix of first-order partial derivatives of the residuals with respect to the camera and 3D point parameters. It captures the sensitivity and the linear relationship between the residuals and the parameters. The partial derivatives are computed using the chain rule of differentiation and the parameterization of the camera and 3D points.

The NLS problem is then solved iteratively using a nonlinear optimization algorithm, such as the Levenberg-Marquardt algorithm, to minimize the objective function. The optimization algorithm updates the camera and 3D point parameters iteratively until convergence is reached. The convergence criterion can be based on the change in the objective function or the change in the parameters.

Once the NLS problem is solved, the hessian matrix is computed, which is a square symmetric matrix with dimensions equal to the total number of camera and 3D point parameters. The hessian matrix represents the curvature of the objective function near the optimum and is a measure of the local stability and accuracy of the estimated parameters.

To obtain the covariance matrix, the hessian matrix is inverted, typically by using a numerical linear algebra library. The inverse of the hessian matrix provides information about the variances, covariances, and correlations of the estimated parameters. The diagonal elements of the covariance matrix represent the variances of the estimated parameters, and the off-diagonal elements represent the covariances and correlations between pairs of parameters. The covariance matrix can be scaled by a factor that depends on the noise model and the weighting of the residuals.

The validity of the covariance matrix can be checked by analyzing its eigenvalues and eigenvectors. A valid covariance matrix should have positive-definite eigenvalues and orthogonal eigenvectors. The eigenvalues represent the variance along the principal axes of the estimated parameters, and the eigenvectors represent the direction of the principal axes. If the eigenvalues are negative or small, it may indicate that the optimization algorithm did not converge to a true minimum or the noise model is incorrect.

In conclusion, extracting the covariance matrix from the bundle adjustment pipeline involves computing the hessian matrix, solving the NLS problem, and inverting the hessian matrix. The covariance matrix provides valuable information about the uncertainty and covariance of the estimated parameters and can be used for various downstream applications.

3.2.1 INS

By rigidly attaching an inertial measurement unit (IMU) to the body of the agent, it is possible to measure the acceleration and angular rotation of the agent; however, the IMU does not directly measure the acceleration and rotation of the agent, but the forces acting upon the IMU. Therefore, this thesis must present a model and or method which will felicitate the bi-directional conversation of the values measured by the IMU (denoted with a subscript m) and the true values (denoted with a subscript t). The equations in this section of the thesis follow from [178].

$$a_m = R_t^T(a_t - g_t) + a_b + a_n \tag{15}$$

$$w_m = w_t + w_b + w_n \tag{16}$$

$$a_t = R_t(a_m - a_b - a_n) + g_t \tag{17}$$

$$w_t = w_m - w_b - w_n \tag{18}$$

where the subscripts n and b denote noise and bias, respectively. a denotes the acceleration and w denotes the rotational velocity. the gravitational force is represented by g in practice this is a constant based on the location of the experiment.

The measurements gained from the IMU act as the inputs to the system of dynamic equations that governs the motion of the agent. The equations are adapted from physics, namely the study of kinematics. Where required the rotation of the agent is expressed using the quaternion format, using the notation q. At other times the rotation is expressed as a rotation matrix R, the chief concern that dictates which of the formats are used is the onset of gimble lock.

To construct the adaption of the system of equations required for this thesis, it must be noted that the position p of the agent is a function of the velocity v and time t, where velocity itself is a function of time and acceleration a. Noting that the bias vectors and gravity are constant, the continuous time system of equations becomes:

$$\dot{p} = v \tag{19}$$

$$\dot{v} = a \tag{20}$$

$$\dot{q} = \frac{1}{2} q \otimes w \tag{21}$$

$$\dot{g} = 0 \tag{22}$$

$$\dot{b_a} = 0 \tag{23}$$

$$\dot{b_g} = 0 \tag{24}$$

This formulation of the system of dynamical equations is not directly applicable to all localisation problems. This is a result of the fact that the formula assumes that the bias vector and the gravity vector are constant. however, a rocket ship may traverse many different regions of space in a single vogue crossing many gravitational fields each with its strength. In addition, high levels of heat exposure may warp or distort the intrinsic camera matrix which is not modeled here as it is assumed that the agent will act in an environment that does not allow for such exposure nor does it recalibrate its lenses.

It should also be noted that the data retrieved from the sensors is captured in discreet time intervals and not as a continuous stream. While the IMU measurement can undergo pre-integration to align with the stereo-image pairs. To my knowledge, there is no known method to interpolate images through time. This lack of interoperability should not be confused with the interoperability of colour chancels in most cameras; as a result, the discreet adaption of the continuous-time system of equations is developed below:

$$p(t) = p(t-1) + v(t-1)\delta t \tag{25}$$

$$v(t) = v(t-1) + a(t-1)\delta t \tag{26}$$

$$q(t) = \frac{1}{2} q(t-1) \otimes w(t-1) \tag{27}$$

$$b_a(t) = b_a(t-1) \tag{28}$$

$$b_g(t) = b_g(t-1) \tag{29}$$

$$g(t) = g(t-1) \tag{30}$$

The equations in this section may define an unstable system if time and acceleration are infinite, however, the practice limitation on the experimentation prevents this from occurring. Finally, by incorporating the sensor measurement model into the discreet time system of dynamical equations, a programmable set of equations governing the INS is derived:

$$p(t) = p(t-1) + \frac{1}{2}((R_t(a_m - a_b - a_n) + g_t))\delta t \tag{31}$$

$$v(t) = v(t-1) + (R_t(a_m - a_b - a_n) + g_t)\delta t \tag{32}$$

$$q(t) = \frac{1}{2} q(t-1) \otimes (w_m - w_b - w_n) \tag{33}$$

$$b_a(t) = 0 \tag{34}$$

$$b_g(t) = 0 \tag{35}$$

$$g(t) = 0 \tag{36}$$

Note that it is not possible to know the actual noise components of the IMU measurements in practice. To rectify this in the practical implementation, the noise components are set to zero. Furthermore, due to the practice constraints of real-world motion, it is known that the acceleration and angular rotation must be continuous in time, thus as the sample rate tends to ∞ the contour of the rotation and acceleration is piecewise smooth. This allows the smoothing of the rates post integration via derivative smoothing with a window size of three.

3.2.2 The Visual Odometry Solution

The VO solution is constructed using the stereo-image pairs received from the camera sensors at each time step. The images then undergo the detection and extraction of feature points that facilitate the spatial and temporal matching of the features in the images. The spatial matching is used to triangulate the 3D location of the feature points. The temporal matches are then used as inputs into the bundle adjustment algorithm alongside the 3D locations, to optimise the estimate of the agent's position based upon an initial estimate derived from a minimal P3P solver.

Upon the receipt of the images, the algorithm rectifies the images. This maps the epipoles of each image to infinity, which then allows the use of epipolar lines to restrict the search regions allowing for an outlier-removal scheme.

In practice, keypoints are extracted in each image using the GoodFeatures2Track algorithm. This was due to the assumption that the reliability of pattern-based keypoint algorithms may become unreliable when considering the multispectral nature of the thesis, and the fact that the GoodFeatures2Track algorithm has been shown to exhibit the appropriate level of robustness and compatibility within an MS-VIO system.

After detection, the keypoints are extracted using the histogram of gradients (HOG) algorithm, in which a 128-dimensional vector is constructed using the gradients in the surrounding region of the keypoint. This vector is then used to match the key points in the stereo-image pair spatially.

This matching then results in a set of key points that have known locations in both images; however, after stereo-rectification the aforementioned test is used to remove any outliers that may have matched erroneously.

The keypoints in the left image for which the 3D locations are known are then tracked into the left image of the next timestep using the KLT tracker. The set of keypoints that have been successfully tracked and the set of keypoints that have been successfully matched are then compared and the set of points that do not appear in both sets are discarded. This process constructs a one-to-one correspondence between the points in the 3D point cloud and the 2D keypoint in the left image of the seconded stereo-pair.

The next sub-process of the algorithm attempts to form an initial estimate of the pose of the agent given a minimal set of 3D-2D correspondences. The actual P3P-solver used in the thesis is the linear P3P-Solver.

The linear solver uses the fact that given any two points, a triangle can be formed between them and the centre of projection. As the locations of the feature points are known, Euclidean geometry will yield the length of the line segment between the two 3D points. After a series of polynomial reductions employing the use of Sylvester's resultant from residue theory and a third point, a single 8^{th}-degree polynomial holds the solution to the algebraic variety; however, this 8^{th}-degree polynomial can be further reduced into a 4^{th}-degree polynomial that can be solved using known techniques such as Galoa Theory.

In practice, as many 4^{th}-degree polynomial as possible are constructed from the final set of accepted feature correspondences. The polynomials are then formulated into a stacked matrix and the SVD is then used twice to solve for the algebraic variety.

The single positive non-complex solution encoded in the algebraic variety is employed as an initial estimate of the pose of the agent. It is typically far from the actual pose of the agent. It is known that without further refinement, the trajectory described by a series of these initial estimates is often disjoint. To ensure that the estimated trajectory is a single continuous trajectory, the initial pose estimate is further refined using the motion-only bundle adjustment algorithm.

The inclusion of the bundle adjustment sub-process increases the computational complexity of the solution considerably; this is due to the minimisation step that operates over each set of stereo-images and all 3D-2D correspondences.

The optimization function required to be minimised by the bundle adjustment process is summed over all N points in the point cloud and all M images to minimise the distance d between the projection of the point cloud onto the image plane and the location of the feature x with image coordinates of i and j. It is often the case that a real-valued scalar λ is used to scale the output. The $P(a, b)$ is the projection of the feature points into the image plane. This equates to minimising the area of various triangles that constitute the errors of the feature matchining.

$$\underset{a_j,b_i}{\text{minimize}} \sum_{i}^{N} \sum_{j}^{M} \lambda d(P(a_j, b_i) - x_{ij})^2 \tag{37}$$

To further improve the robustness of the operation, this step incorporates the RANSAC algorithm with a 99.99 percent confidence level.

This estimate is considered the final estimate of the current iteration of the VO solution and is concatenated with all previous poses to form the currently estimated trajectory of the agent. The rotation and translation estimates of the current iteration may be used to compute the estimated current position of the agent as follows:

$$p_t = p_{t-1} + R_{vo}T_{vo} \tag{38}$$

$$R_t = R_{t-1}R_{vo} \tag{39}$$

Whilst the INS system employs the use of the quaternion representation of the rotation of the agent, the VO solution does not. This is due to the possibility of gimble lock occurring in the INS system, but not in the VO solution. This means that the VO solution update of the pose (the orientation and the position) of the agent is computed solely upon the rotation matrix representation of the agent rotation. In the above system of equations, the subscript vo denotes the optimised output of the visual odometry iteration. The R denotes a rotation matrix and the T the translation.

There exists a further set of optimisations that were utilised in this work, which were derived from the practical constraints of the real world driving task. Given that the landscape in which the agent propagated, had a maximum speed limit imposed by law, the maximum speed of the agent was restricted to this speed. In addition to this the agent could not istantouly reverse its orientation nor was it flipped upside down this enables a restriction to be placed on the angular velocity of the agent. This was also aided by the fact that the maximum horizontal motion in a straight road is given by the distance between the opposing pavements. Given that the model of the car employed as the agent is known so are its default specs. This information was employed to determine the maximum rate of acceleration - by use of the nought to sixty rate - and ultimately lead to the maximum possible linear motion in a given window of frames.

3.2.3 The EKF Fusion

As the agent propagates through the environment, the confidence of each measurement will vary in two distinct ways. Firstly, their confidence in each successive set of measurements will vary as the level of drift increases in the system. Secondly, at each instance in which measurements are taken, there will naturally be some variation in the confidence of the two solutions. The confidence of the visual odometry solution will vary with the reliability of the feature matching whilst the confidence of the INS increases with the noise model of the INS developed from the parameters on the IMU data sheet.

This then suggests that the optimal estimate of the agent position is not that of the INS or the VO solution but some combination of the two. This optimal combination is achieved through the employment of the EKF.

The error state EKF begins by defining some state vector that must be tracked in each time step and optimally estimated by the fusion of the two solutions. In this thesis, the state vector comprises five different vectors and a quaternion, making it a 19×1 column vector.

$$X = [p^T, v^T, q^T, b_a^T, b_w^T, g^T]^T \tag{40}$$

This state vector is then updated through the use of the INS transition matrix. This transition matrix is a linearization of the aforementioned equations around the previous pose of the agent. The filter works by assuming that the true values of the state vector X_t are the sum of the nominal (INS) state X_n and the imperfections of the model both those in the model and those not in it, denoted as X_e.

$$X_t = X_n + X_e \tag{41}$$

$$X_n = [p^T, v^T, q^T, b_a^T, b_w^T, g^T]^T \tag{42}$$

$$X_e = [\delta p^T, \delta v^T, \delta\theta^T, \delta b_a^T, \delta b_w^T, \delta g^T]^T \tag{43}$$

Both the angular rotations w and the angular error $\delta\theta$ are defined locally with respect to the nominal or estimated state, whilst there exists some evidence that globally defined angular rates exhibit better stability. This thesis follows the conventional paradigm as it has been better explored. This formulation also allows for the direct use of the measurements w_m which is not possible in the global definition.

Whilst most of the vectors in the state follow the standard compositions in a real-valued vector space there exist two notable outliers: the quaternion representation of the rotation and the error of the rotation matrix. The composition of the quaternion follows the corresponding appendix; however, the rotational error δR is composed using the exponential of the skew symmetric cross matrix of the angular error $\delta\theta$.

It becomes critical to consider the sensor impact upon the model, namely the quality of the IMU. For this, the exact values can typically be found on the IMU data sheet. The incorporation of the IMU sensor noise requires the integration of the values in the data sheet, which results in:

$$V_i = \sigma_{\tilde{a_n}}^2 \Delta t^2 \tag{44}$$

$$\Theta_i = \sigma_{\tilde{w_n}}^2 \Delta t^2 \tag{45}$$

$$A_i = \sigma_{a_w}^2 \Delta t \tag{46}$$

$$\Omega_i = \sigma_{w_w}^2 \Delta t \tag{47}$$

The jacobian matrices can then be constructed as:

$$F_x = \begin{bmatrix} I & I\Delta t & 0 & 0 & 0 & 0 \\ 0 & I & -R[a_m - a_b]_x \Delta t & -R\Delta t & 0 & I\Delta t \\ 0 & 0 & R^T[w_m - w_b]_x \Delta t & 0 & -I\Delta t & 0 \\ 0 & 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & 0 & I \end{bmatrix} \tag{48}$$

$$F_i = \begin{bmatrix} 0 & 0 & 0 & 0 \\ I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{49}$$

$$Q_i = \begin{bmatrix} V_i & 0 & 0 & 0 \\ 0 & \Theta_i & 0 & 0 \\ 0 & 0 & A_i & 0 \\ 0 & 0 & 0 & \Omega_i \end{bmatrix} \tag{50}$$

Next, in the exact same manner as the traditional EKF, the kalman gain is derived. However, unlike the traditional filter, it is then used to approximate the error state. Note that the update equation for the covariance matrix is symmetric. This is not the only option; however, it was exploited due to its computational stability.

$$K = \frac{PH^T}{HPH^T + R} \tag{51}$$

$$\delta x = K(y - h(x_t)) \tag{52}$$

$$P = (I - KH)P(I - KH)^T + KRK^T \tag{53}$$

The h matrix is derived by finding the transition matrix that converts the observation to the format of the measurement. The partial derivative of h is then employed to derive the extraction matrix H via the chain rule.

$$H \equiv \frac{\partial h}{\partial \delta x} \tag{54}$$

At this point, the error state δx can be added directly to the state via the normal composition, after which it is reset for the next iteration of the filter.

3.3 Experiments

It is the principle objective of this subsection of the PhD thesis to introduce the audience to the results gathered from the experimentation of the course of study outlined within this chapter of the PhD thesis. This is done through the examination of computed trajectories and estimated trajectories measured against the ground truth or the exact trajectory undertaken by the agent in accordance with the measuring apparatus. It should be noted that the precision of the estimated trajectories where accurately given in accordance with the algorithm and measurements of the sensors rigidly attached to the agent, the precision of the trajectory which depicted the actual path of travel transverse by the agent varies in accordance with the precision of the measuring apparatus. Whilst this point might be considered inconsequential, it is actually of considerable note. This is due to the fact that the optitracks measuring apparatus employed to track the true trajectory undergone by the agent possesses the ability to do so to a degree of precision that the sensors rigidly attached to the agent cannot achieve.

Due to this, it is possible to conclude that the lack of precision in the estimated trajectories would always result in elongated drift, which is most evident when algorithmically compensating for the measurements of the sensors. A clear example of this would be the use of integration and double integration employed within the inertial navigation system, which results in elongated drifts. This is in fact a characteristic of the navigational system and not the algorithms of employed within this thesis. As the inertial navigation system requires the use of integration, it introduces an error of integration and due to the lack of precision of the inertial measurement unit in comparison to the ground truth measuring apparatus, and the assumption of averages over sequential time steps, the resulting average is the result of our larger interval of time. This force the navigation system to be far more sensitive to the errors of integration resulting in elongated trajectory estimates.

This particular subsection of the thesis can be further divided into two distinct subsections: the first introduces the audience to both data sets employed in this subsection and clarifies the significant points which arose during the instrumentation of the experiments which is omitted from the algorithmic sections; the second showcases the results of the experimentation and delivers a thorough analysis of the results obtained.

3.3.1 The Data Sets

The two data sets employed within this part of the thesis are exactly those which have been introduced to the audience in the prior sections of the thesis. Although the audience is familiar with these data sets, it is critical to the evaluation of the results presented within this thesis and collected during the course of the instrumentation of the experimentation of this thesis. The audience is made aware of multiple points of clarification concerning these data sets and the implementation of the experimentation upon them without the introduction of any ambiguity, which may be removed from the minds of the audience by further clarification.

The first point of clarification is the purpose behind the use of multiple data sets within this particular aspect of the thesis and not the proceeding sections. It should be noted that as the body of work developed within this thesis is novel, it is often difficult to put into context against pre-existing solutions taken to the extreme, this means that it is often difficult to see and critically evaluate the performance criteria and metrics of the work conducted within this thesis against the pre-existing body of literature developed by the larger research community. In order to alleviate this problematic situation, an open-source data set is used to evaluate the component of the work developed within this thesis, which most closely relates to the well-established work in the larger field. This is done to enable the reader to compare the results between both data sets in this section and be able to contrast the work in this thesis with the larger field, whilst also exploiting the second data set to contrast the work developed with this thesis. In this fashion, it is hoped that the reader is able not only to contrast the various results presented here against each other, but also has the ability to relate the work presented here to external papers and other publications. Unfortunately, as it was not possible to find a pre-existing open-source data set which enabled all the experimentation done within this thesis, it proved impossible to gather these results without the generation of a novel data set.

It should be noted that a substantial amount of time has elapsed during the initiation of the thesis and the conclusion of the PhD program. As a result, the state of the world has not been constant during this. This has led to the purposeful removal of the third sequence of the KITTI data set from this thesis. The dataset previously enabled the use of all data collected during this particular sequence. However, due to its overlap with some portion of the evaluation sequences of the KITTI benchmark, it was later restricted and it was no longer possible to gain from the authors the inertial measurements associated with the sequence. This combined with a laptop failure which resulted in the removal of the inertial measurements associated with the sequence from the machines of the author of this thesis and the lack of its availability online from other sources has forced its removal from this thesis. It should be noted that it is still possible to offer the visual only trajectory estimates for this thesis; however, the author of this thesis believes that to be disingenuous and so has opted to omit this sequence and only provides full and clear results. The evidence of this claim can be presented to the examiner upon request.

3.3.2 The Results

The principal objective of this section of the thesis is to display the results obtained from the practical application of the methodology of the preceding section of this thesis.

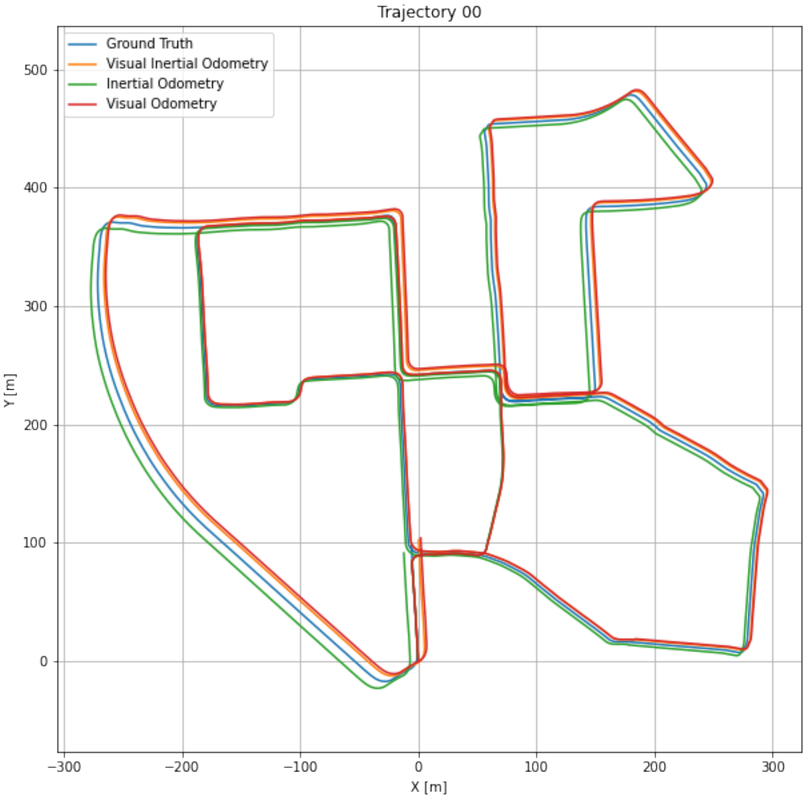

Figure 9: This is a graphical depiction and comparison of the six-degree-of-freedom pose estimation results of three GNSS/GPS alternative navigation systems. Computed upon sequence 00 of the world-renowned KITTI data set.

From Figure 9, it is apparent that the inertial navigation system produces the poorest results, whilst the EKF most closely resembled the ground truth and the visual and dormitory methods are sandwiched between both estimates. This is the typical behavior that can be expected and as such there is little insight to be gained from this sequence.

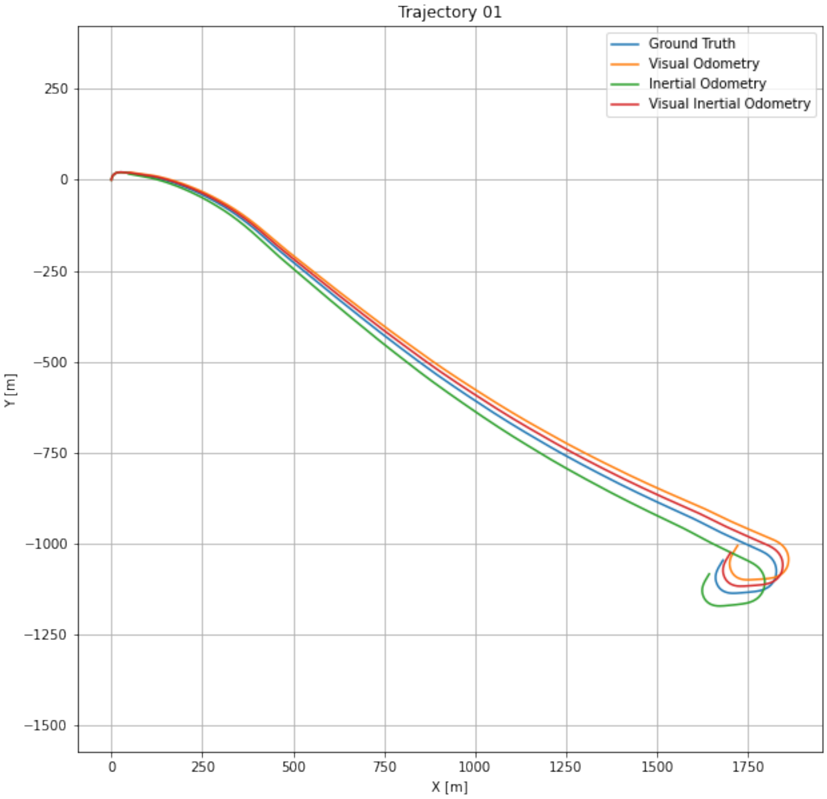

Figure 10: This is a graphical depiction and comparison of the six degrees of freedom pose estimation results of three GNSS/GPS alternative navigation systems. Computed upon sequence 01 of the world renowned KITTI dataset.

In the depiction captured in Figure 10, it is apparent that the inertial navigation system closely resembles the ground truth and the other two systems produce significantly worse estimates of the trajectory. This illustrates the two key points which are worthy of note when comparing these three systems. The first point is that in short trajectories, the inertial navigation system typically is rather close to the ground truth unless some environmental factors cause deviation from the ground truth. An example of such an environmental factor would be the wheel slip phenomenon that is present on the surface of Mars and hounds the Mars rover. Secondly, the worst performance of the visual odometry system is likely due to failure to capture accurate estimations of the series of rotations accurately, this highlights the fact that the fussed system is a weighted average of the other two systems and so may produce worse results than any individual system should the outliers in an estimate of one of the systems be significantly greater than that of the other.

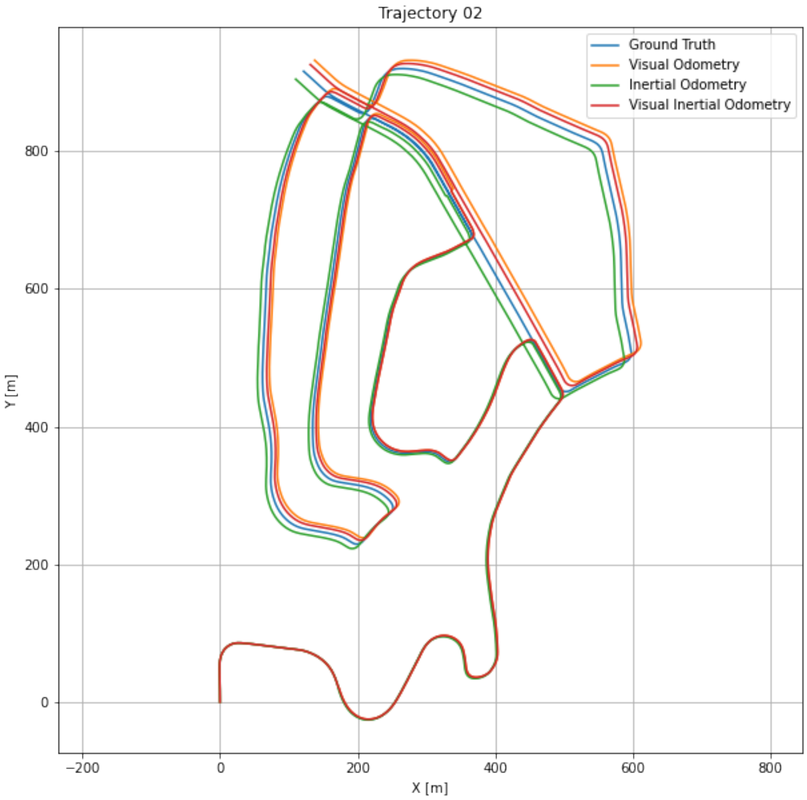

Figure 11: This is a graphical depiction and comparison of the six-degree-of-freedom pose estimation results of three GNSS/GPS alternative navigation systems. Computed upon sequence 02 of the world-renowned KITTI data set.

The illustration of the trajectory estimates within Figure 11 demonstrates the fact that whilst in the beginning stages of the trajectory approximation, all three methods will be initialized with the same conditions and therefore have similar estimates in the initial stages. As time continues to propagate forward, the estimate of the inertia navigation system may begin to have severe deformities within its contour. As a result of this, it can oftentimes look very different from the estimates produced by the other two methods and indeed the ground truth.

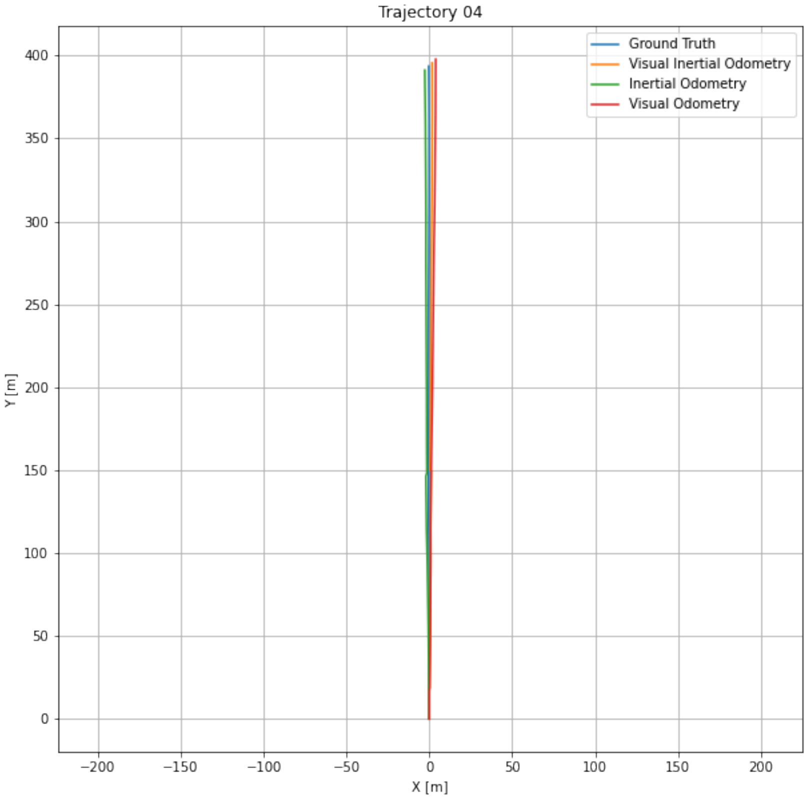

Figure 12: This is a graphical depiction and comparison of the six-degree-of-freedom pose estimation results of three GNSS/GPS alternative navigation systems. Computed upon sequence 04 of the world-renowned KITTI data set.

It is apparent from the depiction within Figure 12 that the short nature of the trajectory enabled the national navigation system to produce the best estimate of the ground truth trajectory. However, this is likely aided by the fact that the visual odometery navigation system seems to have picked up bad feature-matching, which has resulted in drastically different estimates of the rotation and translation parameters, resulting in a trajectory that no longer resembles a ground truth. It further highlights the fact that the filter could compensate for the errors in one of the two trajectory estimates with the trajectory estimate of the other method. This is evident by the fact that the Y value ranges from approximately 100 to 150 meters, the trajectory of the filter has a difficult contour from the visual odometry method and the difference in this contour can easily be explained by the difference in the contour of the inertial navigation system and its own.

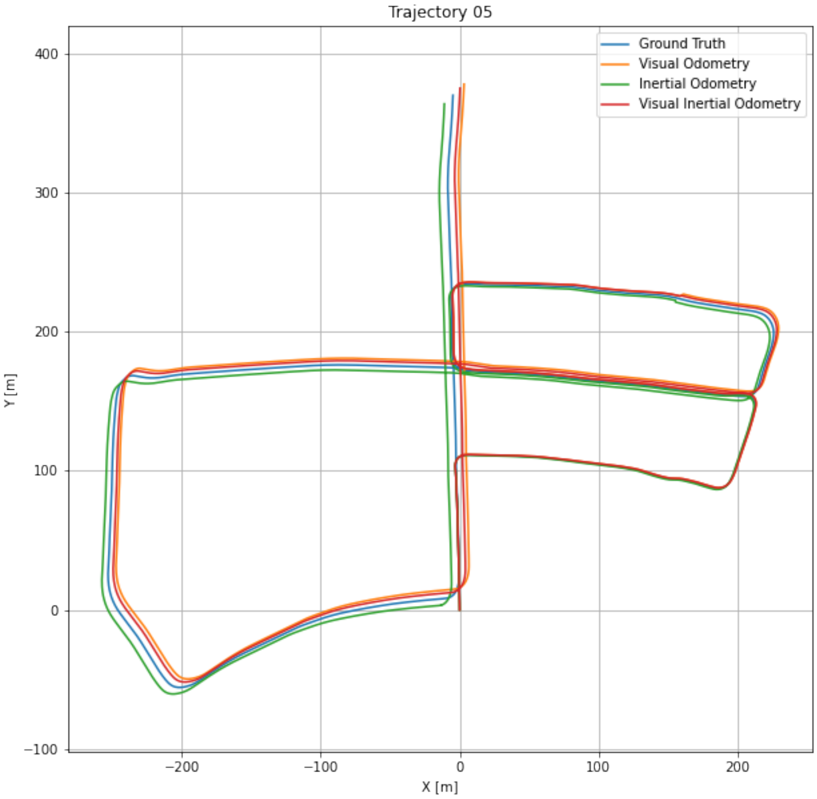

Figure 13: This is a graphical depiction and comparison of the six-degree-of-freedom pose estimation results of three GNSS/GPS alternative navigation systems. Computed upon sequence 05 of the world renowned KITTI data set.

There is little insight to be grounded from the description of Figure 13, which has now already been gained from the previous trajectories of this data set. As such, it is only inserted here for the sake of completeness.

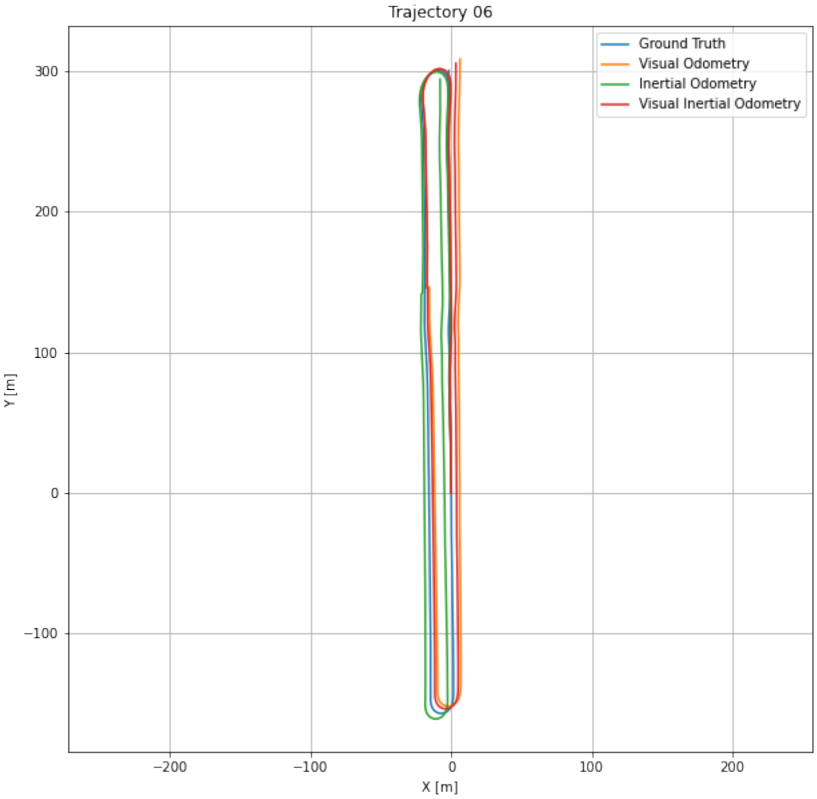

Figure 14: This is a graphical depiction and comparison of the six-degree-of-freedom pose estimation results of three GNSS/GPS alternative navigation systems. Computed upon sequence 06 of the world-renowned KITTI data set.

Figure 14 demonstrates that the inertial navigation system has a heavy reliance on its sesnors, which is not the case with the visual navigation system as such in conditions with the inertial navigation, sensors cannot pick up accurate rotation and translation information and certain features of the trajectory. As evident by the elongated closed loop which is evident on the right-hand side of the ground truth trajectory is no longer closed in the measure navigation estimate. Further to this, whilst the visual navigation system relying on feature-matching has a far different estimate than the fusion of the two, the filter provides the closest estimate with the least error as should be expected.

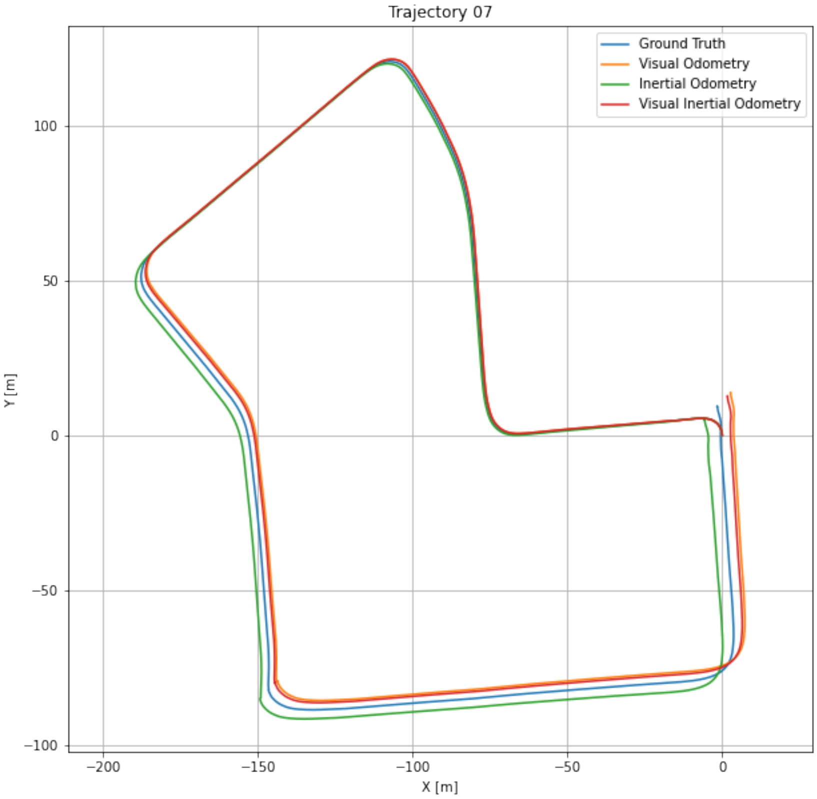

Figure 15: This is a graphical depiction and comparison of the six-degree-of-freedom pose estimation results of three GNSS/GPS alternative navigation systems. Computed upon sequence 07 of the world-renowned KITTI data set.

Figure 15 demonstrates the over a shorter trajectory, with smooth rotation and acceleration in the first and second derivatives. The inertial navigation system often outperforms visual based systems. This is shown to be the case as the immersion navigation system overlaps at the ground truth for a significant proportion of the trajectory, after which it is superimposed upon by either of the remaining two estimation systems.

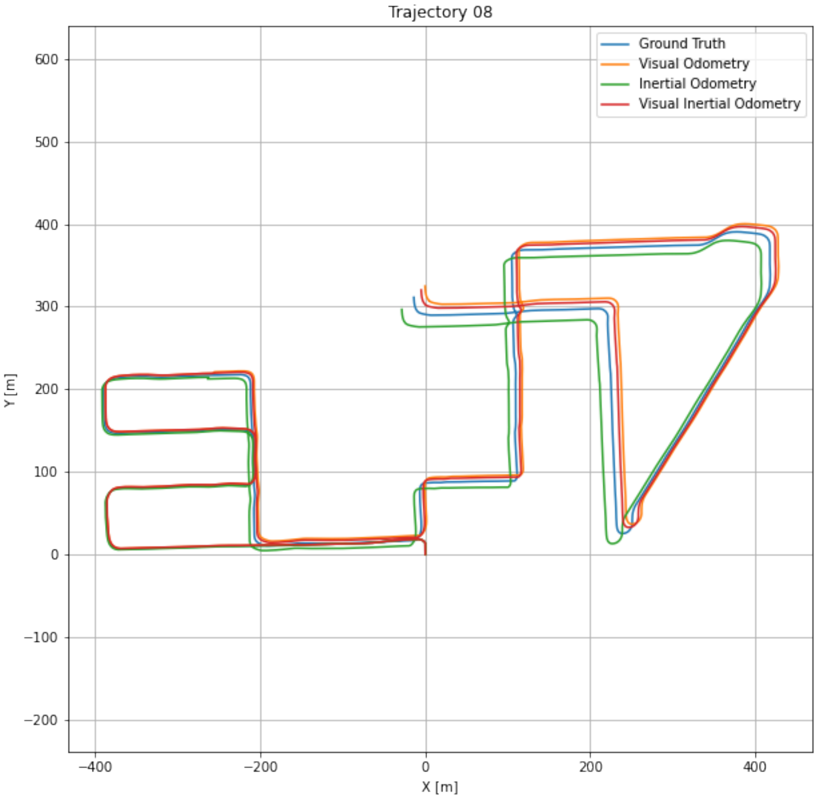

Figure 16: This is a graphical depiction and comparison of the six-degree-of-freedom pose estimation results of three GNSS/GPS alternative navigation systems. Computed upon sequence 08 of the world-renowned KITTI data set.

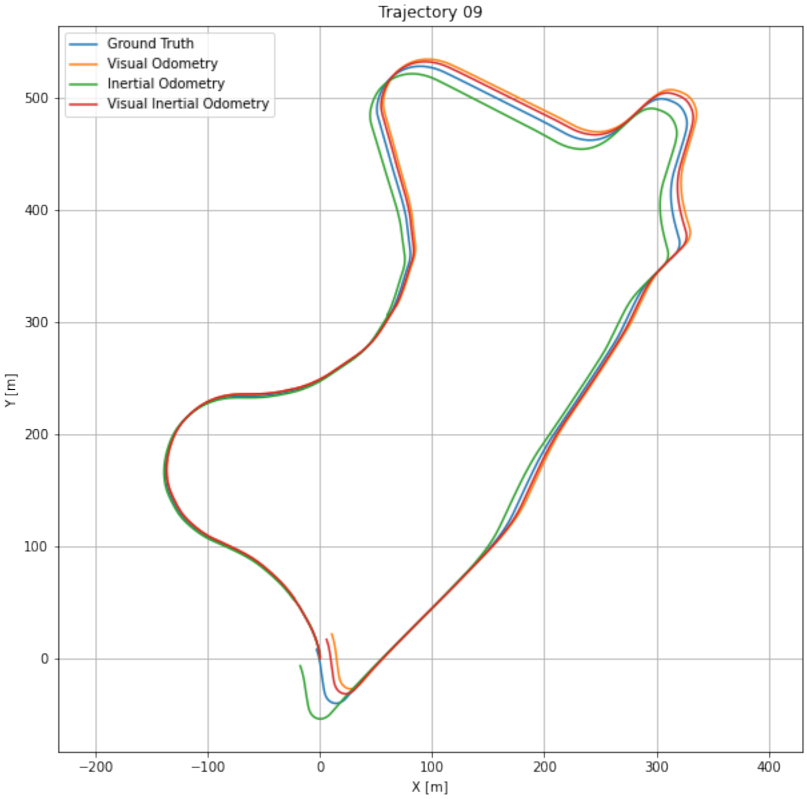

Figure 17: This is a graphical depiction and comparison of the six-degree-of-freedom pose estimation results of three GNSS/GPS alternative navigation systems. Computed upon sequence 09 of the world-renowned KITTI data set.

Figure 17 is unlike any other trajectory currently depicted in the data set. This is due to the fact that it operates in the most expected of ways oscillating between the filter inertial and visual methods in the accuracy of estimation and reliance on the error propagation of each system. In cases in which the visual and inertial methods have relatively lower errors and the filter of the fusion method provides the optimal result; however, in cases in which the visual methods have too great a drift the fusion method suffers as it is being pulled by the worst producer model. As a result, the better preforming model is often the closest to the ground truth.

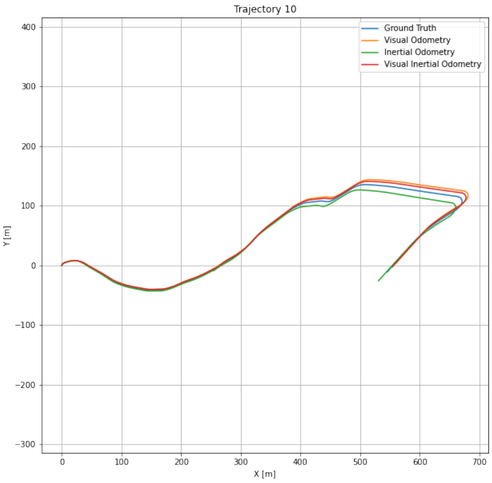

Figure 18: This is a graphical depiction and comparison of the six-degree-of-freedom pose estimation results of three GNSS/GPS alternative navigation systems. Computed upon sequence 10 of the world-renowned KITTI data set.

Figure 18 provides no information in addition to that which has been provided by pre-existing sequences of this data set, as such it is only included here for the purposes of completeness.

The following plots no longer represent the KITTI dataset but are results from the in-house dataset. The title of each plot will indicate which class and therefore difficulty the trajectory belongs to, through the interference tag. The trajectory will be given in the first subset of the title, finally, the presence of external objects is depicted by the final subset of the title.

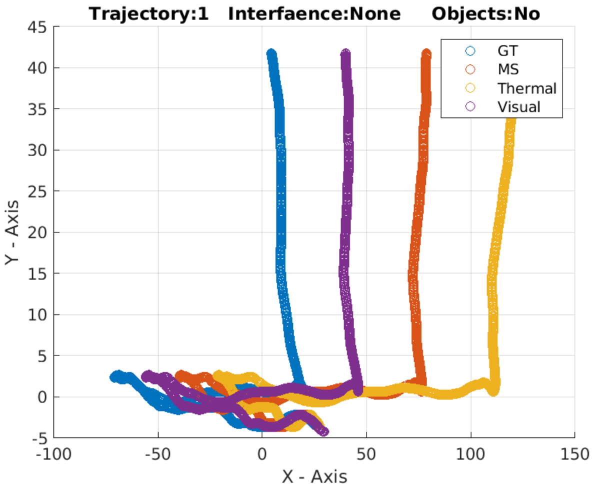

Figure 19: This figure is a graphical depiction of the results of the extended Kalman filter visual-inertial odometry solution. The experiment is run on trajectory number 1 and has interference in the no subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 19 shows that all three methods have a similar contour to each other in terms of their error dynamics. This is likely due to the instrumentation of the experimentation. The motion of the agent was almost perfectly contained in the XY plane and the motion in the Y axis is almost perfectly linear. Due to this, it is understandable that the majority of the error is in the X-axis, as the decoupling of rotation and translation shows that the rotation estimate can produce up to 80% of the errors. It also shows that the thermal EKF is the worst performing as is expected due to the fact the feature-selecting protocol was developed for the visual domain and not optimised for the thermal one.

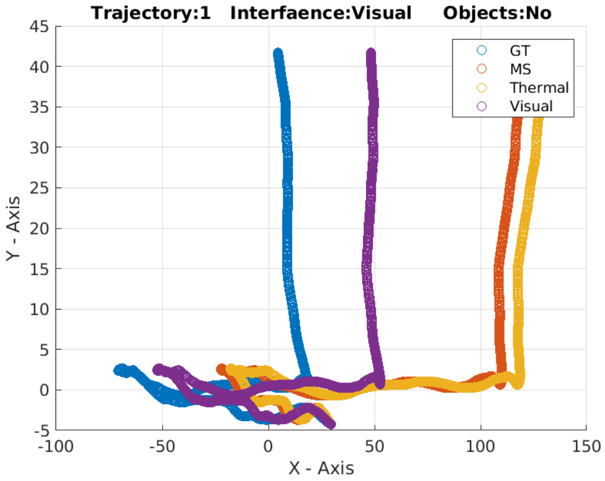

Figure 20: This figure is a graphical depiction of the results of the extended Kalman filter visual-inertial odometry solution. The experiment is run on trajectory number 1 and has interference in the visual subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

It can be seen in Figure 20 that when the stereo filter is run upon a trajectory, one with visual interference and no obstacles being present, the visual modality produces the best result and the multispectral produces significantly worse results than before. This is evidence that the interference caused the visual modality to lose some of the better matching feature points which could previously have been identified across modalities. This then forces the optimization to utilize a significantly worse optimization constraining equation which evaluates to the substantially worse result.

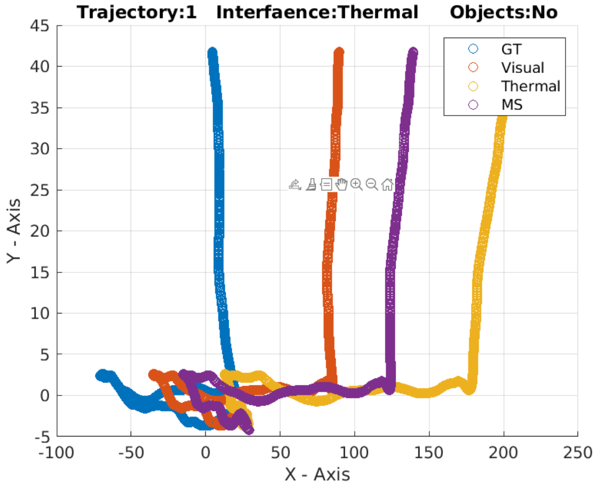

Figure 21: This figure is a graphical depiction of the results of the extended Kalman filter visual-inertial odometry solution. The experiment is run on trajectory number 1 and has interference in the thermal subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 21 shows that the results generated on the first trajectory, with thermal interference and no obstacles present, produce a significantly worse thermal pose estimation as can be deduced from the new scale present on the x-axis. The placement of the interference mechanism within the trajectory suggests that only the portion of the trajectory in which the interference was present suffered from the lack of consistent texture; however, it should also be noted that this had knock on effects on the remainder of the trajectory. Thus, the final position is significantly distinct from that of the previous run.

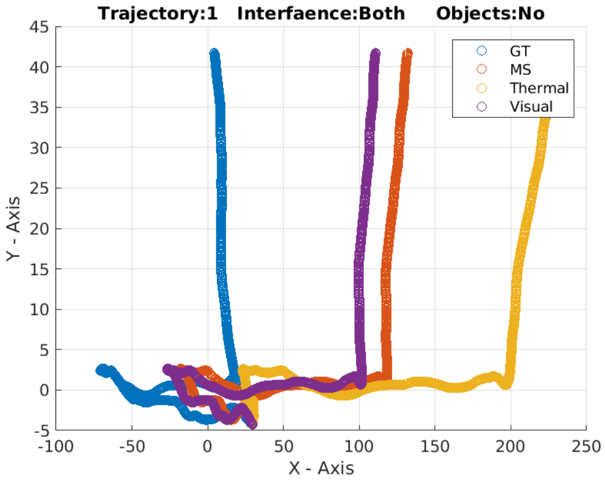

Figure 22: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 1 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 22 conveys facts that upon trajectory, one when both forms of interference or present and no obstacles are present, the thermal mortality suffers far more than the visual and the filter, and therefore closely resembles the visual pose estimate.

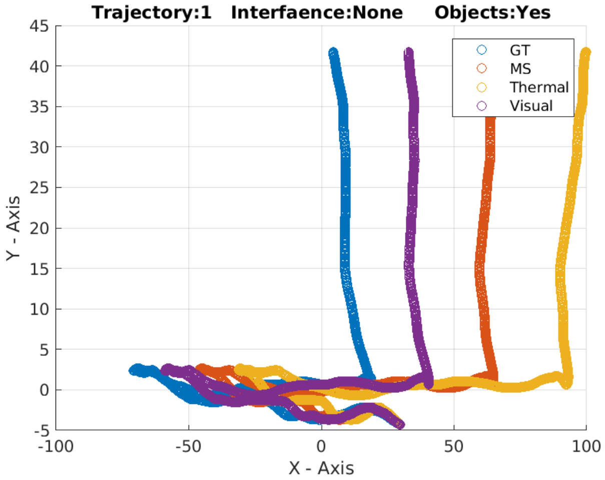

Figure 23: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 1 and has interference in the no subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 23 clearly shows that the introduction of obstacles to the path of the trajectory, with no interference, has improved their result. This is due to the fact that significant proportions of the environment under which the experiment was run were lacking in texture, and therefore identifiable features which could be distinguished and tract across frames. As a result of this the introduction of new texture into the scene seems to have improved the results to some degree.

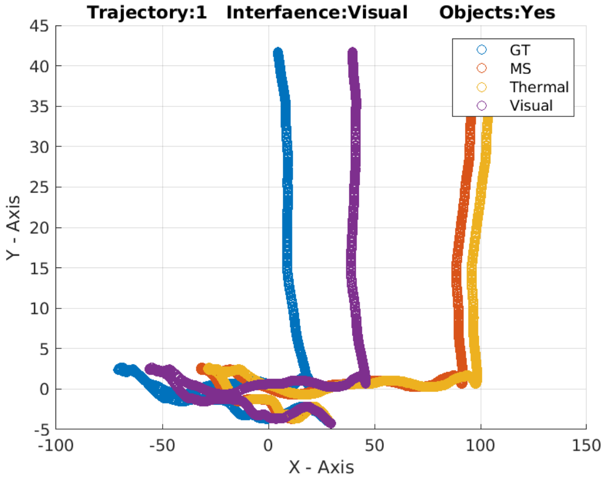

Figure 24: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 1 and has interference in the visual subsection of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 24 only serves to show that the presence of obstacles on the trajectory, with the visual interference, once again reduces the error in all directions and modalities as a result of the heightened texture.

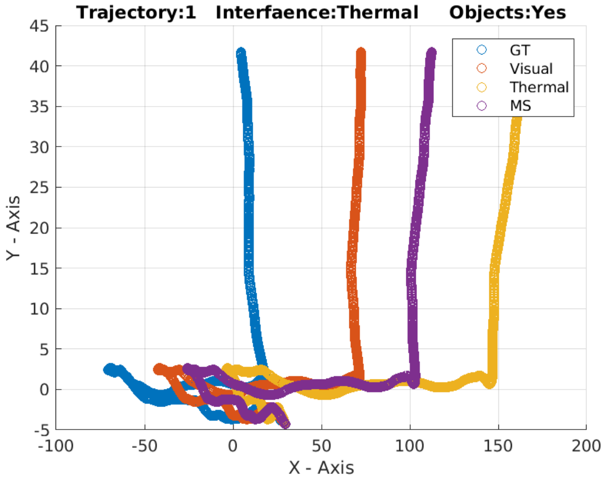

Figure 25: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 1 and has interference in the thermal subsection of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 25 only serves to show that the presence of obstacles on the trajectory, one with the thermal interference, once again reduces the error in all directions and modalities as a result of the heightened texture.

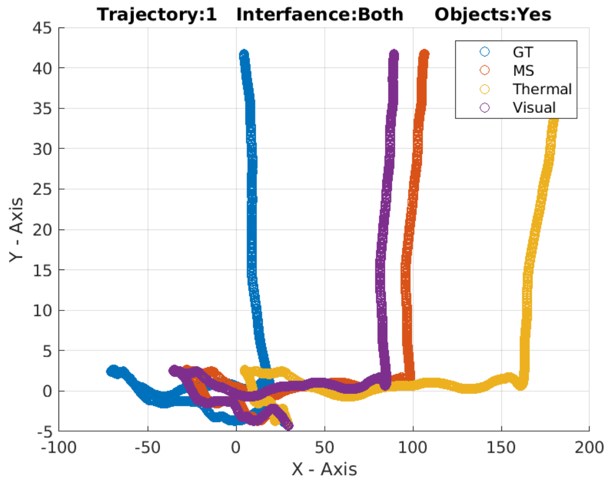

Figure 26: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 1 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 26 only serves to show that the presence of obstacles on the trajectory, with both interferences, once again reduces the error in all directions and modalities as a result of the heightened texture.

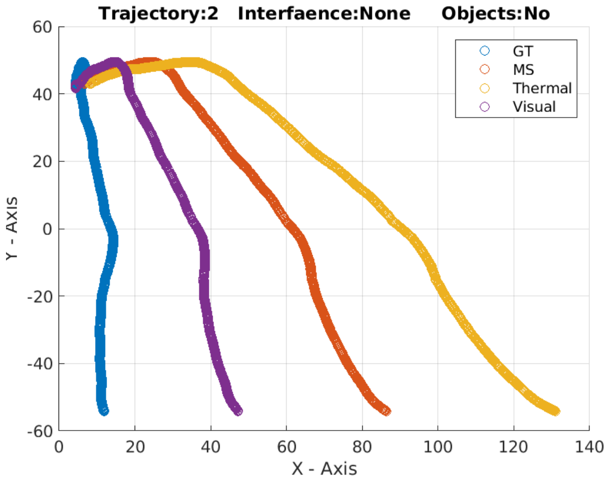

Figure 27: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 2 and has interference in no subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 27 depicts the run of the trajectory with no interference and no obstacles present. In this run, it is possible to see that the ground truth has almost no variation on the x axis relative to the Y; however, this is not the case for the three approximations produced by the three different solutions. The reason for this likely stands from the instability of the robot platform during the practical experimentation instrumentation due to the perturbation along the Z axis of the robot. There was additional angular change to the perceived locations of the feature points and in such a short trajectory with little rotation, this seems to have produced a substantial change in the expected results and the results gained.

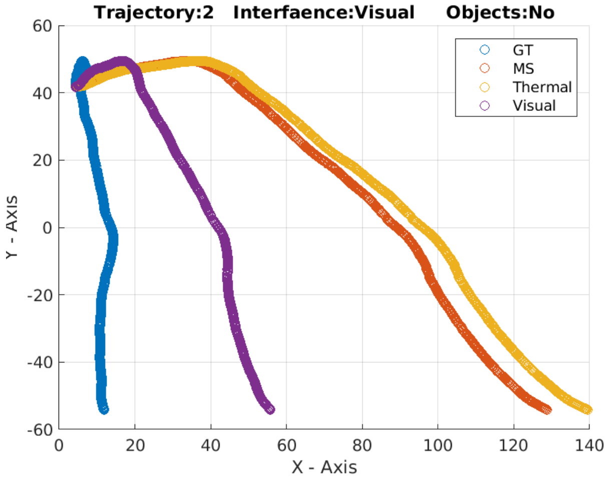

Figure 28: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 2 and has interference in the visual subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 28 depicts a trajectory with visual interference and no obstacles present. In this situation, it can be seen that the additional interference in the visual modality pushes the filter to produce results far closer to the thermal modality, which should be expected due to the four-way matching required by the feature point matching algorithm.

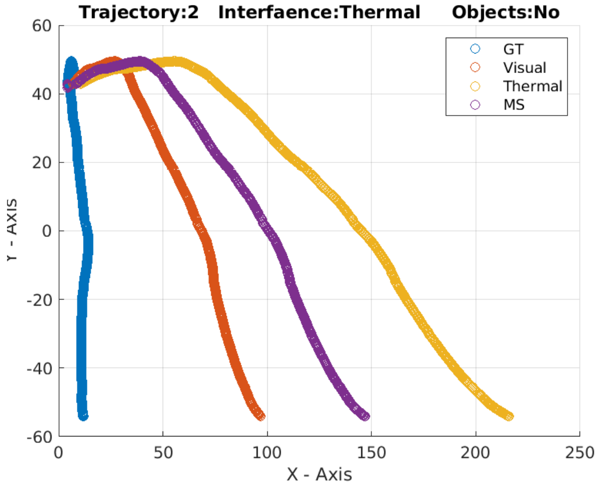

Figure 29: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 2 and has interference in the thermal subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 29 reinforces the conclusions derived from the prior trajectory. This illustration offers little more than reinforcement of the previously found inferences.

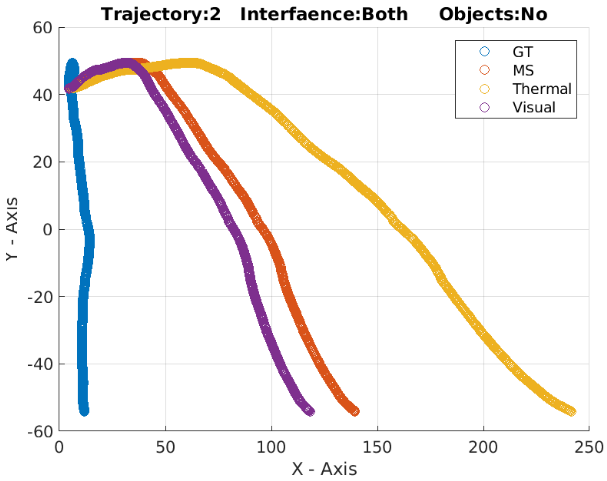

Figure 30: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 2 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 30 demonstrates the fact that the visual interference seems to have a less negative effect than the thermal interference. This seems to stem from the fact that the scene was homogeneous in temperature and so had far fewer features to lose.

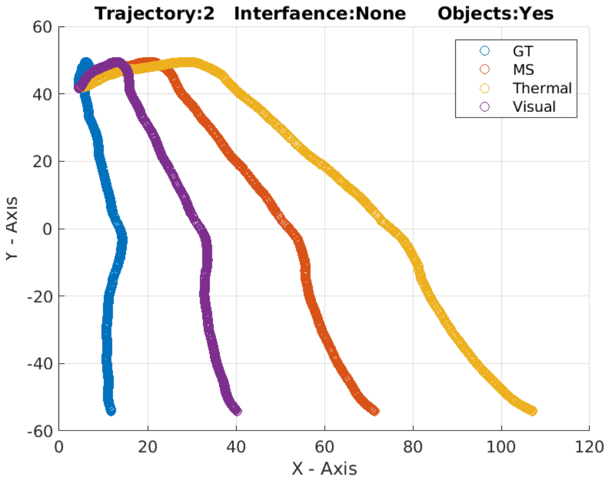

Figure 31: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 2 and has interference in no subsection of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 31 shows that the obstacles have improved the texture in the scene and in the absence of any interference, the filters behave as expected.

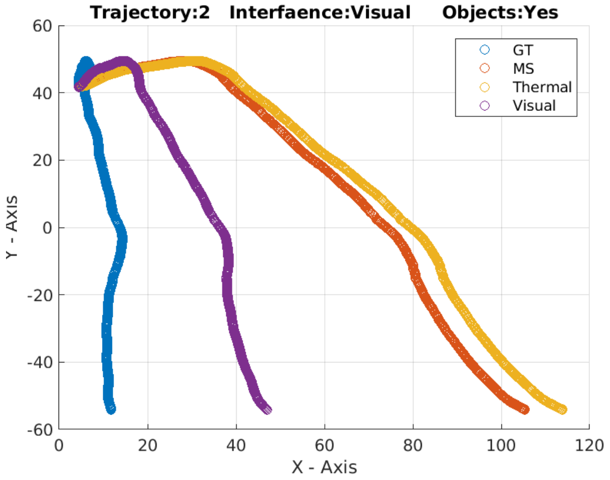

Figure 32: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 2 and has interference in the visual subsection of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 32 shows that the introduction of the obstacles does improve the results; however, it does not completely compensate for the introduction of interference.

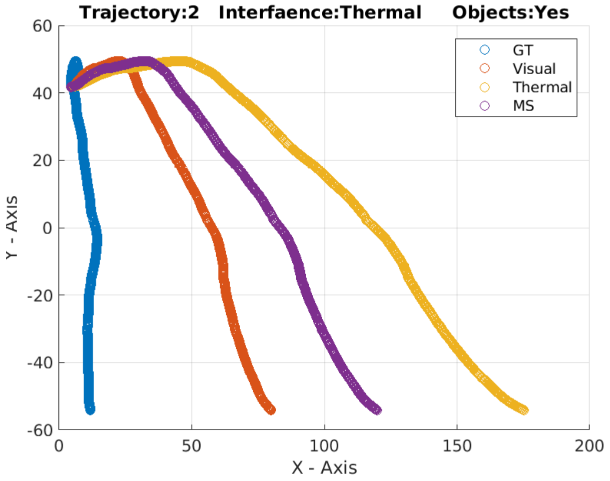

Figure 33: This figure is a graphical depiction of the results of the extended Kalman filter visual-inertial odometry solution. The experiment is run on trajectory number 2 and has interference in the thermal subsection of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 33 demonstrates that the multispectral solution can act as a smoothing filter at times, as its gradient is smoother in places then either of the stereo solutions.

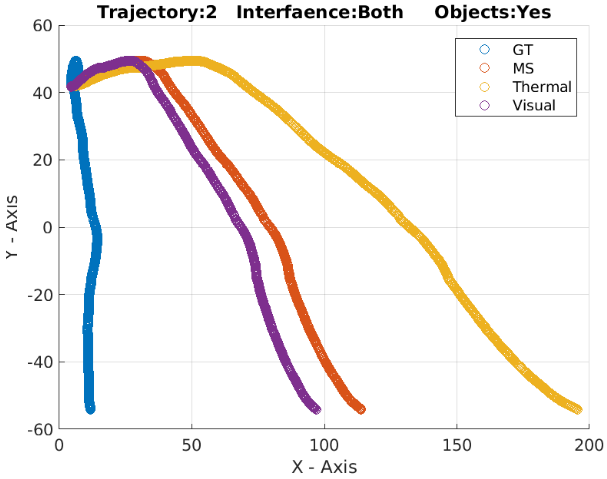

Figure 34: This figure is a graphical depiction of the results of the extended Kalman filter visual-inertial odometry solution. The experiment is run on trajectory number 2 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 34 shows that the obstacles cannot completely offset the introduction of the interference and the thermal interference is far more effective than the visual interference.

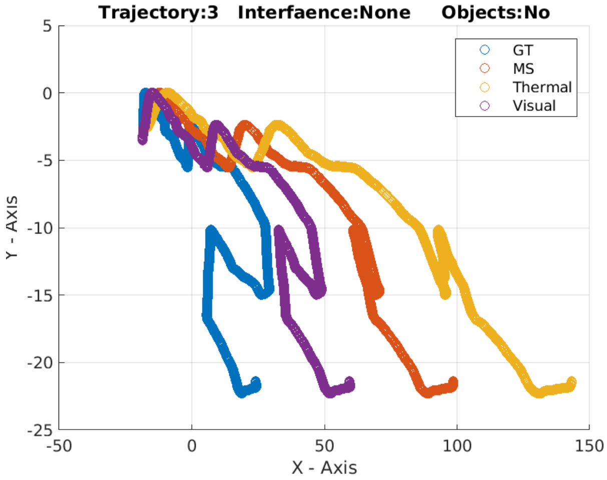

Figure 35: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 3 and has interference in no subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 35 demonstrates the base case of the EKF on trajectory three. It also shows that the compensation of the filter has left the multispectral EKF result to lose the hook shape. This is likely showing that during this step, the features were easily trackable in each modality but far worse to track between modalities.

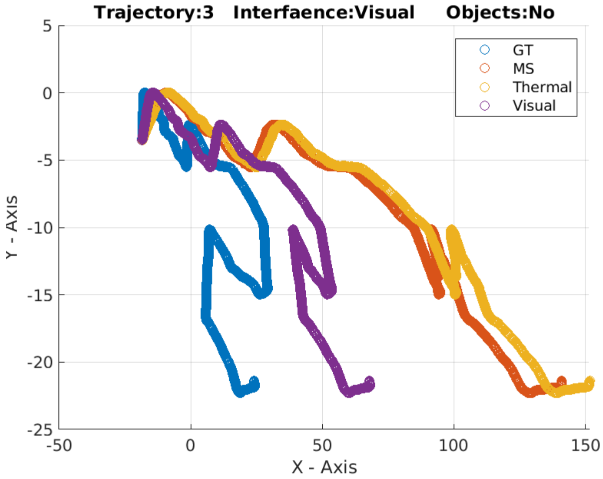

Figure 36: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 3 and has interference in the visual subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 36 shows a rather interesting result and that is that the visual interference is making it hard to match thermal features in the visual image. This means that the features with the strongest matching criteria that were picked up are from the thermal modality.

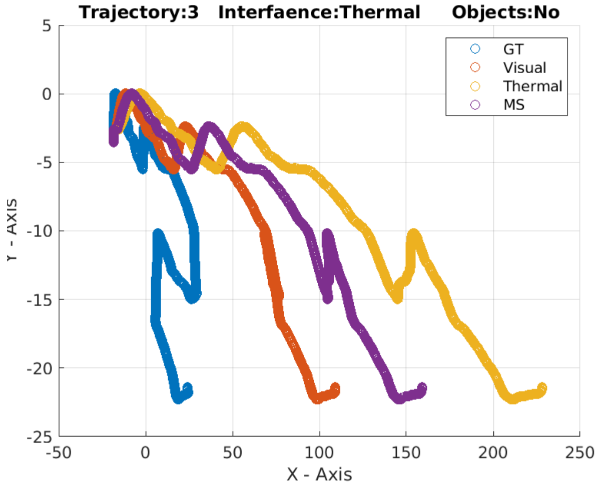

Figure 37: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 3 and has interference in the thermal subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 37 has a flattened loop in the visual image pose estimate, even though the trajectory was only subject to the thermal interface, this is due to the limited control over the experimental environment that the author had. During this trajectory, much of the visual cues were obfuscated by the university staff moving stuff in the background of the environment.

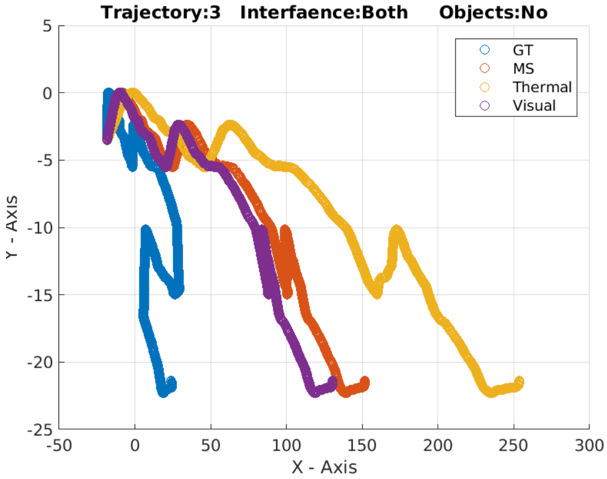

Figure 38: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 3 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 38 is quite interesting as it shows that the visual feature points matching did not track the hooky stick shape in the trajectory, as well as the thermal modality. It also shows that the feature points in the scene were so bad that they resulted in a wrong orientation for the hooky stick.

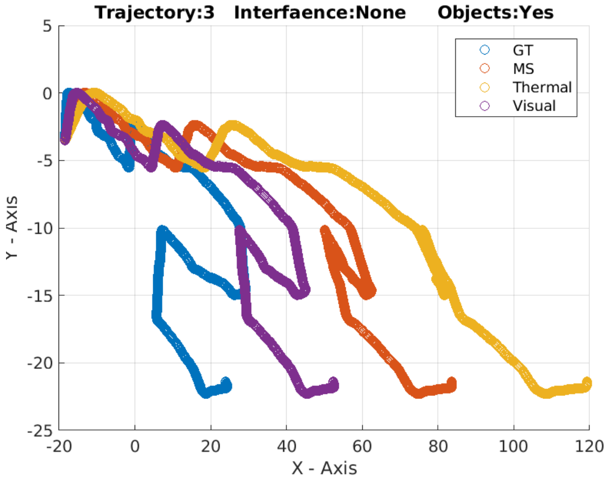

Figure 39: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 3 and has interference in no subsection of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 39 shows an interesting result which is the fact that in the absence of the interference and with the addition of texture from the obstacles means that the visual points were far more clearly represented than the thermal as is shown by the shape and orientation of the hook.

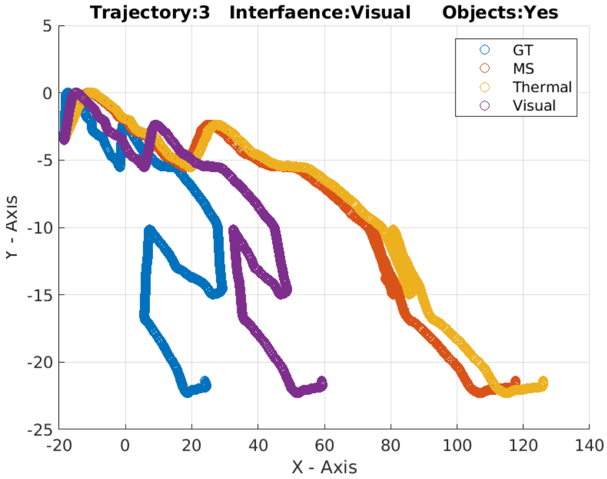

Figure 40: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 3 and has interference in the visual subsection of the electromagnetic spectrum. There were objects present during this experimentation.

In Figure 40, the extra texture was enough to ensure the correct contour of the visual trajectory, but the introduction of the visual interference has made the matching step place more weight on the thermal feature points, which did not track the contour of the ground truth very well.

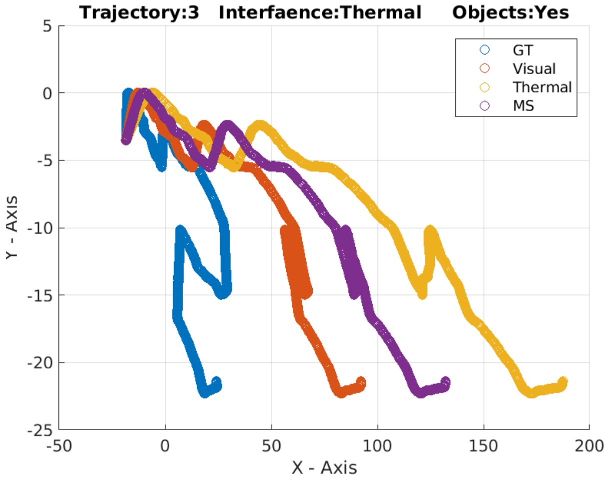

Figure 41: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 3 and has interference in the thermal subsection of the electromagnetic spectrum. There were objects present during this experimentation.

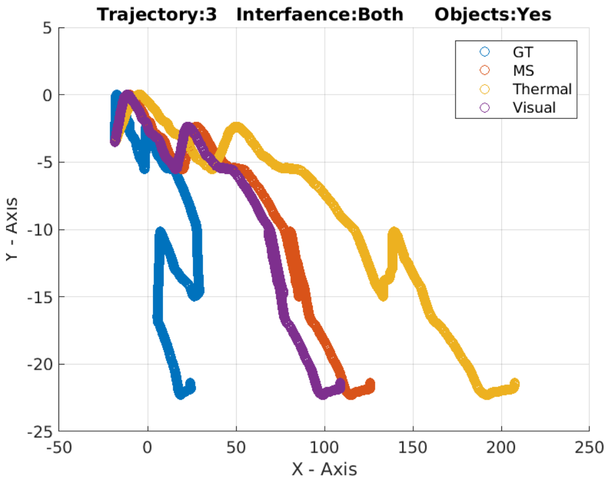

Figure 42: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 3 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Based upon the previous results, Figure 42 can be seen as an encoding of multiple results. This is the expected behaviour.

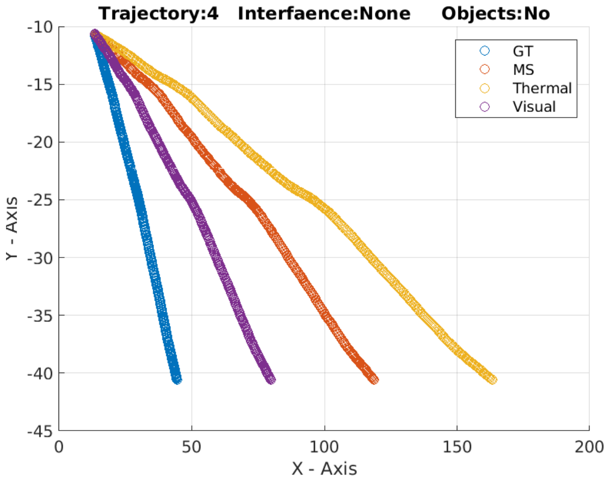

Figure 43: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 4 and has interference in no subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 43 depicts the results of running the test on a short trajectory. Whilst the results performance is consistent with expectations, it can be seen that there will be little rotation in this trajectory.

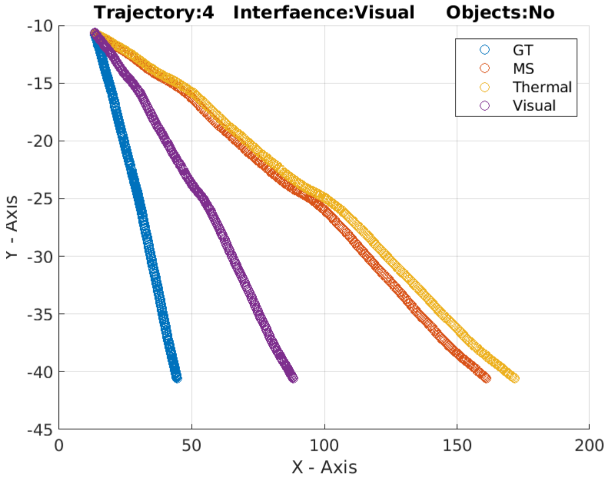

Figure 44: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 4 and has interference in the visual subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 44 introduces the presence of visual interference and as a result, the multispectral solution tends to the stereo thermal solution.

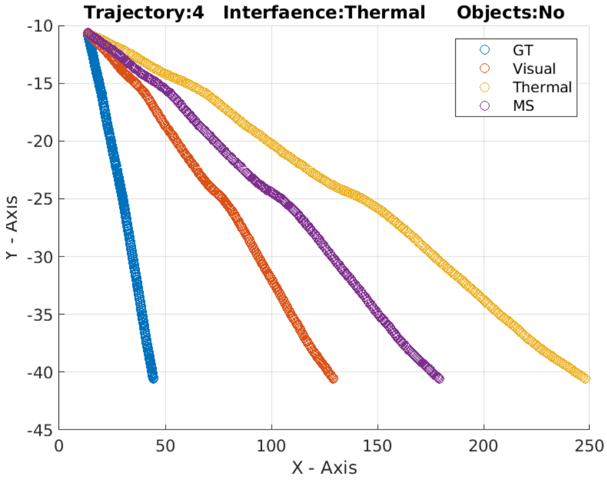

Figure 45: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 4 and has interference in the thermal subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

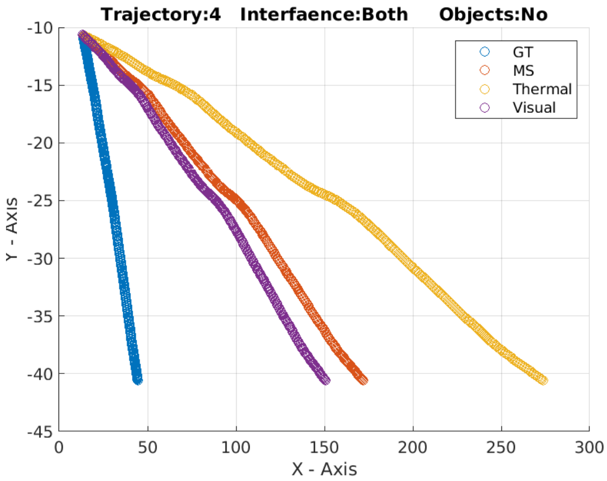

Figure 46: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 4 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 46 has no objects present but has both interfaces present. In such an event, each trajectory seems to have a greater deviation from the ground truth but with no change in their respective rankings. This is the expected behaviour.

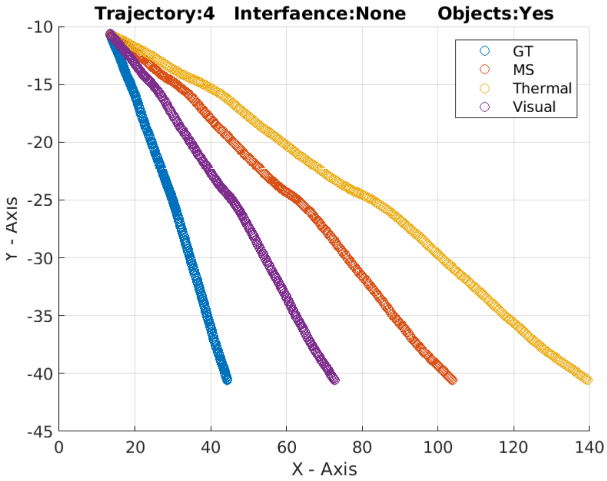

Figure 47: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 4 and has interference in no subsection of the electromagnetic spectrum. There were objects present during this experimentation.

In Figure 47 where objects are present but with no interference of any kind, it is clear that the errors are slightly lower due to the appearance of the extra texture in the scene.

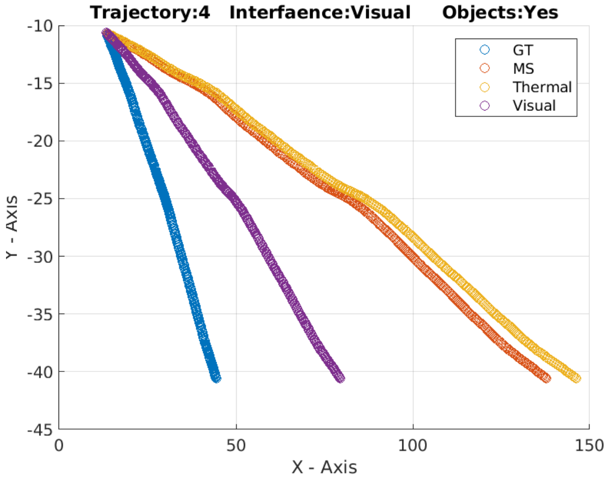

Figure 48: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 4 and has interference in the visual subsection of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 48 once again demonstrates that the introduction of visual interference has pushed the multispectral solution towards the thermal as it has caused no shift in the thermal estimate but has made the visual counterpart(s) worse.

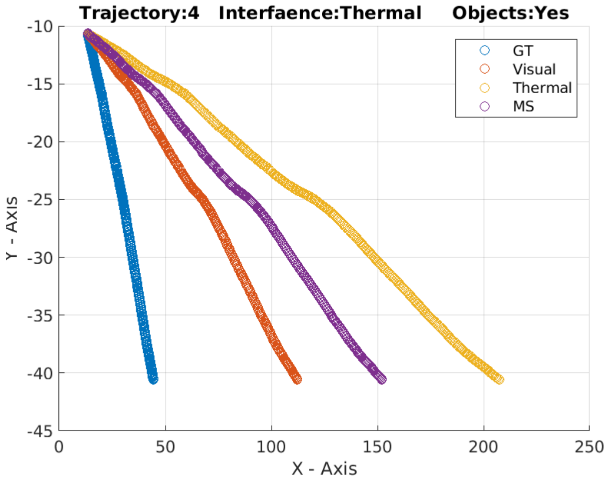

Figure 49: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 4 and has interference in the thermal subsection of the electromagnetic spectrum. There were objects present during this experimentation.

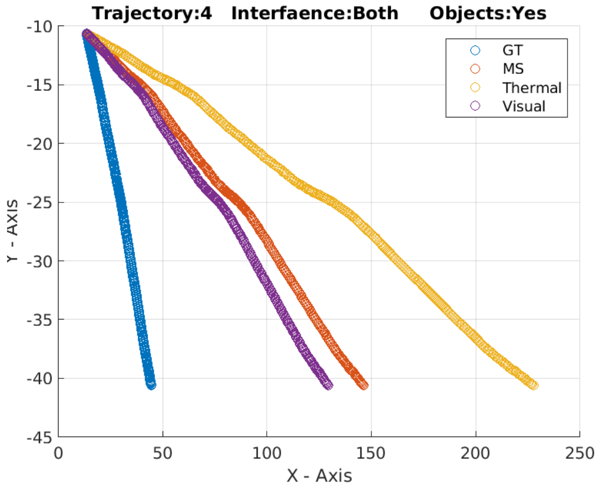

Figure 50: This figure is a graphical depiction of the results of the extended Kalman filter visual inertial odometry solution. The experiment is run on trajectory number 4 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 50 shows no behaviour that has not been presented by previous trajectory plots; however, it does seem to have the multispectral and visual solutions very close to each other. A possible explanation for this could be the four-way matching of feature points skewing to the use of visual feature points and not the thermal due to the effect of the thermal interference being larger than the visual.

3.4 Conclusions and Future Work

It is the principle objective of this section of the thesis to convey the conclusions of the experimental results and theoretical framework presented in this chapter of the thesis and to outline possible future work that could enhance the work done here or provide additional detail on nuance for further development of the field.

In this section of the thesis, it has been found that the use of data fusion methods to compensate for the robustness of a single solution is possible, however, doing so has some unexpected consequences, as well as some expected ones. It is also the findings of this section that the lack of texture within the location in which the practical instrumentation took place results in some rather unusual behaviour. In most circumstances it is known that the introduction of obstacles could often reduce the estimation qualities of the navigational solution; however, the lack of texture in the region actually results in a benefit to the navigational systems estimation qualities, should such objects be present.

The introduction of interference in either or both of the modalities often leads to the exact behaviour, which is expected; however, at times it produces some rather unexpected phenomena. This is due to the fact that one of the two modalities seems to have resulted in the creation of an influence on the interference and the use of a four-way matching protocol in the matching feature points across modalities often can skew the output.

To the world, these findings are of interest, yet a second experimental operation may be required. It is a belief of the author of this thesis that the world could benefit from redoing this experimentation in an environment with significantly more texture and on a robot that does not suffer from such perturbations in the Z axis. This is due to the fact that both of these problems have led to the observation of phenomena which is normally considered to be rare with considerable frequency within this experimentation. Whilst this is somewhat desirable for the thesis, it is not descriptive of the behaviour of the system under the normal operating conditions.

In conclusion, the work done here provides great insight and showcases many results which are often rare in less controlled environments, acts as a proven list of the existence of such phenomena and so may be considered valuable. It is also highlighted that due to the inadequacies of the practical limitations of this project, the considerable reason to redo the practical experimentation.

4 Deep Odometry

It is a principle objective of this section to introduce the audience to the body of work this thesis has produced in accordance with the notion of a deep neural network-based navigational system pose estimator. It should be noted that this is really different from the latest sections of the thesis, which may use similar technology (artificial intelligence) to reach their conclusions; however, they have a different primary aim. As the primary goals are sufficiently distinct from that of this section of the thesis, it was desirable to construct them as individual sections, thereby allowing for a progression of the narrative of the thesis.

The primary works in this section of the thesis consist of the construction of a novel artificial-intelligence-based navigational system which exploits the use of optical flow to estimate the trajectory of the agent to which it is applied and a generalizable framework by which great optimization can be taken for arbitrary multi-modality or multispectral deep-learning-based end-to-end solutions. Although these works are to be taken in conjunction, due to their novelty and application they merit standing alone; as such, they are taken to be subcomponents of a single section and not a single component of a single section of this thesis.

Recent advancements in data-driven computer vision tasks have allowed learning-based methods to explore monocular computer vision tasks without explicitly applying geometric theory. Such methods can address challenges associated with classic monocular VO problems, including feature extraction, depth estimation and data association. While Machine Learning techniques exist [179], Deep Learning methods tend to produce more satisfactory results automatically and we therefore primarily utilize this technology in monocular computer vision work.

Optic flow-based algorithms also gain much consideration. CL-VO [180] employs a cascade optical flow network for more detailed flow estimation and introduces Curriculum Learning strategy to perform bounded pose regression, while DROID-SLAM [181] iteratively optimizes camera poses and depth by employing a recurrent update based on optical flow estimation network RAFT [182], where Dense Bundle Adjustment layer leverages geometric constraints to increase accuracy and robustness without retraining, enabling this monocular method to handle stereo or RGB-D input without retraining. In [183], learn the latent space of optical flow which yields motion estimation constrained by sequential images.

4.1 Deep Multispectral Inertial Odometry

It is the principle objective of this section of the thesis to introduce the audience to the central concepts behind the construction of the deep visual inertial navigation system constructed in this thesis and to convey the results and experimental implementations of the artificial intelligence based model. It is also desirable to express the novelty of this approach in the employment of optical flow. This section of the thesis is also used as a simple illustrated example of the best complexity of employing different sampling configurations from the electromagnetic spectrum in a single end-to-end navigation solution, which arises from the substantial number of permutations which can and do emerge from even severely limited sensor configuration. An example of such a severely limited sensor configuration would be those exploited in this thesis in which only one or two modalities are used to sample the three-dimensional phenomena present in the scene and present them in a two-dimensional pictorial representation known colloquially as an image.

4.1.1 Motivation

Given the disastrous findings of the last chapter, it is the principal objective of this chapter to attempt to develop a multispectral VIO model that will not suffer the same drop in performance. To achieve this objective, a study of the electromagnetic spectrum reveals that there exist seven broad classifications of electromagnetic spectra that can be labelled as a modality, as each of the seven spectra may be combined with each other. There are 42 possible pairings; however, for the purposes of multispectral VO, counting only unique pairings reduces the count to 21.

Given that there exist 21 unique pairings, this thesis suffers from hardware limitations and so must attempt to develop rigorously a modular approach to multispectral visual odometry, which may be adapted to other modality pairs with relative ease. Following this, the thesis tests the viability of the approach upon the visual-thermal MS combination and compares the findings to both the stereo-thermal and stereo-visual solutions.

This approach of deriving the 21 unique pairings can be applied to each of the 7 layers of possible fusions, from single modalities on the 1^{st} level to the single possible fusion on the 7^{th} level. It should also be noted that it is possible to construct an image-based odometry system from non-EM based imagery and so there must be some form of reconciliation between the EM and non-EM based (or hybrid) solutions. This would mean that this chapter of the thesis provides to the tree of human knowledge:

- The reconciliation of the field of visual odometry to all possible EM-based imagery, conforming to the pinhole camera model.
- Conclusive proof that the portability of deep-learning based models enables them to adapt to modalities other than the one they were trained on.
- A novel taxonomy of newly reconciled fields, inclusive of all hybrid methods.

4.1.2 Methodology

To provide the ability to adapt to all EM-based modalities, it was determined that a modular approach would work best, as the ability to replace components would result in not having to retrain the entire model each time the modality was altered.

To this end, it is possible to deduce that there must be some layer of abstraction in the model, which results in large portions of the model being employed (without retraining) regardless of modality. This would suggest that there must exist some sort of separation between the feature extraction steps and the feature-matching stage inherent in the model.

The requirement of backpropagation to determine the correct kernel values for the feature extraction components of the model further stipulated that the feature extraction process be completely independent of the remainder of the model whilst also being independent of the training of other modalities.

These factors combined to force the model components that enable the extraction of features to be modular and indeed self-contained models, which regardless of the modality of the input must produce the same output modality. This requirement of a consistent output modality imposed a daunting restriction on the project. This restriction forces the model to be able to convert different encodings into a single universal feature map stack; which is not trivial when considering that each feature extraction module would have to learn to extract a set of feature points from image-based encodings of the photons in a scene or the distribution of heat in a scene, and then proceed to convert this feature set into a novel output scheme that can remain sufficiently stable to enable the feature-matching pipeline. This would mean the prevention of visual/thermal artefacts in the output modality.

The feature maps must then be sufficiently consistent to prevent the forced retraining of the model, i.e. the only effect the change of modality should have would be to alter the weights of certain finite subsets of the model. This further requires that the proportion of the model that extracts the feature set from the image planes must have a single architecture, reducing drastically the possibility of optimisation.

It is a fact that the segregation of the KITTI data set scoreboard demonstrates the limitation of each sensor class. It was deemed prudent to enable a module that would initially incorporate a single IMU signal to enable data fusion but remain sufficiently versatile to enable more abstract data fusion in further revisions. To this end, it was determined that prior processing of the IMU signal may enable the system to learn richer features from the IMU signal, whilst removing noise and smoothing the signal.

Due to the desire of the model to enable the portability of traditional visual odometry techniques to other modalities, it proved necessary to attempt to encode geometric solutions to the visual odometry problem into the model and re-train a proportion of that model on a different modality, whilst preventing the model from moving to a different geometric solution. Due to the black-box nature of deep-learning models, the significance of this task could not be underestimated.

The mathematical solution to the visual odometry problem selected for the encoding is one based on optical flow. The central idea is that the optical flow of the scene is an encoding of the motion of the agent. The homography is employed to encode the motion of the scene between images in a monocular image stream; however, the motion captured by the homography is the complement of the motion undergone by the camera sensor rigidly attached to the agent.

Whilst it may seem trivial to exploit this relationship to extract the motion of the scene and then proceed to decode the motion into the motion of the agent. The presence of distortions to the optical flow of the scene significantly complicates this problem.

The existence of the distortions in the optical flow motion field of the scene arises due to the existence of moving objects such as people and vehicles in the scene, which do not follow the motion of the scene but move in accordance with some other objective. Such objects are termed actors and the motion generated from the actors are known as local motion fields. This reduces to the addition of another outlier detection and removal scheme into the model.

Due to the fact that the black-box nature of deep networks prohibits the exploration of the inner workings of the model, it is not possible to directly constrain the model to the desired mathematical solution. However, the universal learnability of the model enables the model to learn any relationship in existence. This suggests that forcing the model inputs to contain only a single possible solution to the problem would result in the model learning this relationship.

The exploitation of this suggestion forces the construction of both carefully crafted models and data streams in a single solution. This results in Figure 51.

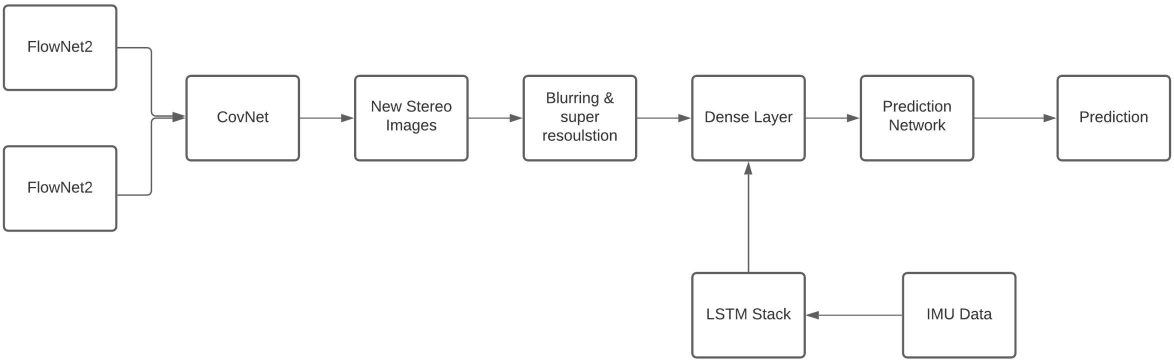

Figure 51: A Flowchart of the constructed model.

The FlowNet2 model can detect the local motion field generated by the agents in a scene and develop a tightly fitting mask overlaying the motion field. This enables the removal of the local motion fields. Later on, however, the model also maps the input stream into a novel output image which is independent of the image input modality. This results in the FlowNet2 models acting both as self-contained models and as interchangeable modules that can be trained on each modality independently and results in an output belonging to the same image space as all other modalities.

The new stereo image pair generated by the FlowNet2 models essentially map the input images from each modality to a consistent image space with each local motion field labelled. This new stereo image pair acts as the only input seen by the motion removal network that removes the local motion fields the output of which is even more pixelated than the output of the FlowNet2 networks originally. In order to prevent this from hindering the learning of the prediction layer, a superresolution CNN is stacked on top of the motion removal network.

The output of the superresolution CNN is flattened into a single dense layer. The dense layer is then concatenated with the six dense neurons representing the output of the processed IMU signal. The IMU signal is processed through a stack of three identical LSTM units, each with 1000 hidden nodes, before resulting in a six-neuron representation of the signal which is trained to correspond to the rotation and the motion of the sensor.

This stacked dense vector is the input to the final prediction module, which through backpropagation jointly trains the motion removal, superresolution, LSTM, and prediction modules; however, the FlowNet2 models are trained independently from the rest of the model. This can be viewed as the stacking of a series of models where the inputs to the first set of models (FlowNet2 instances) are the captured images and the input to the second stack are the FlowNet2 generated images, whilst the input to the 3^{rd} model is the stacked output of the 2^{nd} set of images and the processed IMU signal, enabling some degree of modularity.

The systematic removal of all the local motion fields forces the model to learn the appropriate encoding as there exists no other possible encoding that could produce a consistent estimate of the agent's motion. This is enforced through the weight optimisation step.

The architecture of the network is represented in Figures 11-12, with the notable exception of the FlowNet2 component which is identical to that of the original paper.

Layer	Kernel	Padding	Stride	Channels
Conv 1	11	3	1	64
Conv 2	7	2	1	128
Conv 3	5	2	1	256
Conv 3-1	3	1	1	256
Conv 4	3	1	1	512
Conv 4-1	3	1	1	512
Conv 5	3	1	1	512
Conv 5-1	3	1	1	512
Conv 6	3	1	1	1024

Table 2: A review of the beginning modules of the model.

Layer	Kernel	Padding	Stride	Channels
Conv 1	11	1	1	1024
Conv 2	7	1	1	2048
Conv 3	5	1	1	4500
Conv 4	3	1	1	5000
Conv 5	3	1	1	5500
Conv 6	3	1	1	5750
Conv 7	3	1	1	5820
Dense 1	5820	-	-	3000

Table 3: A review of the middle module of the model.

Layer	Kernel	Padding	Stride	Channels
Dense 1	-	-	-	4000
Dense 2	-	-	-	3000
Dense 3	-	-	-	2000
Dense 4	-	-	-	1500
Dense 5	-	-	-	1000
Dense 6	-	-	-	500
Dense 7	-	-	-	6

Table 4: A review of the final modules of the model.

4.1.3 Training

To train the model upon the KITTI data set, it is imperative to note that the training process required a labelled data set; thus, only the first 11 sequences of the data set were employed. The images were normalised to have a constant standard deviation of 1 and a mean of 0 over the whole data set and each channel.

Post-normalisation data is then ready to be utilised in the training pipeline. When considering the splitting of the data set into a training, validation, and test sets, it has been shown that there are two methods of doing so: splitting the sequences into each set individually, and selectively placing the data set sequences into each class. The choice of method is known to have a considerable impact on the final results of the model. Due to this, each submodel was trained individually to ensure convergence to the optimal results.

Due to the fact that the model requires at each stage the output of the previous section, the submodules were naturally trained in order. The first compound, the FlowNet2 model, was trained exactly as outlined in the original paper, as subsequent results based upon the model have shown no substantial improvement upon this method of training the model. Previous work has attempted both methods and indeed showed that the change would not benefit the results.

After the successful training of the FlowNet2 models, the next submodel to be trained is the motion-removal submodel. The motion-removal model was trained against a novel data set, developed in Matlab. The labelled data was the result of replacing the local image fields identified by the FlowNet2 model with the static scenery left behind after the motion of the agent. This novel data set was used as a novel image stream with a one-to-one correspondence with the frames of the output of the FlowNet2 model. Each of the two streams has the same dimensions and, when flattened, allows for the creation of a custom convolutional neural network.

The second submodel was trained for 1000 epochs and reached an acceptable performance of around 800 epochs into training; however, it seems there is still some room for further optimisation of the model as the validation results did not completely level off. However, this model took over two months of training on an HPC GPU cluster to reach 1000 epochs over the whole data set. The average results of the training set produced a 3% error using the squared sum of deviations error loss function. The data set for this model was taken by splitting the proportions of each individual sequence into each of the sets, using a 70:20:10 ratio.

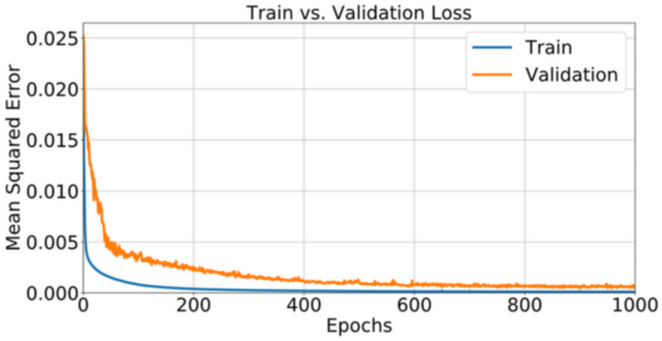

Figure 52: A graphical depiction of the training and validation loss of the first submodule.

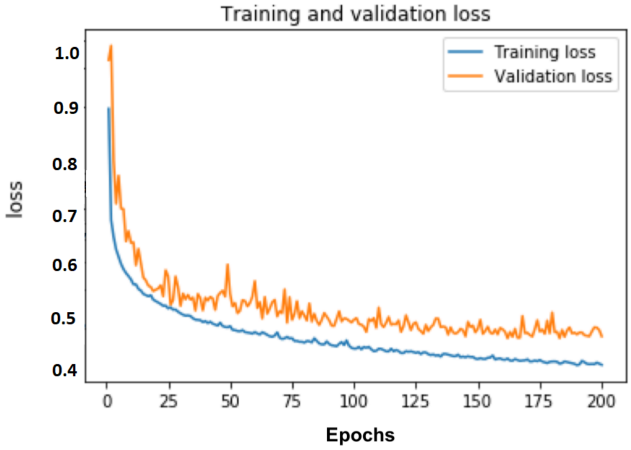

Figure 53: A graphical depiction of the training and validation loss of the seconed submodule.

At this stage, the processing of the IMU has to be considered, in addition to its incorporation into the prediction layers. The IMU measurements are combined into a single six-dimensional vector, which is passed through a set of 3 LSTMs stacked on top of each other with identical architectures. Each LSTM module has a hidden size of 100 neurons and a random weight initialisation. The output of the LSTM stack and the second submodel are stacked into a single one-dimensional dense layer, which is then used as the input of the final prediction submodel. The output of the final submodel and therefore the whole model is compared against the ground truth values obtained from the data set. The backpropagation of this submodel is also employed to train the LSTM stack.

The final prediction model was trained using a data set splitting of selectively allocating the sequences to each of the three sets. This process was inspired by the separation and normalisation schemes employed in [245]. This particular submodule was trained for 200 epochs and encountered some problems during the training process. Due to the integration with the LSTM stack, this module encountered the exploding gradients problem, which forced the decline in the learning rate from 0.01 to 0.001 and the introduction of a dropout of 0.2% on each layer. This combined with the introduction of gradient clipping to 10% of the actual change when the gradient exceeded unity removed the exploding gradient problem.

To train this submodel, the invention of a novel optimization metric was required. Inspired by the results of the study of the decoupling of translation and rotation estimation, the optimization metric is a linear combination of the SSD error of both the translational error and the rotational error respectively. The hyperparameter used to construct the metric lambda is used to set the weighting of the rotational error. From the results of the decoupling study, it was noted that this value should not force the weighted sum to be bound in the inclusive range of 0 and 100% of the total error. The optimized value of the parameter is 2.

4.1.4 Results

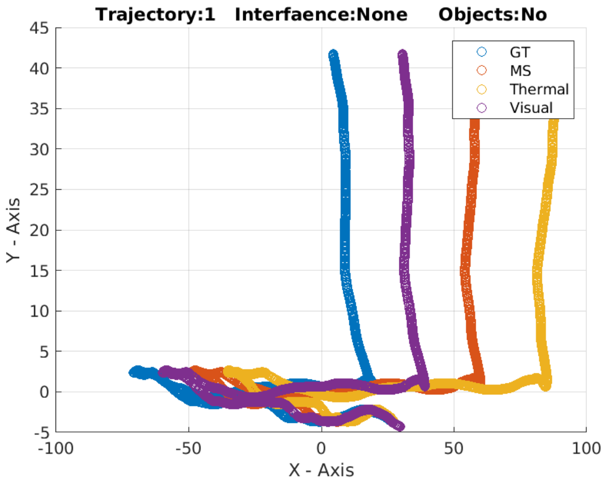

Figure 54: The experiment is run on trajectory number 1 and has interference in no subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 54 demonstrates the results of the deep visual inertial navigation solution on trajectory one with no interface or objects present. The results are desirable as the spread between the three solutions is fairly narrow and the multispectral solution is almost halfway between the single modality trajectories showcasing the ability of the kernel optimisation to select better features than the traditional feature matching pipeline.

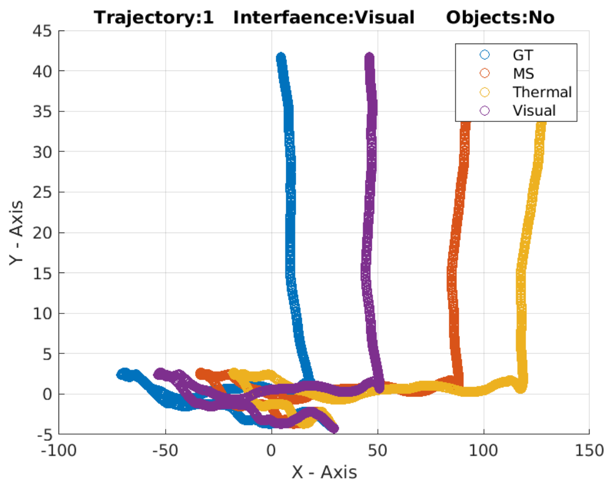

Figure 55: The experiment is run on trajectory number 1 and has interference in the visual subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 55 is of considerable interest as it showcases the fact that the ability of the deep-learning method to select better feature points often makes it susceptible to diversions from the training data in extreme fashion. This is shown by the fact that the visual interface has affected the stereo thermal run.

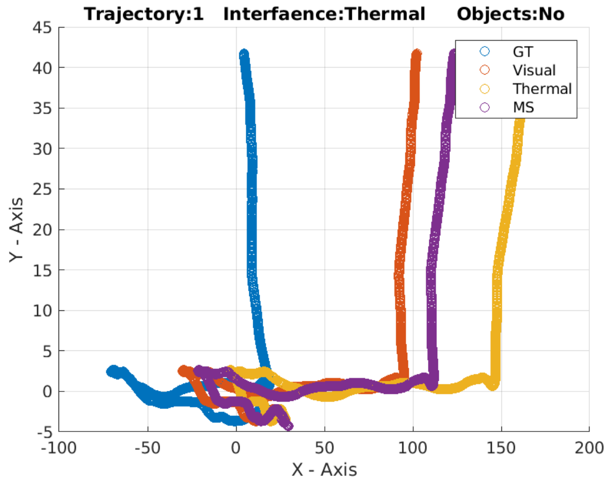

Figure 56: The experiment is run on trajectory number 1 and has interference in the thermal subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 56 should be carefully viewed as the legend key is substantially different from the subsequent image and may lead to misinterpretation. On careful examination, it can be seen to follow the expected behaviour. It should be noted that the results of this section are tested over the same data as various other solutions in this thesis, but it is the objective of each such test to prove the novelty of the approach and not to improve the accuracy of the previous solution.

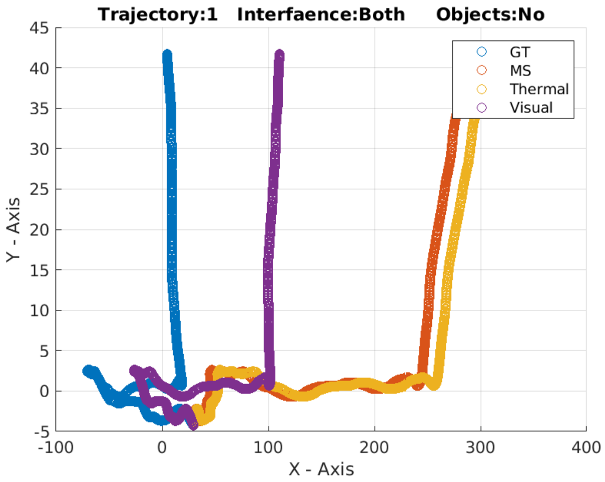

Figure 57: The experiment is run on trajectory number 1 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 57 shows the results of both interfaces on trajectory one without obstacles. This shows that the AI model has a far harder time with the thermal modality in the presence of the interference than the visual. It also shows that the multispectral solution closely follows the thermal, suggesting the models converge to a local minimum or the architecture needs to be redone.

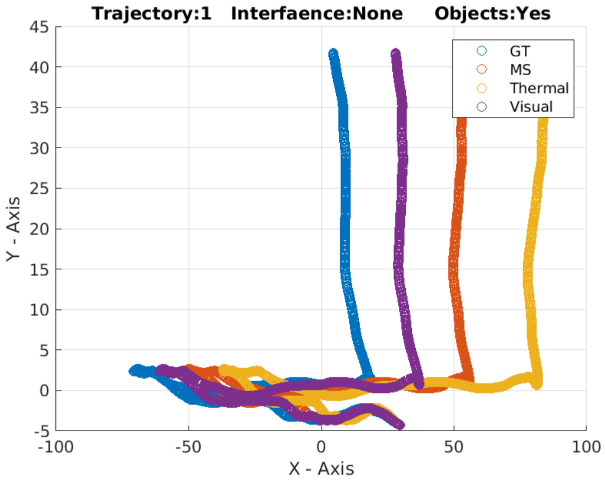

Figure 58: The experiment is run on trajectory number 1 and has interference in no subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 58 demonstrates that even the AI-based solution produces better results with the presence of the obstacles, proving further evidence for the lack of texture in the scene.

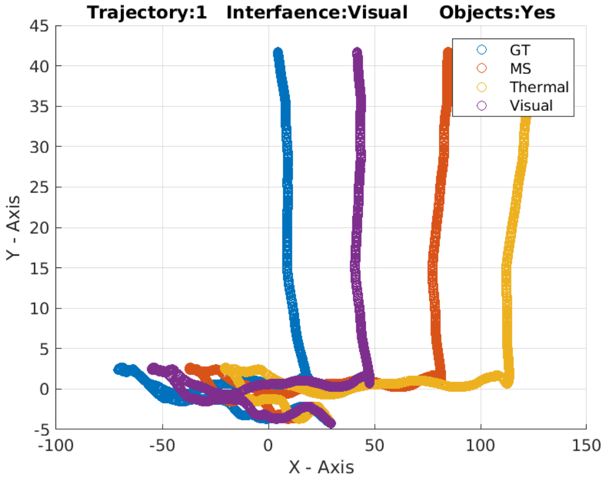

Figure 59: The experiment is run on trajectory number 1 and has interference in the visual subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 59 demonstrates a worsening of the visual and multispectral versions of the model and not the thermal. This suggests that the AI model can sometimes isolate the visual interface in the visual pose estimation and sometimes it leaks over to the thermal modality. This is likely due to the distance from the light source and the temperature of the light source. It is unfortunate but the modeling of such phenomena is out of the scope of this thesis.

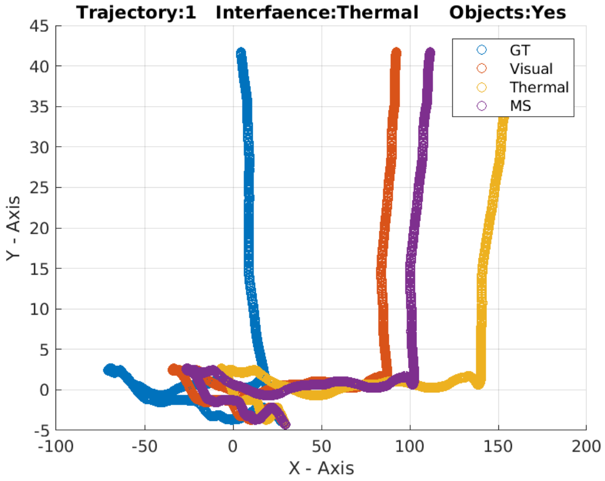

Figure 60: The experiment is run on trajectory number 1 and has interference in the thermal subsection of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 60 demonstrates the expected results namely that the introduction of the thermal interface will push the multispectral solution closer to the visual trajectory estimate.

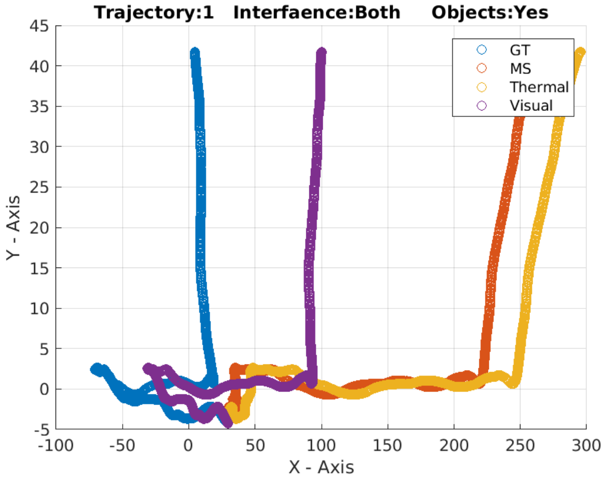

Figure 61: The experiment is run on trajectory number 1 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 61 demonstrates that the introduction of the interface worsens the trajectories of all three navigation solutions, but once again the thermal and the multispectral are the most affected. Further to this is the fact the introduction of the obstacles has widened the distance between the thermal and multispectral solutions. This is probably due to the fact that the introduction of new features in the low-texture region affects both base modalities, resulting in the visual modality having a larger impact on the multispectral when considering that the interface affects the thermal modality far worse than the visual.

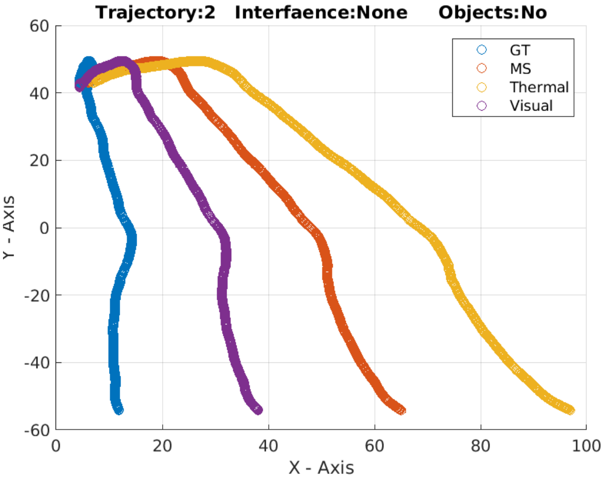

Figure 62: The experiment is run on trajectory number 2 and has interference in no subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 62 is clearly showing that the highly tuned hyperoptimised kernels of the AI model cannot closely match the ground truth. The fact that this phenomenon is present in this small trajectory in both the classical EKF and AI solutions is further evidence of the low texture in the scene.

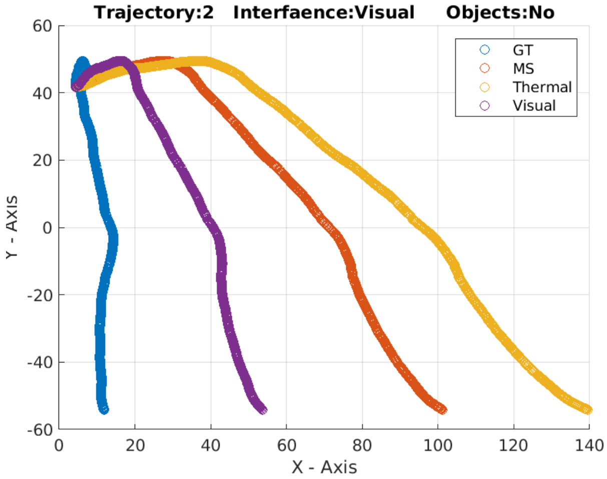

Figure 63: The experiment is run on trajectory number 2 and has interference in the visual subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 63 highlights the fact that the introduction of the visual interface may have worsened all the trajectories; however, the contour of the trajectory estimates did not change.

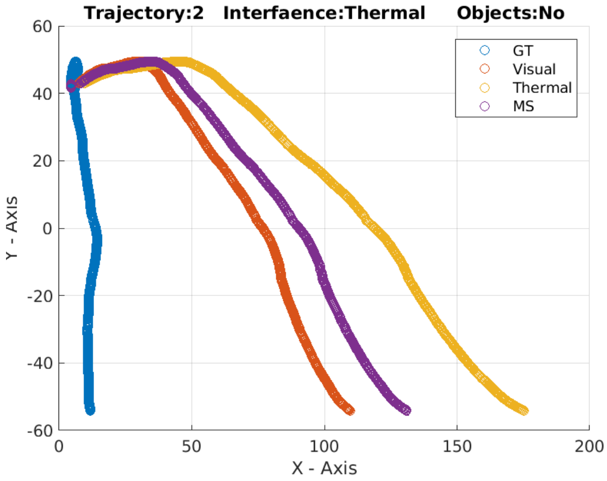

Figure 64: The experiment is run on trajectory number 2 and has interference in the thermal subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 64 seems to have a different trajectory contour to the previous figure, however, this is a visual illusion given by the scale of the trajectory. This comparison also shows that the introduction of both interfaces will cause significant errors in the pose estimation.

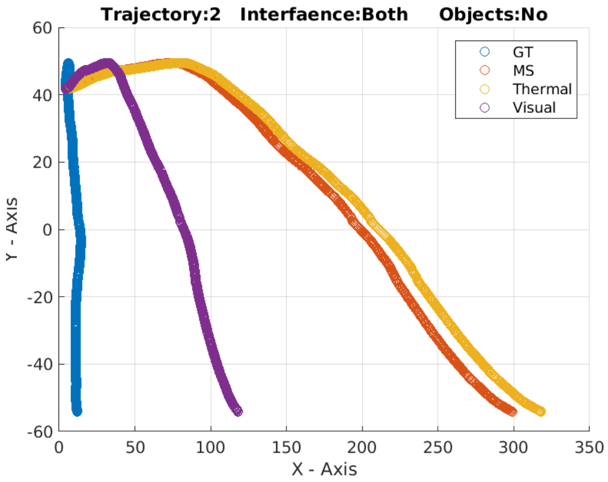

Figure 65: The experiment is run on trajectory number 2 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 65 shows that as the error in the x axis of the three solutions tends to ∞, the ground truth appears to tend to a straight line. This is a perceptive trick of the scale of the X axis, which increases to express the errors of the three navigational solutions.

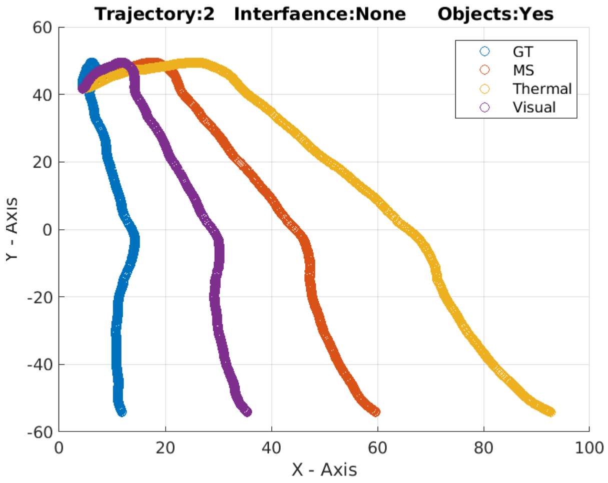

Figure 66: The experiment is run on trajectory number 2 and has interference in no subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 66 once again demonstrates that the introduction of the obstacles actually helps the solutions estimate the trajectory of the agent.

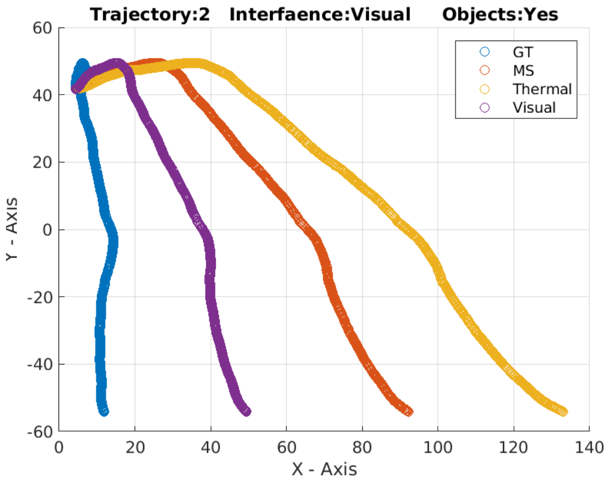

Figure 67: The experiment is run on trajectory number 2 and has interference in the visual subsection of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 67 demonstrates no result which would be of interest due solely to the fact that no deductions or inferences can be drawn from it in excess of theories provided by other trajectories.

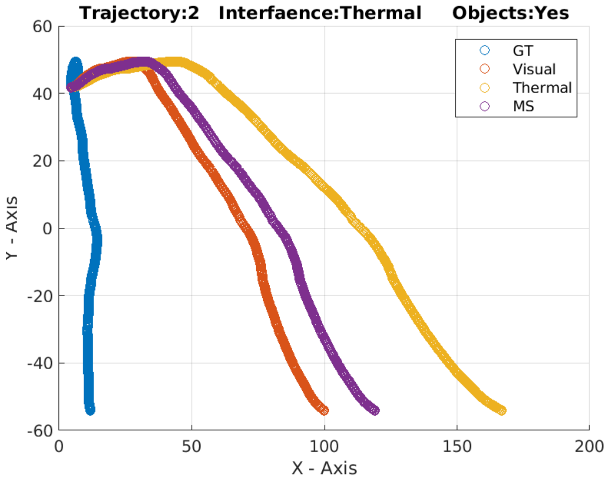

Figure 68: The experiment is run on trajectory number 2 and has interference in the thermal subsections of the electromagnetic spectrum. There were objects present during this experimentation.

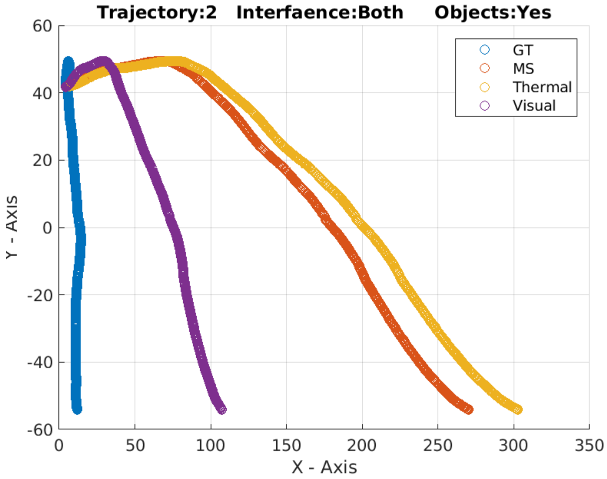

Figure 69: The experiment is run on trajectory number 2 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 69 has a very interesting feature and that is the point of divergence between the thermal and multispectral trajectories. Given that the gap between the two appears to increase as a function of the propagation of the agent, it is possible that the visual feature became more prominent later in the trajectory. This would mean that the effects of the interface are not constant in time and vary over the course of the trajectory. This is explainable by the distance between the agent and the source of the interference.

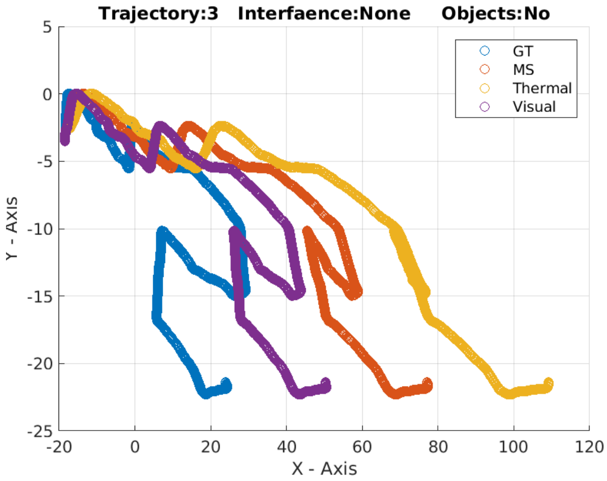

Figure 70: The experiment is run on trajectory number 3 and has interference in no subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 70 demonstrates that the thermal image stream is not very good. This is due to the practical set-up and the fact that the FLIR cameras get relatively hot if they are run continuously and interfere with their own image capture. This can be seen here.

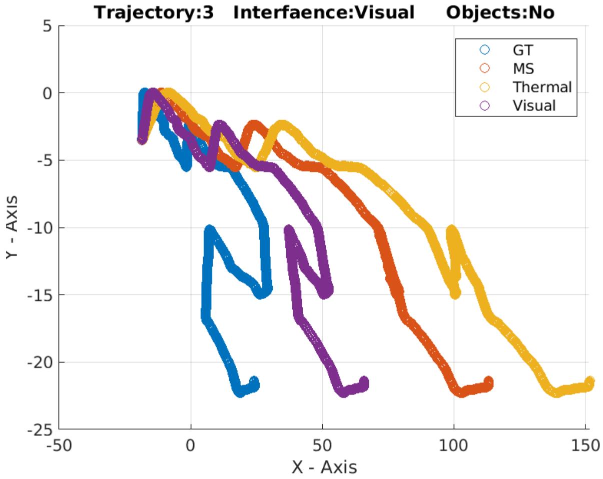

Figure 71: The experiment is run on trajectory number 3 and has interference in the visual subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 71 does not allow the multispectral estimate to produce the hook, clearly identifying that the feature points corresponding to the features in the three-dimensional world around the hook are not easily matched between modalities even with the use of AI.

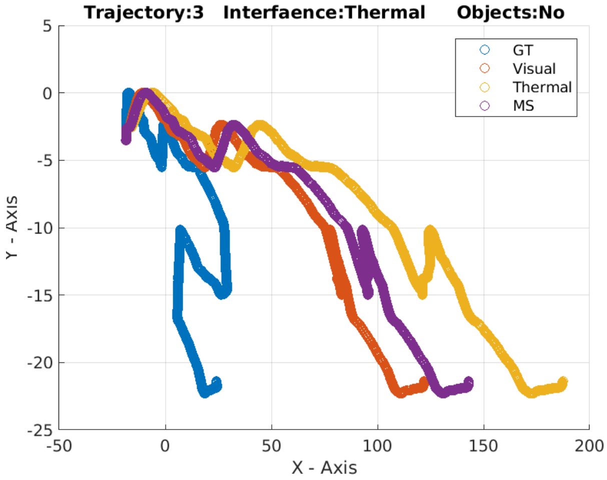

Figure 72: The experiment is run on trajectory number 3 and has interference in the thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

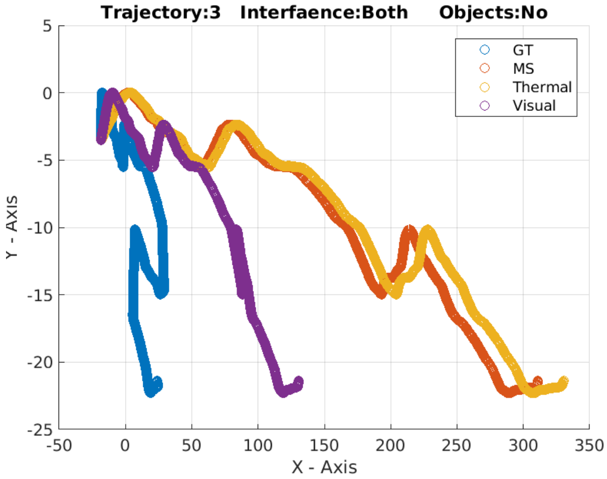

Figure 73: The experiment is run on trajectory number 3 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 73 is rather interesting in the fact that the use of the interference has caused both the visual solution to lose the distinctive hook shape in its contour and force the multispectral to follow the thermal closely in addition to the increase in the magnitude of error, highlighting the sensitivity of the AI model.

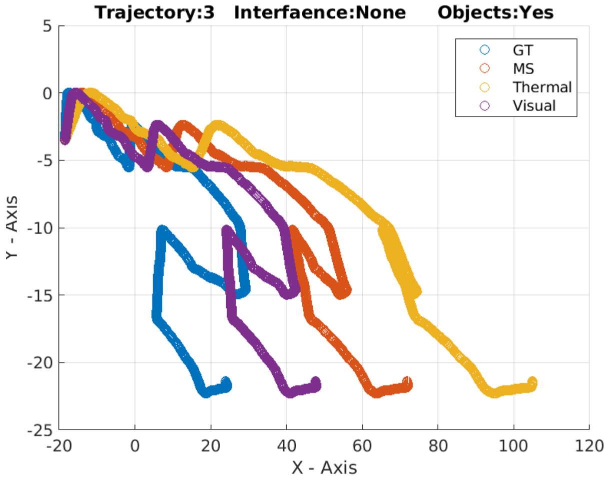

Figure 74: The experiment is run on trajectory number 3 and has interference in no subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 74 illustrates the fact that whilst the hook portion of the trajectories contour is hard to track, the additional heat affecting the thermal sensors made it far harder to identify in the thermal modality. The introduction of the additional texture allowed the visual and multispectral solutions to retain the hook, suggesting that the features in the proximity of the hood are more easily seen in the visual domain and not the thermal.

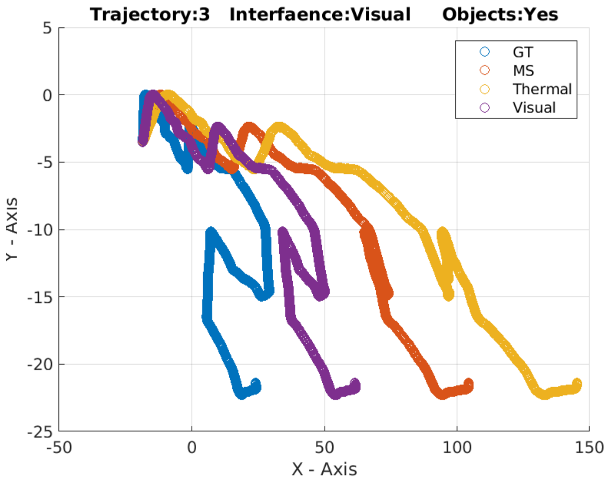

Figure 75: The experiment is run on trajectory number 3 and has interference in the visual subsection of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 75 demonstrates the fact that the introduction of the visual interface has dampened the contour of the visual trajectory and lessened its hook shape. The additional heat from the interference sources appears to compensate for the errors in the thermal modality. The multispectral solution now does not seem to have a hook anywhere as distinctive as it should. This is probably because it was collapsed from the thermal and visual changes.

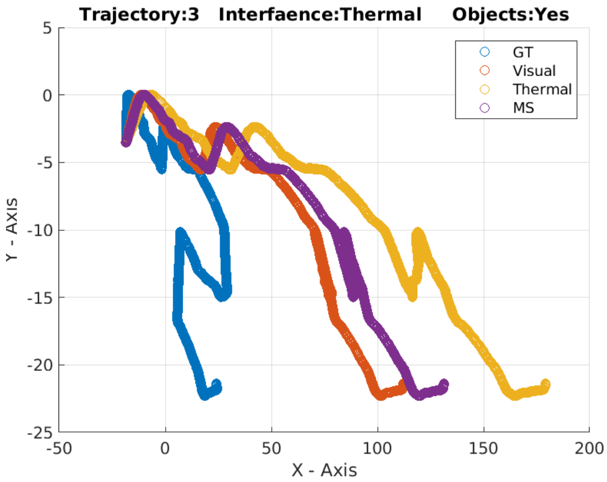

Figure 76: The experiment is run on trajectory number 3 and has interference in the thermal subsection of the electromagnetic spectrum. There were objects present during this experimentation.

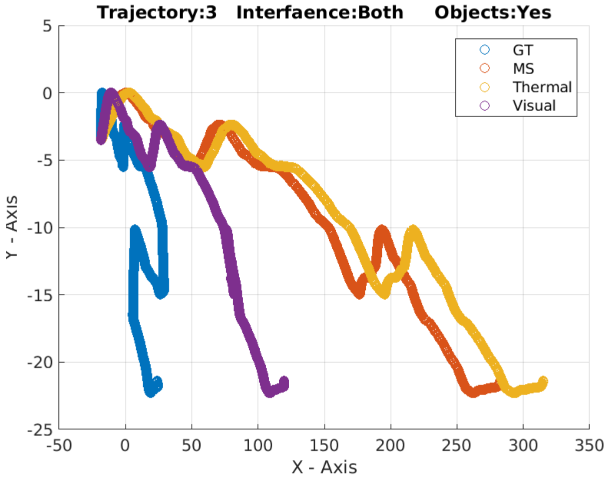

Figure 77: The experiment is run on trajectory number 3 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 77 continues to demonstrate the fact that having both interference types on at once greatly changes the results. Whilst the loss of the hook in the visual trajectory is expected, the fact that the thermal and multispectral retain the hook is quite baffling.

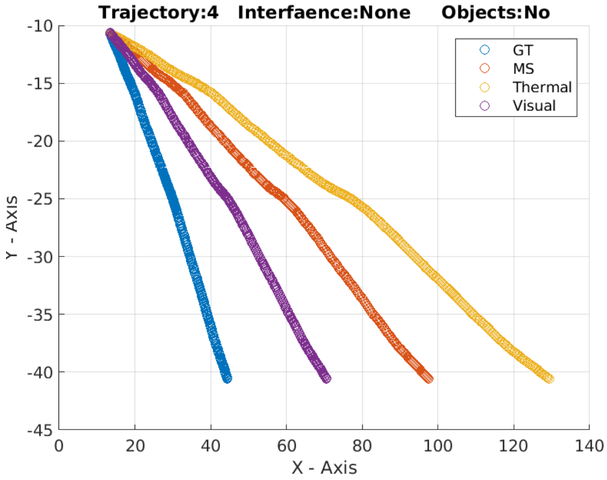

Figure 78: The experiment is run on trajectory number 4 and has interference in no subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 78 has no point of interest as the trajectory is familiar from the previous chapter.

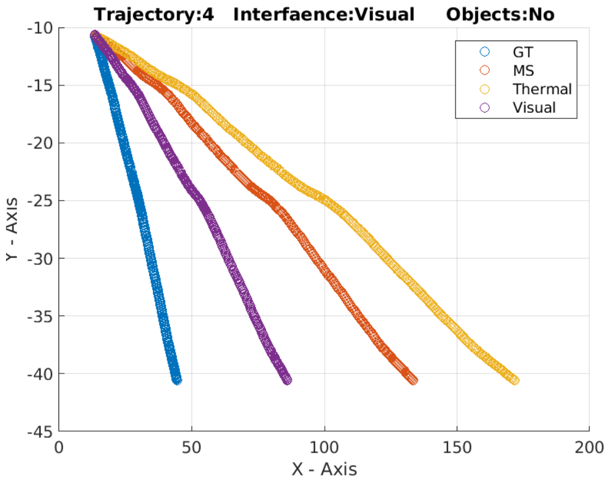

Figure 79: The experiment is run on trajectory number 4 and has interference in the visual subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 79 has little to no discernible change in the contours of the trajectories from Figure 78. However, the range of the x axis shows that the errors are far larger with the added interference.

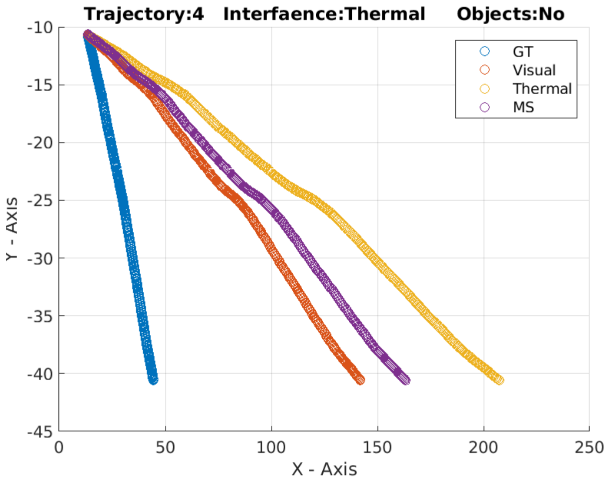

Figure 80: The experiment is run on trajectory number 4 and has interference in the thermal subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

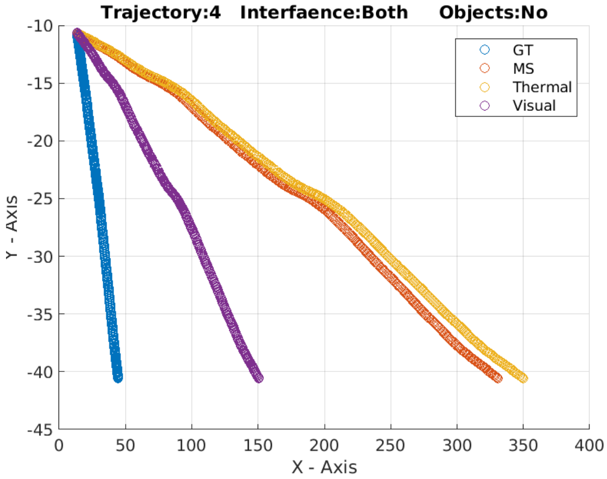

Figure 81: The experiment is run on trajectory number 4 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 81 presents no behaviour that is not apparent in previous trajectories, but the error range is substantially larger. The unique part of this image is that the error rate grows far slower than the size of the X-axis showing the thermal and multispectral being far closer than they actually are.

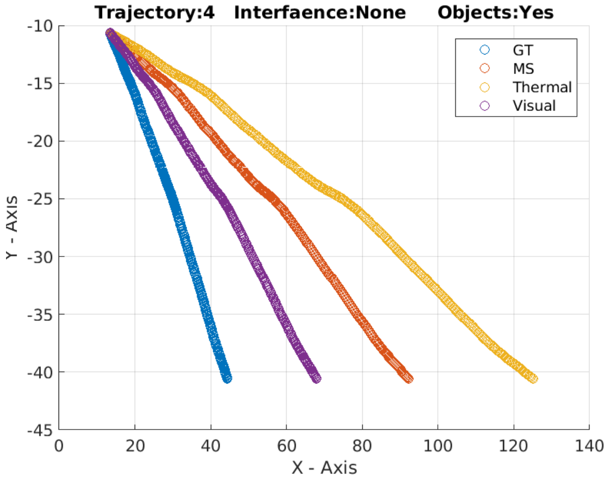

Figure 82: The experiment is run on trajectory number 4 and has interference in no subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 82 has a lower error than its counterpart, suggesting that the presence of textured objects was once again useful to the pose estimation process.

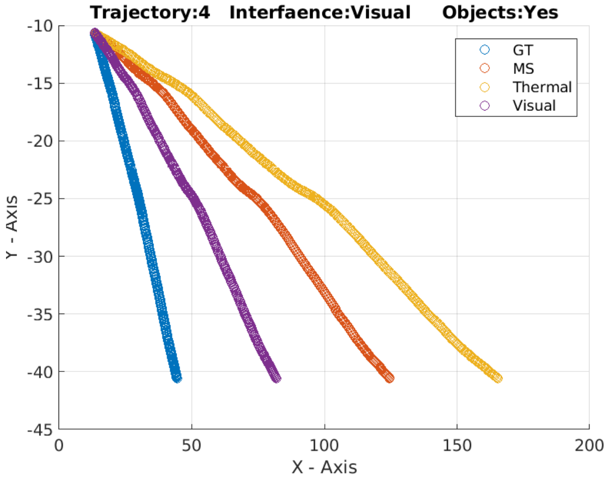

Figure 83: The experiment is run on trajectory number 4 and has interference in the visual subsection of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 83 has resulted in the ground truth appearing to be a perfectly straight line. This serves to identify the strange point of this AI method and can only truly pay their part in larger errors.

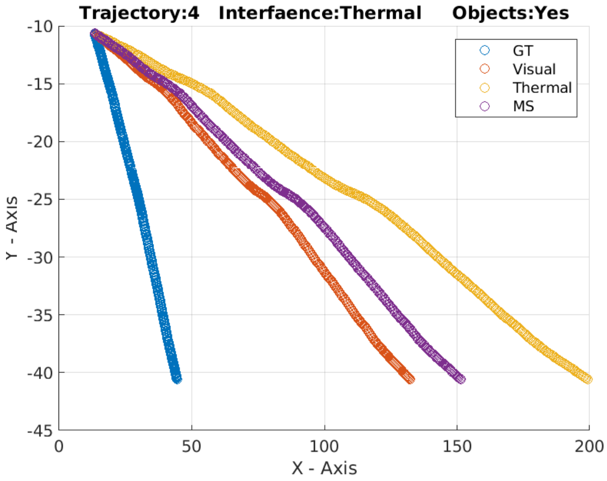

Figure 84: The experiment is run on trajectory number 4 and has interference in the thermal subsection of the electromagnetic spectrum. There were objects present during this experimentation.

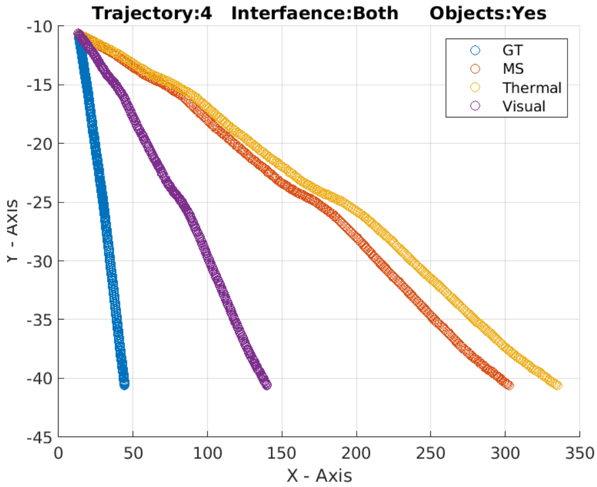

Figure 85: The experiment is run on trajectory number 4 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 85 demonstrates that it's possible to widen the gaps between the trajectory estimates when both interference types are present within the trajectory.

4.1.5 Conclusions and Future Work

It is the principle objective of this section of the thesis to convey the conclusions of the experimental results and theoretical framework present in this chapter thesis, and to outline possible future work that could enhance the work done here or provide additional details on nuance for further development of the field.

This section of the thesis has proceeded to develop a method that can allow for both visual and thermal-based image processing in order to estimate each of the agents. It does also extended this to the multi-spectral domain.

A point of view of the results suggests that the interference plays a significant role in the quality of the pose estimation. The presence of objects can greatly alter the magnitude of the errors whilst the presence of interference can directly alter the contour of the trajectory to which it applies.

Furthermore, it has become clear that the combination of various things such as the sensors heating up, the addition of the interferences, and the textureless region can lead to results which are far from expected.

Future work in this line of research may be well advised to consider the addition of tri-sensor configurations which would enable not just thermal and visual modalities but also an additional modality and construct a repetition of this work for such a case. Further, it is apparent that the combination of three unique imaging sensors would have such a great impact on the literature and perhaps even a large number of sensors should be explored. This, though, leads directly to an exhaustive search of all possible combinations of sensors and the analysis of the results for them.

It will also be of considerable interest for future work to examine how the inertial navigation system could be used as a regulatory device or be regulated by the sensors of a particular modality. Such a line of work would likely require the introduction of various forms of sampling and smoothing in order to ensure a method by which such things can be investigated is achievable.

4.2 Multispectral Error Elimination

It is the objective of this chapter of the thesis to develop upon the foundations developed from the deep multispectral inertial odometry model in such a way that further expands on the idea of abstraction in order to provide a method to optimise multispectral models. Section 4.2.1 provides a detailed account of the motivation behind the work done, and how it enables the transition of the feature matching into an automated multispectral setup by exploitation of the backpropagation algorithm that is omnipresent in the deep-learning based solutions that currently represents the sate-of-the-art in the field. Section 4.2.2 provides a detailed explanation of the step-by-step procedure employed during the course of the study, whilst the results of the experiments and the implementation are described in Section 4.2.3. Finally, Section 4.2.4 concerns the final findings of the study and the possible directions of future work.

4.2.1 Motivation

Given that data fusion methods such as the kalman filter can be employed to fuse the estimates of different navigation systems and can be converted into a similar AI based solution, it becomes natural to determine if the nature of the transition from traditional to AI based solution may be exploited to further reduce the drift present in the system.

In order to generalize this method, to all forms of AI based deep-learning solutions, it is vital to employ the use of some method or mechanism present within all AI-based solutions. Given the classical view of deep-learning architectures, it is immediately apparent that the architecture of the solutions may vary from solution to solution and thus cannot be exploited in this fashion. Similarly, the data set will vary and so should not be used in this fashion. It should also be noted that the use of the data set or any limitations upon the data set imposed by such a method would defeat the purpose of its construction. As such, the only resulting mechanism which can be exploited and is present within all AI based structures such as deep-learning networks is the backpropagation algorithm.

It is not immediately clear if the backpropagation algorithm could have been exploited to determine the method by which the reduction of multispectral drift could be reduced from the system; however, given the nature of the kernel optimization which results from the backpropagation algorithms employment, it was hoped that it may be possible to optimize the kernels in such a fashion that they converge to a single feature-detection scheme which does not degrade with modality and may be extendable to other combinations of electromagnetic wavelengths.

If it was indeed possible to exploit the backpropagation algorithm in such a fashion, then could it also be possible to exploit the found solution, if it exists, to further reduce the adaption required by the system or model to cope with external sensor changes or changes in the sense of configuration? This was likely impossible however desirable. It became clear that any solution must likely have multiple rounds of training in order to be adaptable to a change in the modality of the image-forming sensor(s).

The method by which this was accomplished is depicted in the proceeding section. The section proceeding that contains the results of the experimentation used to verify these assessments of the method. Finally, this chapter concludes with a summary of the findings resulting from the work and experimentation, in addition to a theorized set of future works that may be used to build upon the works of this chapter.

4.2.2 Methodology

The central problem that arises from the conversion of visual odometry solutions into multispectral visual odometry solutions is the introduction of multispectral drift. This is known to arise due to the features of the three-dimensional scene being mapped into different points or regions, in each of the modalities resulting in two very different two-dimensional image planes for two different modalities. It also arises due to some features only being present in a single modality. Such features are known as modality-specific artefacts. For example, due to the fact that normal optical images detect light or illumination levels within a scene, and thermal imagery is based upon the heat levels present in the scene or variation in temperature, there is often a drastic difference between optical and thermal depictions of a single three-dimensional scene. This framework addresses such artefacts by focusing on the formation of feature points present in both modalities. The construction of the framework shows that it is always possible to make such deep-learning based visual odometry solutions to focus on such formations. In order to ensure that the modality-specific artefacts of the multispectral image streams do not interfere with the investigation of the multispectral feature point-matching framework, the abstraction scheme compartmentalizes the project into a series of smaller deep-learning models. This compartmentalization enables the earlier models in the pipeline to construct an abstract representation of the two input image streams, which enables the models later in the pipeline to only have knowledge of the abstracted images. By carefully constructing this abstraction, it is possible to only consider the desired properties of the input streams during the prediction step employed in this work, which is a two-stage process. The first stage detects the locations of the sift feature points in the image play. The second stage generates a blank image with dimensions consistent with the corresponding input stream, upon which the locations of the feature points are superimposed. This ensures that the only information contained in the abstracted images is the location of the feature points in the corresponding time step. Reference [2] suggests the existence of various robustness problems inherited in this approach; as such, a second abstraction is also developed. This obstruction is identical to the first in all but one aspect - the obstruction is based on the edge map of the input image, not the location of the feature points within the image.

The central idea of this framework is to employ the stereo nature of all such multispectral deep-learning solutions through the use of abstraction of compartmentalization, to enable the already present backpropagation mechanism to optimally adapt the model to form a single modality (i.e. stereo visual clothes) to a multispectral solution, thereby removing the drift attributable to the multispectral feature-matching problem.

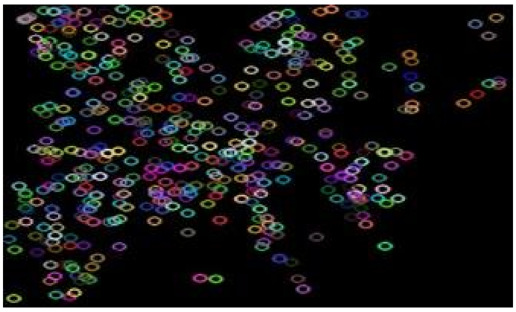

Figure 86: A illustration of the abstraction of a visible image into an encoding of the feature points in the scene.

Figure 86 presents the abstraction from a visible image to an abstraction representation of the feature points in the image. Figure 86 shows that the results of the abstraction form a modality-specific-artifact-free image.

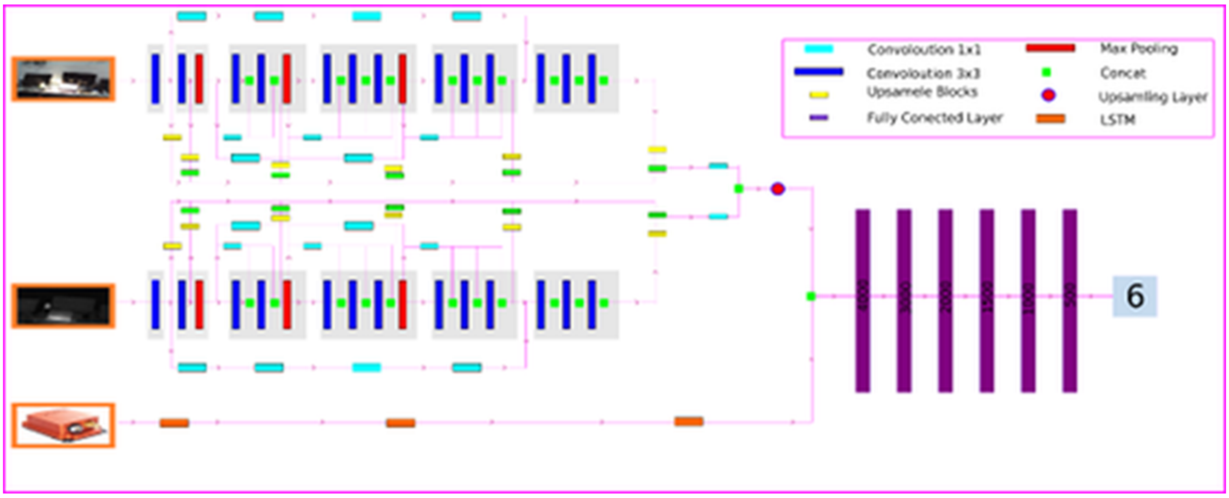

Figure 87: The complete architecture of the entire framework.

Figure 87 introduces the complete architecture of the entire framework. The framework consists of three input streams, two image streams, and one inertial input stream. The existence of the inertial data stream is not required by the framework but is employed here as regulation on the pose estimation. The generic framework consists of a prediction model - in this instance, a fully connected dense model - which leverages the abstraction generated by two identical models, known as obstruction models. Note that the two identical models refer to a single deep-learning based model architecture. As the final prediction model is always acting upon a stereo input produced by the two identical abstraction models, it proves feasible to initiate the parameters of the two abstraction models to be identical. In this manner, the modality initially acts upon a stereo inertial input consisting of two image streams of the same modality. At this point, the model is not subject to the multispectral feature-matching problem as both of the images are often of the same modality. The model is optimized through backpropagation to produce an acceptable trajectory estimate through adequate training.

As the model now works as desired in the production of a valid pose estimation of the egomotion of the agent, the entire model is subject to weight freezing except for one of the two obstruction models. The unfrozen model is then subject to a change of its input streams from the visible modality to a thermal one. The model at this point can be considered as an unoptimized multispectral visual odometry solution. The increase in the era of the model can now be directly attributed to the change in modality. The entire model is then retrained using ground truth pose measurements. Due to the frozen weights, only the thermal abstraction model has the parameters updated. The original detector networks were trained with label data as a supervised learning method, using classically generated abstracted images. The retraining of the thermal abstraction model is done through the backpropagation of the pose estimation error and not the image error.

The construction of the proposed framework into its modular components enables the employment of multiple distinct loss functions. This allowed for the initial detectors to be trained via a custom loss function and then enabled the thermal abstraction network to be updated from the custom loss function of the final prediction model. This is critical as there are currently no acceptable methods for the generation of multispectral feature point images or edge maps. The employment training scheme bypasses such concerns by retraining the thermal network. This then forces the thermal network to generate the abstracted images minimizing the pose estimation. While this framework would not produce the optimal feature points or edge detector networks in the secondary modality, it will eliminate the multispectral drift out of the model's final pose estimation. Note that this elimination is likely not optimal due to the local minima present in the backpropagation optimization step; however, it is likely far more accurate than a single multispectral network would be without its employment.

The basic obstruction model is not the subject of this work. This is in fact critical to the innovation of this work as this shows that the framework may be made to work on visual odometry solutions that will not construct a specifically for it and so is generalizable to some degree. However, as the data set is different, it proved necessary to train the model for a different number of epochs. The main adaption is the addition of an upsampling layer at the end to produce a flattened tensor of 1,500 neurons. This is done post the concatenation of the two obstructions to fuse them into a single prediction model converting feature points and edge map locations into a six-degree-of-freedom pose.

The output of the obstruction networks are flattened and concatenated into a single dense vector. This dense vector is then appended to buy the 1,000 dense node output of the LSTM stack. The LSTM stuck consists of three identical unidirectional single LSTMs with a hidden size of 1,000. The weights of these are initialised as random samples from a Gaussian with zero mean and a standard deviation of unity. Table 5 demonstrates the design of the final sub modules.

Layer	Kernel	Padding	Stride	Channels
Dense 1	-	-	-	4000
Dense 2	-	-	-	3000
Dense 3	-	-	-	2000
Dense 4	-	-	-	1500
Dense 5	-	-	-	1000
Dense 6	-	-	-	500
Dense 7	-	-	-	6

Table 5: A review of the final modules of the model.

The proposed framework is developed as a multispectral deep visual inertial odometry solution and incorporates the inertial measurement unit data to regulate the final pose estimation. In particular, the sensor is utilized to estimate the rotation rate and the acceleration. The ability of the LSTM model to surmount the problems of traditional recurrent networks such as the exploding and vanishing gradient problems is

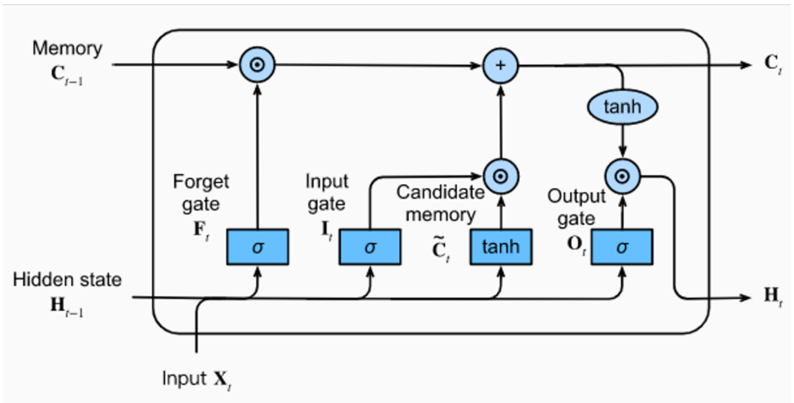

Figure 88: The internal gated layout of the LSTM.

the prime reason that it has frequently been exploited to include the inertial data into deep visual or double tree solutions. The LSTM's ability to learn the long-term dependencies is owed to its gated design. The gated network determines which sectors of the previously hidden state should be kept or discarded in the current iteration and go into the LSTM allowing you to record the previous information without alteration. This in combination with the full state of the non-gated part of the LSTM enables a model to learn long-term dependencies. In addition, LSTMs encode the motion of the agent as perceived by the inertial sensors thereby normalizing data training. This is due to the international gated nature of the LSTM in contrast to the fully connected layers in the final prediction model. This study uses three stacked unidirectional LSTMs within which the key characteristic is the hidden layer size which was optimized to be $1,000$ neurons. Figure 88 is the internal gated design of the LSTM model with X and + symbols referring to element-wise multiplication and addition respectively. The sigmoid and $tanH$ are activation functions.

Training first begins with the isolated visual abstraction network (this section of the work presents the training of the edge maps, noting that the only distinction between the training of the edge maps and the feature points is the labelled vector at this stage). The obstruction network is trained for 395 epochs over the visual images of the first two sequences of the data set which are augmented with random visible images augmented from flicker. It was decided that training the model over the whole indoor data set would hinder the validity of the test; as such, only the first two sequences were adopted. However, the length of the sequence proved not to be sufficient to train the model and the data set augmentation was adopted. This training is subject to the minimization of the following loss function:

This is a binary cross-entropy loss function that represents the labelling of each pixel J in an image I as either a point on an edge or not. This is done through the results coming from the canney edge detector via the abstracted images. This then naturally leads to the construction of two sets I_+ and I_-, which represent the points on the edge and the pixels not on an edge, respectively. The inclusion of the beta parameter ensures that the loss generated by each pixel only informs the minimization once and corresponds to the correct loss summation. The minimization of this loss ensures that each picture is in the correct set and therefore that the output is the correct edge map. The same concept can be applied to feature point abstractions.

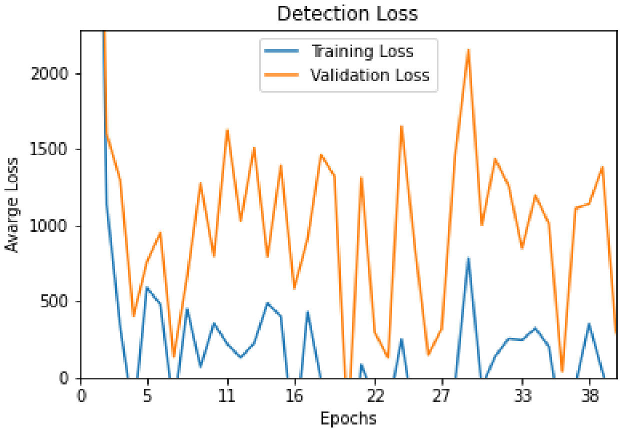

Figure 89: A graphical depiction of the validation and training loss of the model.

The abstraction network was initially trained for 397 epochs and only reached a usable state at the 391^{st} epoch, suggesting that the data set had some anomalies. The validation and training loss of the model is presented in Figure 89. The large swings in the abstraction loss optimization depicted in Figure 89 suggest that the model will benefit from the introduction of a gradient clipping to get the last closer to a typical exponential decay curve. However, this did not prevent the model from converging to a usable set of parameters. This network had some trouble learning this is due to the lack of data required by the architecture of the network. To fix this we should have had far more data with no interference added.

Figure 89 indicates that the original model is challenged with the random data set augmentation in addition to the move from a clean data set design specifically for edge detection purposes. It also demonstrates an insufficient clean up of the employed data set. Further research is required to clarify the loss proportion which is attributable to the model versus the data cleaning process.

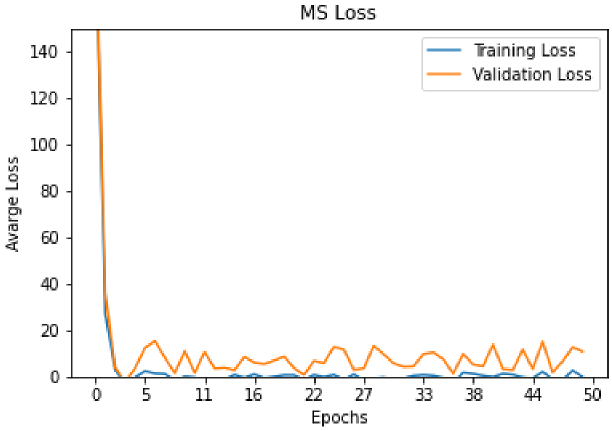

Figure 90: A graphical depiction of the training and validation loss associated with the MS loss training loop.

Figures 89 and 90 show that the training of the final prediction submodule is critical as it enables the output of the abstraction network to be optimised jointly. This optimisation resulted in a new loss level that is far closer to the ideal than the obstruction model. This demonstrates a rapid yet substantial decrease in the loss when retaining the thermal network.

The training of the entire model like the retraining of the thermal network is subject to the following loss function:

$$L = \|\hat{T} - T\|_2 + \lambda\|\hat{R} - R\|_2 \tag{55}$$

This loss function is a minimization in the L2 Norm of the total translational error T and a scale quantity of the rotational R error, with the symbol $denotingthegroundtruth.\lambda$ is included to account for the fact that rotational estimation error is the main source of drifting the solution, requiring some scaling in the loss function, the value of the parameter that proved optimal is 20.

4.2.3 Results

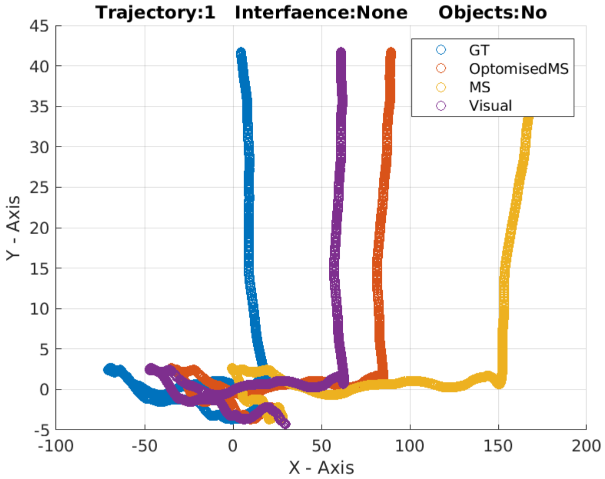

Figure 91: The experiment is run on trajectory number 1 and has interference in no subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 91 is the base case of the algorithm run on the first trajectory. It clearly demonstrates that the framework works as the optimised multispectral trajectory is closer to the ground truth than the normal multispectral trajectory. It also shows that the method may have failed to eliminate 100% of the additional feature matching error as the visual trajectory outperforms both multispectral solutions. This is attributable to the local minimum of the training loop.

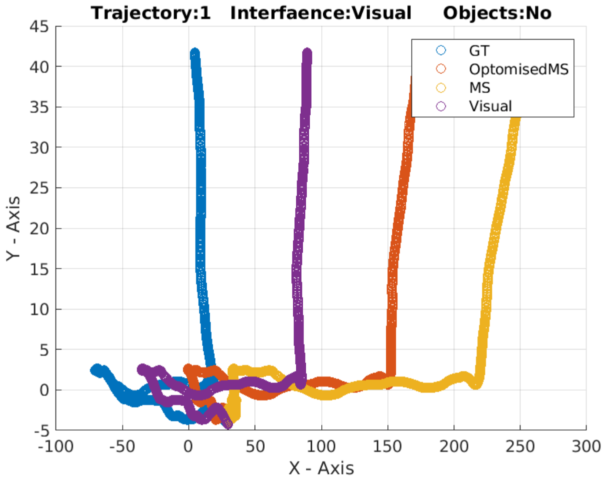

Figure 92: The experiment is run on trajectory number 1 and has interference in the visual subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 92 demonstrates that the addition of visual interference has worsened all trajectory estimates. This is the expected behaviour as all three trajectories incorporate the visible image stream.

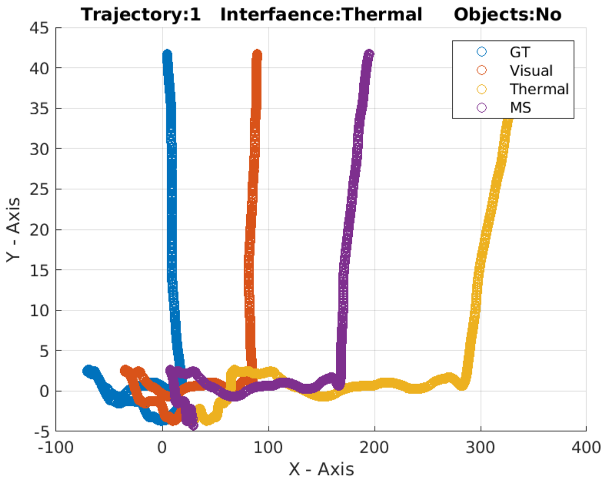

Figure 93: The experiment is run on trajectory number 1 and has interference in the thermal subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

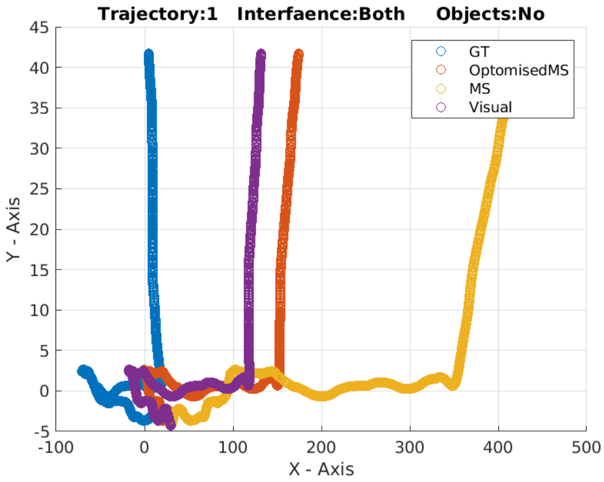

Figure 94: The experiment is run on trajectory number 1 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 94 demonstrates that the ability to remove the multispectral feature-matching error helps prevent a single modality from completely describing the error of solution when interference is present.

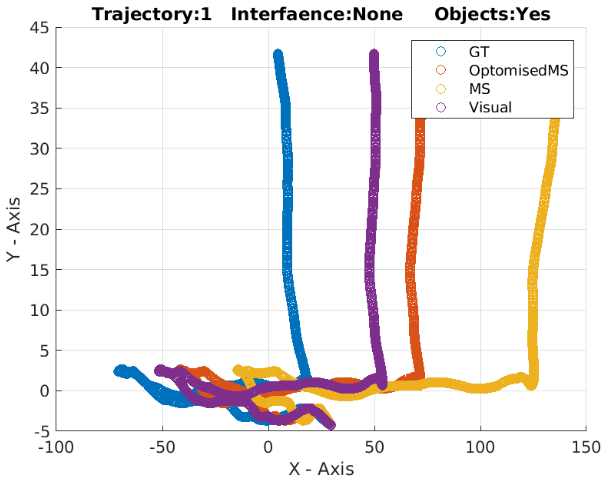

Figure 95: The experiment is run on trajectory number 1 and has interference in no subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 95 once again demonstrates that the introduction of additional texture in the scene is helpful for the optimisation of the pose estimation.

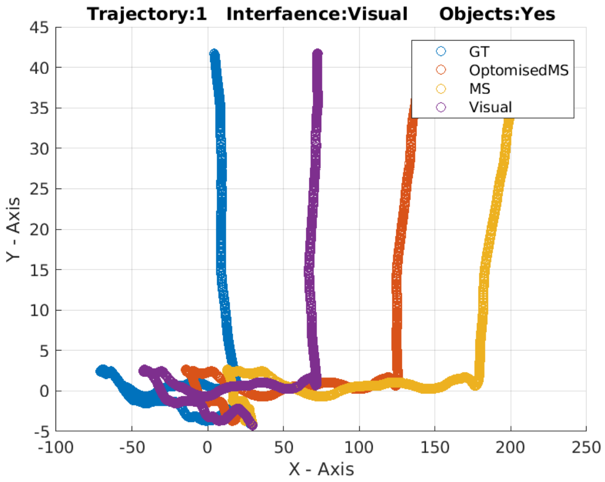

Figure 96: The experiment is run on trajectory number 1 and has interference in the visual subsection of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 96 demonstrates the fact that visual interference worsens all three trajectories, by obfuscating the better matching feature points, and the additional texture is not enough to compensate for this.

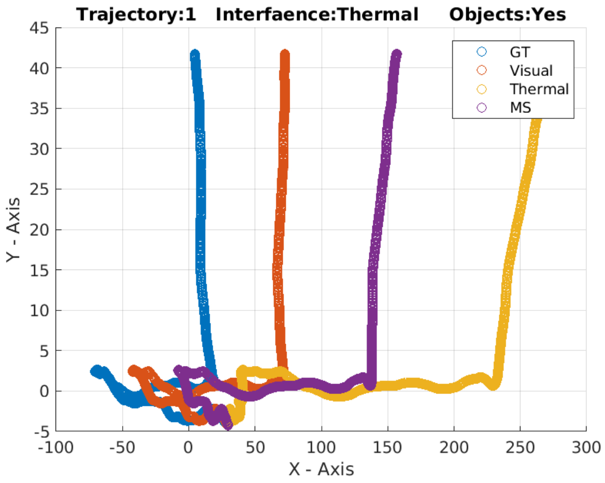

Figure 97: The experiment is run on trajectory number 1 and has interference in the thermal subsections of the electromagnetic spectrum. There were objects present during this experimentation.

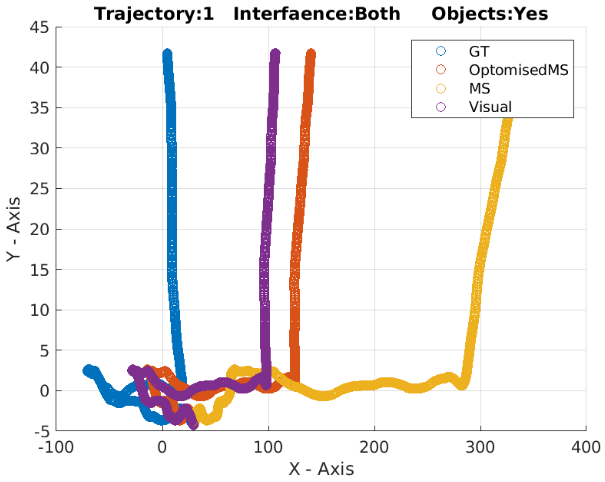

Figure 98: The experiment is run on trajectory number 1 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 98 demonstrates the fact that the thermal modality is the most susceptible to the interface, but the visual modality can be used to limit the additional error from the thermal interference.

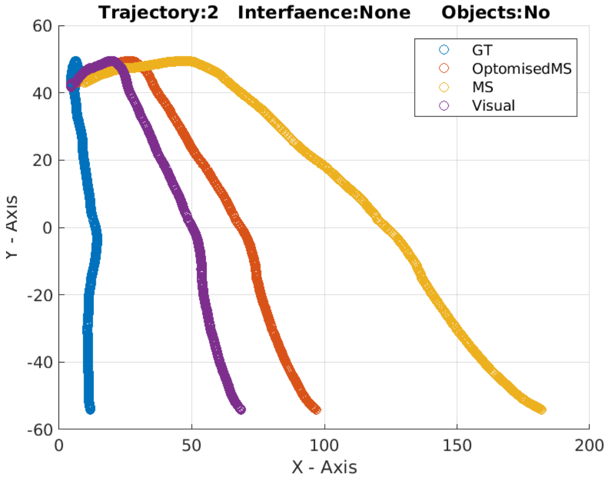

Figure 99: The experiment is run on trajectory number 2 and has interference in no subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 99 has a unique point of interest in the arc portion of the multispectral trajectory contour which seems to be brought in once optimised. This suggests that the framework can help to optimise the translation portion of the error but cannot be used to produce such insights into the rotational error.

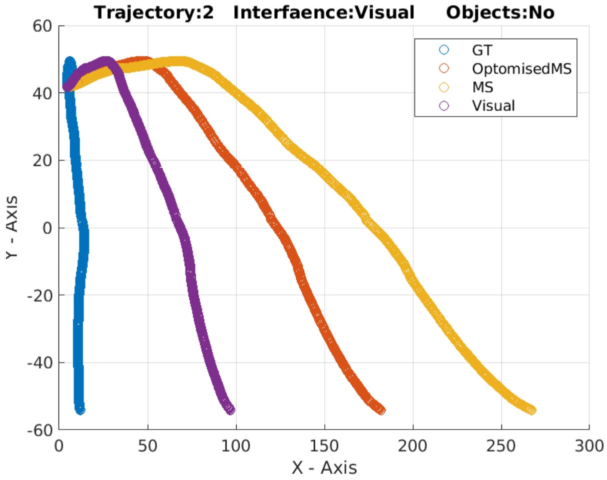

Figure 100: The experiment is run on trajectory number 2 and has interference in the visual subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 100 contributes no new insights which have not been presented by previous trajectories, and is included here solely for the purposes of completeness.

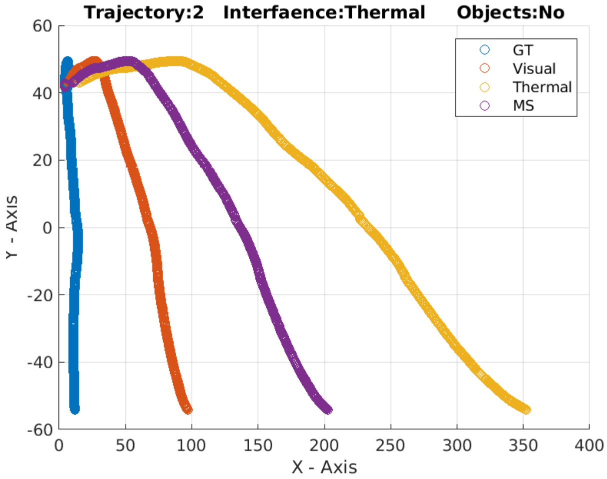

Figure 101: The experiment is run on trajectory number 2 and has interference in the thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

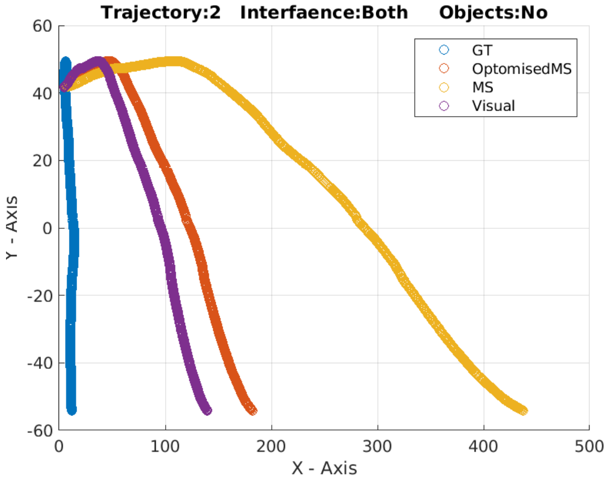

Figure 102: The experiment is run on trajectory number 2 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 102 shows a relatively small distance between the optimised multispectral solution and the visual solution in comparison to the thermal solution, highlighting once again that the thermal interference is significantly more prominent than the visual.

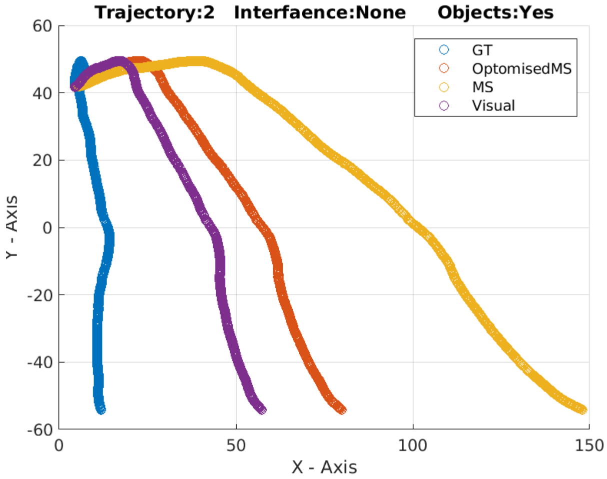

Figure 103: The experiment is run on trajectory number 2 and has interference in no subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 103 once again highlights the relatively low texture in the scene, through the heightened performance of the systems in the presence of the obstacles.

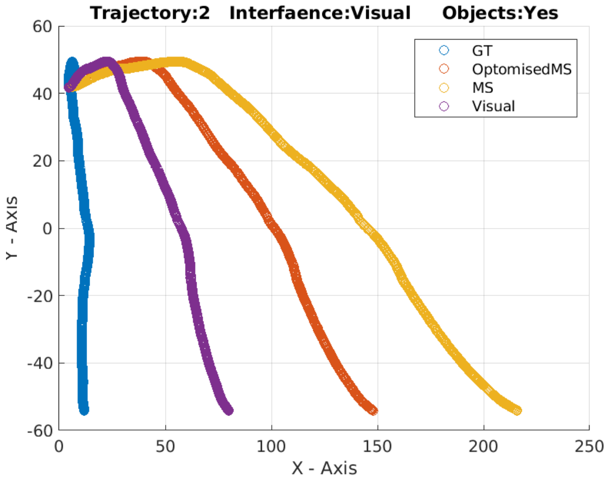

Figure 104: The experiment is run on trajectory number 2 and has interference in the visual subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 104 demonstrates the fact that the visual interface affects all the trajectories for the worse. This is shown through the lacklustre trajectory estimates.

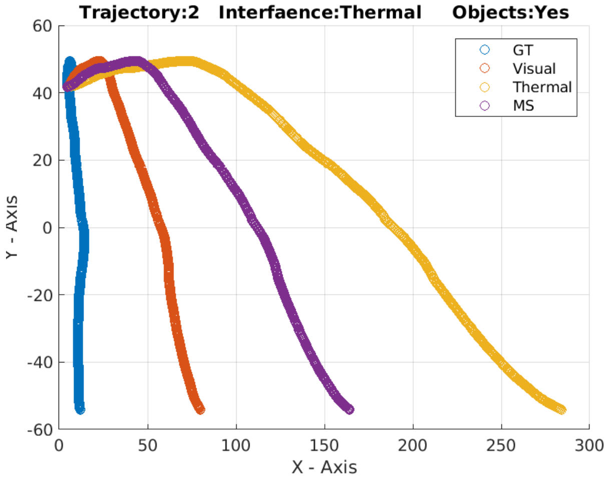

Figure 105: The experiment is run on trajectory number 2 and has interference in the thermal subsections of the electromagnetic spectrum. There were objects present during this experimentation.

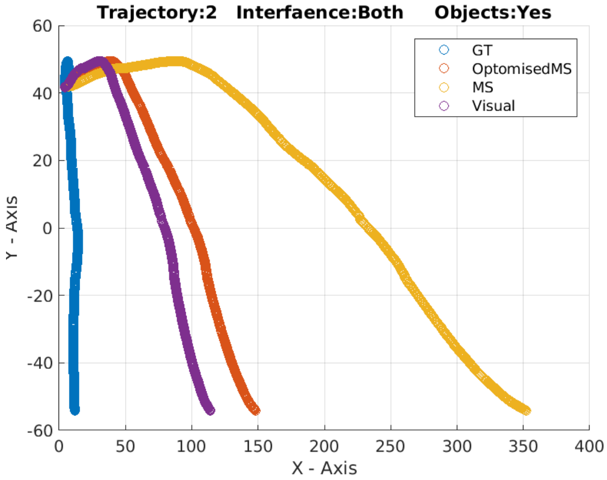

Figure 106: The experiment is run on trajectory number 2 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 106 demonstrates that the thermal modality is far more sensitive to interference than the visual modality. This may have something to do with the instrumentation of the interference and its relative effect on the image plane as measured in the number of pixels.

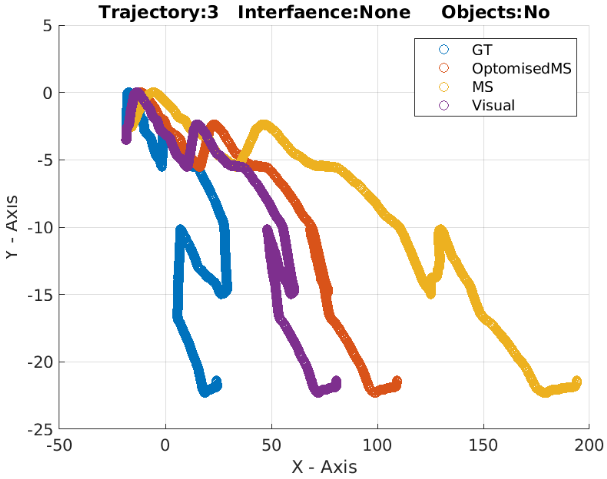

Figure 107: The experiment is run on trajectory number 3 and has interference in no subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 107 is the base case of trajectory three under this framework and highlights some rather interesting behaviour, namely the fact that the optimisation of the multispectral solution has a different contour than the other trajectories. This is likely to do with the architecture of the model and its hyperparameter optimisation during the second round of training. This may be possible to deal with through pruning.

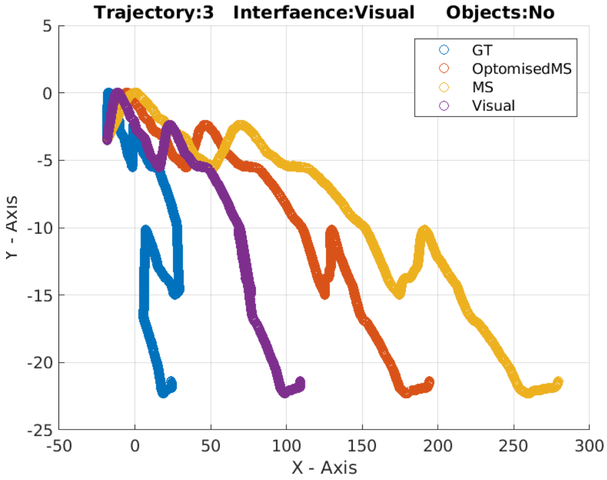

Figure 108: The experiment is run on trajectory number 3 and has interference in the visual subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 108 produces some rather interesting behaviour, namely that the visual trajectory lost the hook shape from its contour whilst the optimised multispectral trajectory gained it. The change in the visual trajectory is explainable by the introduction of the visual interference, but the change in the optimised multispectral trajectory is far more interesting. This is attributable to the kernel optimisation.

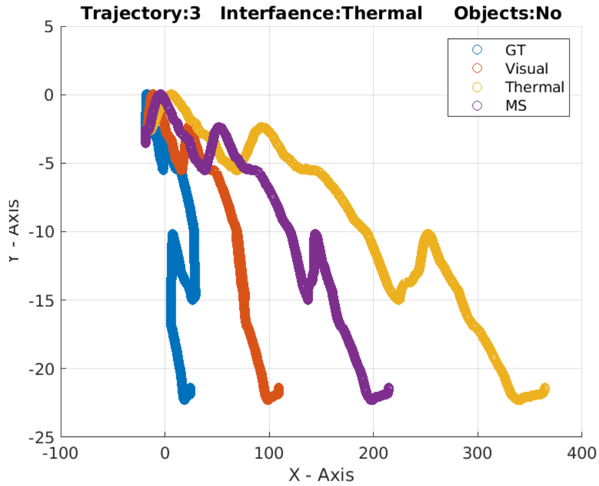

Figure 109: The experiment is run on trajectory number 3 and has interference in the thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

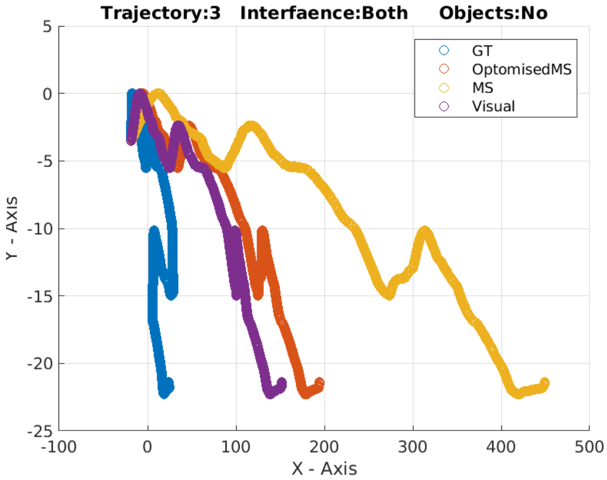

Figure 110: The experiment is run on trajectory number 3 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 110 demonstrates that the thermal interference is so much more significant than the visual interference and as a result the contour of the optimised multispectral leans further into the visual trajectory estimate.

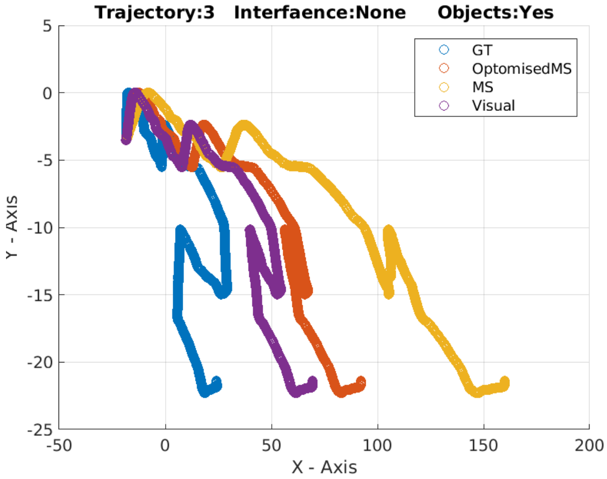

Figure 111: The experiment is run on trajectory number 3 and has interference in no subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 111 demonstrates that the presence of the obstacles improves the trajectory estimate due to the additional texture in the environment.

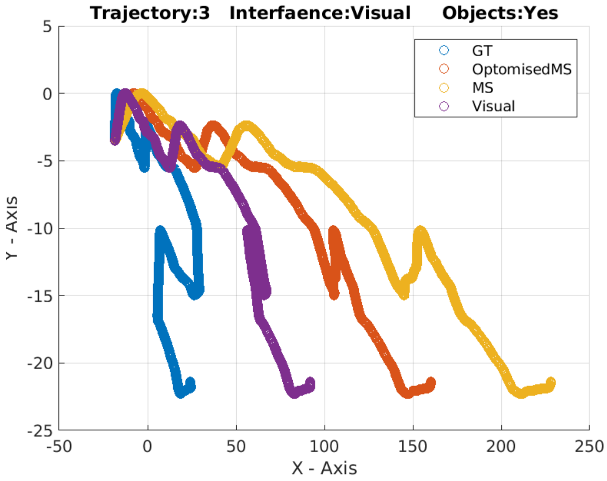

Figure 112: The experiment is run on trajectory number 3 and has interference in the visual subsection of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 112 produces little to no phenomena worthy of consideration, with the notable exception of the fact that the optimised multispectral trajectory is at one point out of phase with the thermal and has an enlarged version of the loop from the visual trajectory. This is a quirk of the optimisation process that is present in the use of deep-learning models, but that would not be possible in the EKF as the kalman gain K would optimise for a point somewhere between the two trajectories. Noting that this assumes two single modality systems were fused together with an EKF, or similar such filter, which is not the case in this section of the thesis.

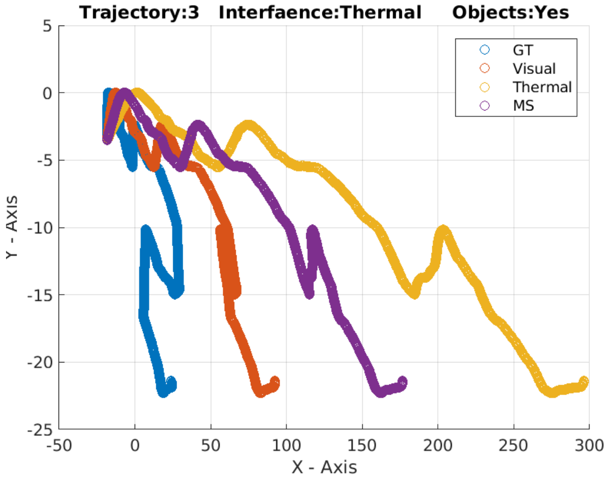

Figure 113: The experiment is run on trajectory number 3 and has interference in thermal subsection of the electromagnetic spectrum. There were objects present during this experimentation.

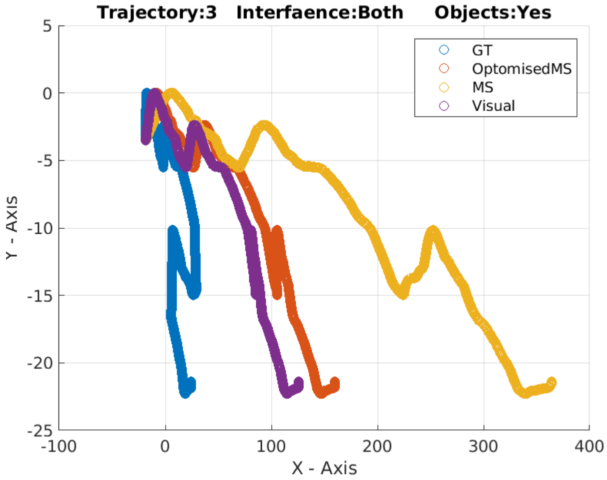

Figure 114: The experiment is run on trajectory number 3 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 114 shows the expected behaviour with the notable exception of the thermal interfaces causing such a large error in the multispectral solution and elongating its trajectory to the point that the three peaks are now out of phase.

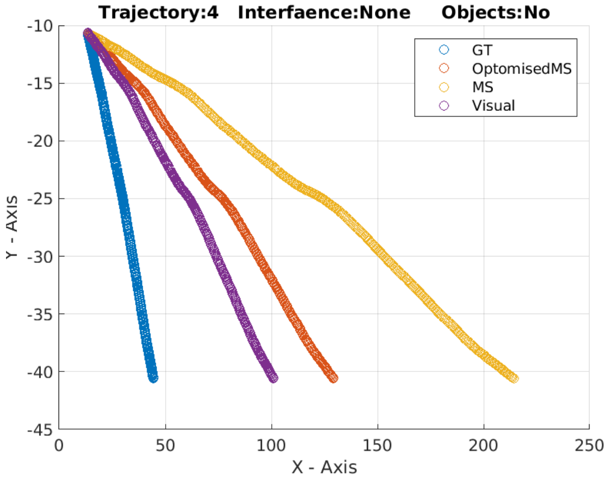

Figure 115: The experiment is run on trajectory number 4 and has interference in no subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 115 is the base case of trajectory four under this section of the thesis. Unfortunately, it has no behaviour of interest.

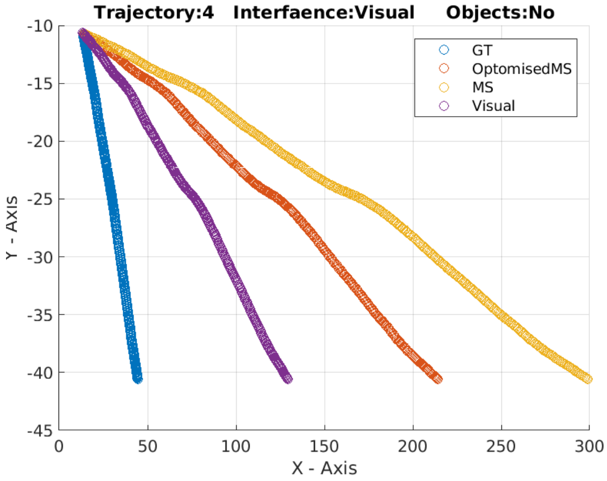

Figure 116: The experiment is run on trajectory number 4 and has interference in the visual subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 116 depicts a disproportional increase in the error of the multispectral trajectories due to the visual interference. This suggests that the presence of the visual interference makes it much harder to match feature points between modalities.

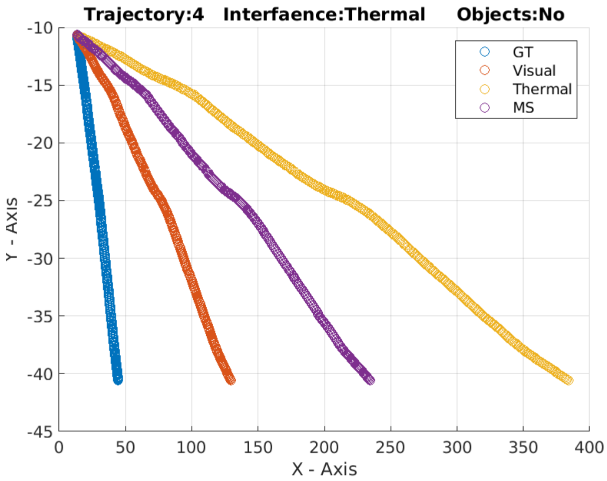

Figure 117: The experiment is run on trajectory number 4 and has interference in the thermal subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

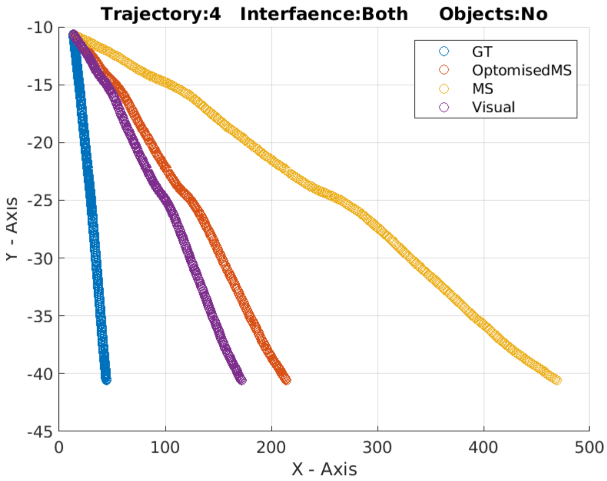

Figure 118: The experiment is run on trajectory number 4 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 118 is evidence for the fact that the optimised kernels are better at matching across modalities in the presence of visual interference than they are in the presence of thermal interference.

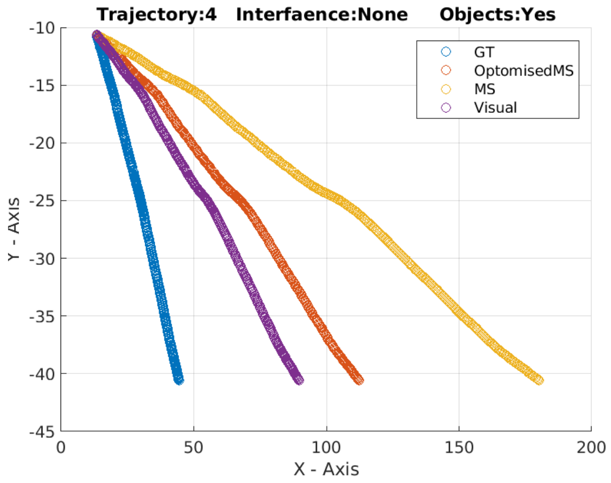

Figure 119: The experiment is run on trajectory number 4 and has interference in no subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 119 demonstrates the fact that the introduction of the obstacles has increased the performance of the three systems, which is evidence for the lack of texture in the scene.

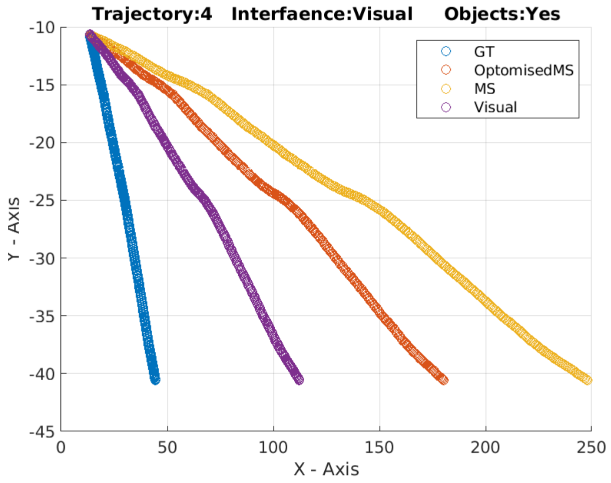

Figure 120: The experiment is run on trajectory number 4 and has interference in the visual subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 120 shows little phenomena of interest. This is due to the fact that much of this phenomenon has been depicted by the previous trajectories.

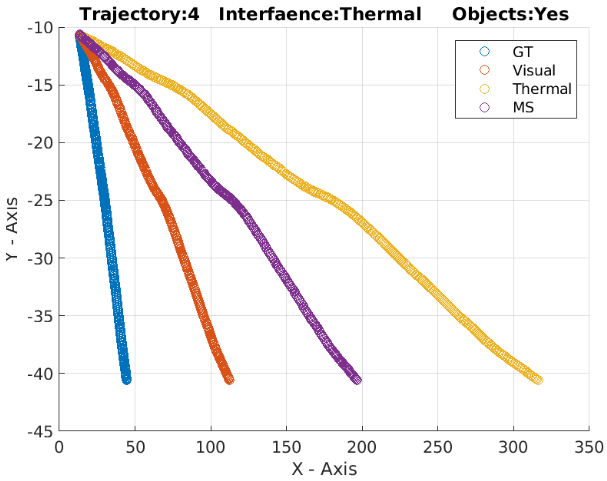

Figure 121: The experiment is run on trajectory number 4 and has interference in the thermal subsection of the electromagnetic spectrum. There were objects present during this experimentation.

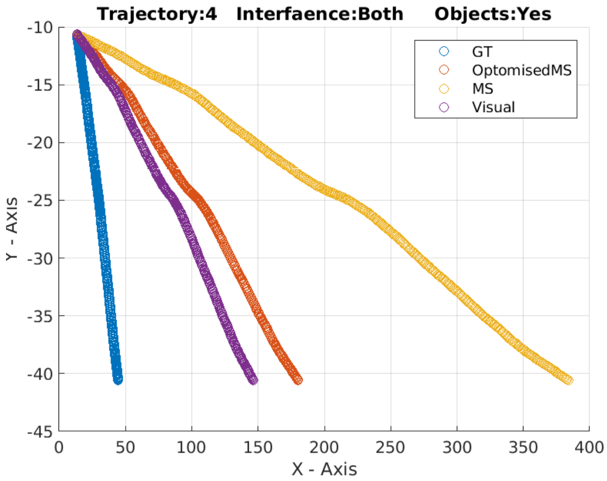

Figure 122: The experiment is run on trajectory number 4 and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 122 depicts a substantial increase in the multispectral error due to the presence of both types of interference although the presence of the obstacles helps minimise the error with the addition of extra texture.

4.2.4 Conclusions and Future Work

It is the principle objective of this section of the thesis to convey to the audience the conclusions of the experimental results and theoretical framework present in this chapter thesis and to outline possible future work that could enhance the work done here or provide additional detail on nuance for further development of the field.

The work in this section of the thesis has developed a generalizable framework by which it is possible to optimize the kernels of a pre-existing neural network architecture compliant with the framework by exploitation of the pre-existing backpropagation algorithm to minimize the drift resulting from matching feature points between modalities. The results of this section have proven that the framework needs to be adapted to incorporate procedures by which to prevent local minimization of the objective function as this can produce rather staggering results, leaving behind much of the drift that the framework attempts to eliminate.

Furthermore, some of the more exotic behaviours expressed by the experimental results suggest that the framework can be enhanced with the addition of differentiable search and pruning of the various deep-learning models post-optimization. This would remove noise-enhancing neurons, resulting in a better pose estimation but would also result in a highly sensitive model.

The current existing model of the framework has been empirically shown to be really sensitive to the introduction of interference and the texture within the scene. Whilst this is also true of existing none AI base solutions, it seems that each round of optimization increases the sensitivity of the resultant model to the interference. For this reason, it is suggested that some work be developed to construct novel methods of enhancing the robustness of the AI-based neural network. Of particular interest is the use of generator and discriminator-based algorithms such as those exploited within GANs.

To conclude the body of work developed in this chapter has enabled the construction of a generalizable framework which can minimize the multimodality feature point matching error and has proven its technical feasibility. This body of work requires further adaptation before it can produce state-of-the-art results.

5 AutoEncoder-Based Multispectral Visual Odometry

It is the principle objective of this chapter to introduce the audience to the work conducted during the autoencoder-based experimentation of this thesis and the results that it has produced. The purpose of this chapter of the thesis is to attempt to construct the ability to employ a monocular visual odometry process on multiple modalities of images in a simultaneous fashion to limit the computational requirements of visual odometry algorithms when applied to multispectral solutions. As it is known that the employment of stereo-visual odometry will introduce modality-specific artefacts that strictly increase the drift of the system, this chapter attempts to encode the key features of multiple modalities into a single latent space. This is done through the employment of various artificial intelligence methods. Due to the novelty of this approach, it is not known what the ideal latent space is and how it should be exploited, in addition to this it is not known how to generate such a latent space but given the ability of deep learning models to learn any mathematical relationship this chapter exploits deep learning based latent space generation solutions.

5.1 Large Code Based Odometry

5.1.1 Motivation

It is the principle objective of this section of the thesis to introduce the audience to the motivation of the author in the development of the following body of work. It should be noted that the motivations for this work relate chiefly in the matters identified in previous sections of this thesis.

The previous sections of this thesis have uncovered several shortcomings with the current algorithms and deep learning architectures, some of which will force asymptotic lower bounds on the drift of the system which far exceed the onset of drift in the system without these problems. The first clear example of such a problem is the inability to guarantee that the same feature points exist in images depicting the same scene which have been captured with different sensory apparatuses that are reliant upon a different section of the electromagnetic spectrum. Whilst this has been the key topic of research for much of this thesis, an idea that has yet to be examined is the additional complexity associated with solving this problem.

The next problem which arises from the additional computational complexity cost is the size of the hardware required. The additional complexity makes it difficult to utilize miniature processing chips which could be attached to micro aerial vehicles. Given the trend of monotonically increasing computational power in smaller devices, this problem will likely stop existing time. The monetary cost version of this problem would persist indefinitely, irrespective of the existence of the computational complexity problem.

Another problem that arises quite naturally from the employment of multiple sensors is the inability to utilize monocular visual odometry techniques. Monocular visual odometry naturally surfers from scale ambiguity, however, there exist various solutions to this including several artificial intelligence models, thus it may be desirable to employ artificial intelligence-based solutions designed for monocular use cases upon multispectrum solutions.

This list of principle concerns led to the motivation for a single solution that rendered all these problems obsolete. It is the primary objective of this line of research to motivate a possible solution to these problems. The results of which are quite promising.

5.1.2 Methodology

It is the principle objective of this section of the thesis to describe to the audience in detail the methodology associated with the instrumentation of the experimentation corresponding to the work undergone to examine the feasibility of autoencoder-based visual odometry.

Given that there are many objectives of this particular line of research, the methodology first had to consider doing each individually and then if it was possible to combine the resultant benefits into a single solution. The idea of pursuing each objective independently of the rest quickly became obsolete, this was due to the fact that some sort of optimal latent space representation of the data which encodes both the thermal and visual imagery, would likely result in lower computational requirements and allow it to be

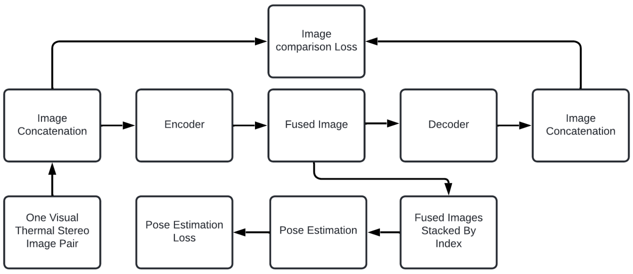

Figure 123: A graphical depiction of the overall design of the model.

utilized by monocular micro aerial vehicles.

This idea quickly presented many problems, the greatest technical problem came from analysing the principal component analysis protocol. Due to the fact that the principal component analysis requires the construction of mutually orthogonal basis vectors to form a vector space in n-dimensions, it was difficult to imagine how this would directly result in a two-dimensional image. The second problem arose from the dilemma which follows from the first, if a direct solution to transform the output of the principal components into an image befitting the scope of this thesis was not practical to formulate, then could an indirect solution possibly be feasible? It became apparent that an indirect solution to this problem would substantially increase the computational complexity of the solution and therefore may encumber micro aerial vehicles from utilizing the solution, thus an alternative method of encoding is required. Many alternative forms of latent space encoding were examined however either the mathematical underpinning of the formulations or the resultant output disqualified them from further consideration.

Finally, the idea of encoding images of two modalities into a single image hyper-optimized for the problem of pose estimation led to the employment of autoencoders. The idea became to concatenate and transform the two images - of different modalities - into the required input for a pre-existing autoencoder model, then the model would generate an encoding before decoding it into an approximation of the original input. Then the loss function could be selected to minimize the loss between the inputs and outputs of the model resulting in a lower dimensional tensor containing the core information encoded within the concatenated images. Then removing the latter part of the autoencoder and replacing it with the final part of the DeepVIO model may facilitate a second round of back propagation-based optimization. The second round of optimization would result in three components:

- An encoder network that constructs the latent space representation;
- The latent space encoding;
- The decoder network that converts the latent space encoding into a pose estimation.

Figure 123 provides a graphical depiction of the workflow of the system, from which it is trivial to identify that the application of this system may be exported to a wide range of autoencoder models. It is due to this that the success of the approach is likely to monotonically increase as a function of time.

Due to the limited scope of this thesis, the Alexnet and Inception networks were used as the backbone of the autoencoder. Both models had significant changes made to their architecture in order to better fit the research, the last two layers of each model were replaced with the final layers of the deepVIO model. The final layers of the deepVIO model were previously specified within this thesis, however, for the purposes of clarity the required table has been reproduced in Table 6.

Layer	Kernel	Padding	Stride	Channels
Dense 1	-	-	-	4000
Dense 2	-	-	-	3000
Dense 3	-	-	-	2000
Dense 4	-	-	-	1500
Dense 5	-	-	-	1000
Dense 6	-	-	-	500
Dense 7	-	-	-	6

Table 6: A review of the final modules employed in the decoder networks of both autoencoders.

Figure 124 provides a tensor depiction of the Alexnet-inspired model. This model does not use the entire Alexnet, but opts to take the main convolutional kernel structure and employ it to form the neck of the autoencoder. This is due to concerns with the full Alexnet model that would have prevented its use as an autoencoder. The first such concern is the use of the dense layers at the end. Due to the symmetric nature of the autoencoder, it would be impossible to have a single dense layer utilised as the code. The number of layers in the model was also a concern as it could have led to the introduction of the vanishing gradient problem. This is alleviated in the autoencoder as both the encoder and decoder networks can be optimised as stand-alone modules post initially being trained together with a random weight initialisation. This in effect allows both models to have half the effective length of the AlexNet model, preventing the bashing gradient.

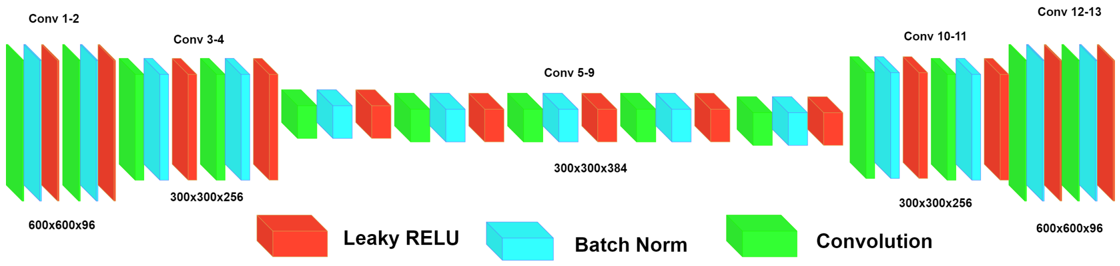

Figure 124: A tensor representation of the AlexNet-inspired autoencoder.

The architecture of the network is simply a series of convolutional layers based on the dimensions and stride values of the original paper [184]. Each of the layers employs batch normalisation and the leaky relu activation function. The use of batch normalisation is employed to improve the training time of the network, whilst the leaky relu function diminishes the probability of the onset of the vanishing gradient problem. The central portion of the network is simply a dense layer which acts as the code portion of the autoencoder network.

Both models were initially trained by taking the SSIM [185] error between the input tensor and the output of the decoder module. By ensuring that the tensors are trained against each other, it was possible to ensure that the code is a latent space representation of the input tensor. This itself is problematic as the latent space may encode the identity function, this is prevented by the dimensionality of the encoding.

The Inception network is renowned for the vast number of improvements it made to Alexnet [186]. This included the widening of the layers, due to the fact that not all the image information was being propagated into the later stages of the model, the Inception network employed the use of a series of filters with kernels of varying size inside each Inception block and a max pooling layer. The resultant output of these layers were combined into a single output that was then fed to the later layers. Another significant development is the employment of small kernel filters to significantly reduce the computational cost of the model. An example of this is a 1 or 3 kernel before a convolution kernel. Notably, this type of convolution occurs after a max pooling not immediately before. Figure 125 depicts the width of the model Inception blocks.

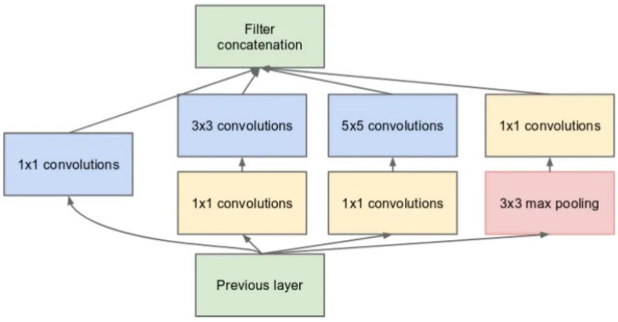

Figure 125: A graphical depiction of the multi-scale blocks of the Inception network.

The width of the Inception network portion is given by Figure 125 but the architecture of the encoder portion is given by Figure 126. The decoder is just the encoder reversed.

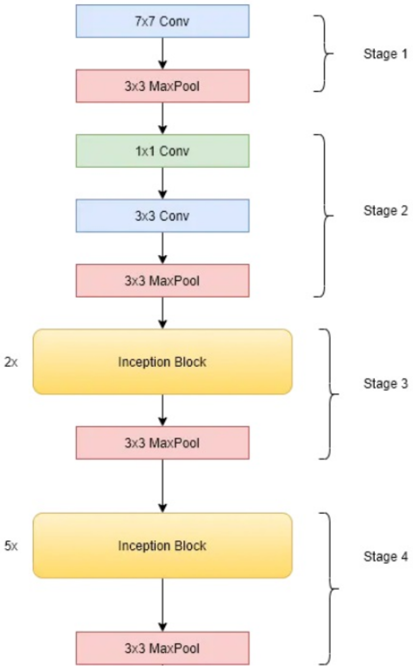

Figure 126: A graphical depiction of the pipeline of the Inception portion of the autoencoder network.

Both models were subject to the same loss function when optimizing for the trajectory estimation. This lost function is the same function that was used to optimize some of the previous models in this thesis. It was utilized as it is known to work in such scenarios. In the following loss function, the optimized value of the λ hyperparameter is 15.

$$L = \|\hat{T} - T\|_2 + \lambda\|\hat{R} - R\|_2 \tag{56}$$

The training of both models had very similar outputs in their respective loss optimizations. This is evidenced by figure 127 and figure 128 which depict the loss for the two models respectively. It can be seen that neither of the two models was truly the best choice for this type of experimentation as both loss functions show some sign of being coerced or forced into a local minimum that was taken to be practically acceptable. The lack of true convergence, as is typically produced in such graphs, highlights this fact and suggests that further research should be aimed at selecting better base architectures to work with or utilizing more mathematically sophisticated loss functions. These networks do not appear to be learning, to counter this future work should attempt to better normalise the input and attempt to do a different standardisation of the input data.

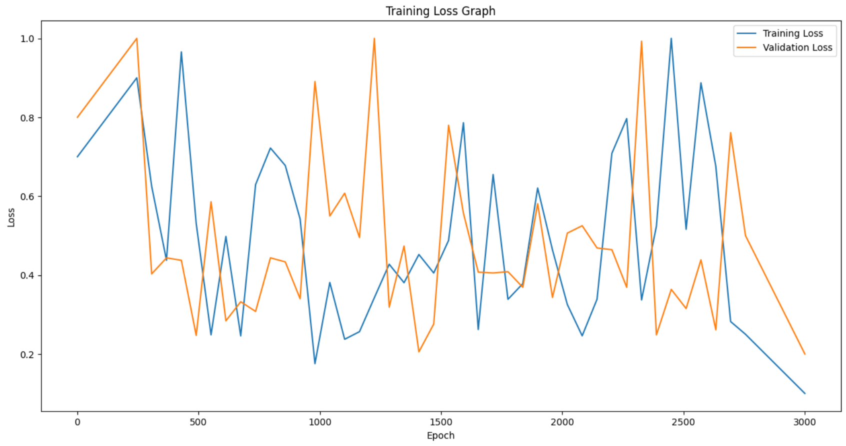

Figure 127: This figure is a graphical depiction of the evolution of the training loss function and the validation loss function of the Alexnet deep learning model.

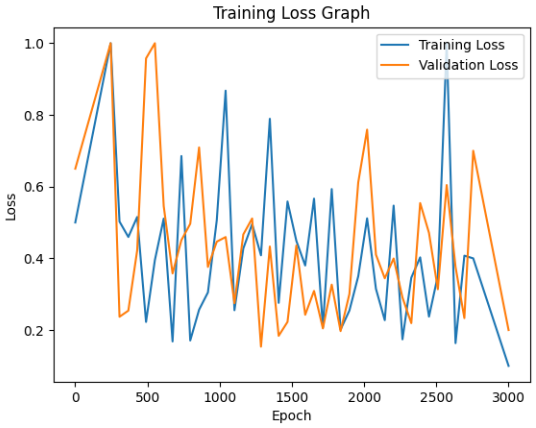

Figure 128: This figure is a graphical depiction of the evolution of the training loss function and the validation loss function of the Inception network deep learning model.

5.1.3 Results

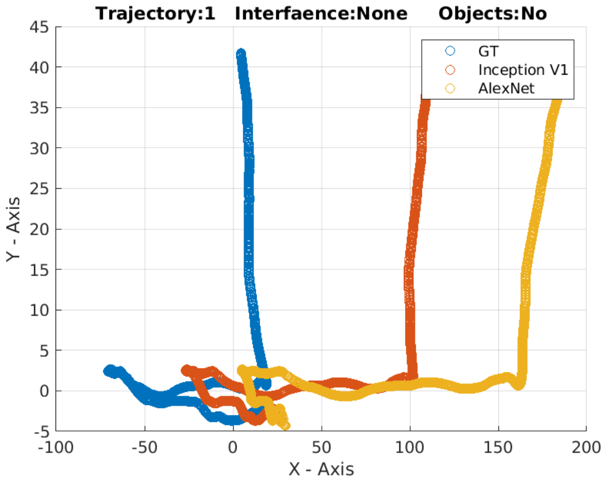

Figure 129: The experiment is run on the 1 trajectory and has interference in no subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 129 is the base case for the testing of the autoencoder based models. It demonstrates a substantially worse rotation and transnational error than any other model in this thesis. This is likely due to the incompatibility between the autoencoder structure and the final layers employed to extract the pose estimation. As demonstrated by Figure 128 the training of the final model is not typically demonstrating the need for a better architecture.

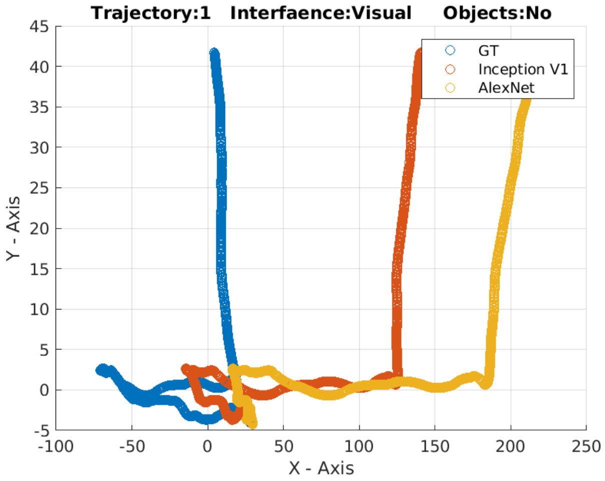

Figure 130: The experiment is run on the 1 trajectory and has interference in the visual subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 130 is illustrative of the fact that both models do not adapt to the introduction of the visual interference. This suggests that the interference in the latent space is far more effective than in the normal image plane. This suggests that it may be possible to improve this.

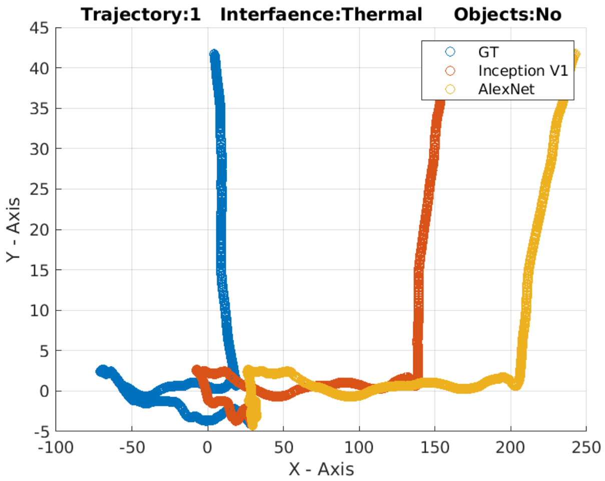

Figure 131: The experiment is run on the 1 trajectory and has interference in the thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

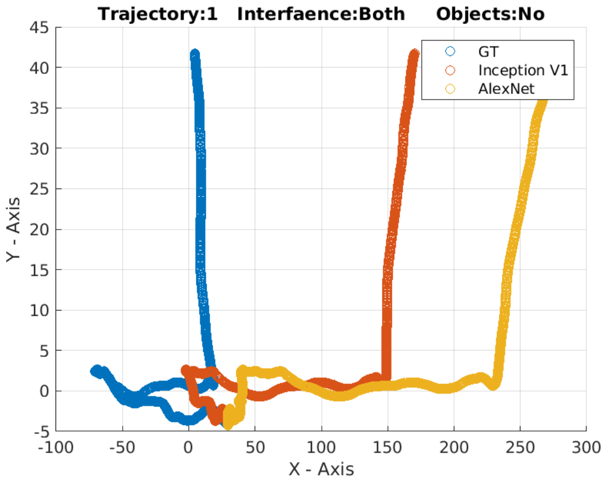

Figure 132: The experiment is run on the 1 trajectory and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 132 clearly shows that the employment of both interference sources is far more impactful on the autoencoder based solution than the deepVIO solution. This may be correctable by adjusting the shape of the latent space.

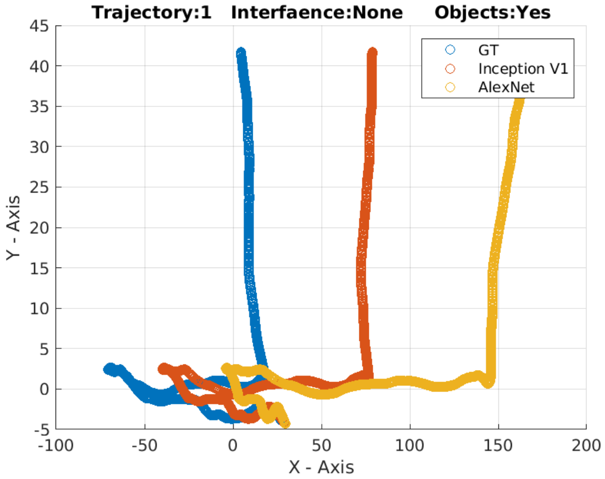

Figure 133: The experiment is run on the 1 trajectory and has interference in no subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 133 once again shows that the employment of enhanced texture in the scene improves the performance of the system. This suggests that texture is mapped into the latent space well.

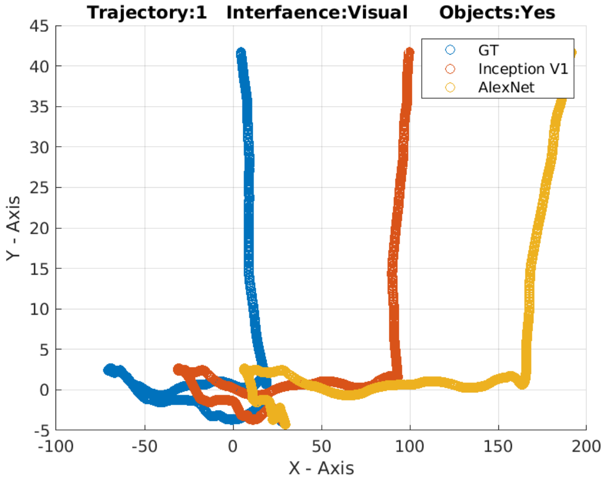

Figure 134: The experiment is run on the 1 trajectory and has interference in the visual subsection of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 134 demonstrates that the introduction of the enhanced texture in the scene improves the performance of the solution, however, the absolute improvement that it makes is far less than the improvement caused by the introduction of the texture enchancing obstacles in the DeepVIO model.

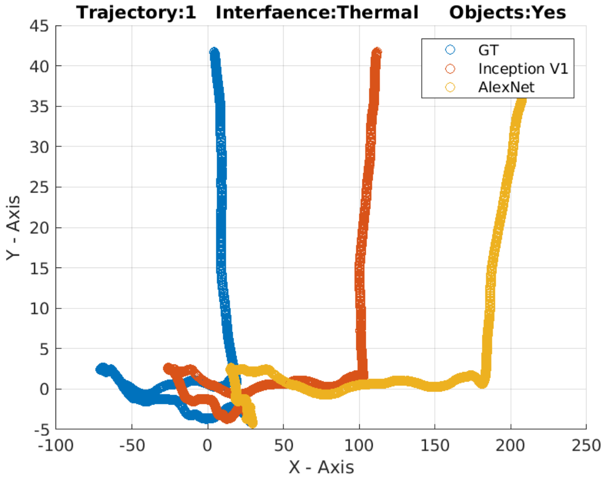

Figure 135: The experiment is run on the 1 trajectory and has interference in the thermal subsection of the electromagnetic spectrum. There were objects present during this experimentation.

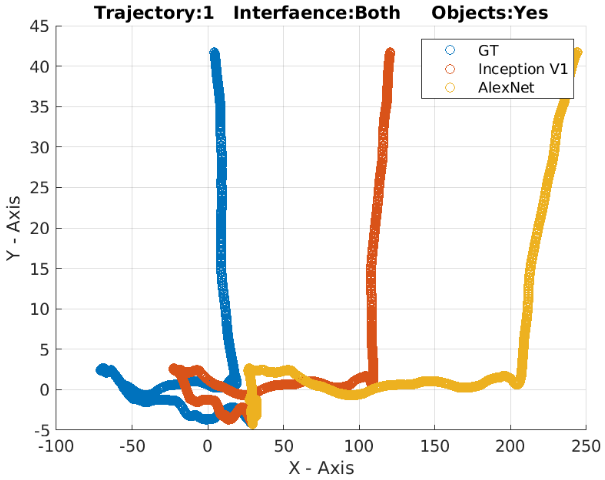

Figure 136: The experiment is run on the 1 trajectory and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 136 once again demonstrates that the employment of the texture enhancing obstacles is more effective on the latent space method than the direct pose estimation methods. It is unclear if this is due to the suboptimal training of the models or the structure of the latent space this represents a novel avenue of further study.

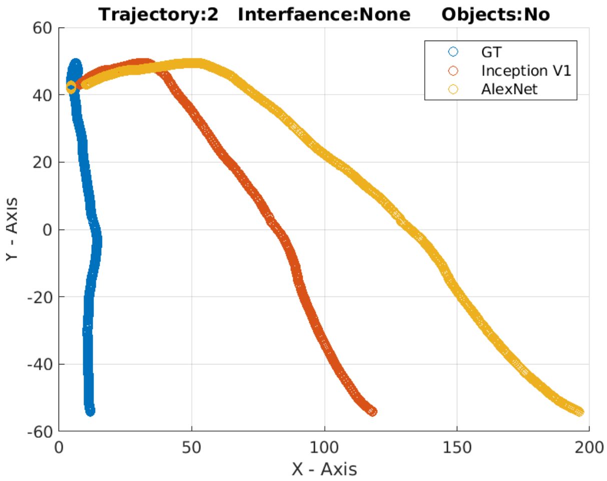

Figure 137: The experiment is run on the 2 trajectory and has interference in no subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 137 once again shows that the rotational error at the offset of both models' estimated trajectory is extremely large. This is similar to all the previous solutions developed in this thesis as such it is evidence that the environment at this point of the trajectory is desperately lacking in feature points that could accurately demonstrate a rotational estimate.

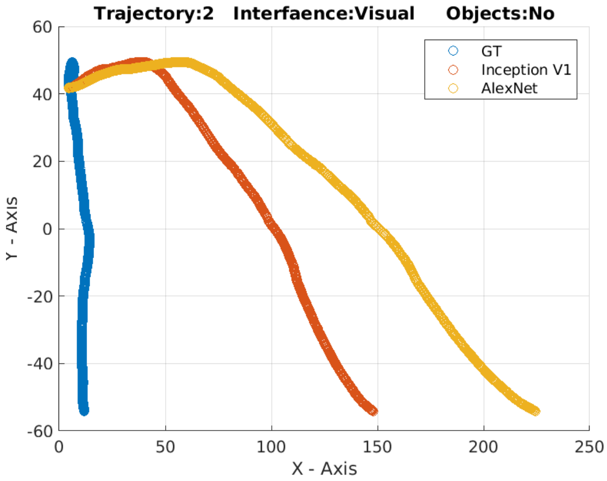

Figure 138: The experiment is run on the 2 trajectory and has interference in the visual subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 138 demonstrates that the addition of visual interference negatively affects the trajectory estimates of both models. The interesting thing is that this interference, in the absence of obstacles, affects each model on this trajectory approximately equally.

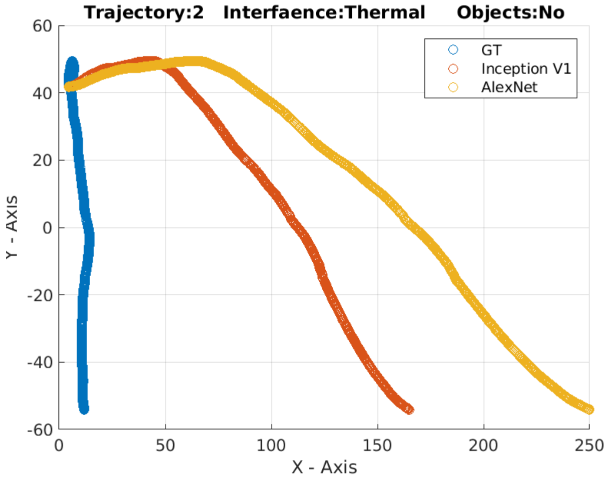

Figure 139: The experiment is run on the 2 trajectory and has interference in the thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

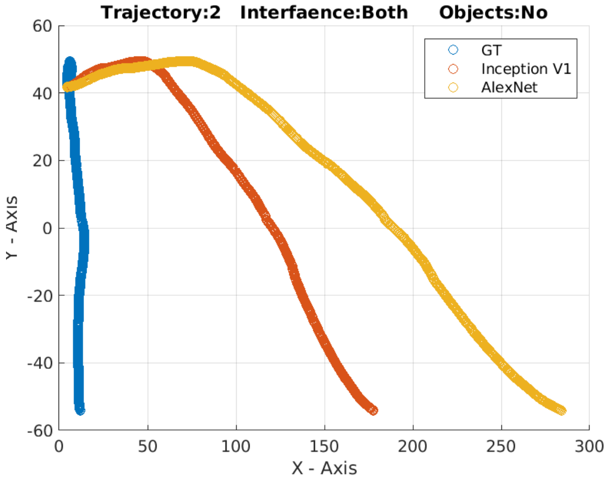

Figure 140: The experiment is run on the 2 trajectory and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 140 is illustrative of the fact that the introduction of both interferences at the same time has a far larger impact on the pose estimation than the employment of a single interference type. Whilst this is common to many of the novel solutions developed during the course of this thesis, it is not as apparent as it is upon the autoencoder models. This seems to be due to the mapping of the images into the reference space.

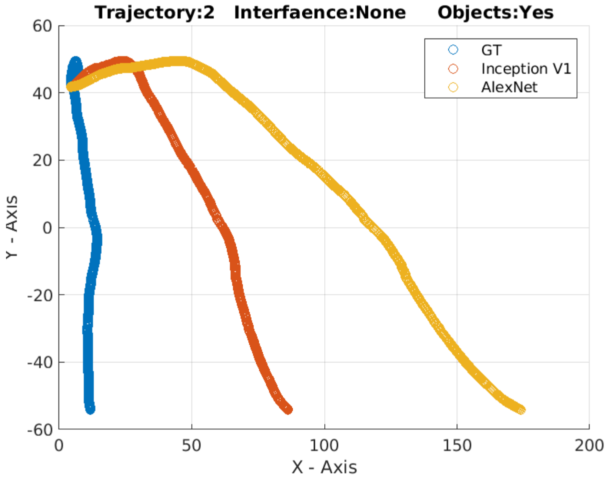

Figure 141: The experiment is run on the 2 trajectory and has interference in no subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 141 is further evidence for the fact that the introduction of the extra texture through the introduction of the obstacles minimises the drift of the estimated trajectory.

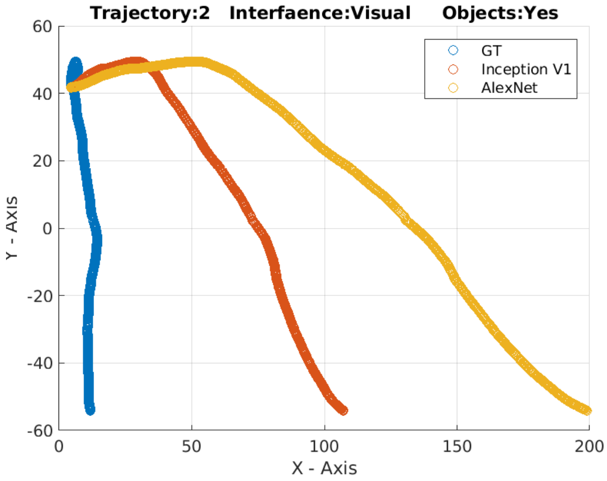

Figure 142: The experiment is run on the 2 trajectory and has interference in the visual subsection of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 142 further demonstrates the negative consequences of the introduction of the visual interference.

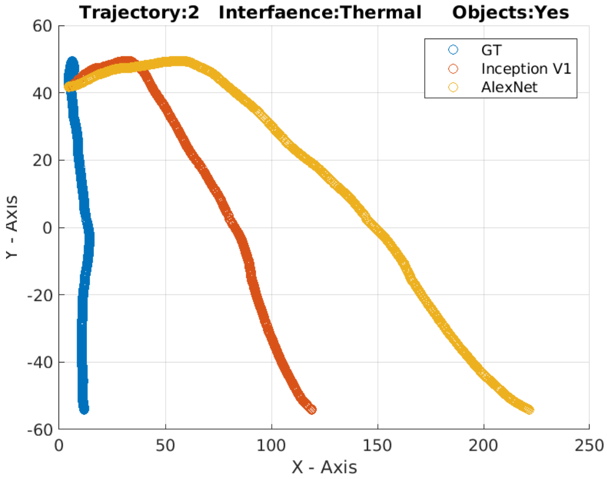

Figure 143: The experiment is run on the 2 trajectory and has interference in the thermal subsection of the electromagnetic spectrum. There were objects present during this experimentation.

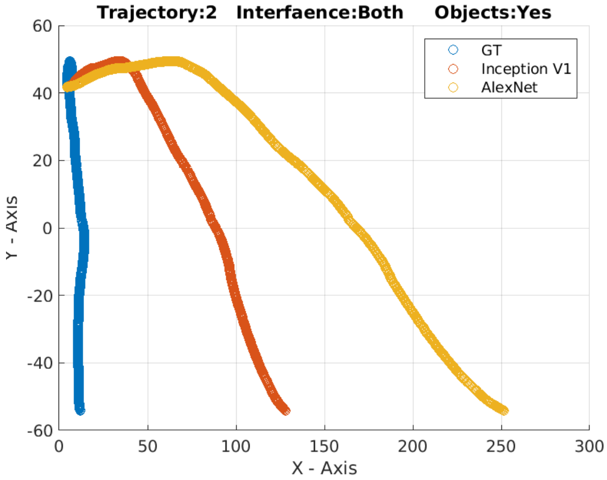

Figure 144: The experiment is run on the 2 trajectory and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 144 once again depicts the fact that the use of both forms of interference is significantly worse for the trajectory estimation than the employment of any one form of interference. It also demonstrates the fact that the employment of obstacles aid to minimise the increase in the drift arising from the employment of the inference.

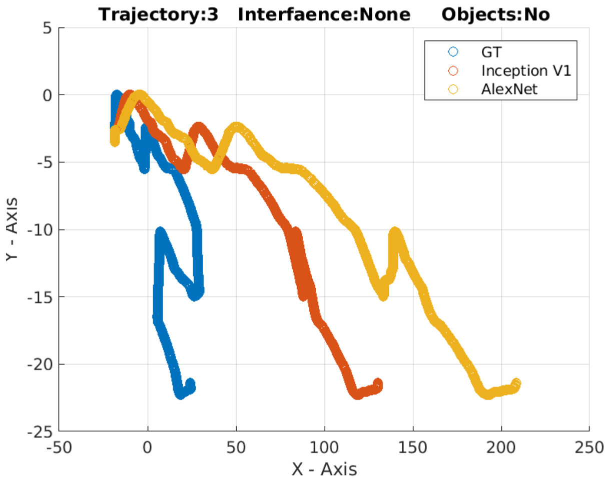

Figure 145: The experiment is run on the 3 trajectory and has interference in no subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 145 is the base case of the 3rd trajectory results based upon the autoencoder latent space theory. It clearly demonstrates that the AlexNet model is better at tracking the contour of the trajectory, but the Inception model produces an estimated trajectory closer to the ground truth.

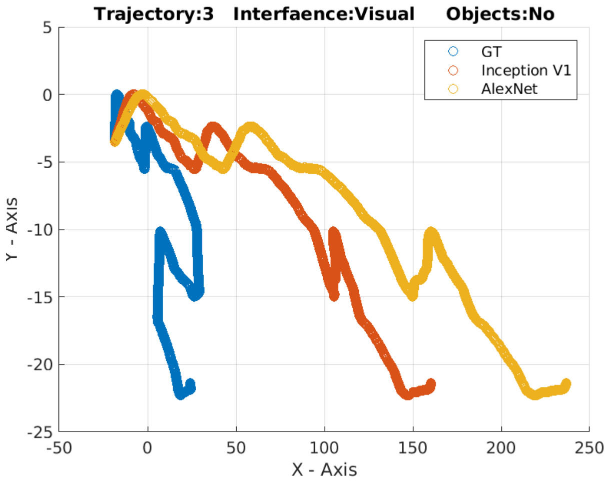

Figure 146: The experiment is run on the 3 trajectory and has interference in the visual subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 146 demonstrates the subsequential degradation of the model's performance once the visual interference is introduced.

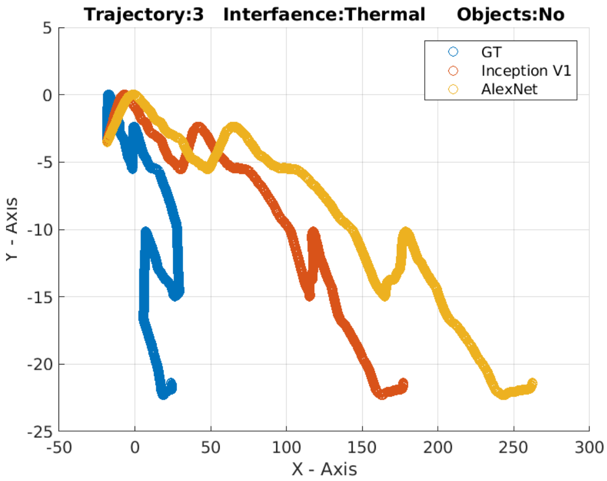

Figure 147: The experiment is run on the 3 trajectory and has interference in the thermal subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

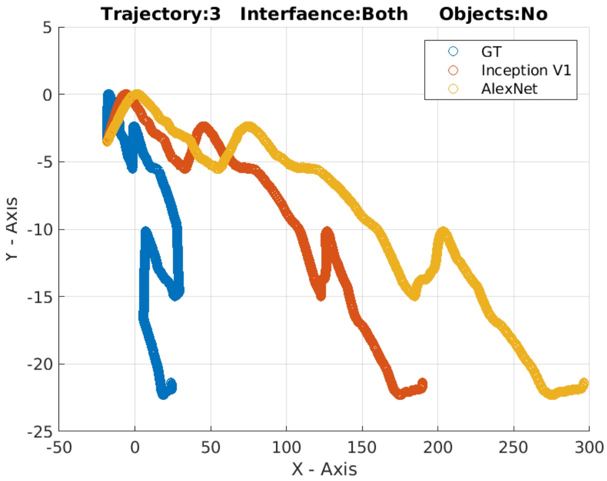

Figure 148: The experiment is run on the 3 trajectory and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 148 demonstrates the fact that the introduction of both types of interference is responsible for a substance degradation in the performance of the models, far in excess of the use of a single type of interference.

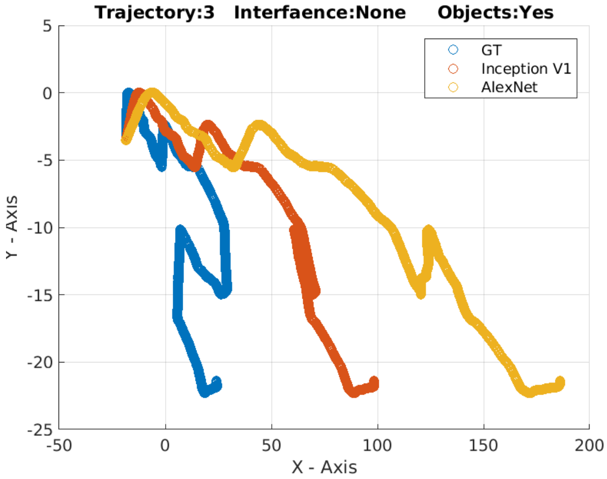

Figure 149: The experiment is run on the 3 trajectory and has interference in no subsection of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 149 demonstrates the fact that the enhancement of the texture in the scene, through the introduction of the obstacles, improves the results of the models.

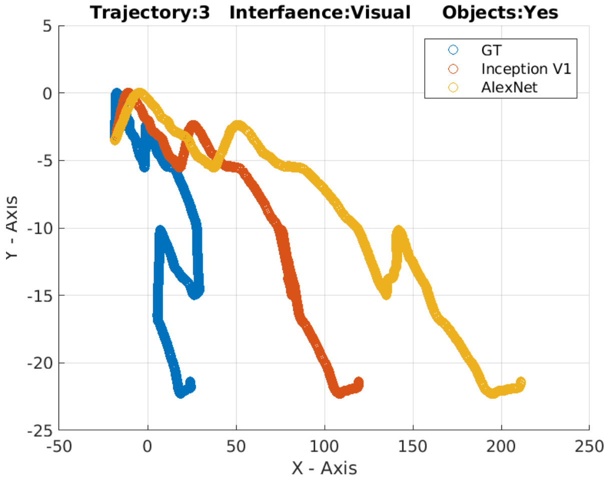

Figure 150: The experiment is run on the 3 trajectory and has interference in the visual subsection of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 150 once again demonstrates the fact that the visual interference degrades the performance of the model.

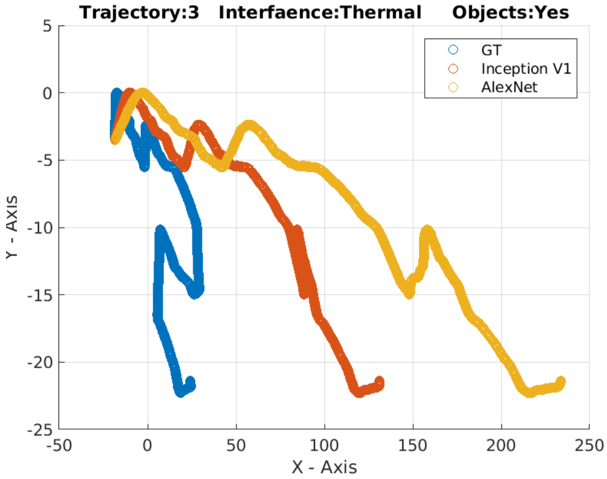

Figure 151: The experiment is run on the 3 trajectory and has interference in the thermal subsection of the electromagnetic spectrum. There were objects present during this experimentation.

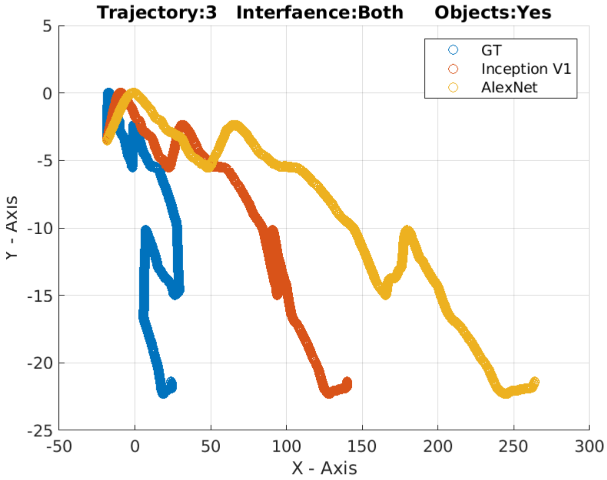

Figure 152: The experiment is run on the 3 trajectory and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 152 conveys clearly the fact that having both types of interference present during the data capture for this trajectory enhanced the drift of both solutions. It further demonstrates the fact that the introduction of the enhanced texture aided in reducing the increased drift resulting from the introduction of the interference but fails to mitigate it completely.

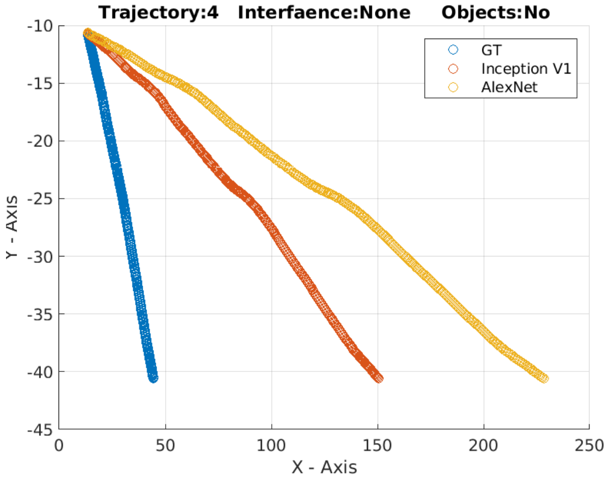

Figure 153: The experiment is run on the 4 trajectory and has interference in no subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 153 is the base case for the fourth trajectory. It has no interference and no obstacles. The results garnered from this depiction are in line with the previous three base cases as is expected.

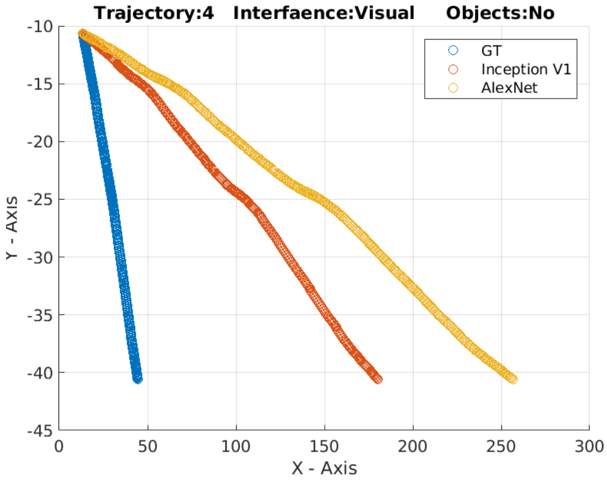

Figure 154: The experiment is run on the 4 trajectory and has interference in the visual subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 154 is indicative of the fact that the introduction of visual interference is in opposition to the quality and accuracy of the pose estimate generated by both solutions.

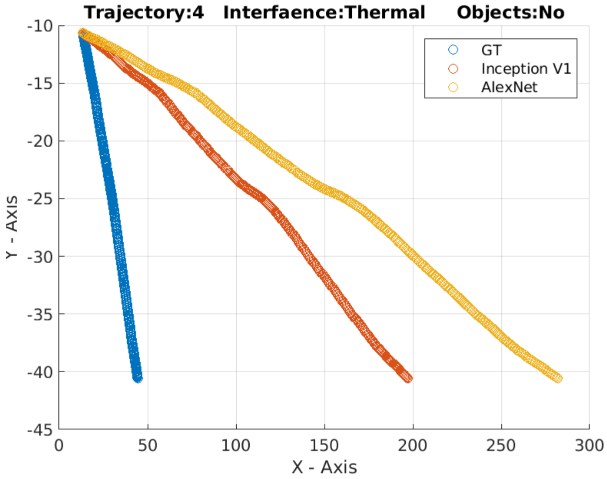

Figure 155: The experiment is run on the 4 trajectory and has interference in the thermal subsection of the electromagnetic spectrum. There were no objects present during this experimentation.

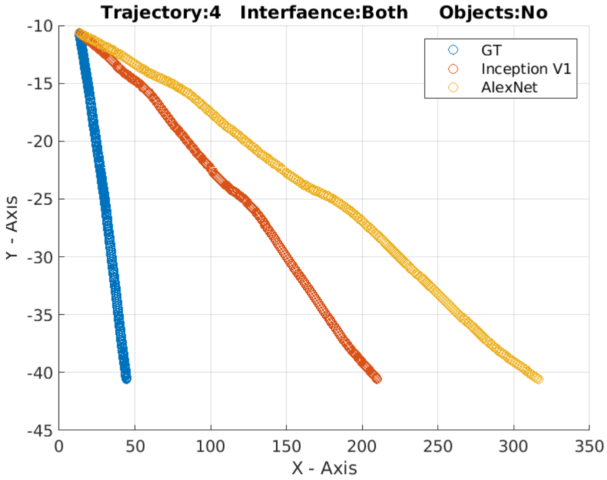

Figure 156: The experiment is run on the 4 trajectory and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were no objects present during this experimentation.

Figure 156 is evidence of the fact that the introduction of both forms of interference is significantly worse for the pose estimation of both solutions than any single form of interference.

Figure 157 is evidence for the fact that the introduction of the texture enhancing obstacles enhances the pose estimation of both models, in comparison to the base case.

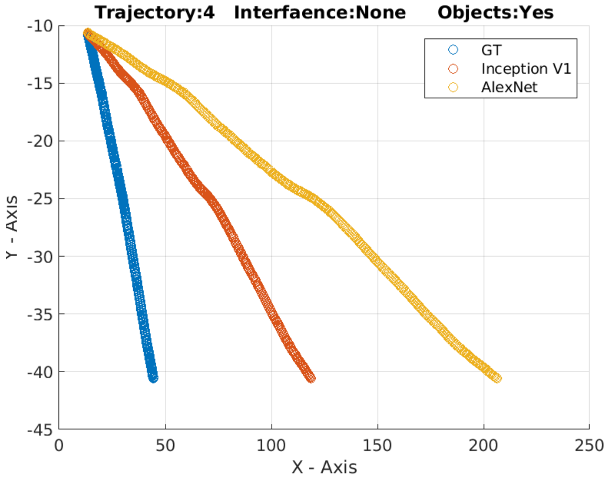

Figure 157: The experiment is run on the 4 trajectory and has interference in no subsections of the electromagnetic spectrum. There were objects present during this experimentation.

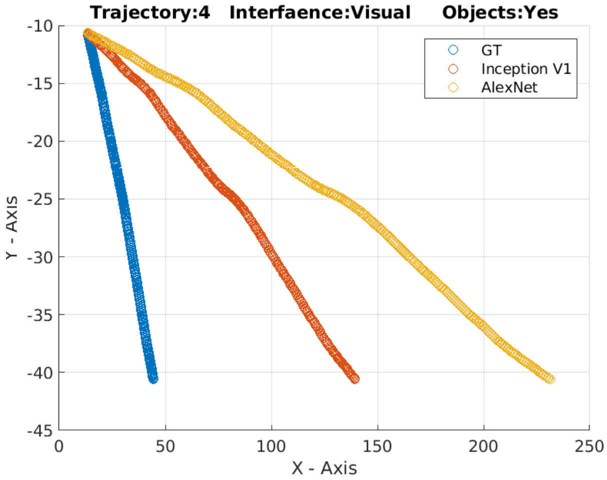

Figure 158: The experiment is run on the 4 trajectory and has interference in the visual subsection of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 158 concludes that the introduction of the visual interference results in sustainably worse trajectory estimation by both models than without it. It further demonstrates that the introduction of the texture enhancing obstacles improves the performance of the models but the final results are still worse than without the introduction of visual interference.

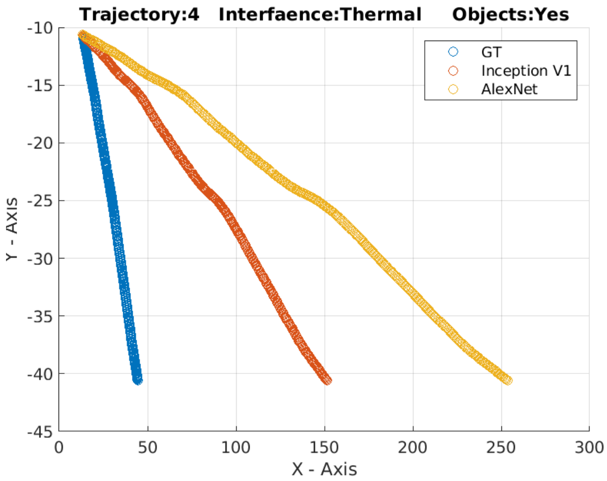

Figure 159: The experiment is run on the 4 trajectory and has interference in the thermal subsection of the electromagnetic spectrum. There were objects present during this experimentation.

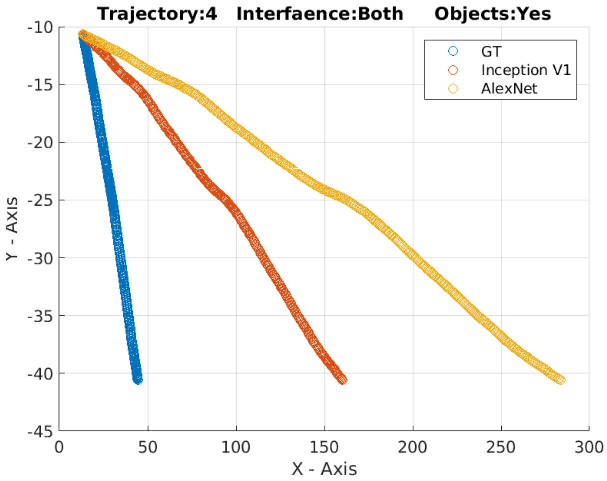

Figure 160: The experiment is run on the 4 trajectory and has interference in the visual and thermal subsections of the electromagnetic spectrum. There were objects present during this experimentation.

Figure 160 is clear in dictating that the introduction of both types of interference degrades the pose estimation of both models significantly more than the introduction of any single type of interference. It also demonstrates that the introduction of the obstacles can mitigate some of the enhanced drift produced by the two types of interference, but not negate it completely.

5.2 Combined Image Plane Odometry

The purpose of employing more than one modality in a solution is robustness but this comes at the expense of speed, by combining multiple modalities into a single image both robustness and speed can be achieved in offline applications. The image fusion taxonomy has three levels: pixel level, feature level and decision level.

Within the taxonomy pixel level methods can be subdivided into space domain based methods and transform domain based methods. Space domain methods employ a range of techniques from weighted averages [187] to block-based fusion strategies [188]. Examples of transform domain include multi-scale decomposition based fusion methods, methods based on the Laplacian pyramid, discrete wavelet transform [189], crvelet transform [190], non-subsampled contourlet [191] and non-subsampled shearlet methods [192], [193].

5.2.1 Motivation

It is the principal objective of this section of the thesis to advance the work done with visual odometry based on large code encoders. This is done by attempting to exploit the code in any fashion that could enable traditional visual odometry techniques to work.

As it is possible to conduct deep visual odometry, it should also be possible to conduct traditional visual odometry on the code, however, the existing feature point detection and matching algorithms are defined for a single modality image this is therefore also something that could be improved. To be clear, this section of the thesis views deep learning methods that operate on images of a single modality as traditional, the non-traditional method is the application of deep learning-based techniques optimised for fused images which are in themselves optimised for the visual odometry process. The fusion of thermal and visual images has already been done independently of autoencoder based visual odometry [194]. The resultant images did enable visual odometry via the methods of multiple view geometry [194]. The problem is that this work did not address the optimization of the fused images for the visual odometry pipeline, nor did the preexisting work exploit the use of autoencoders, which is an obvious next step. Further to this it does not appear that the authors of this work had any desire to reconcile their work with the novel taxonomy introduced at the start of this thesis. This is done through the development of autoencoder based fused abstracted images which in this end-to-end deep learning pipeline may be replaced with those of a different pair of modalities.

The central motivation for applying the traditional visual odometry techniques on the code generated by the encoder is purely the computational complexity of the solution and the hardware requirements of the solution. By allowing for the development of custom cameras, it may be possible to use the traditional solutions on vastly smaller agents.

It should be noted that the fact that the code represents more than one modality may possibly be extended to several or possibly or modalities in a single code. this would have the effect of being able to sample all EM-based phenomena in a single code.

5.2.2 Methodology

A neural network based on a variational autoencoder is employed to construct a latent space representation of the fused image, in a similar fashion to the last section. This autoencoder is employed in order to tightly control the latent space shape, which was not done in the previous section, the control of the latent space enables the code to mimic the number of channels found in either of the modalities - in this work, the three colour channels of the RGB images is employed.

As a single tensor of one visual and one thermal image, the images can be seen in Figures 161-162, is constructed as a concatenation of the two images along the colour channels, the tensor dimension is the width by the length by the 2 channels. This includes one channel from the colour images and the single channel representing the heat intensity in the thermal image. The original thermal and visual images must have the same width and length for this to occur. During this step the image with the higher resolution is downsampled to match the dimentional of the lower one. This step is done prior to stacking.

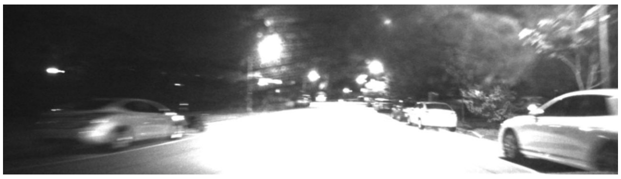

Figure 161: An example of the thermal image used to make up the concatenated tensor.

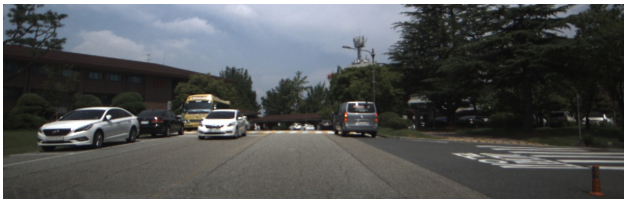

Figure 162: An example of the visual image used to make up the concatenated tensor.

This tensor is then propagated through the neural network that outputs the fused image. It should be noted that the network includes multiple modules and skip connections in order to ensure the employment of most of the key data in the original image.

The first convolutional layer of the image processing modules outputs a single tensor with a shape of $128x128x1$ for each of the visual and thermal images. Then another two convolutional blocks are used to half the length and width of the images and double their depth each time. All three of these convolutions employ $3x3$ kernels. This is used with a stride of 2 and padding of 1. Due to the presence of the vanishing gradient problem, batch normalisation is used at each convolution and the activation function is set to be the leaky relu function. This is due to the fact that the model fails to train under the relu activation function as the gradients vanish around the 50^{th} epoch.

The output of the convolutional networks is a stack of feature maps for each of the two original images each of which has an associated mean and variance which follows the Gaussian distribution, through the normal laws of the Gaussian distribution it is then possible to combine these values into a single mean and standard deviation that represents the latent space of the variational autoencoder.

The one-dimensional vector of the latent space is then transformed into a $32x32x256$ tensor and is then deconvoluted a second time into a tensor with a shape of $64x64x256$. This is then added element-wise to the contamination of the second convolution of each of the original images. After this, a deconvolution is used to half the tensor depth and double the remaining dimensions.

This deconvolutional process is repeated a second time with the contamination of the first convolutional layer as opposed to the second. This is then finalised by another convolutional layer that employs the sigmoid activation function and maps to the original input images.

Due to the complex nature of the network, a sophisticated loss function is necessary:

$$\lambda_a \epsilon_{MSE} + \lambda_b \epsilon_{SSIM} + \epsilon_{VAE} \tag{57}$$

The ϵ_{MSE} represents the pixel-wise difference between the fused image and the thermal image. The structural similarity of the fused image and the visible image is accounted for by the ϵ_{SSIM} term which is given as $1 - SSIM(visible, fused)$. Finally the ϵ_{VAE} is the required loss function that constrains the autoencoder latent space. This is a function of the mathematical derivation of the variational auto-encoder and cannot be removed.

The fused image network FuseNet is far too sophisticated to be explained in a concise manner, as a result, multiple figures and text will be employed to explain the minutia of the network. The first thing to note is that the FuseNet can be considered to be a large black-box model with one output and two input tensors. As displayed in Figures 163. This blacbox can be expanded into a series of smaller blacboxes connected through a variety of connection methods, the relationships between these submodels can be visually depicted as in Figure 164. The node-wise decomposition of the reparametrization layer can be seen in Figure 165, this easily demonstrated the considerable complexity of the model.

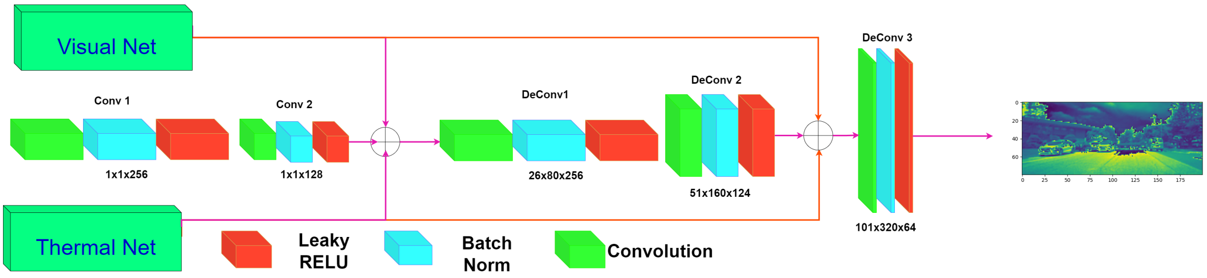

Figure 163: A graphical depiction of the FusionNet black box model.

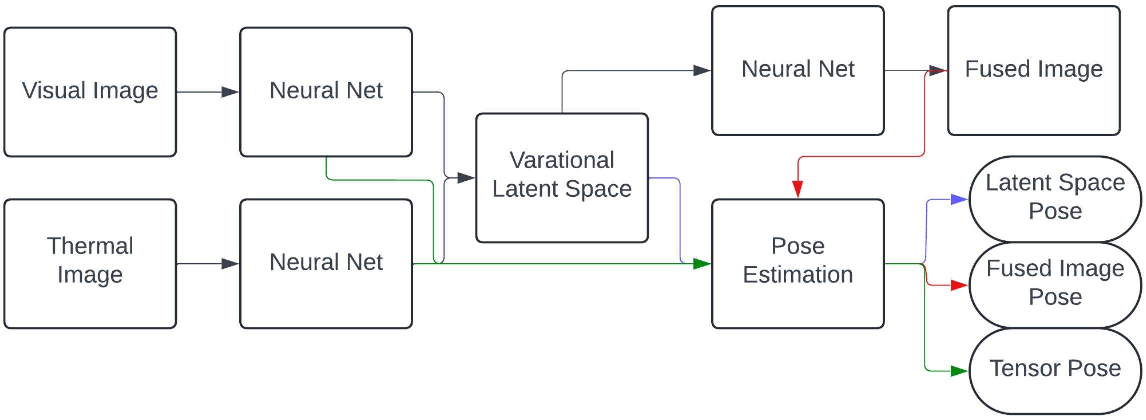

Figure 164: A graphical depiction of the node wise breakdown of the solution.

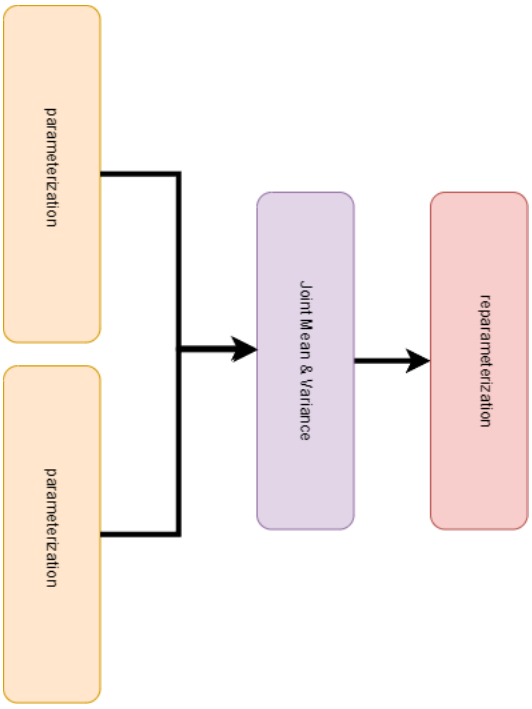

Figure 165: A node wise breakdown of reparamtisation layer.

As noted in Figure 164 the sophistication of the FuseNet is not in its modules, but rather the well-designed connections between them. It is the careful construct that prevents the loss of the gradient and ensures that latent space is truly representative of both modalities. It is further responsible for ensuring that the final pose estimation component of the modal has access to all the information present in the scene.

The fused image is then used as an input into a distinct visual odometry neural network (VoNet) that takes as input two sequential fused images and ultimately predicts the six degrees of freedom pose. This model provides conclusive proof that the new taxonomy of visual odometry is valid, as it never views the original single modality images and is able to derive the pose from the fused image. In order to ensure that the model did not learn to predict the pose from an alternative data source, the model was restricted in its input domain to only take the fused images and so is not allowed to become a visual-inertial model. The model is quite simple in its design, it consists of three aspects:

- A series of convolutional layers
- A single flattened layer
- A series of dense layers

The convolutional layers are originally used to extract the deeper features in the fused images, which are prevalent in both images. This is effective as the batch normalisation prevents the onset of the vanishing gradient problem and the leaky relu function used after each batch normalisation also prevents the loss of the gradient. The flattening layer then converts the output of the convolutional layers into a form compatible with the seven dense layers that convert the features extracted by the convolutional layers into a pose estimate. Table 7 shows the model layers.

Layer	Kernel	Padding	Stride	In Channels	Out Channels
Conv 1	5	0	4	2	64
Conv 2	3	0	4	64	128
Conv 3	3	0	4	128	256
Flattern	0	0	0	–	7680
Dense 1	0	0	0	7680	4000
Dense 2	0	0	0	4000	3500
Dense 3	0	0	0	3500	3000
Dense 4	0	0	0	3000	2000
Dense 5	0	0	0	2000	1500
Dense 6	0	0	0	1500	500
Dense 7	0	0	0	500	6

Table 7: A high level overview of the image fusion model.

In line with the original purpose of employing the use of the varational autoencoder, a 3^{rd} neural network is constructed that takes as input the sequential latent tensors and predicts the six degree of pose estimation from them. The model takes as input a stack of two sequential pytorch tensors which are the outputs of the reparamterisation of the latent layer in the fusion network. Each such output is known as the latent vector of the spatial pair of visual and thermal images. It should be noted that the fusion network operates in the spatial dimension whilst the odometry networks work on both the spatial and temporal networks.

The LatentNet is defined to be a simple sequential neural network consisting of a flattering layer and seven dense layers, it was ultimately modeled upon the VoNet, this was due to the fact that the output of the two models should be identical once trained (in theory) however as the input of the two models varies in type the convolutional layers of the VoNet are no longer required. Furthermore, the output of the flatten layer is also distinct in the model as the LatentNet only employs the layer to constrict the stack of latent tensors into a single row, whilst the VoNet requires the layer to convert the feature maps into dense nodes. It should further be noted that as the convolutional layers are not required in the LatentNet, the network does not employ the use of batch normalisation or the relu/leaky relu activation function. The model layers are presented in Table 8.

Layer	In Channels	Out Channels
Flattern	–	512
Dense 1	512	4000
Dense 2	4000	3500
Dense 3	3500	3000
Dense 4	3000	2000
Dense 5	2000	1500
Dense 6	1500	500
Dense 7	500	6

Table 8: A high level overview of the LatentNet layers.

Due to the effectiveness of U-Net [195], it is also useful to construct a convolution neural network that allows for the processing of a combined convolutional block. The block itself is normally combined horizontally across the model, however, if the inflection point is taken to be the latent tensor then this method would bias the block to either the visual or thermal domain. This is due to the fact that prior to the inflection point each of the vertically aligned blocks is the output of a single modality whereas any block post the inflection point is a product of both modalities. To rectify this the combined block is taken vertically not horizontally and only considers the subset of the model prior to the inflection point. As the third block is the closest to the inflection point it must have sufficient signal to train the latent tensor and as the LatentNet converges upon training it can be deduced that the third block is sufficient for this purpose. Taking into account the fact that the prior two blocks have significantly larger noise in them, the third blocks are the optimal choice.

The final design of the ConvNet, the network used to convert the combined convolutional blocks into a six-degree of freedom pose estimate, is depicted in Table 9. It should be noted that, unlike the LatentNet, this network does require some convolutional layers to process the input.

Layer	Kernel	Padding	Stride	In Channels	Out Channels
Conv 1	5	0	4	2	64
Conv 2	3	0	4	64	128
Conv 3	3	0	4	128	256
Flattern	0	0	0	–	7680
Dense 1	0	0	0	7680	4000
Dense 2	0	0	0	4000	3500
Dense 3	0	0	0	3500	3000
Dense 4	0	0	0	3000	2000
Dense 5	0	0	0	2000	1500
Dense 6	0	0	0	1500	500
Dense 7	0	0	0	500	6

Table 9: A high level overview of the ConvNet model.

5.2.3 Results

It is the principal objective of this section of the thesis to report the training and results of the various models described in the preceding section. This includes the loss and validation graphs of the various models and the optimised hyperparameters of each model. This section also contains various remarks on the training of the model and the inferences or deductions that can be made from them.

The vast quantity of hyperparameter combinations that exist for each model during the training loop, makes it extremely difficult to understand the shape of the last optimisation function. in order to better understand the function many combinations of hyperparameters were tested using the grid search method and the inbuilt tensorboard module.

it should be noted that the results of such optimisations, convey without question that the models are chaotic in nature and subject to stream sensitivity in regard to hyperparameter changes. would slate variation in the hyperparameters associated with the training group and a model the results may prohibit the convergence of the model's training loop.

Any comparison or study of these last functions should rightfully begin with the first model in the sequence which would be the thermal and visual image fusion model. This model was quite difficult to train which resulted in a rather unexpected result. this was the fact that due to the lack of intensity in the thermal image pixels, the SSIM loss function resulted in a division by zero error preventing the optimisation of the model's internal parameters.

In order to diagnose this new model was tested with a different data set which had a different intensity of its thermal imagery, the model did train on that set and produced an unbounded loss which is problematic as no clear trajectory could be deduced from the model. In order to account for this and ensure that the model produced some form of output the SSIM loss component of the last function was removed from the loss function.

Whilst the last function now has in effect two terms, resulting from the removal of the SSIM term, it is still possible to graphically depict the product of both terms. it should be noted that whilst the individual terms greatly differ in the contours of the training graphs, the combined graph is far more standardised. This is evident in 166.

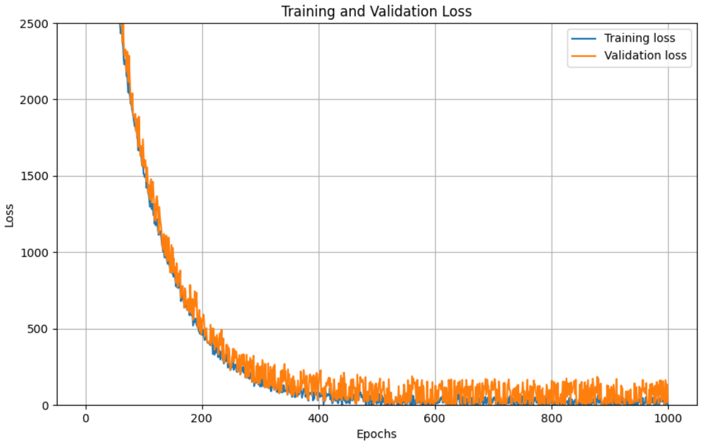

Figure 166: A graphical depiction of the training results of the FuseNet.

As can be deduced from the various plots in Figure 166 any change in hyperparameters does to some degree alter the contour of the loss function, however, its primary effect is the determination of the rate of convergence and whether in practice the training process would converge. it may not be apparent in Figure 166 that by selecting certain open parameter combinations the graph will diverge to Infinity. this is due to the omission of such trajectories from the graph in order to prevent scale ambiguity which would render most information untenable.

it can be inferred that a considerable proportion of the variation in any trading results of the model is contributable to the mean squared error component of the loss function. The remaining proportion of the variation in the loss is the result of the VAE component of the loss.

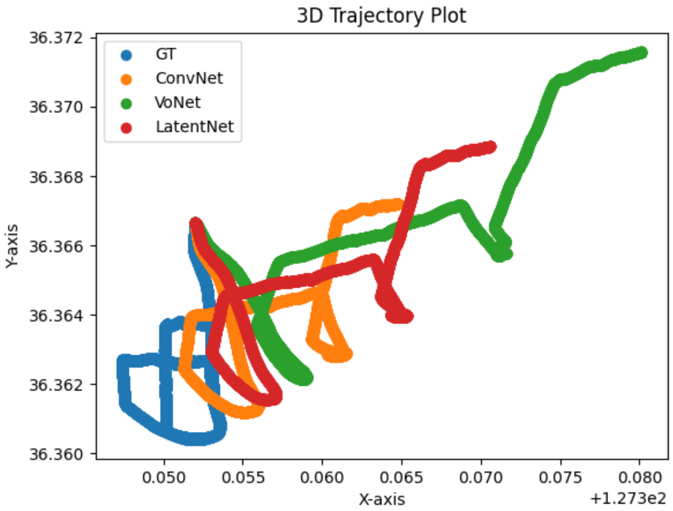

Figure 167: A graphical depiction of the models' results on the 0^{th} trajectory of the dataset.

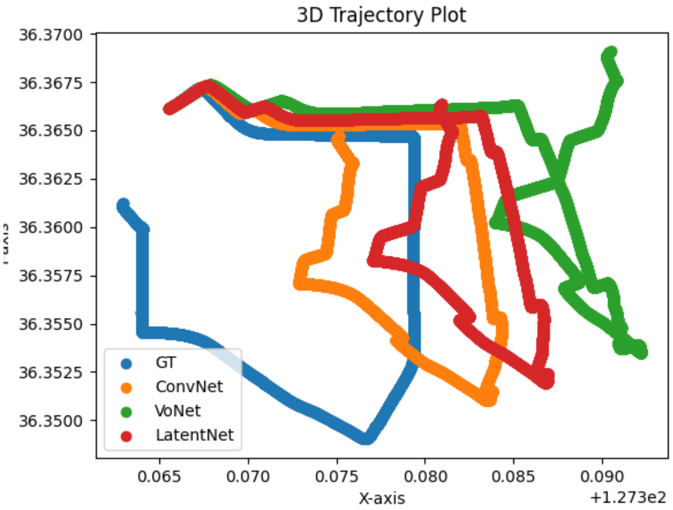

Figure 168: A graphical depiction of the models' results on the 1^{st} trajectory of the dataset.

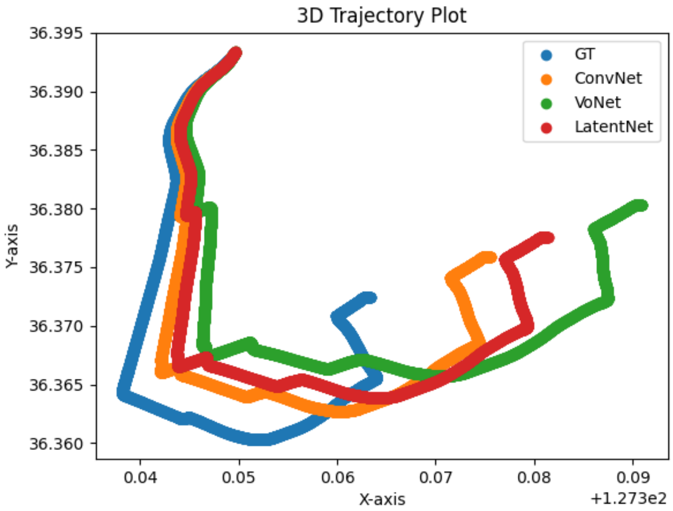

Figure 169: A graphical depiction of the models' results on the 2^{rd} trajectory of the dataset.

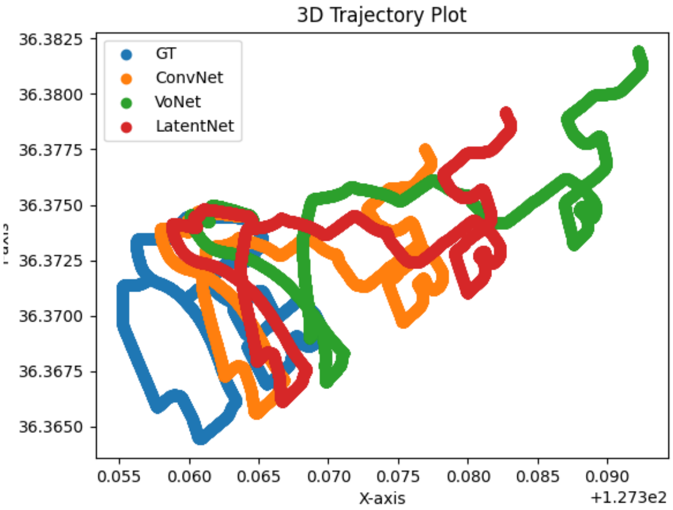

Figure 170: A graphical depiction of the models' results on the 3^{rd} trajectory of the dataset.

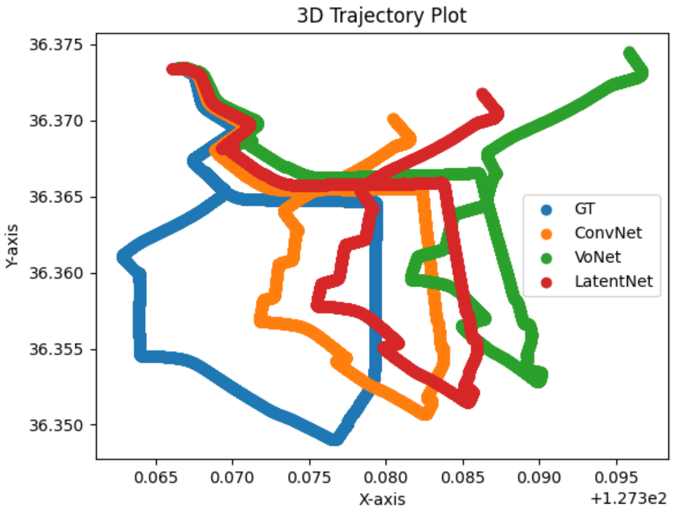

Figure 171: A graphical depiction of the models' results on the 4th trajectory of the dataset.

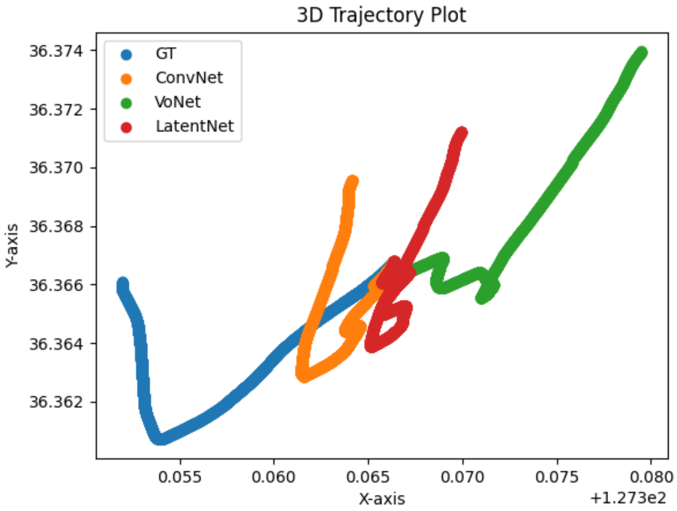

Figure 172: A graphical depiction of the models' results on the 5th trajectory of the dataset.

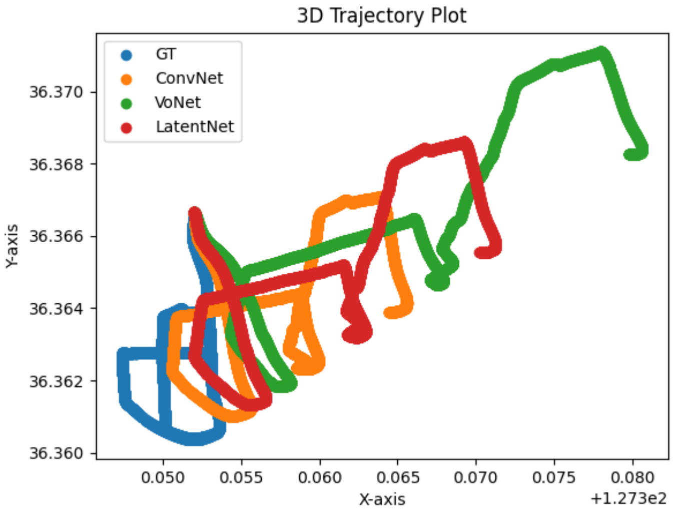

Figure 173: A graphical depiction of the models' results on the 6^{th} trajectory of the dataset.

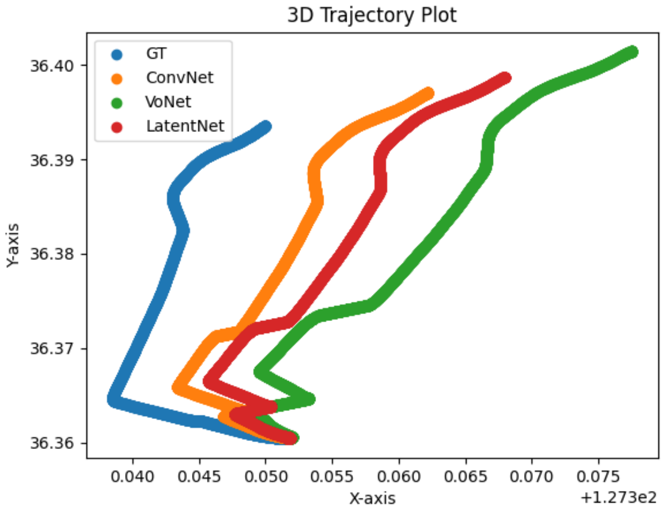

Figure 174: A graphical depiction of the models' results on the 7^{th} trajectory of the dataset.

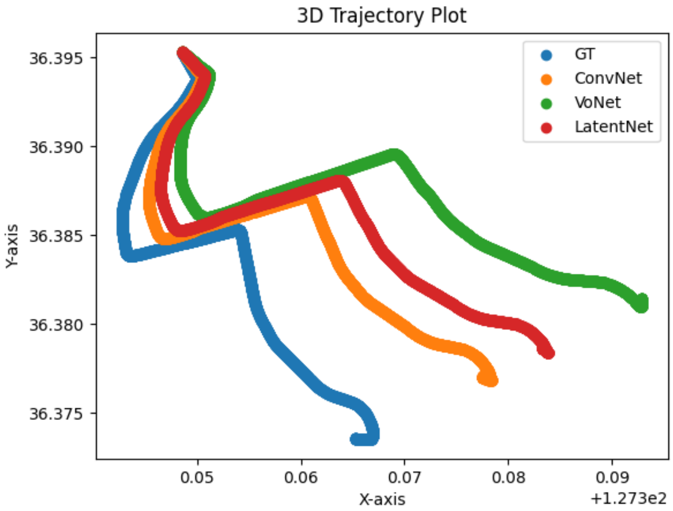

Figure 175: A graphical depiction of the models' results on the 8^{th} trajectory of the dataset.

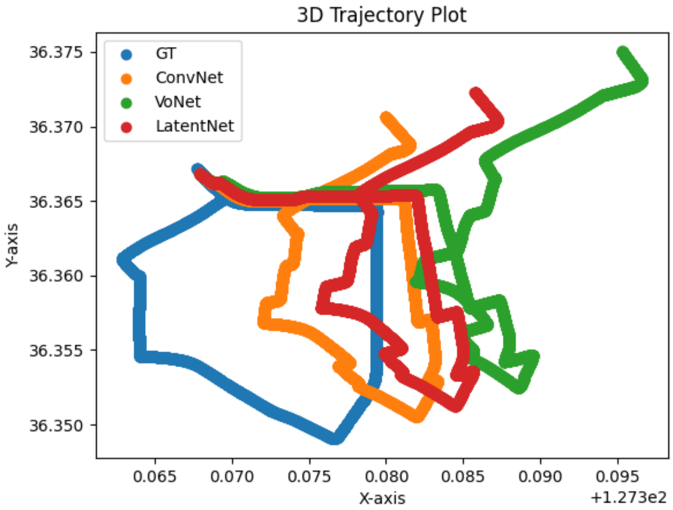

Figure 176: A graphical depiction of the models' results on the 9^{th} trajectory of the dataset.

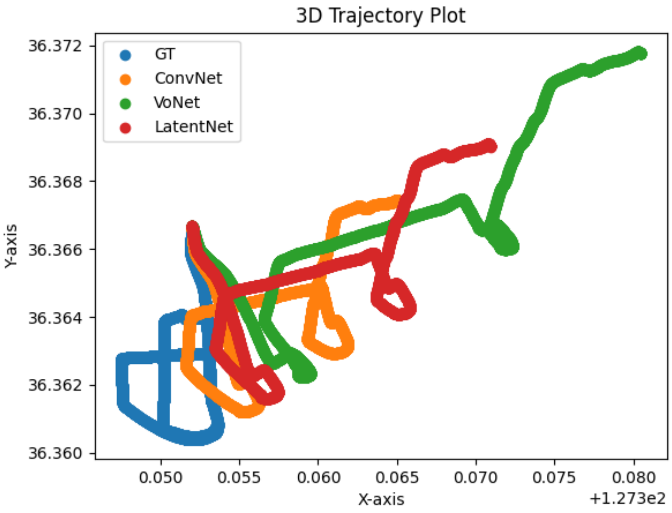

Figure 177: A graphical depiction of the models' results on the 10^{th} trajectory of the dataset.

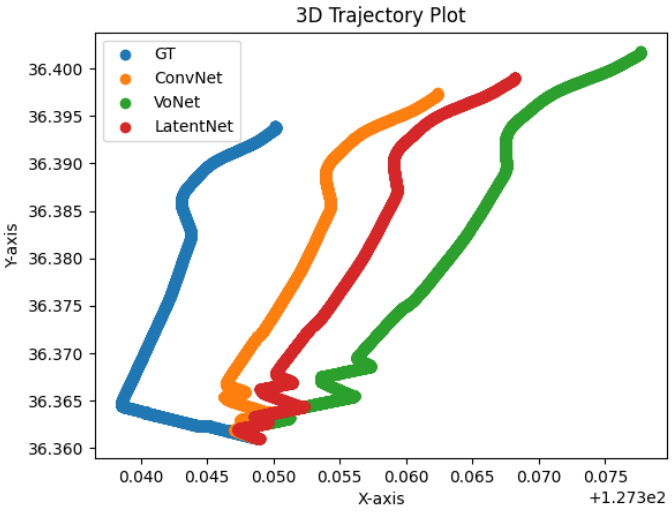

Figure 178: A graphical depiction of the models' results on the 11^{th} trajectory of the dataset.

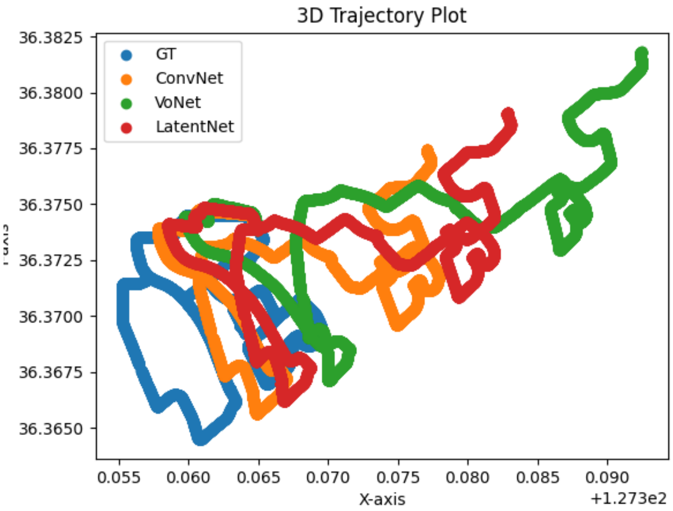

Figure 179: A graphical depiction of the models' results on the 12^{th} trajectory of the dataset.

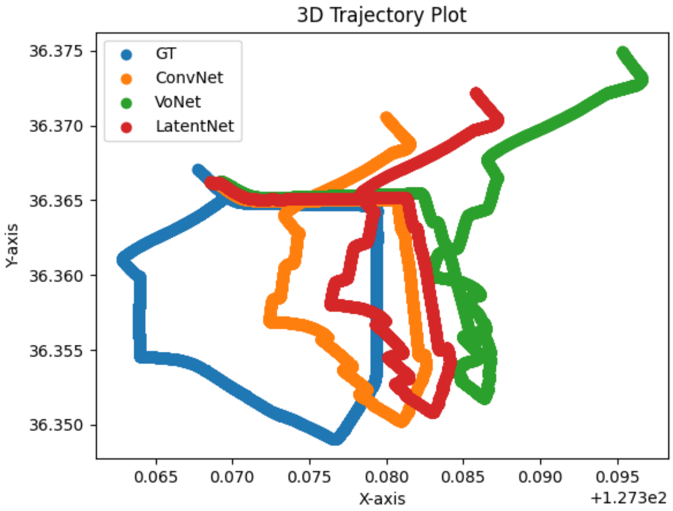

Figure 180: A graphical depiction of the models' results on the 13^{th} trajectory of the dataset.

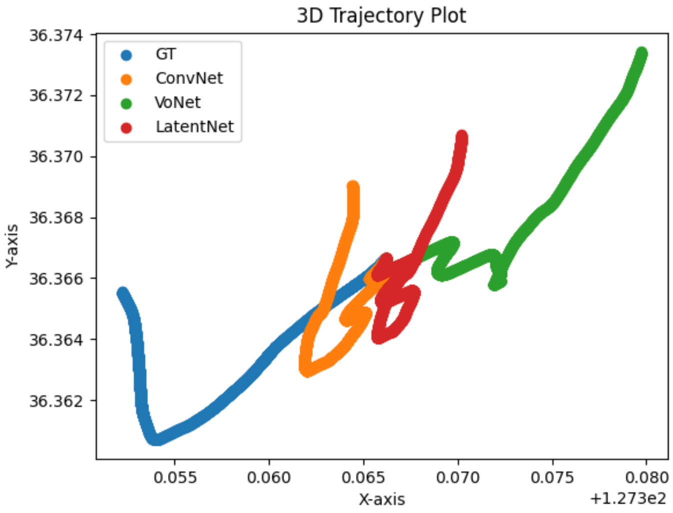

Figure 181: A graphical depiction of the models' results on the 14^{th} trajectory of the dataset.

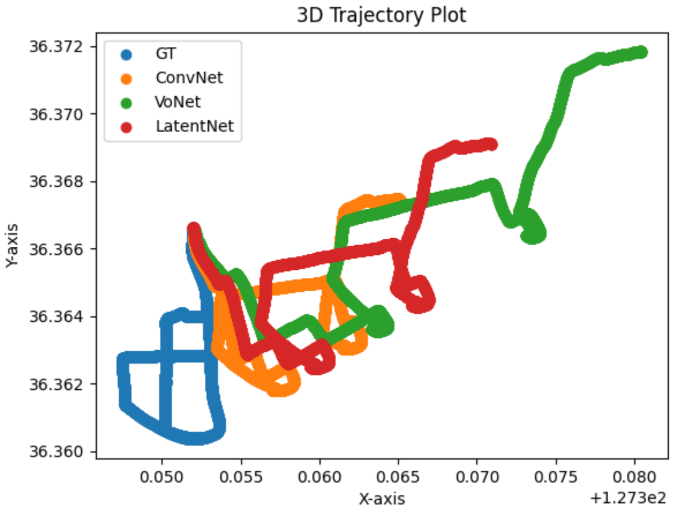

Figure 182: A graphical depiction of the models' results on the 15^{th} trajectory of the dataset.

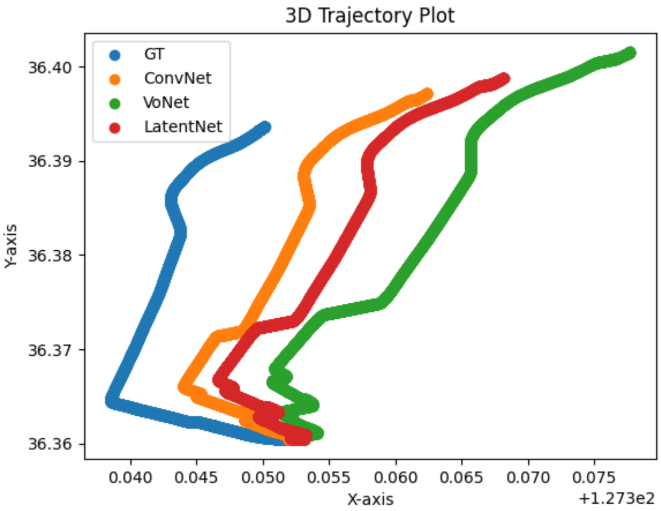

Figure 183: A graphical depiction of the models' results on the 16^{th} trajectory of the dataset.

5.3 Latent Space Exploration

The movement of a three-dimensional scene observed with a camera produces changes to the 2D projected image plan(s). Estimating it requires computing a projection of actual motion onto an image plane [196]. To achieve this goal, algorithms rely on the brightness constancy constraint. A two-dimensional displacement field that shows the apparent movement of patches of pixels classified by brightness in sequential images is known as optical flow[197]. An ideal optical flow field ideally features dense displacement vectors that map all possible points from one image onto their corresponding locations in another image [198]. Gibson first proposed this concept. It should be noted that some of the pixels in the first image may not exist in the second depending on the motion undergone by the agent between the two images.

Techniques based on feature matching attempt to track image features from successive images that are sparse yet discriminative [196], without creating any ambiguous areas; as a result, the computed flow field is sparse but robust. Discriminative features, such as corners and edges, as well as low-contrast features like flat regions, can be matched to determine optical flow [199]. Due to the black-box nature of neural networks, it may not be possible to prove this is the key to their success in the field, but the convolution of their kernels does suggest that this is the case.

5.3.1 Motivation

It is the principal objective of this section to convey to the audience the results of the analysis of the various images and the latent space of those images. In the proceeding sections, a series of new image combinations were made possible on a single reference frame in such a presentation that a natural comparison may be drawn.

It is the purpose of this section to make such a comparison of the various image types, chiefly the section is concerned with a comparison of the fused image with both thermal and visual images. As part of this comparison, it is natural to wonder if the ability to construct optical flow images is present in the fused stream and if any phenomena are present in the optical flow image stream of the fused images if it's possible to extract.

As a result of these questions, this section of the thesis attempts to ask the following series of questions:

- Can an Optical Flow image stream be extracted from the fused images?
- Can a visual odometry pipeline be used on the Optical Flow images if they can be extracted?
- What is the difference between the visual and thermal images?

- What is the difference between the fused image and each of the original images?
- To what extent are the differences between the original images present in the fused image?

5.3.2 Methodology

it is the principal objective of this section of the thesis to discuss the methodology by the results of this section of the thesis where ascertained. Most of this methodology is already discussed to some degree in other sections of the thesis. This is due to the fact that this section of the thesis attempts to apply a similar system as the optical flow system in Chapter 4.1 to the fused images generated in the previous section.

Figure 184 depicts the evolution training and evaluation losses from fine-tuning the optical flow model on the fused images. This was problematic as the optical flow diagrams were made by mapping the optical flow of the visual images onto the fused image plane.

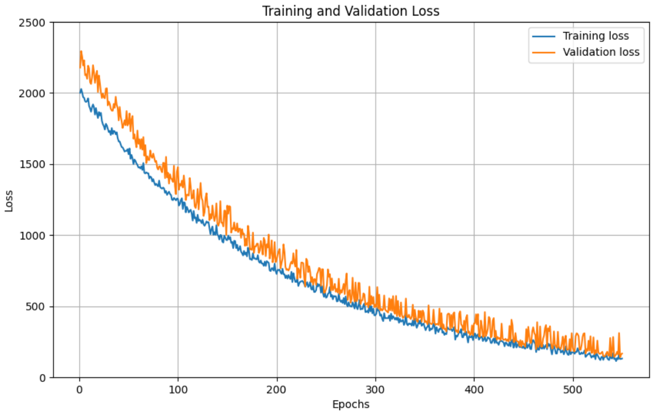

Figure 184: A graphical depiction of the finetuning loss.

The mapping process involved visually attempting to identify correlations between the fused images and the visual images and then passing a selection of visual images to the original model to gain the visual image flow maps. The regions of the flow maps that could be utilised to describe the image space between two sequential fused images. This was not an exact approach as the fused images incorporated both thermal features, which did not exist in the visual images and regions that did not get perfectly mapped into the visual optical flow.

The thermal images were also placed into the optical flow model, the outcome of which is representative of the optical flow of the heat propagation in the scene. The outcome of these images where then overlayed on the mappings of the visual optical flow (with a masking that only enabled the features in the fused images which are not present in the visual mapping to be overlayed). By ensuring that there was a one-to-one correspondence between the visual image the thermal images and the fused images, the optical flow images generated could be considered a valid representation of the fused image optical flow. These images were then used as the domain and codomain to fintune the optical flow model.

In order to determine whether or not an optical flow image stream can be extracted from the fused image space, the application of the fine-tuned optical model to the fused image dataset would prove this. As all the image sequences have had the model produce an optical flow stream without any missing images, this is indeed possible but may need to have substantial improvements in quality.

As the optical flow images can be extracted from the fused images and then generate a valid pose estimate, it is possible to conclude that the odometry pipeline can be applied to the fused image optical flow space. The remaining questions are not yet answerable via the work done here and must be pursued as future work.

5.3.3 Results

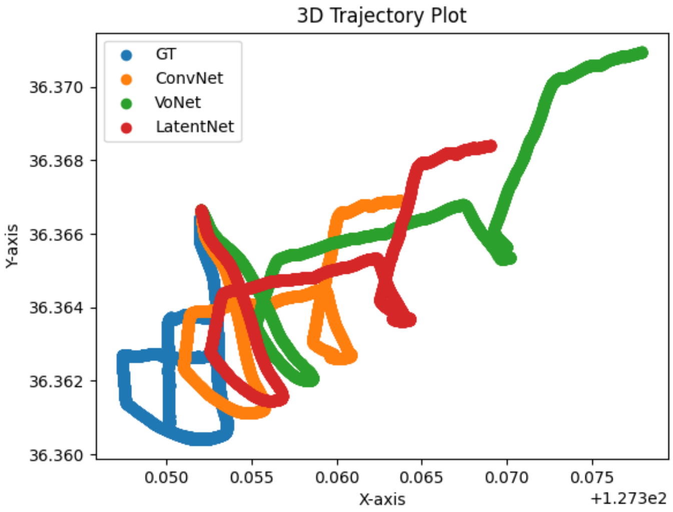

Figure 185: A graphical depiction of the models' results on the 0^{th} trajectory of the dataset.

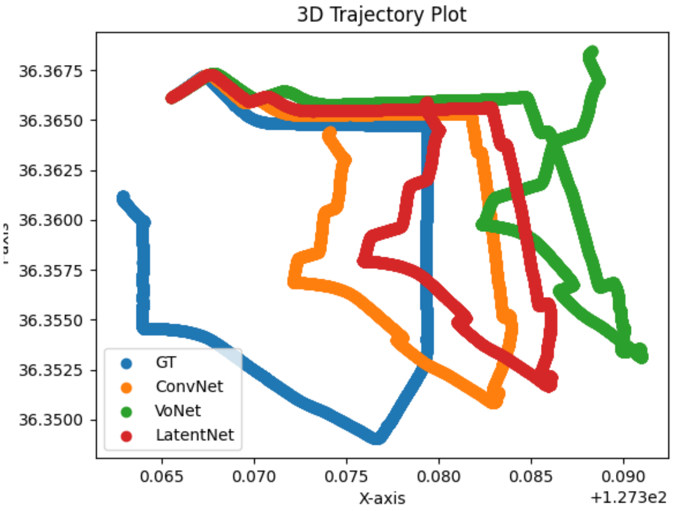

Figure 186: A graphical depiction of the models' results on the 1^{st} trajectory of the dataset.

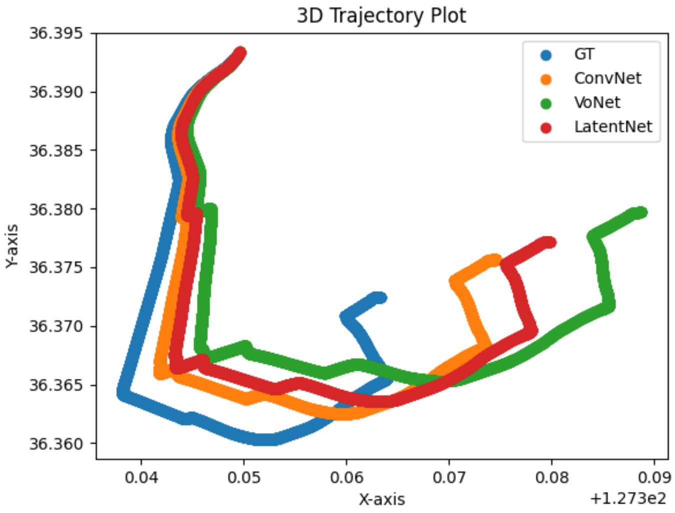

Figure 187: A graphical depiction of the models' results on the 2^{nd} trajectory of the dataset.

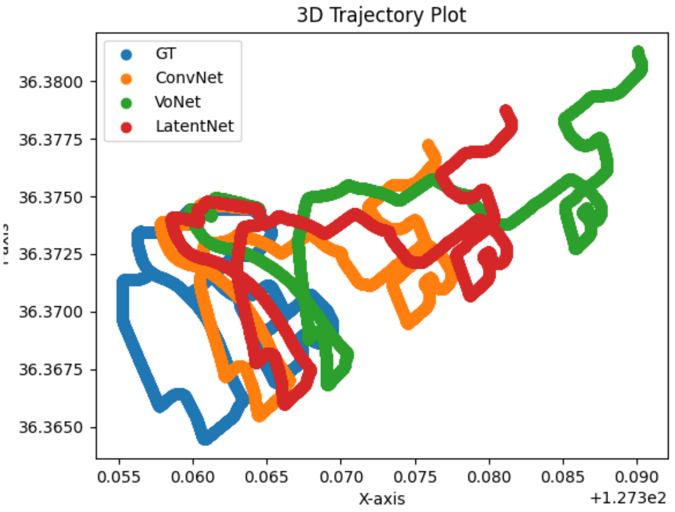

Figure 188: A graphical depiction of the models' results on the 3^{rd} trajectory of the dataset.

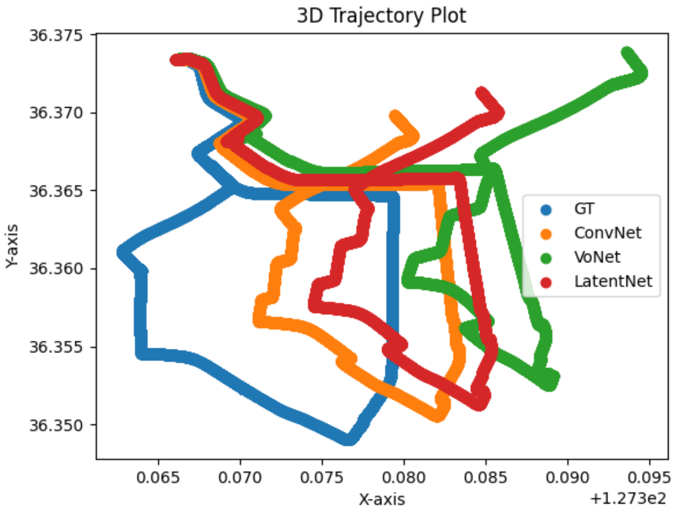

Figure 189: A graphical depiction of the models' results on the 4^{th} trajectory of the dataset.

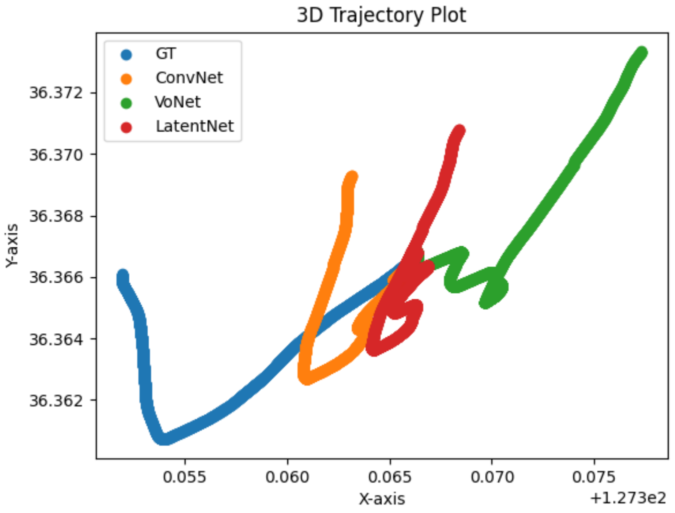

Figure 190: A graphical depiction of the models' results on the 5^{th} trajectory of the dataset.

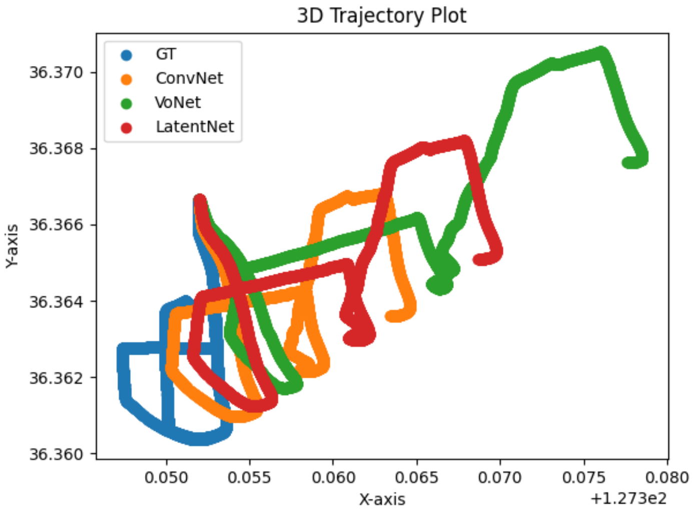

Figure 191: A graphical depiction of the models' results on the 6^{th} trajectory of the dataset.

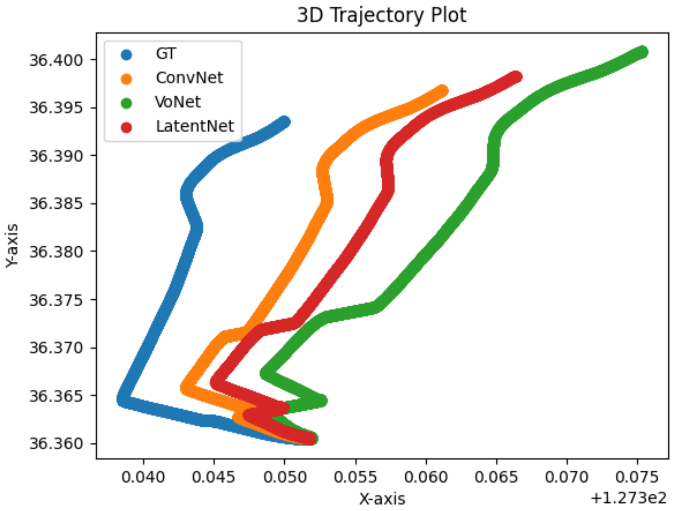

Figure 192: A graphical depiction of the models' results on the 7^{th} trajectory of the dataset.

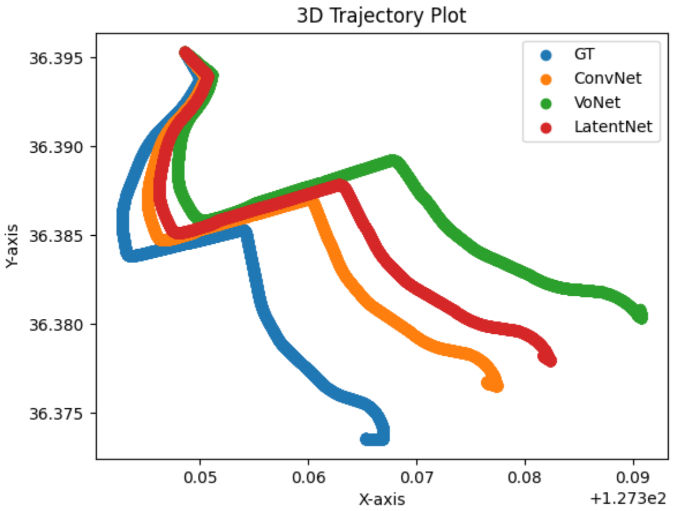

Figure 193: A graphical depiction of the models' results on the 8^{th} trajectory of the dataset.

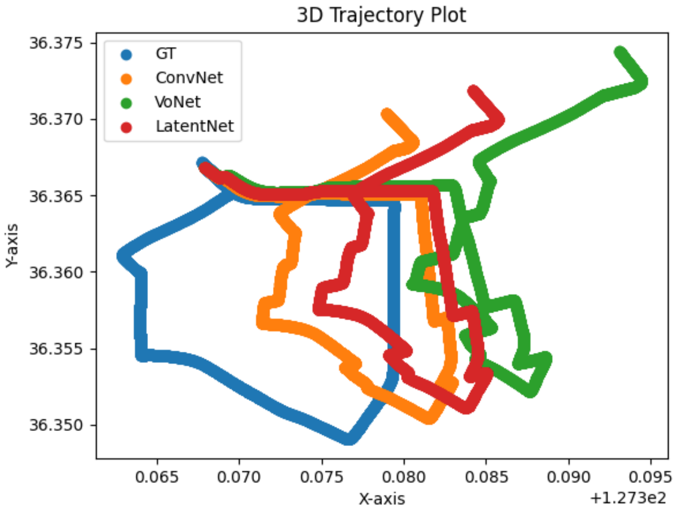

Figure 194: A graphical depiction of the models' results on the 9^{th} trajectory of the dataset.

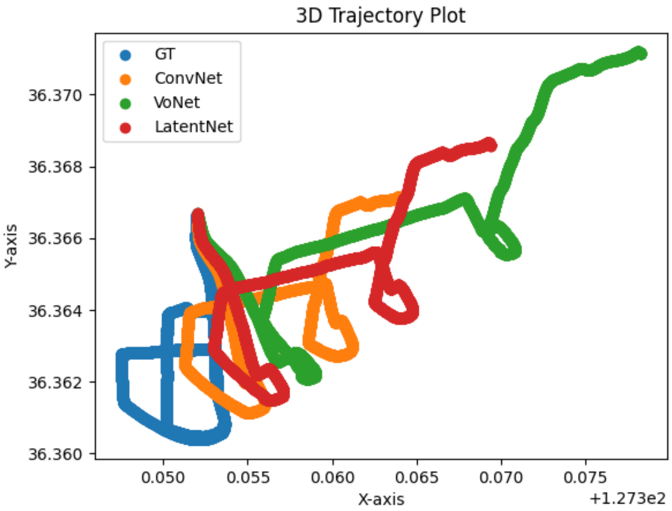

Figure 195: A graphical depiction of the models' results on the 10^{th} trajectory of the dataset.

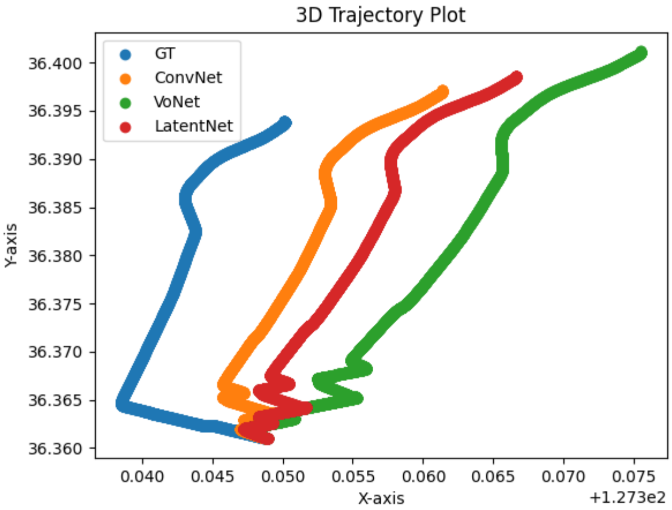

Figure 196: A graphical depiction of the models' results on the 11^{th} trajectory of the dataset.

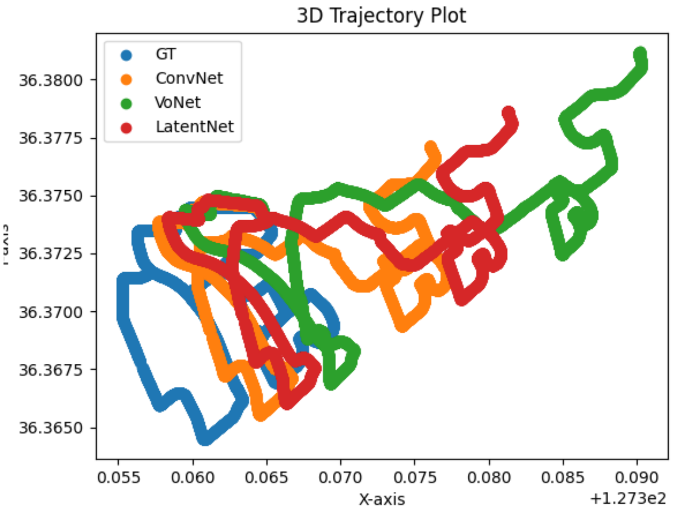

Figure 197: A graphical depiction of the models' results on the 12^{th} trajectory of the dataset.

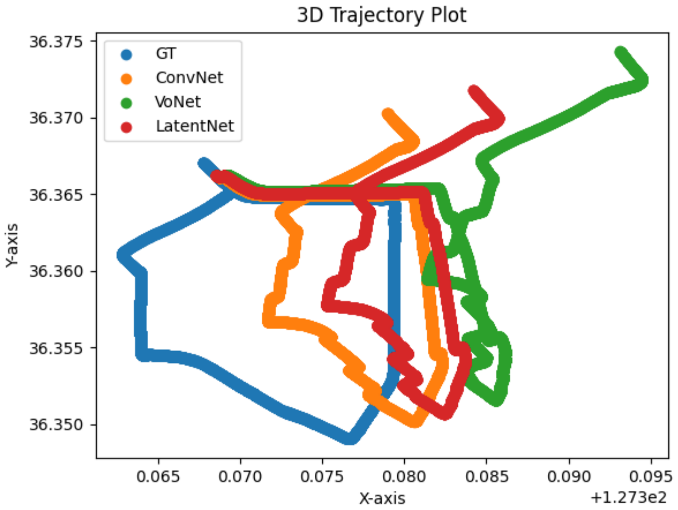

Figure 198: A graphical depiction of the models' results on the 13^{th} trajectory of the dataset.

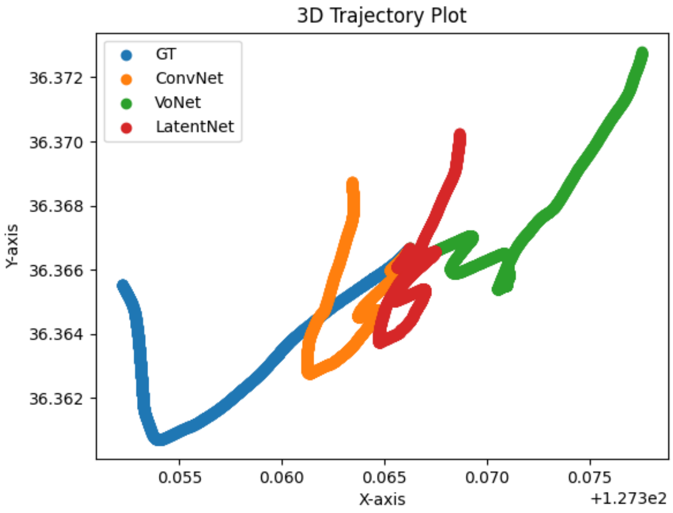

Figure 199: A graphical depiction of the models' results on the 14^{th} trajectory of the dataset.

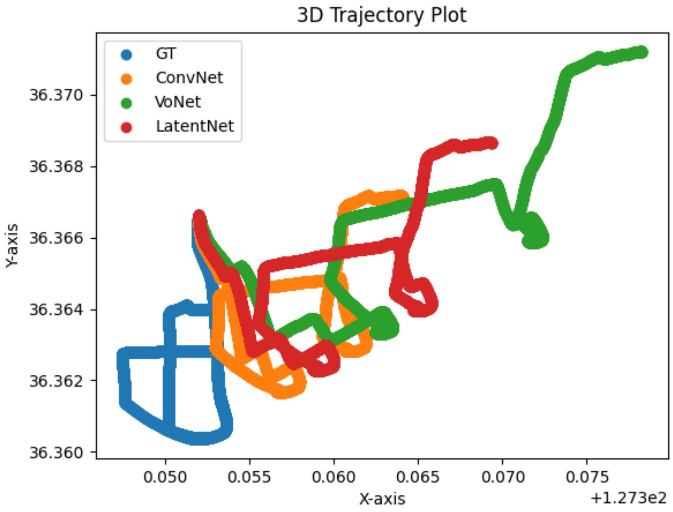

Figure 200: A graphical depiction of the models' results on the 15^{th} trajectory of the dataset.

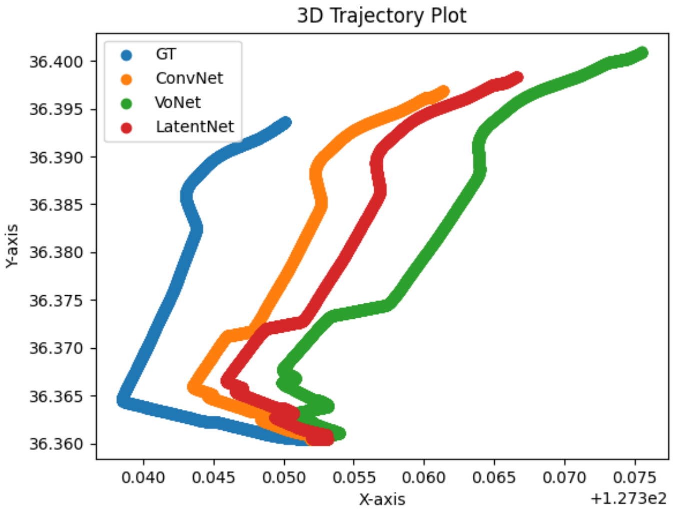

Figure 201: A graphical depiction of the models' results on the 16^{th} trajectory of the dataset.

5.4 Conclusions and Future Work

It is the principal objective of this section of the thesis to conclude the preceding section and inform the reader of the possible direction of the future work which may be built upon it.

This section has demonstrated the ability to fuse multiple modalities into a single image using an autoencoder which can then be utilised for pose estimation. It is shown that there exists a definitive contradiction between image reconstruction and pose prediction. This section of the thesis has also demonstrated the ability to employ a non-image-based latent space representation of multiple modalities that can subjected to regularisation. This section has also shown that the optical flow can be applied to the latent images.

The ability to compute pose estimation on the fused images raises questions regarding the characteristics of the abstract camera representation. This requires a lot of theoretical mathematical development to understand the abstract camera and how they vary with modality.

The ability to translate the novel mathematical framework to a set of traditional visual odometry techniques is also of interest, however, given that AI-based solutions have provided superior results this may never become practically useful. This is due to the fact that hand-crafted feature kernels cannot compare with AI-optimised kernels.

It would further be useful to have the ability to employ the fused image systems when one of the two cameras becomes unusable. for example a thermal or visual decay. This may not be possible as the fused image is multispectral by definition, but the decay of one modality may result in a pure monocular system.

Significant attention must be paid to the employment of optical flow on the fused images. This is due to the fact that the brightness consistency constraint may not hold in theory if the pixels of the input plane are subject to variations in multiple modalities. It may also be possible to construct a multichannel fused image which could aid in preventing this from destroying the system during periods in which the abilities of a modality are forced to decay.

It would be of interest if by splitting the image planes into subsections preexisting autoencoders and GANs may be used to construct a fused image plane where the dimensionality of the input image or the desired latent presentation is no longer a limit on the network architecture.

The employment of modern AI architecture would be useful in constraining the systems for example representing the image space as an embedding problem would enable the training of the scene elements in a latent space, however, this latent space would be vastly different from that of large langue models and their word embeddings. Whilst word embeddings optimise for semantic relationships between words, the latent space in question would optimise for temporal consistency of objects.

In conclusion, this chapter of the thesis has demonstrated some outstanding visual odometry work on latent image representations in a fused image. This has opened up a lot of novel avenues of research some of which are outlined in this section. Due to the novelty of these research questions, they lack a sound mathematical basis on which to operate. This two is a possible avenue of future work.

6 Conclusion

6.1 Overview

It is the principal objective of this section of the doctoral thesis to conclude the afore-presented body of scholarly work and convey to the audience the significance of its conclusion. This is done in addition to the secondary objective of providing some possible ideas of continuation for this work and the significance of the novelty of those approaches. This of course necessitates the construction of a review of the entire thesis as opposed to the individual sections of the thesis which were presented within their own respective sections.

6.2 Summary and Discussion

It is the principal objective of this section of the thesis to convey to the audience the conclusion of the doctoral thesis and both summarize and discuss the benefits of having undertaken this thesis to the wider research community.

To summarise the novel worker comprised in this thesis begins with a novel date of fusion application of a pre-existing solution, in order to test its robustness in novel situations. This is closely followed by the faithful encoding of the solution into a deep learning framework and built upon by the construction of a generalizable deep learning framework that exploits the pre-existing and omnipresent backpropagation algorithm in order to minimize drift between modalities. The next novel innovation is the use of autoencoder networks to compress multi-modality sensor configuration and their pipelines into a single-modality solution. This is all wrapped up in a novel taxonomy that enables a new perspective by which to view the field of visual odometry.

Whilst the novel taxonomy presented here may not appear to be the most directly applicable to the development of new ideas within the field, it is by far the most significant contribution as not only does it enable researchers to ask novel questions which may not have otherwise been posed it also enables stronger links to other areas of research which may be a benefit to the advancement of the field. Two of these are directly apparent in the following sections of this thesis.

The employment of data fusion to enable a stereo thermal kalman filter, filled in a few missing patches in the pre-existing body of work comprising the literature. This is due to variations upon this already being done such as the visual variation, however, the contribution did not end there as the same algorithm was employed with both modalities and in a multimodality configuration a new type of comparison became possible further enriching the knowledge about this particular solution.

The ability to faithfully encode an optical flow-based six degree of freedom pose estimation problem solution into a deep learning network, is beneficial to the development of the field as it enables the field to directly incorporate any abundance made in tangential areas such as human pose detection.

The next great innovation in this doctor of the thesis was the ability to develop a generalizable framework that exploits the omnipresent existence of backpropagation within neural networks to minimize the difference in the drift of multi-spectral visual or odometry solutions. This is highly beneficial as it is omnipresent in all deep learning and solutions to the problem of visual navigation as such it can be readily applied and improve the robustness of such methods minimizing the trade-off in accuracy.

Finally, the thesis concludes with the ability to represent multiple modalities into a single image that can then be exploited to produce the final estimate of the trajectory of the agent. This is very impressive as it enables the ability to directly employ sensors which sample in multiple modalities thereby reducing the task of multi-spectral Visual navigation to a monocular task. As a computational complexity reduction given that depth perception algorithms especially those that utilize machine learning acquire prevalence and high performance results for monocular tests, it becomes self-evident that this is a complexity reducing matter that greatly extends the feasibility of most Solutions.

6.3 Future Work

It is the principal objective of this section of the thesis to outline a plan by which the novel findings of this thesis can be extended. The most notable extension is building on the novel work done in this thesis using autoencoders.

The ability to perform visual odometry on the code generated by the autoencoders is conclusive evidence that the dimensionality of the data required by the odometry process is a small subset of the image space. This suggests that different encodings of the same scene can be generated by varying the modality of the sensor configuration. This would suggest that the various encodings could be overlaid to analyse the similarities and differences between them. This would allow for a far richer understanding of the structural differences of the various modalities. This would then suffice as the premise to develop a series of rich and diverse scenes that would be sampled by various sensors, enabling the development of a novel contextual relationship study that may shape the field's view of the symbiosis of the various modalities.

This relationship of the encodings could also be explored through the exploitation of the support vector machines and T-SNE which would enable an N-dimentional clustering of the data. It is hoped that this could be exploited to generate novel feature-matching methods between the modalities. It is further hoped that the encoding space can benefit from latent factor analysis in the form of principal component analysis, further reducing the size of the dataset required for the odometry process. This could be considered a method of compressing the dataset if it is technically feasible. As such a study relating the size of the compression factor to the loss in the accuracy of the pose estimate will also be required.

Whilst the novel findings of this thesis are often generated from the employment of neural networks, this thesis has not exploited turning to optimise the networks. This could lead to further improvement in the multispectral error elimination by backpropagation, as it could be utilised to allow for a more flexible generalisable framework that could enable selective pruning of both networks to realise a non-symmetric solution that optimises for the distortion of the feature sets projection onto the image plane in each of the two modalities.

There is also ample opportunity to develop a feature description and matching solution that operates not on the modalities specific images but on the abstracted representation space or if it exists the latent factor space associated with the encoding space. It should also be possible to do this for the encoding space itself.

The limited scope of the thesis could also be expanded to include the use of loop closer which would give rise to a study concerning the ability to accurately match key frames across the various spaces. An example of this would be to develop a robust matching method between the encoding space and the modality specific space(s).

In conclusion, there are many advances that can be made to extend the novel work done in this thesis in a non-cooperative environment. The extensions to the work done on autoencoders is by far the most fascinating of the aforementioned advances, however, the list of possible extensions presented here is far from exhaustive.

Bibliography

[1] D. Nistér, O. Naroditsky, and J. Bergen, “Visual odometry for ground vehicle applications”, *Journal of Field Robotics*, vol. 23, no. 1, pp. 3–20, 2006.

[2] C. Harris, M. Stephens, *et al.*, “A combined corner and edge detector”, in *Alvey vision conference*, Citeseer, vol. 15, 1988, pp. 10–5244.

[3] D. Nistér, “A minimal solution to the generalised 3-point pose problem”, in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR*, vol. 1, 2004.

[4] M. Buczko and V. Willert, “Flow-decoupled normalized reprojection error for visual odometry”, in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 2016, pp. 1161–1167.

[5] K. Yan, H. Tian, E. Liu, R. Zhao, Y. Hong, and D. Zuo, “A decoupled calibration method for camera intrinsic parameters and distortion coefficients”, *Mathematical Problems in Engineering*, vol. 2016, 2016.

[6] T. Mouats, N. Aouf, L. Chermak, and M. A. Richardson, “Thermal stereo odometry for uavs”, *IEEE Sensors Journal*, vol. 15, no. 11, pp. 6335–6347, 2015.

[7] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, “Robust visual inertial odometry using a direct ekf-based approach”, in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, IEEE, 2015, pp. 298–304.

[8] U. Orguner and F. Gustafsson, “Storage efficient particle filters for the out of sequence measurement problem”, in *11th International Conference on Information Fusion*, IEEE, 2008, pp. 1–8.

[9] M. R. U. Saputra, P. P. de Gusmao, C. X. Lu, *et al.*, “Deeptio: A deep thermal-inertial odometry with visual hallucination”, *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1672–1679, 2020.

[10] R. Kümmerle, G. Grisetti, H. M. Strasdat, K. Konolige, and W. Burgard, “G2o: A general framework for graph optimization”, *2011 IEEE International Conference on Robotics and Automation*, pp. 3607–3613, 2011.

[11] G. Grisetti, R. Kümmerle, C. Stachniss, and W. Burgard, “A tutorial on graph-based slam”, *IEEE Intelligent Transportation Systems Magazine*, vol. 2, no. 4, pp. 31–43, 2010.

[12] S. Wang, R. Clark, H. Wen, and N. Trigoni, “Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks”, in *2017 IEEE international conference on robotics and automation (ICRA)*, IEEE, 2017, pp. 2043–2050.

[13] D. Scaramuzza and F. Fraundorfer, “Visual odometry [tutorial]”, *IEEE robotics & automation magazine*, vol. 18, no. 4, pp. 80–92, 2011.

[14] D. Nistér, O. Naroditsky, and J. R. Bergen, “Visual odometry”, *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 1, pp. 652–659, 2004.

[15] H. P. Moravec, *Obstacle avoidance and navigation in the real world by a seeing robot rover.* Stanford University, 1980.

[16] L. Matthies and S. Shafer, “Error modeling in stereo navigation”, *IEEE Journal on Robotics and Automation*, vol. 3, no. 3, pp. 239–248, 1987.

[17] C. F. Olson, L. H. Matthies, H. Schoppers, and M. W. Maimone, "Robust stereo ego-motion for long distance navigation", in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, IEEE, vol. 2, 2000, pp. 453–458.

[18] L. H. Matthies, *Dynamic stereo vision.* Carnegie Mellon University, 1989.

[19] C. Harris and J. Pike, "3d positional integration from image sequences", *Image and Vision Computing*, vol. 6, no. 2, pp. 87–90, 1988, 3rd Alvey Vision Meeting, ISSN: 0262-8856. DOI: `https://doi.org/10.1016/0262-8856(88)90003-0`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/0262885688900030`.

[20] J.-M. Frahm, P. Fite-Georgel, D. Gallup, *et al.*, "Building rome on a cloudless day", in *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, Springer, 2010, pp. 368–381.

[21] M. J. Hannah, *Computer matching of areas in stereo images.* Stanford University, 1974.

[22] H. P. Moravec, "Towards automatic visual obstacle avoidance", in *International Joint Conference on Artificial Intelligence*, 1977.

[23] W. Förstner, "A feature based correspondence algorithm for image matching", *ISPRS ComIII, Rovaniemi*, pp. 150–166, 1986.

[24] C. F. Olson, L. H. Matthies, M. Schoppers, and M. W. Maimone, "Rover navigation using stereo ego-motion", *Robotics and Autonomous Systems*, vol. 43, no. 4, pp. 215–229, 2003.

[25] A. Milella and R. Siegwart, "Stereo-based ego-motion estimation using pixel tracking and iterative closest point", in *Fourth IEEE International Conference on Computer Vision Systems (ICVS'06)*, IEEE, 2006, pp. 21–21.

[26] A. Howard, "Real-time stereo visual odometry for autonomous ground vehicles", in *2008 IEEE/RSJ international conference on intelligent robots and systems*, IEEE, 2008, pp. 3946–3952.

[27] Y. Cheng, M. W. Maimone, and L. Matthies, "Visual odometry on the mars exploration rovers-a tool to ensure accurate driving and science imaging", *IEEE Robotics & Automation Magazine*, vol. 13, no. 2, pp. 54–62, 2006.

[28] M. Maimone, Y. Cheng, and L. Matthies, "Two years of visual odometry on the mars exploration rovers", *Journal of Field Robotics*, vol. 24, no. 3, pp. 169–186, 2007.

[29] P. Besl and H. McKay, "A method for registration of 3-d shapes. ieee trans pattern anal mach intell", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 14, pp. 239–256, Mar. 1992. DOI: `10.1109/34.121791`.

[30] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd, "Real time localization and 3d reconstruction", in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, IEEE, vol. 1, 2006, pp. 363–370.

[31] D. Nistér, "An efficient solution to the five-point relative pose problem", *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 6, pp. 756–770, 2004.

[32] J.-P. Tardif, Y. Pavlidis, and K. Daniilidis, "Monocular visual odometry in urban environments using an omnidirectional camera", in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2008, pp. 2531–2538.

[33] L. Kneip, R. Siegwart, and M. Pollefeys, "Finding the exact rotation between two images independently of the translation", in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, Springer, 2012, pp. 696–709.

[34] B. Horn and B. Schunck, "Determining optical flow", *Artificial Intelligence*, vol. 17, pp. 185–203, Aug. 1981. DOI: 10.1016/0004-3702(81)90024-2.

[35] F. Fraundorfer, D. Scaramuzza, and M. Pollefeys, "A constricted bundle adjustment parameterization for relative scale estimation in visual odometry", in *2010 IEEE International Conference on Robotics and Automation*, IEEE, 2010, pp. 1899–1904.

[36] N. Sünderhauf, K. Konolige, S. Lacroix, and P. Protzel, "Visual odometry using sparse bundle adjustment on an autonomous outdoor vehicle", in *Autonome Mobile Systeme 2005: 19. Fachgespräch Stuttgart, 8./9. Dezember 2005*, Springer, 2006, pp. 157–163.

[37] K. Konolige, M. Agrawal, and J. Solà, "Large-scale visual odometry for rough terrain", in *International Symposium of Robotics Research*, 2007.

[38] J.-P. Tardif, M. George, M. Laverne, A. Kelly, and A. Stentz, "A new approach to vision-aided inertial navigation", in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nov. 2010, pp. 4161–4168. DOI: 10.1109/IROS.2010.5651059.

[39] A. I. Comport, E. Malis, and P. Rives, "Accurate quadrifocal tracking for robust 3d visual odometry", in *Proceedings 2007 IEEE International Conference on Robotics and Automation*, IEEE, 2007, pp. 40–45.

[40] H. Javidnia and P. Corcoran, "Accurate depth map estimation from small motions", in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2453–2461.

[41] D. E. Jacobs, J. Baek, and M. Levoy, "Focal stack compositing for depth of field control", *Stanford Computer Graphics Laboratory Technical Report*, vol. 1, no. 1, p. 2012, 2012.

[42] H. Lin, C. Chen, S. B. Kang, and J. Yu, "Depth recovery from light field using focal stack symmetry", in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3451–3459.

[43] S. Suwajanakorn, C. Hernandez, and S. M. Seitz, "Depth from focus with your mobile phone", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3497–3506.

[44] P. Kellnhofer, T. Ritschel, K. Myszkowski, and H.-P. Seidel, "Optimizing disparity for motion in depth", *Computer Graphics Forum*, vol. 32, Jul. 2013. DOI: 10.1111/cgf.12160.

[45] P. Kellnhofer, P. Didyk, K. Myszkowski, M. M. Hefeeda, H.-P. Seidel, and W. Matusik, "Gazestereo3d: Seamless disparity manipulations", *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–13, 2016.

[46] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross, "Nonlinear disparity mapping for stereoscopic 3d", *ACM Transactions on Graphics (TOG)*, vol. 29, no. 4, pp. 1–10, 2010.

[47] C. Wang and A. A. Sawchuk, "Disparity manipulation for stereo images and video", in *Stereoscopic Displays and Applications XIX*, SPIE, vol. 6803, 2008, pp. 473–484.

[48] R. Jiang, R. Klette, and S. Wang, “Statistical modeling of long-range drift in visual odometry”, in *Computer Vision–ACCV 2010 Workshops: ACCV 2010 International Workshops, Queenstown, New Zealand, November 8-9, 2010, Revised Selected Papers, Part II 10*, Springer, 2011, pp. 214–224.

[49] D. Allan and H. Hellwig, “Time deviation and time prediction error for clock specification, characterization, and application”, in *IEEE 1978 Position Location and Navigation Symposium*, 1978, pp. 29–36.

[50] T. S. Hoon, P. Singh, and S. K. Ayop, “Working with logarithms”, *Malaysian Education Dean’s Council Journal*, vol. 6, no. 6, pp. 121–129, 2010.

[51] R. Horaud, M. Hansard, G. Evangelidis, and C. Ménier, “An overview of depth cameras and range scanners based on time-of-flight technologies”, *Machine vision and applications*, vol. 27, no. 7, pp. 1005–1020, 2016.

[52] J. A. Hall, “Can solid-state imaging devices replace television camera tubes?”, *Optical Engineering*, vol. 16, no. 3, pp. 224–232, 1977.

[53] J. C. Mullikin, L. J. van Vliet, H. Netten, F. R. Boddeke, G. Van der Feltz, and I. T. Young, “Methods for ccd camera characterization”, in *Image Acquisition and Scientific Imaging Systems*, Spie, vol. 2173, 1994, pp. 73–84.

[54] N. Waltham, “Ccd and cmos sensors”, *Observing Photons in Space: A Guide to Experimental Space Astronomy*, pp. 423–442, 2013.

[55] S. Lee, “Engineering novel semiconductors for advanced infrared sensors”, Ph.D. dissertation, University of Washington, 2023.

[56] I. Bloch, “Information combination operators for data fusion: A comparative review with classification”, *IEEE Transactions on systems, man, and cybernetics-Part A: systems and humans*, vol. 26, no. 1, pp. 52–67, 1996.

[57] D. L. Hall and J. Llinas, “An introduction to multisensor data fusion”, *Proceedings of the IEEE*, vol. 85, pp. 6–23, 1997.

[58] D. Smith and S. Singh, “Approaches to multisensor data fusion in target tracking: A survey”, *IEEE transactions on knowledge and data engineering*, vol. 18, no. 12, pp. 1696–1710, 2006.

[59] A. A. Goshtasby and S. G. Nikolov, “Image fusion: Advances in the state of the art”, *Information Fusion*, vol. 8, pp. 114–118, 2007.

[60] I. Corona, G. Giacinto, C. Mazzariello, F. Roli, and C. Sansone, “Information fusion for computer security: State of the art and open issues”, *Information Fusion*, vol. 10, no. 4, pp. 274–284, 2009.

[61] H. Wache, T. Voegele, U. Visser, *et al.*, “Ontology-based integration of information-a survey of existing approaches.”, in *Proceedings of the IJCAI’01 Workshop on Ontologies and Information Sharing, Seattle, Washington, USA, Aug 4-5*, 2001.

[62] G. L. Rogova and V. Nimier, “Reliability in information fusion: Literature survey”, in *Proceedings of the seventh international conference on information fusion*, vol. 2, 2004, pp. 1158–1165.

[63] J. Yao, V. V. Raghavan, and Z. Wu, “Web information fusion: A review of the state of the art”, *Information Fusion*, vol. 9, no. 4, pp. 446–449, 2008.

[64] L. Klein, *Sensor and Data Fusion Concepts and Applications* (Tutorial Text Series). SPIE, 1999, ISBN: 9780819432315. [Online]. Available: `https://books.google.gr/books?id=nxMfAQAAIAAJ`.

[65] H. Boström, S. F. Andler, M. Brohede, *et al.*, *On the definition of information fusion as a field of research*, 2007.

[66] E. Waltz, "Data fusion for c3i: A tutorial", *Command, Control, Communications Intelligence (C3I) Handbook*, pp. 217–226, 1986.

[67] B. Dasarathy, *Decision Fusion.* IEEE Computer Society Press, 1994, ISBN: 9780818644528. [Online]. Available: `https://books.google.gr/books?id=7q9QAAAAMAAJ`.

[68] I. Goodman, R. Mahler, and H. Nguyen, *Mathematics of Data Fusion* (Theory and Decision Library B). Springer Netherlands, 1997, ISBN: 9780792346746.

[69] M. M. Kokar, J. A. Tomasik, and J. Weyman, "Formalizing classes of information fusion systems", *Information Fusion*, vol. 5, no. 3, pp. 189–202, 2004.

[70] S. J. Julier and J. K. Uhlmann, "A non-divergent estimation algorithm in the presence of unknown correlations", in *Proceedings of the 1997 American Control Conference*, IEEE, vol. 4, 1997, pp. 2369–2373.

[71] A. Makarenko, A. Brooks, T. Kaupp, H. Durrant-Whyte, and F. Dellaert, "Decentralised data fusion: A graphical model approach", in *2009 12th International Conference on Information Fusion*, IEEE, 2009, pp. 545–554.

[72] J. Uhlmann, S. Julier, H. Durrant-Whyte, *et al.*, "A culminating advance in the theory and practice of data fusion, filtering and decentralized estimation", *Technical Report, Covariance Intersection Working Group (CIWG)*, 1997.

[73] K.-C. Chang, C.-Y. Chong, and S. Mori, "On scalable distributed sensor fusion", in *11th International Conference on Information Fusion*, IEEE, 2008, pp. 1–8.

[74] A. Marrs, C. Reed, A. Webb, and H. Webber, "Data incest and symbolic information processing, united kingdom defence evaluation and research agency", Tech. Rep, Tech. Rep., 1999.

[75] S. P. McLaughlin, R. J. Evans, and V. Krishnamurthy, "Data incest removal in a survivable estimation fusion architecture", in *Proceedings of the International Conference on Information Fusion*, 2003, pp. 229–236.

[76] L. Y. Pao and M. Kalandros, "Algorithms for a class of distributed architecture tracking", in *Proceedings of the 1997 American Control Conference*, IEEE, vol. 3, 1997, pp. 1434–1438.

[77] S. P. McLaughlin, R. J. Evans, and V. Krishnamurthy, "A graph theoretic approach to data incest management in network centric warfare", in *7th International Conference on Information Fusion*, IEEE, vol. 2, 2005, 8–pp.

[78] T. Bréhard and V. Krishnamurthy, "Optimal data incest removal in bayesian decentralized estimation over a sensor network", in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, IEEE, vol. 3, 2007, pp. III–173.

[79] W. Khawsuk and L. Y. Pao, "Decorrelated state estimation for distributed tracking of interacting targets in cluttered environments", in *Proceedings of the 2002 American Control Conference*, IEEE, vol. 2, 2002, pp. 899–904.

[80] D. Macii, A. Boni, M. De Cecco, and D. Petri, "Tutorial 14: Multisensor data fusion", *IEEE instrumentation & measurement magazine*, vol. 11, no. 3, pp. 24–33, 2008.

[81] M. B. Hurley, "An information theoretic justification for covariance intersection and its generalization", in *Proceedings of the Fifth International Conference on Information Fusion*, IEEE, vol. 1, 2002, pp. 505–511.

[82] L. Chen, P. O. Arambel, and R. K. Mehra, “Estimation under unknown correlation: Covariance intersection revisited”, *IEEE Transactions on Automatic Control*, vol. 47, no. 11, pp. 1879–1882, 2002.

[83] W. Niehsen, “Information fusion based on fast covariance intersection filtering”, in *Proceedings of the Fifth International Conference on Information Fusion*, IEEE, vol. 2, 2002, pp. 901–904.

[84] D. Franken and A. Hupper, “Improved fast covariance intersection for distributed data fusion”, in *7th International Conference on Information Fusion*, IEEE, vol. 1, 2005, 7–pp.

[85] Z. M. Durovic and B. D. Kovacevic, “Qq-plot approach to robust kalman filtering”, *International Journal of Control*, vol. 61, no. 4, pp. 837–857, 1995.

[86] M. Kumar, D. P. Garg, and R. A. Zachery, “A method for judicious fusion of inconsistent multiple sensor data”, *IEEE Sensors Journal*, vol. 7, no. 5, pp. 723–733, 2007.

[87] Y. Bar-Shalom, H. Chen, and M. Mallick, “One-step solution for the multistep out-of-sequence-measurement problem in tracking”, *IEEE Transactions on aerospace and electronic systems*, vol. 40, no. 1, pp. 27–37, 2004.

[88] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems* (Artech House radar library). Artech House, 1999, ISBN: 9781580530064. [Online]. Available: `https://books.google.gr/books?id=lTIfAQAAIAAJ`.

[89] Y. Bar-Shalom, “Update with out-of-sequence measurements in tracking: Exact solution”, *IEEE Transactions on aerospace and electronic systems*, vol. 38, no. 3, pp. 769–777, 2002.

[90] M. Mallick, S. Coraluppi, and C. Carthel, “Advances in asynchronous and decentralized estimation”, in *IEEE Aerospace Conference Proceedings*, IEEE, vol. 4, 2001, pp. 4–1873.

[91] K. Zhang and X. R. Li, “Optimal update with out-of-sequence measurements for distributed filtering”, in *Proceedings of the Fifth International Conference on Information Fusion*, IEEE, vol. 2, 2002, pp. 1519–1526.

[92] K. Zhang, X. R. Li, and Y. Zhu, “Optimal update with out-of-sequence measurements”, *IEEE Transactions on Signal Processing*, vol. 53, no. 6, pp. 1992–2004, 2005.

[93] M. Mallick, J. Krant, and Y. Bar-Shalom, “Multi-sensor multi-target tracking using out-of-sequence measurements”, in *Proceedings of the Fifth International Conference on Information Fusion*, IEEE, vol. 1, 2002, pp. 135–142.

[94] S. Challa and J. A. Legg, “Track-to-track fusion of out-of-sequence tracks”, in *Proceedings of the Fifth International Conference on Information Fusion*, IEEE, vol. 2, 2002, pp. 919–926.

[95] L. A. Zadeh, “Review of a mathematical theory of evidence”, *AI Magazine*, vol. 5, no. 3, p. 81, Sep. 1984. DOI: `10.1609/aimag.v5i3.452`.

[96] M. C. Florea, A.-L. Jousselme, É. Bossé, and D. Grenier, “Robust combination rules for evidence theory”, *Information Fusion*, vol. 10, no. 2, pp. 183–197, 2009.

[97] R. R. Yager, “On the dempster-shafer framework and new combination rules”, *Information sciences*, vol. 41, no. 2, pp. 93–137, 1987.

[98] P. Smets, “The combination of evidence in the transferable belief model”, *IEEE Transactions on pattern analysis and machine intelligence*, vol. 12, no. 5, pp. 447–458, 1990.

[99] J. Dezert, "Foundations for a new theory of plausible and paradoxical reasoning", *Information and Security*, vol. 9, pp. 13–57, 2002.

[100] F. Voorbraak, "On the justification of dempster's rule of combination", *Artificial Intelligence*, vol. 48, no. 2, pp. 171–197, 1991.

[101] J. K. Uhlmann, "Covariance consistency methods for fault-tolerant distributed data fusion", *Information Fusion*, vol. 4, no. 3, pp. 201–215, 2003.

[102] S. Maskell, "A bayesian approach to fusing uncertain, imprecise and conflicting information", *Information Fusion*, vol. 9, no. 2, pp. 259–277, 2008.

[103] S. Challa, T. Gulrez, Z. Chaczko, and T. Paranesha, "Opportunistic information fusion: A new paradigm for next generation networked sensing systems", in *7th international conference on information fusion*, IEEE, vol. 1, 2005, 8–pp.

[104] C. Wu and H. Aghajan, "Model-based human posture estimation for gesture analysis in an opportunistic fusion smart camera network", in *IEEE Conference on Advanced Video and Signal Based Surveillance*, IEEE, 2007, pp. 453–458.

[105] A. D. Tafti and N. Sadati, "Novel adaptive kalman filtering and fuzzy track fusion approach for real time applications", in *3rd IEEE Conference on Industrial Electronics and Applications*, IEEE, 2008, pp. 120–125.

[106] X. Huang and S. Oviatt, "Toward adaptive information fusion in multimodal systems", in *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers 2*, Springer, 2006, pp. 15–27.

[107] M. M. Kokar, K. Baclawski, and H. Gao, "Category theory-based synthesis of a higher-level fusion algorithm: An example", in *9th International Conference on Information Fusion*, IEEE, 2006, pp. 1–8.

[108] V. Nimier, "Introducing contextual information in multisensor tracking algorithms", in *Advances in Intelligent Computing—IPMU'94: 5th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems Paris, France, July 4–8, 1994 Selected Papers 5*, Springer, 1995, pp. 595–604.

[109] B. Yu and K. Sycara, "Learning the quality of sensor data in distributed decision fusion", in *9th International Conference on Information Fusion*, IEEE, 2006, pp. 1–8.

[110] F. Delmotte, L. Dubois, and P. Borne, "Context-dependent trust in data fusion within the possibility theory", in *1996 IEEE International Conference on Systems, Man and Cybernetics. Information Intelligence and Systems*, IEEE, vol. 1, 1996, pp. 538–543.

[111] S. A. Sandri, D. Dubois, and H. W. Kalfsbeek, "Elicitation, assessment, and pooling of expert judgments using possibility theory", *IEEE transactions on fuzzy systems*, vol. 3, no. 3, pp. 313–335, 1995.

[112] R. Haenni and S. Hartmann, "Modeling partially reliable information sources: A general approach based on dempster–shafer theory", *Information fusion*, vol. 7, no. 4, pp. 361–379, 2006.

[113] Z. Elouedi, K. Mellouli, and P. Smets, "Assessing sensor reliability for multisensor data fusion within the transferable belief model", *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 1, pp. 782–787, 2004.

[114] E. J. Wright and K. B. Laskey, "Credibility models for multi-source fusion", in *9th International Conference on Information Fusion*, IEEE, 2006, pp. 1–7.

[115] U. Shin, J. Park, and I. S. Kweon, "Deep depth estimation from thermal image", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1043–1053.

[116] M. James and C. McMichael, "Micro air vehicles–toward a new dimension in flight", *US DAPPA/TTO Report*, 1997.

[117] F. Santoso, M. Liu, and G. Egan, "Linear quadratic optimal control synthesis for a uav", in *12th Australian International Aerospace Congress, AIAC12, Melbourne, Australia*, 2007.

[118] F. Santoso, M. Liu, and G. Egan, "H2 and h_∞ robust autopilot synthesis for longitudinal flight of a special unmanned aerial vehicle: A comparative study", *IET Control Theory & Applications*, vol. 2, no. 7, pp. 583–594, 2008.

[119] M. Liu, G. K. Egan, and F. Santoso, "Modeling, autopilot design, and field tuning of a uav with minimum control surfaces", *IEEE Transactions on Control Systems Technology*, vol. 23, no. 6, pp. 2353–2360, 2015.

[120] M. A. Garratt and J. S. Chahl, "Vision-based terrain following for an unmanned rotorcraft", *Journal of Field Robotics*, vol. 25, no. 4-5, pp. 284–301, 2008.

[121] M. A. Garratt and A. Cheung, "Obstacle avoidance in cluttered environments using optic flow", in *Australian Conference on Robotics and Automation*, 2009.

[122] M. A. Garratt, A. J. Lambert, and H. Teimoori, "Design of a 3d snapshot based visual flight control system using a single camera in hover", *Autonomous Robots*, vol. 34, pp. 19–34, 2013.

[123] M. Garratt and J. Chahl, "An optic flow damped hover controller for an autonomous helicopter", in *Proceedings of the 22nd International UAV Systems Conference, Bristol, UK*, 2007, pp. 16–18.

[124] W. E. Green, P. Y. Oh, and G. Barrows, "Flying insect inspired vision for autonomous aerial robot maneuvers in near-earth environments", in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, IEEE, vol. 3, 2004, pp. 2347–2352.

[125] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman filter: Particle filters for tracking applications*. Artech house, 2003.

[126] L. Smith and V. Aitken, "The auxiliary extended and auxiliary unscented kalman particle filters", in *2007 Canadian Conference on Electrical and Computer Engineering*, IEEE, 2007, pp. 1626–1630.

[127] G. Evensen, *Data Assimilation: The Ensemble Kalman Filter*. Springer Berlin Heidelberg, 2006, ISBN: 9783540383017. [Online]. Available: `https://books.google.gr/books?id=VJ2oOecHhOYC`.

[128] A. Budhiraja, L. Chen, and C. Lee, "A survey of numerical methods for nonlinear filtering problems", *Physica D: Nonlinear Phenomena*, vol. 230, no. 1-2, pp. 27–36, 2007.

[129] K. Y. Leung, F. Inostroza, and M. Adams, "An improved weighting strategy for rao-blackwellized probability hypothesis density simultaneous localization and mapping", in *2013 International Conference on Control, Automation and Information Sciences (ICCAIS)*, IEEE, 2013, pp. 103–110.

[130] D. J. Allerton and H. Jia, "A review of multisensor fusion methodologies for aircraft navigation systems", *The Journal of Navigation*, vol. 58, no. 3, pp. 405–417, 2005.

[131] P. Zhan, D. W. Casbeer, and A. L. Swindlehurst, “A centralized control algorithm for target tracking with uavs”, in *Conference Record of the Thirty-Ninth Asilomar Conference on Signals, Systems and Computers, 2005.*, IEEE, 2005, pp. 1148–1152.

[132] Y. Zhai, M. B. Yeary, J. P. Havlicek, and G. Fan, “A new centralized sensor fusion-tracking methodology based on particle filtering for power-aware systems”, *IEEE Transactions on Instrumentation and Measurement*, vol. 57, no. 10, pp. 2377–2387, 2008.

[133] N. A. Carlson, “Federated filter for distributed navigation and tracking applications”, in *Proceedings of the 58th Annual Meeting of the Institute of Navigation and CIGTF 21st Guidance Test Symposium*, 2002, pp. 340–353.

[134] D. Simon, *Optimal State Estimation: Kalman, H_∞, and Nonlinear Approaches.* Wiley, 2006, ISBN: 9780470045336. [Online]. Available: `https://books.google.gr/books?id=UiMVoP%5C_7TZkC`.

[135] S. Chen, Y. Li, and N. M. Kwok, “Active vision in robotic systems: A survey of recent developments”, *The International Journal of Robotics Research*, vol. 30, no. 11, pp. 1343–1377, 2011.

[136] H. Lim, J. Lim, and H. J. Kim, “Real-time 6-dof monocular visual slam in a large-scale environment”, in *2014 IEEE international conference on robotics and automation (ICRA)*, IEEE, 2014, pp. 1532–1539.

[137] P. Gohl, M. Burri, S. Omari, *et al.*, “Towards autonomous mine inspection”, in *Proceedings of the 2014 3rd International Conference on Applied Robotics for the Power Industry*, IEEE, 2014, pp. 1–6.

[138] F. Santoso, M. A. Garratt, and S. G. Anavatti, “Visual–inertial navigation systems for aerial robotics: Sensor fusion and technology”, *IEEE Transactions on Automation Science and Engineering*, vol. 14, no. 1, pp. 260–275, 2016.

[139] J. Stowers, M. Hayes, and A. Bainbridge-Smith, “Altitude control of a quadrotor helicopter using depth map from microsoft kinect sensor”, in *2011 IEEE International Conference on Mechatronics*, IEEE, 2011, pp. 358–362.

[140] F. Santoso, M. A. Garratt, M. R. Pickering, and M. Asikuzzaman, “3d mapping for visualization of rigid structures: A review and comparative study”, *IEEE Sensors Journal*, vol. 16, no. 6, pp. 1484–1507, 2015.

[141] F. Wang, J. Cui, S. K. Phang, B. M. Chen, and T. H. Lee, “A mono-camera and scanning laser range finder based uav indoor navigation system”, in *2013 International Conference on Unmanned Aircraft Systems (ICUAS)*, IEEE, 2013, pp. 694–701.

[142] L. Yu, Q. Fei, and Q. Geng, “Combining zigbee and inertial sensors for quadrotor uav indoor localization”, in *2013 10th IEEE International Conference on Control and Automation (ICCA)*, IEEE, 2013, pp. 1912–1916.

[143] P. Corke, “An inertial and visual sensing system for a small autonomous helicopter”, *Journal of robotic systems*, vol. 21, no. 2, pp. 43–51, 2004.

[144] L. Meier, P. Tanskanen, L. Heng, G. H. Lee, F. Fraundorfer, and M. Pollefeys, “Pixhawk: A micro aerial vehicle design for autonomous flight using onboard computer vision”, *Autonomous Robots*, vol. 33, pp. 21–39, 2012.

[145] D. Abeywardena, Z. Wang, S. Kodagoda, and G. Dissanayake, “Visual-inertial fusion for quadrotor micro air vehicles with improved scale observability”, in *2013 IEEE International Conference on Robotics and Automation*, IEEE, 2013, pp. 3148–3153.

[146] T. Mouats, N. Aouf, and M. A. Richardson, “A novel image representation via local frequency analysis for illumination invariant stereo matching”, *IEEE Transactions on Image Processing*, vol. 24, no. 9, pp. 2685–2700, 2015.

[147] S. Vidas, R. Lakemond, S. Denman, C. Fookes, S. Sridharan, and T. Wark, “An exploration of feature detector performance in the thermal-infrared modality”, in *2011 International Conference on Digital Image Computing: Techniques and Applications*, IEEE, 2011, pp. 217–224.

[148] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors”, *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, pp. 1615–30, Nov. 2005.

[149] K. Mikolajczyk, T. Tuytelaars, C. Schmid, *et al.*, “A comparison of affine region detectors”, *International journal of computer vision*, vol. 65, pp. 43–72, 2005.

[150] N. Sünderhauf, K. Konolidge, T. Lemaire, and S. Lacroix, “Comparison of stereovision odometry approaches”, in *Workshop Planetary Rovers, IEEE International Conference on Robotics and Automation (ICRA05)*, 2005.

[151] S.-H. Jung, J. Eledath, S. Johansson, and V. Mathevon, “Egomotion estimation in monocular infra-red image sequence for night vision applications”, in *2007 IEEE Workshop on Applications of Computer Vision (WACV’07)*, IEEE, 2007, pp. 8–8.

[152] R. Chatila and J. Laumond, “Position referencing and consistent world modeling for mobile robots”, in *Proceedings. 1985 IEEE International Conference on Robotics and Automation*, IEEE, vol. 2, 1985, pp. 138–145.

[153] T. Mouats, N. Aouf, A. D. Sappa, C. Aguilera, and R. Toledo, “Multispectral stereo odometry”, *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 3, pp. 1210–1224, 2014.

[154] K. Owens and L. H. Matthies, “Passive night vision sensor comparison for unmanned ground vehicle stereo vision navigation”, *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings*, vol. 1, 122–131 vol.1, 2000.

[155] T. B. Schon and J. Roll, “Ego-motion and indirect road geometry estimation using night vision”, in *2009 IEEE Intelligent Vehicles Symposium*, IEEE, 2009, pp. 30–35.

[156] A. Rankin, A. Huertas, L. Matthies, *et al.*, “Unmanned ground vehicle perception using thermal infrared cameras”, in *Unmanned Systems Technology XIII*, Spie, vol. 8045, 2011, pp. 19–44.

[157] K. Hajebi and J. S. Zelek, “Structure from infrared stereo images”, in *2008 Canadian Conference on Computer and Robot Vision*, IEEE, 2008, pp. 105–112.

[158] S. J. Krotosky and M. M. Trivedi, “On color-, infrared-, and multimodal-stereo approaches to pedestrian detection”, *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 4, pp. 619–629, 2007.

[159] T. Luhmann, J. Piechel, and T. Roelfs, “Geometric calibration of thermographic cameras”, *Thermal infrared remote sensing: sensors, methods, applications*, pp. 27–42, 2013.

[160] P. Engström, H. Larsson, and J. Rydell, “Geometric calibration of thermal cameras”, *Proc SPIE*, Oct. 2013.

[161] S. Vidas, R. Lakemond, S. Denman, C. Fookes, S. Sridharan, and T. Wark, “A mask-based approach for the geometric calibration of thermal-infrared cameras”, *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 6, pp. 1625–1635, 2012.

[162] J. S. Zelek, M. Holbein, K. Hajebi, D. C. Asmar, and D. Cheng, “Ir depth from stereo for autonomous navigation”, in *Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XVI*, SPIE, vol. 5784, 2005, pp. 316–330.

[163] J. Harguess and S. Strange, “Infrared stereo calibration for unmanned ground vehicle navigation”, in *Unmanned Systems Technology XVI*, SPIE, vol. 9084, 2014, pp. 276–283.

[164] T. Mouats and N. Aouf, “Fusion of thermal and visible images for day/night moving objects detection”, in *2014 Sensor Signal Processing for Defence (SSPD)*, IEEE, 2014, pp. 1–5.

[165] C. Papachristos, F. Mascarich, and K. Alexis, “Thermal-inertial localization for autonomous navigation of aerial robots through obscurants”, in *2018 International Conference on Unmanned Aircraft Systems (ICUAS)*, IEEE, 2018, pp. 394–399.

[166] C. Papachristos, S. Khattak, and K. Alexis, “Autonomous exploration of visually-degraded environments using aerial robots”, in *2017 International Conference on Unmanned Aircraft Systems (ICUAS)*, IEEE, 2017, pp. 775–780.

[167] S. Vidas and S. Sridharan, “Hand-held monocular slam in thermal-infrared”, in *2012 12th International Conference on Control Automation Robotics & Vision (ICARCV)*, IEEE, 2012, pp. 859–864.

[168] P. V. K. Borges and S. Vidas, “Practical infrared visual odometry”, *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, pp. 2205–2213, 2016.

[169] C. Brunner, T. Peynot, T. Vidal-Calleja, and J. Underwood, “Selective combination of visual and thermal imaging for resilient localization in adverse conditions: Day and night, smoke and fire”, *Journal of Field Robotics*, vol. 30, no. 4, pp. 641–666, 2013.

[170] F. Burian, P. Kocmanova, and L. Zalud, “Robot mapping with range camera, ccd cameras and thermal imagers”, in *2014 19th International Conference on Methods and Models in Automation and Robotics (MMAR)*, IEEE, 2014, pp. 200–205.

[171] 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)*, IEEE, 2017, pp. 5682–5687.

[172] E. Emilsson and J. Rydell, “Chameleon on fire—thermal infrared indoor positioning”, in *2014 IEEE/ION Position, Location and Navigation Symposium-PLANS 2014*, IEEE, 2014, pp. 637–644.

[173] D. Borrmann, H. Afzal, J. Elseberg, and A. Nüchter, “Mutual calibration for 3d thermal mapping”, *IFAC Proceedings Volumes*, vol. 45, no. 22, pp. 605–610, 2012.

[174] M. Kamel, T. Stastny, K. Alexis, and R. Siegwart, “Model predictive control for trajectory tracking of unmanned aerial vehicles using robot operating system”, *Robot Operating System (ROS) The Complete Reference (Volume 2)*, pp. 3–39, 2017.

[175] Z. Zhang, “A flexible new technique for camera calibration”, *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[176] M. Gschwandtner, R. Kwitt, A. Uhl, and W. Pree, “Infrared camera calibration for dense depth map construction”, in *2011 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2011, pp. 857–862.

[177] P. Furgale, J. Rehder, and R. Siegwart, “Unified temporal and spatial calibration for multi-sensor systems”, in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2013, pp. 1280–1286.

[178] J. Solà, *Quaternion kinematics for the error-state kalman filter*, 2017. arXiv: `1711.02508 [cs.RO]`. [Online]. Available: `https://arxiv.org/abs/1711.02508`.

[179] R. Roberts, H. Nguyen, N. Krishnamurthi, and T. Balch, "Memory-based learning for visual odometry", in *2008 IEEE International Conference on Robotics and Automation*, IEEE, 2008, pp. 47–52.

[180] M. R. U. Saputra, P. P. De Gusmao, S. Wang, A. Markham, and N. Trigoni, "Learning monocular visual odometry through geometry-aware curriculum learning", in *2019 international conference on robotics and automation (ICRA)*, IEEE, 2019, pp. 3549–3555.

[181] Z. Teed and J. Deng, "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras", *Advances in neural information processing systems*, vol. 34, pp. 16 558–16 569, 2021.

[182] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow", in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, Springer, 2020, pp. 402–419.

[183] Y. Huang, B. Zhao, C. Gao, and X. Hu, "Learning optical flow with r-cnn for visual odometry", in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2021, pp. 14 410–14 416.

[184] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, ISSN: 0001-0782. DOI: `10.1145/3065386`. [Online]. Available: `https://doi.org/10.1145/3065386`.

[185] J. Nilsson and T. Akenine-Möller, "Understanding ssim", 2020. arXiv: `2006.13846 [eess.IV]`. [Online]. Available: `https://arxiv.org/abs/2006.13846`.

[186] C. Szegedy, W. Liu, Y. Jia, *et al.*, *Going deeper with convolutions*, 2014. arXiv: `1409.4842 [cs.CV]`. [Online]. Available: `https://arxiv.org/abs/1409.4842`.

[187] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art", *Information Fusion*, 2017.

[188] A. A. Goshtasby and S. G. Nikolov, "Guest editorial: Image fusion: Advances in the state of the art", *Information Fusion: Special Issue on Image Fusion: Advances in the State of the Art*, vol. 8, pp. 114–118, 2007.

[189] G. Pajares and J. M. De La Cruz, "A wavelet-based image fusion tutorial", *Pattern recognition*, vol. 37, no. 9, pp. 1855–1872, 2004.

[190] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone, "Remote sensing image fusion using the curvelet transform", *Information fusion*, vol. 8, no. 2, pp. 143–156, 2007.

[191] J. Adu, J. Gan, Y. Wang, and J. Huang, "Image fusion based on nonsubsampled contourlet transform for infrared and visible light image", *Infrared Physics & Technology*, vol. 61, pp. 94–100, 2013.

[192] M. Yin, W. Liu, X. Zhao, Y. Yin, and Y. Guo, "A novel image fusion algorithm based on nonsubsampled shearlet transform", *Optik*, vol. 125, no. 10, pp. 2274–2282, 2014.

[193] L. Liu, M. Song, Y. Peng, and J. Li, "A novel fusion framework of infrared and visible images based on rlnsst and guided filter", *Infrared Physics & Technology*, vol. 100, pp. 99–108, 2019.

[194] J. Poujol, C. A. Aguilera, E. Danos, B. X. Vintimilla, R. Toledo, and A. D. Sappa, "A visible-thermal fusion based monocular visual odometry", in *Robot 2015: Second Iberian Robotics Conference: Advances in Robotics, Volume 1*, Springer, 2016, pp. 517–528.

[195] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation", in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, Springer, 2018, pp. 3–11.

[196] A. Bruhn, "Variational optic flow computation: Accurate modelling and efficient numerics", *Department of Mathematics and Computer Science, Saarland University, Saarbrücken, Diss*, 2006.

[197] B. K. Horn and B. G. Schunck, "Determining optical flow", *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[198] J. J. Gibson, "The perception of the visual world.", 1950.

[199] O. Mac Aodha, A. Humayun, M. Pollefeys, and G. J. Brostow, "Learning a confidence measure for optical flow", *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 5, pp. 1107–1120, 2012.