HETEROGENOUS MULTIMODAL SENSOR FUSION VIA CANONICAL CORRELATION ANALYSIS AND EXPLAINABLE AI

by

ASAD VAKIL

A thesis submitted in partial fulfillment of the
requirements for the degree of

ELECTRICAL AND COMPUTER ENGINEERING

2020

Oakland University
Rochester, Michigan

Thesis Advisory Committee:

Jia Li, Ph.D., Chair
Daniel Aloi, Ph.D.
Manohar Das, Ph.D.
Robert Ewing, Ph.D.

ACKNOWLEDGMENTS

I would like to thank Oakland University for affording me the unimaginable opportunity to complete my study here, despite all the challenges which had frustrated all my efforts. I would especially like to express my deep and sincere gratitude to my research advisor, Professor Jia Li, whom for without her patience, guidance, and understanding I most definitely would not have made it this far, especially with my thesis. I also remain grateful and thankful to the aid of Dr. Erik Blasch and Professor Robert Ewing and the opportunities and insights they have provided. And of course, this research would not be possible without the support of the AFOSR grant FA9550-18-0287.

To my fellow researchers within the Li research group, especially Jennifer Liu, your help and insights have been invaluable to my work. To my friends and peers who have helped me thus far and supported me when I needed it most. And to Person of Interest, which is always a reminder of the many intricacies involved in machine learning, the potential heights that can possibly be reached in this field, the importance of turning off Bluetooth on your phone, and a slightly grimmer reminder that a benevolent AI is nothing but fiction. Naturally, as I would quite literally not be here without them, I would like to also thank my parents for their love and unconditional support. And their constant reminders that its way better to go through graduate school now, rather than 10 years down the road after having started a family and working full time. We may yet just have another Dr. Vakil in this family, I am working on it.

ABSTRACT

HETEROGENOUS MULTIMODAL SENSOR FUSION VIA CANONICAL CORRELATION ANALYSIS AND EXPLAINABLE AI

by

ASAD VAKIL

Adviser: Professor Jia Li, Ph.D.

The integration of heterogenous sensor modalities to facilitate multi-modal information fusion has many possible approaches and outcomes. Passive RF data, constructed from the In-phase and Quadrature component (I/Q) data processed via histogram and enhanced electro optical (EO) data via dense optical flow (DOF) are further enhanced with Canonical Correlation Analysis (CCA) in order to achieve object detection and tracking. In order to determine the impact of the P-RF histograms and the noticeable improvement of Canonical Correlation Analysis, Explainable AI is implemented in order to determine the weight and impact the canonical variates provide to the fusion model.

TABLE OF CONTENTS

TABLE OF CONTENTS—Continued

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

EO	"Electro Optical"
P-RF	"Passive Radio Frequency"
CCA	"Canonical Correlation Analysis"
CED	“Canny Edge Detection”
CM	“Confusion Matrix”
DOF	“Dense Optical Flow”

CHAPTER ONE

INTRODUCTION

1.1 Overview of Multimodal Fusion

In the information age, there exist many methods of collecting information on an environment or subject through sensing. Pressure, radar, acoustic, chemical, electromagnetic, thermal, proximity, and optical methods are but a start for the long list of the many methods by which information on a specific environment or target can be collected. Each of these independent modalities have their own individual strengths and weaknesses in information collection inherent to the method of sensing. Regardless of whether or not the modality is active or passive in nature, absolute or relative, or vulnerable to different forms of outside interference, or capable of providing different level of depth to the information obtained, the ability to utilize the information efficiently is invaluable for the reliable, credible, and robust performance of a system. To enhance robustness, sensor fusion, also referred to as information fusion or multi-view learning, seeks to reduce uncertainty by integrating multiple sources of information.

To better describe the improvement of performance that multimodal sensor fusion provides, the terms representation and completeness are the best descriptors of the advantages of fusion. Information obtained throughout the fusion process has an abstract level higher than the original individual input data. This combined representation allows

for a richer semantic and higher resolution on the data when compared to that of each initial source of information. Furthermore, by bringing in new information to the current knowledge of the environment, sensor fusion provides a more thorough view for decision making and classification.

If individual sensors only provide information that remains independent of other sensors, then by combining them into a more coherent space the fusion will provide an overarching view of the environment. The system's ability to discern patterns and identify discriminating features based on the information provided is increased due to this more comprehensive view. Even if the information is redundant or concordant the accuracy will still improve from this view. That being said the number of sensors employed is a factor in the cost analysis of whether or not a multi-sensor system is better than a single sensor system. An appropriate criterion has to be implemented in order to accurately assess the reliability and performance of the whole system. While it may not be possible to completely eliminate uncertainty in decision making or classification, the ability to produce more reliable and accurate results from an emergent view can be extremely beneficial. Pooling together information from other sources and views provides enhanced reliability, extended parameter coverage, and improved resolution to a system.

As an emerging field, sensor fusion research is far from achieving the level of the human mind when it comes to analyzing different sources of data simultaneously. There are a number of problems that can arise from feeding multiple sources of information continuously into a system. Data association, sensor uncertainty, and the management of the input data are the more common problems. In many cases these problems might arise

due to the inherent ambiguity of certain sensors, or due to factors such as device noise or even the ambiguity of the environment in which the information is being collected. The number of potential challenges will always be dependent on the application as will the optimal approach for using information that is based on that particular application or environment. The hallmark of a robust system of sensor fusion is the ability to handle such uncertainties and provide consistent results based on the information collected in that environment.

Most sensors do not directly generate a signal from an external phenomenon but are instead best summarized as a culmination of conversion steps. Because of the changes that occur, the output read by the user may deviate from the actual input the system receives, which can impact factors and characteristics of the system such as accuracy, precision, resolution, and sensitivity. Typically, these individual characteristics for different sensors can be managed before the fusion process in order to improve the performance of the system. Dynamic characteristics, however, may vary between changes in the input. The speed and frequency of response, settling time, and sensor lag are issues that might lead to inherent errors due to the sensor fusion architecture. How those problems are addressed will be important for the system's ability to properly utilize the different sensor inputs.

The methods of using the input information are not necessarily only limited by what type of sensor is providing the data. In the real world, a single object may have multiple, different representations in different domains. This is especially true for systems with heterogenous modalities, as innately different sensors have the potential to provide completely different views, providing different information on the environment

and the target that a single modality might not be able to use. Exploiting these different representations in different domains allows the information to potentially be used to achieve a number of goals. For applications such as detection, classification, and tracking via multi-view learning methods, such as machine learning and similar algorithms, how different representations of that object are used can be integral to multimodal fusion. One particular method that will be focused on in this thesis is the application of canonical correlation analysis (CCA) with respect to multi-modal fusion.

Canonical correlation analysis is a statistical method for reducing dimensionality between a pair of datasets, jointly learning a shared subspace between the two sources of data (aka views) and is considered a method of achieving multi-view learning. Canonical correlation analysis has been used in the field of sensor fusion in recent years [1] for a number of applications. A statistical method that focuses on maximizing the correlation between different sets of data has a versatile number of applications and is a tool that other fusion models can take capitalize on in order to improve overall performance. This can especially be useful when the exact relationship between different sensors is not completely known and might be more difficult to intuitively exploit.

This is especially the case in passive radio frequency (P-RF) and electro optical (EO) sensor fusion. Due to the nature of P-RF data, using the radio frequency data for applications that do not reconstruct relative position via methods such as radar, can be quite difficult to use. While there are a wide variety of methods to utilize EO data, even after synchronizing two modalities it is not always apparent what method should be used to achieve the goal of fusing the two sources of information together. The use of a powerful tool used in the fields of image enhancement [2] and cognitive radio [3]

naturally makes experimenting with the use of Canonical Correlation Analysis (CCA) for EO and P-RF sensor fusion highly desirable for object detection and tracking.

1.2 Machine Learning and Information Fusion

The integration of raw data and information fusion is a prevalent and necessary step in the information age. There exists a vast range of classical probabilistic data fusion techniques, but the inherent versatility of deep learning provides many potential benefits to many fields. When using nonlinear data or when prior knowledge does not necessarily provide an approach, the potential that deep learning has with methods such as unsupervised or end-to-end learning is extremely appealing. While the application of machine learning is far from an infallible problem solver, there are a number of situations in which the method is a viable solution.

Machine learning, under the right circumstances, has an almost endless amount of potential. While there are a number of applications that seem far and out of reach of what is available currently, there are many mundane and everyday examples of what deep learning can accomplish. In everyday utility and entertainment applications, passive AI can be seen, learning, and making decisions based on the input the user provides. From suggested videos on YouTube, Netflix, or Spotify to product recommendations or auto-fill suggestions from Google or iPhone or Android, it's frighteningly clear just how versatile and retentive the results can be. The ability to utilize particular algorithms to make computer systems "learn" by using given data without specific programming is what defines machine learning as a whole. Creating computer systems that can see, know, learn, and predict the world like a human being and having the end goal of eventually reaching the singularity.

Machine learning is traditionally divided into three classes based on its attributes for learning, unsupervised, supervised, and semi-supervised learning. If the information is completely labeled out, with every input having a labeled output, the algorithm seeks to construct a model to map the input to the output, this is referred to as supervised learning. Typically, supervised learning deals with applications such as classification or regression, includes methods such as Support Vector Machines (SVM) or Artificial Neural Networks (ANN). For unsupervised learning, algorithms are forced to extract features and patterns themselves, seeking similarity or distance of data inputs. Finding associations with their internalized heuristics, these clustering methods are referred to as unsupervised learning. An example of unsupervised learning would be K-means clustering. Typically, these applications seek to reduce the complexity of a problem or aid in matters such as feature selection. Finally, with regards to what lies between the two, semi-supervised learning is the halfway point between the two approaches, starting off with a series of labeled data points as well as some data points that are not known. The end goal of the semi-supervised model is to classify some of the unlabeled data using the labeled information set. These models seek to accomplish that which supervised models do, to predict a target value for a specific input data set. Applications for semi-supervised models include fields such as speech analysis or web content classification.

While deep learning applications and the versatility in what kind of problems they can solve sound great, there are certain matters that need to be addressed. Efficiency, quality, stability, robustness, extensibility, and even aspects such as privacy and transparency are important to consider. The run of the mill neural network requires a lot of input, and if a data model fusion cannot use its resources economically this becomes

an issue when the dataset available isn't particularly large. Determining how to best utilize the fusion network, whether its preprocessing data to improve the model's ability to find discriminating details and what kind of impact it has on information accuracy is an important consideration. The level at which fusion is implemented might also impact the data. Is the information lost from preprocessing worth the feature extraction? Is implementing data/low level fusion worth it? Or for the particular application is it better to simply pool together assessments of the data and implement decision level fusion instead? If an underlying environmental factor is changed, is the fusional quality still ensured? More importantly is the data fusion model capable of being improved or expanded beyond its initial training?

While there are many potential applications for data fusion and machine learning, ranging from fault detection, navigation, multi-target tracking, classification, discussing potential pitfalls for such fusion models is important. Data imperfection, data inconsistency, data confliction, data alignment, data heterogeneity, and the actual location of fusion are the best descriptors of the typical issues machine learning based fusion can run into. Data captured by sensors is often imprecise, and it is the incomplete and uncertain nature of such modalities that makes sensor fusion so appealing. Data inconsistency is a major issue for fusion models if they are incapable of distinguishing the reasons that cause the noises, or unable to remove or suppress such noise. With regards to systems that utilize belief functions, such as Dempster-Shafer theory, if problems that should be treated independently are erroneously integrated with each other, a representation error can occur, known as data confliction. For data captured from different sensors with different frames, if the information is not properly registered or

correlated with each other, then there is the issue of data alignment, the less than harmonious synchronization of information coming from different modalities.

Similar to data alignment, is data heterogeneity. If data of different types or formats have lower quality due to missing values, or high data redundancy or are ambiguous or just plain untruthful, this can be an issue in fusion architecture. Sometimes this error is syntactic in nature, in that the data sources are not communicating with each other. Other times the heterogeneity mismatch might be conceptual in nature, having differences in modeling the same domain of interest. Terminology might also be an issue after solving a syntactic error, with different variations in names when referring to the same entity from different data sources. Finally, after all of the issues with the individual modalities and communication and alignment are solved, there might still be the issue of fusion location. Data fusion might occur in a centralized system or locally in independent nodes, choosing between trading off matters such as reducing communication burden for data accuracy. Choosing how to balance fusion cost and fusion quality is an important choice to consider when designing a fusion model.

At the end of all of this, there remains one final important question to ask. Most methods of deep learning can produce impressive results. But using methods of deep learning such as neural networks essentially makes most of the process something of a blackbox. These hidden layers and nodes are not necessarily designed to be transparent and knowing what discriminating factors are being used to achieve its objective is important to consider. Which leads to the question of how does one know what weights or factors are being used for decision making? And why should that be an important factor for a fusion model?

1.3 The Argument for Explainable AI

The success of deep learning and neural networks has had a major impact in both research and industrial applications. Machine learning has reached a point in which its large impact and potential consequences cannot be ignored. Now, does the inability to differentiate between a cat and a dog sound particularly important? Probably not life threatening, at most maybe a threat to someone's GPA. But consider applications such as autonomous driving or applications in the medical field. It only takes even a momentary dysfunction in a computer vision algorithm for a car accident to occur or a single false negative in disease detection to massively impact a patient. This begs an important question about interpretability and explainability of machine learning algorithms. Who is accountable when things go wrong? How did things go wrong? If things are not going wrong, then do we know why? Can we leverage characteristics or inputs to further improve performance?

There are some who might argue that a "right to explanation" is unnecessary, sometimes even harmful, potentially even stifling innovation. For example, back in 2017 when the European Union announced its General Data Protection Regulation (GDPR) in order to ensure the right to obtain an explanation and a right to opt out, there were those who argued against such measures. These arguments largely consisted of points such as such regulations would do little to help actual consumers and instead would slow down the development and use of AI in Europe. Such critics were quick to point out that holding developers to such a standard would be unnecessary and even unfeasible. And to a degree such detractors would not be wrong. Fundamentally, many algorithms used in machine learning are not easily explainable. The output of a deep neural network is made

of many layers of computations, and the complexities between the input and what aspect of the input activates the neurons is difficult to determine in many cases. However, considering some of the uses such as algorithmic trading [4], medical diagnosis, and autonomous vehicles [5] and the major impact they can have on their consumers, it is clear that models that handle such important applications should have some level of interpretability.

When using a machine learning algorithm, there are a number of important things to consider when troubleshooting the results. It is not enough to look at the accuracy score and assume that the objective has been reached successfully. When looking at the results that the model has troubles with, it is important to ask why did the model output this prediction? Does that prediction align with the domain logic of the application that the model is supposed to solve? Or is the model treating irrelevant features that have more weight when it misidentifies an input? Also important is determining if the model's behavior is consistent among different subsets of data. If it is behaving differently when in theory the model should not, then it is important to compare and contrast the results in order to identify where the behavior is varying. It may be that the model is bias towards a specific feature that is exclusive to that part of the dataset, and if so then the bias should be removed as to avoid negatively impacting the performance. Determining what influences the predictions is important in order to reach a more desirable outcome for the model.

There are other aspects to interpretability that can also impact the design and training of such models. Due to the nature of how training data is setup or what information is provided, the model may be learning to search for features that might not

necessarily be relevant when compared with real data. For example, model trained to recognize the area of a hotel property based on image inputs might misclassify a swimming pool area with that of the bathroom. Without any form of explainability, it becomes difficult to determine how and why this error occurred, but with interpretability the reason might be obvious. With the application of Gradient Class Activation Maps, the heatmaps might indicate that the bias that caused the model to misclassify the swimming pool image as a bathroom image was in fact the metallic rails both areas share. Machine learning algorithms have been proven not to be immune to bias [6], and by using explainability and visualization techniques it becomes possible to reduce the uncertainty that comes with training a model that's essentially a blackbox.

1.4 Thesis Contribution to the Current State of Knowledge

EO/RF sensor fusion, canonical correlation analysis, and explainable AI are currently popular fields of study., Many of the applications lie in the field of detection and tracking. This master's thesis aims at researching the impact and effect of passive RF modalities when fused with EO data via canonical correlation analysis and explainable AI. While other research in the field uses RF modalities in a number of ways, such as human detection or activity recognition via RFID [7] or Wi-Fi [8], our research focuses on the use of passive RF data in the form of raw I/Q data. In the past, our group has had success in using I/Q data for a number of applications including human detection [9], and my previous work [10] used I/Q data with EO input in order to achieve detection and tracking for vehicles within a DIRSIG simulation.

For the purposes of this papers contributions to the related fields, on the theoretical aspect, the use of passive RF data in sensor fusion, specifically raw I/Q data is

not something that has come up in the related literature. With regards to the use of explainable AI with raw I/Q data, to the best of my knowledge there has not been any literature on the impact and how the data is used. While in terms of methodology, the use of Deep CCA via LSTM [11] and greedy algorithms are not unheard of for fusion and explainable AI respectively, their use in detection and tracking of autonomous vehicles with raw I/Q data has not been explored to the best of my knowledge.

In this thesis, the use of real-world data is used to produce a set of experiments to gauge the effectiveness of P-RF histograms generated from the I/Q data and dense optical flow enhanced images with canonical correlation analysis for tracking and detection of different vehicles. The design of the experiments is such that the explainable AI fusion model is compared with a fusion model that uses canonical correlation analysis and the same data input in order to achieve the same objective. With the use of greedy algorithms to recreate and provide transparency to the fusion process, the results of the research provide insights into the effectiveness of the three types of data input with respect to the prediction and decision-making process.

CHAPTER TWO

LITERATURE REVIEW

2.1 The Implementation of CCA in Sensor Fusion

Canonical Correlation Analysis was first published back in 1936 by Harold Hotelling [12], and has seen an exponential use in recent years [13]. Originally proposed for association between arithmetic potentials the work quickly grew to finding the best predictors among linear functions via maximizing the correlation coefficient between two sets. Following decades of development, and a soaring number of applications since the turn of the 21st century, Canonical Correlation Analysis has grown to be a popular application for many topics of interest.

There are several categories that CCA finds itself being used in. Multiview CCA, probabilistic CCA, deep CCA, kernel CCA, discriminative CCA, sparse CCA and locality preserving CCA. Multiview CCA is essentially an expansion of CCA from being limited to two-view processing into three or more views, finding the canonical variates for multiple data channels. Typically this can be divided into the following subcategories, pairwise correlation [14], zero order correlation [15], and high order correlation [16]. Pairwise correlation is defined by creating a common subspace between two or more groups of data and measuring all the possible pairwise correlations. Zero order correlation is achieved by pushing all individual views into a common representation while in comparison high order correlation is calculated directly among all the views by analyzing tensor. Essentially the zero order correlation approach pushes individual views to approach a latent identical variable. Rather than estimating pairwise correlation the

Tensor CCA approach, on the other hand, simultaneously calculates all views correlation in the high order level by constructing a covariance tensor.

Probabilistic CCA is an approach that produces a model that provides a probabilistic interpretation for a classical algebraic solution. Latent Variable Model [17] and Bayesian CCA [18]. Latent variable models interpret the canonical variate solutions from a probabilistic view. With this model, the two view observations are generated by a common latent variable. The Bayesian CCA probabilistic model adds conjugate prior distribution on the latent variable model. By doing so it can infer posterior distribution in a Bayesian manner. The Bayesian treatment provides a model that is robust towards smaller sample sizes and is easier for the extension and modification by changing the distribution assumptions. This approach does however suffer from the difficulty in producing covariance matrices for high dimensional settings, especially when the sample size is small.

Naturally, with the importance neural network models bring to machine learning, there is a series of CCA applications that use deep learning. Deep CCA comes in many popular forms, typically using a Deep Neural Network (DNN), Autoencoders, or Convolutional Neural Networks (CNN) [19]. The advantages of this approach are that compared to approaches such as Kernel CCA or locality preserving CCA, this nonlinear approach to using CCA is not restricted by a predefined kernel or local information. Rather than relying on handcrafted features, Deep CCA and similar approaches seek more complex nonlinear associations between two views observations by passing the information through a neural network. While the neural network approach provides a more flexible nonlinear mapping approach using a parametric method, the efficiency of

the correlation process is not necessarily guaranteed. Kernel CCA [20] as the name implies uses the kernel trick, in order to map the nonlinear data into a higher dimensional Hilbert feature space in which the data shows a strong linearity. In kernel methods, data is represented as functions or elements in the reproduced kernel Hilbert spaces. For Kernel CCA, the use of the kernel trick allows the implicit nonlinear transformation of the data from a high dimensional input space to be solved and extract a nonlinear relationship between the two views. Traditionally Kernel CCA is split between the regularization and non-regularization approaches in order to solve ill-conditioned matrices (coefficient matrix due to a small change in the constant coefficient resulting in a large change in the solution) or ill-posed (large condition number). Regularizations approaches the ill-posed Langrangean by simply adding a regularization term such as ridge-type regularization or Laplacian regularization, while non-regularization transfers the ill-condition problem to a different method, such as Principal Component Analysis.

Discriminative CCA [21] applications are widely split between global and local discriminative methods. Discriminative CCA utilizes the label information into the CCA framework in a way such that it learns the correlation matrix and minimizes within-class separation while between-class variation is maximized. Global Discriminative CCA seeks to maximize intra-class similarity while minimizing inter-class closeness, while Local Discriminative CCA on the other hand defines the discriminative information in terms of local neighbors. This approach allows the local variation to potentially avoid the impact of outliers on performance. Sparse CCA [22] discovers interpretable associations in a high dimensional Multiview array. The method applies a penalty function to overcome the issue of variables that are too highly correlated, making covariance matrices

potentially unstable or undefined. The approach is typically divided between element and group level Sparse CCA. Element level Sparse CCA applies the penalty function on individual variables in order to reduce some of the elements of the solution to zero. Group level Sparse CCA on the other hand applies the penalty function to a group of structural information based on the input data, providing potentially more meaningful results in practice.

Locality Preserving CCA [23] is an application of CCA that is based off of the assumption that if data points are closed in the input high dimensional space, then following their projection into low dimensional space the data points should still be close. Using CCA as the basis for determining correlation between neighbors, the trivial correlations are removed. Locality Preserving CCA is traditionally divided between the deleting and adding strategy variations. In the deleting strategy, the local structure is preserved by eliminating redundancy information between weakly correlated datapoints. For the adding strategy, the local association is highlighted by strengthening intrinsic connection between strongly correlated samples.

Given the major impact and diversity in the methods used for multi-modal fusion, the appeal of CCA becomes obvious for high dimensional data. The applications for CCA range heavily in fusion, advertisement classification, phoneme classification, facial image matching, multi-label classification, multi-annotations fusing, word embedding, object and text classification, and even fields such as emotion or human action recognition are all various applications of CCA that require multiple sources of data. For this reason, especially given the somewhat ambiguous nature of the passive RF data used in experimentation, the application of CCA is utilized.

2.2 Passive Radio Frequency Data and Electro Optical Fusion

When it comes to modalities that involve radio frequency (RF) and electro-optical (EO) sensors, the focus for fusion research has traditionally been on active RF sensors and applications. Whether it is the fusion of Synthetic Aperture Radar (SAR) and multi-spectral images [24] via random forest classifier or double weighted decision level neural network fusion scheme [25], many EO/RF sensor fusion applications normally use active RF in one form or another. Doppler radar and imaging radar (e.g., side-looking airborne radar), as well as other similar active RF sensors, are well suited for tracking a moving target, a combination that is highly desirable when successfully fused with EO input. The combined view and the exploitation of the two types of sensor modalities still has room for improvement [26] however. RF based modalities excel in providing range, angular, and spectral resolution of information from RF modalities and the benefits of combining that data with higher spatial resolution of EO based sensors is extremely desirable for detection and tracking of targets in a number of environments.

There are a number of RF based modalities that are used in applications such as tracking, proximity, localization, and detection. While many EO modalities are intuitively easier for humans to understand and to implement for similar applications, unlike RF modalities, RF approaches to such applications are less susceptible to outer factors such as ocular interference. RF-based sensors are not limited or obscured by factors like visual interference from natural phenomenon such as fog, clouds, snow, or any other form of weather that would otherwise normally interfere in the collection of EO data. In addition to this, RF based sensors can provide repetitive coverage over a wide geographical area, and in doing so can determine the precise distance and velocity of a target.

While most research is focused on active RF applications, there are a number of advantages for the implementation of passive RF modalities such as passive radar or RFID over the use of active methods. Passive RF modalities are difficult to detect, typically having lower power requirements and have lower costs than the ones associated with the construction and usage of active radar, and are harder to implement countermeasures against, such as jamming and spoofing which can corrupt the collection of RF-based modalities and transmitted imagery. Combining the two modalities improves the overall reliability and have been implemented in a few applications for target detection, estimation, and tracking, summarized below in table 1.

For the research in [27], support vector machine is used as a final method of classification in order to achieve sense-and-avoid for unmanned aircraft. The purpose of the fusion is to use two complementary instruments, passive radar, and an EO/IR system to not only the detection of aircraft, but also the identification of the model and relative threat to the unmanned aircraft. The architecture for fusion first preprocesses the thermal and visible images, isolating the propulsion and aircraft before extracting the characteristics. These features are then correlated with the relative distance and

Table 1. Tracking via EO/Passive RF Sensor Fusion

Input Data	Method	References
Passive Radar and EO/IR sensor input	Unmanned Aircraft Vehicle sense and avoid application using SVM classifier	[27]
Full Motion Video and Passive RF	Sheaf-based heterogeneous sensor fusion using passive RF collected via Doppler Radar and Full Motion Video for target detection and tracking.	[24]
Full Motion Video and Passive RF	Joint Manifold Learning based heterogeneous data fusion approach to form a joint sensor data manifold for vehicle detection and tracking.	[28]
Full Motion Video and Passive RF	Deep learning approach using feature manifold representations for multi-object tracking and detection.	[29]
Full Motion Video and Passive RF	Autoencoder based Dynamic Deep Directional-unit network to achieve unsupervised upstream sensor fusion for the detection and tracking of vehicles.	[30]

orientation of the radar return, and subsequently to create a multispectral aircraft signature, which is used as an input for the SVM classifier.

The use of an autoencoder-based dynamic deep directional-unit network [30] was capable of learning compact, abstract feature representations from the high dimensional spatiotemporal data of full motion video and I/Q data for the purposes of event behavior characterization. The architecture exploits the access to elements of interest within regions of interest using temporal tracking and supervised classification before being fed into a decentralized supervised discrimination layer that applies Bayesian program learning in order to implement upstream multi-modal data fusion. Among the network's achievements in this paper, a notable benefit of the approach is that the network is capable of reconstructing missing modalities given the observed signatures.

Other research into achieving EO/RF fusion for vehicle tracking and detection using Full Motion Video and P-RF include joint manifold learning [31], sheaf based approach with its data [32], SVM classifier [27]. In [31] and [32], the use of simulation data is used for the primary method of training and testing, while in [27] real data collected from Daytona Beach International Airport is used. In [31], the use of a joint manifold learning fusion approach is used for the mixed simulation data. The use of digital imaging and remote sensing image generation in a DIRSIG dataset provides video measurements and three distributed RF sensors. The intrinsic low-dimensional data, the 2D images of the vehicles, are extracted by manifold learning algorithms from high dimensional data by implementing a linear transformation of the vehicle positions. The RF data is similarly handled by manifold learning, and then the implementation of linear regression is used for tracking. These results were compared with a number of methods,

such as maximally collapsing metric learning or neighborhood preserving embedding, calculating position errors with respect to the ground truth after implementing noise.

Finally, in [32], the use of simulated multi-sensor data is used to locate a moving emitter. The method of fusion implemented is Sheaf Theory, a tool for systematically tracking locally defined data attached to the open sets of a topological set. For the purposes of implementing sensor fusion, the data samples and model of data are used as the inputs of a sheaf based fusion architecture. The model of the data is used to construct the sheaf while the data samples are converted into samples for partial assignment. The outputs of the two are then used to search over the global sections using the optimizer, before using the results to report values over the stalks. During testing, the observed stalks for each sensor measure the real time offset and complex I/Q samples for each RF sensor while the EO data is collected, keeping the xy-location of each detected pixel for the video input. For the modeled stalks, which provide a comparison for each pair of sensors, the time offset relative to the video detection and the time aligned I/Q samples for each group of RF sensors are tracked. The third vertex tracks the true location of the ground truth, the emitter and transmitted signal.

2.3 Explainable AI

As explainable AI is an emerging field in research and industry, there has yet to be a widely adopted standard, let alone a widely adopted method of quantifying interpretability for explaining models. Even discussing methods of how to quantify such approaches is quite a task in of itself, as there are a number of different classifications for such types of interpretability. To even begin discussing the topic of explainable AI, the most important thing to do is to define interpretability.

There are a number of definitions of interpretability or explainability. For domains that deal heavily in images for example, interpretability might be defined as being able to map the predicted class into a domain that the human user might be able to make sense of [33]. In an ideal system, one might even define interpretability as a reasonable explanation as to why a collection of features contributed to the decision-making process, or at least determining how much weight the decision-making process gave to said features [34]. Whichever definition of interpretability one might subscribe to, given the lack of a widely adopted standard, so long as the method provides insights that can answer questions regarding how and why the model performs in the way that it does, that is a method that provides some level of transparency.

From saliency maps to activation maximization, there are a number of methods by which interpretability can be achieved. The distinctions between these types of methods are typically twofold, described as either Ante-hoc or Post-hoc, local or global, or model specific or model agnostic [35]. Ante-hoc and Post-hoc describe an intrinsically interpretable model from different approaches. Ante-hoc systems provide explanations from the beginning of the model. These types of systems enable one to gauge how certain a neural network is about its predictions. An example of an Ante-hoc system includes the Bayesian Rule List [36], a generative model that yields posterior distribution over decision lists consisting of a series of if-then-statements. Other examples can include visualization methods, saliency mask, rule extraction, and even neurons activation. Post-hoc techniques entail creating explainability into a model based on its outcome, marking the part of the input data responsible for the final decision. Like Ante-hoc techniques, this also can include visualization and saliency mapping, but also uses methods such as

gradients or feature importance. These methods are more easily applied in comparison to different models but say less about the whole model in general.

Similar to but not quite the same descriptors are local and global interpretability. Local provides explanations for only each single prediction while global explains the logic for the whole system, from input to every possible outcome. Methods such as Grad-CAM [37] are an example of local interpretability systems, using global average pooling and heat maps of a pre-softmax layer in order to determine the regions of an image responsible for prediction. Lastly, model specific and model agnostic describe the usability of different aspects of the system, with model agnostic being indifferently usable while model specific is tied to a particular type of blackbox or data.

For methods that are Post-hoc oriented solutions to explainability, there exists a number of methods for image based neural networks. Visualizations [38], gradients [39], activation maximizations [40], deconvolutions [41], and decomposition [42] are common approaches. Visualizations techniques will typically use tools such as generative models or saliency maps in order to determine activations produced on each layer of a trained CNN or DNN after processing an image or video. The visualization of the key neurons or neuron layers highlights the responsible features that lead to a maximum activation or the highest possible probability of prediction. Gradients and variants of guided backpropagation similarly emphasize the important unit changes, drawing to attention sensitive features or input data areas.

With these techniques it can potentially be possible to produce artificial prototype class member images that maximize the neuron activation or class confidence values. Deconvolution, sometimes referred to as inverting DNNs [43], can be applied to create

special typical inputs or parts of an input. These special inputs are created to fit the desired output of the network, producing a special layer or single unit in order to recreate the results. Finally, decomposition, also known as isolation, transfer, or limitation of portions of networks can provide further insights into which way single parts of the architecture influence the output layer. Methods such as Automatic Rule Extraction [44] and Decisions Trees are similar to the application of Deep Taylor Decomposition.

CHAPTER THREE

DESIGN AND METHODOLOGY

3.1 The ESCAPE Dataset

In 2019, Air Force Research Laboratory (AFRL) and Michigan Tech Research Institute (MTRI), released their Experiments, Scenarios, Concept of Operations, and Prototype Engineering data set (ESCAPE) [45]. The dataset is a versatile toolkit of different sensor modalities and scenarios that include, infrared (IR), full motion video (FMV), passive RF data, acoustic, seismic, and active radar imagery data. This information is collected via a number of sources, with the majority of the data being collected by portable or preexisting towers that remain stationary during testing. The more mobile sources of data are collected by repurposed DJI M600 Small Unmanned Air Systems (SUAS) Vertical Take Off and Landing (VTOL) aircraft, designated as Echo and Sierra respectively. The two carry payloads that include a Ettus B200 software defined radio receiver and LP0965 Log Periodic Antenna for the collection of RF data, a GPS, an FLIR Vue Pro-R Radiometric thermal camera and Basler ace acA3800-14uc camera for Infrared (IR) and EO data.

The primary advantage of using the dataset is not only the number of options for each of its scenarios but also the design of the ESCAPE dataset. The multi-source data collection comes with a number of vantage points from which information on various ground targets can be used for data fusion research. The design of the dataset enhances the complexity and opportunity of such research by increasing the number of modalities available in addition to outdoor experimental irregularities. In this dataset various ground

vehicles are witnessed by available sensors as leaving the observed scene, then potentially reemerging, thus “escaping” detection and tracking.

There is a total of five different types of ground vehicles used in the dataset, a gas motor Gator utility vehicle, a diesel motor Gator utility vehicle, a pickup truck, a panel van, and a stake rack truck. It should be noted that between the Gator vehicles the diesel powered gator had different acoustical and seismic signatures due to the nature of its propulsion system, despite how relatively similar the John Deer vehicles look compared to the trucks or vans. These five vehicles are the primary focus of the ESCAPE dataset, and by design are always the aforementioned targets of the dataset.

For the multimodal heterogeneous EO/P-RF sensor fusion research presented in this paper the raw RF data is preprocessed to obtain I/Q histograms with respect to the time. The histograms are then aligned with the simulated EO data for the purposes of detecting and discriminating between the different vehicles in each scenario. This P-RF data comes from three sources, designated as points 11, 12, and 13 for the MTRI P-RF sensors used to collect the I/Q Data. While the ESCAPE dataset has a total of nine scenarios, for the purposes of experimentation, the research presented in this thesis uses scenarios 1, 2C, and 2D. These scenarios were picked specifically for a number of reasons. Based on results from earlier research with the dataset, the EO input provided by the two SUAS produced less than acceptable results in terms of accuracy, but also found that the MTRI EO sensor designated as 04 provided the optimal results.

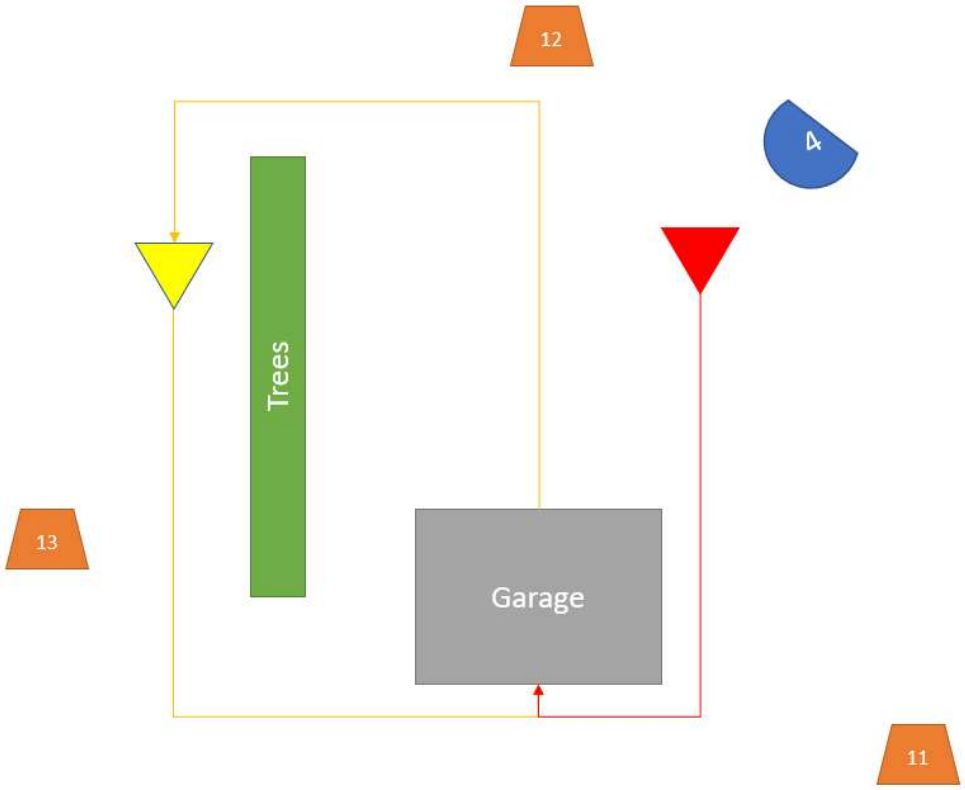

Figure 1. Scenario 1

For scenarios 1, 2C, and 2D the MTRI EO 04 sensor optimally covers the majority of the vehicle interactions that matter for those scenarios, especially for 2D. Finally, that particular blend of scenarios is ideal in that each of the individual scenarios has a different number of vehicles to test robustness of system. It is undesirable to unevenly spread training data as to avoid bias in decision making, and in doing so the most difficult of the three scenarios (2D) can be approached without bias.

With regards to the first two scenarios used, Scenario 1 and Scenario 2C, the setup for the scenarios is straightforward. There are three stationary P-RF sensors, the MTRI 11, 12, and 13, and the single source of EO information, designated as 4. The

vehicles in Scenario 1 and Scenario 2C move, heavily obscured by the tree line, the garage, and the limited stationary angle of the camera from MTRI EO 04. In both of these scenarios a vehicle will move from behind the tree line, almost completely obscured during this time, and then enter the garage. In Scenario 1 this provides the illusion that the stake rack truck simply made a loop around the track when in fact the target was switched with the one that was behind the tree line.

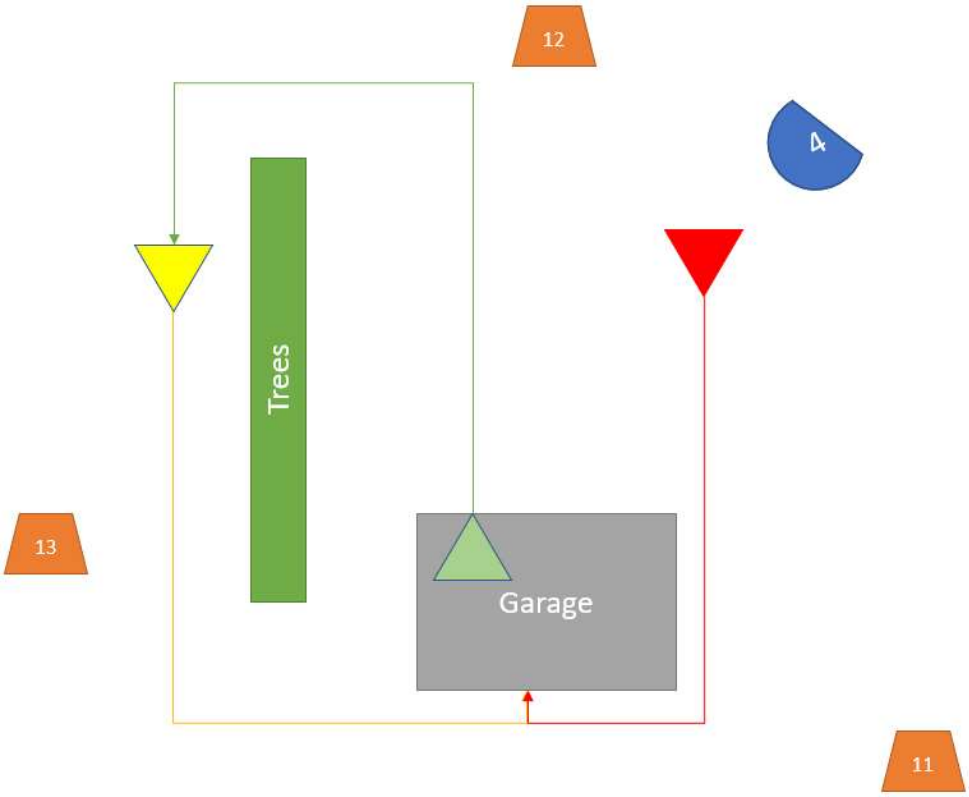

Figure 2. Scenario 2C

In Scenario 2C, there is a similar switch, though this one is more complex as the vehicle emerging from the garage is in fact a completely different vehicle than the pickup truck that was driving behind the tree line. Unlike in Scenario 1 the gas motor Gator utility vehicle has next to no bearing on the main target's "escape", a red herring that creates the illusion of their only being two vehicles in this scenario instead of three. These vehicles are captured through the MTRI EO 04 briefly, with the pickup truck in the background behind the tree line having the least time "on screen".

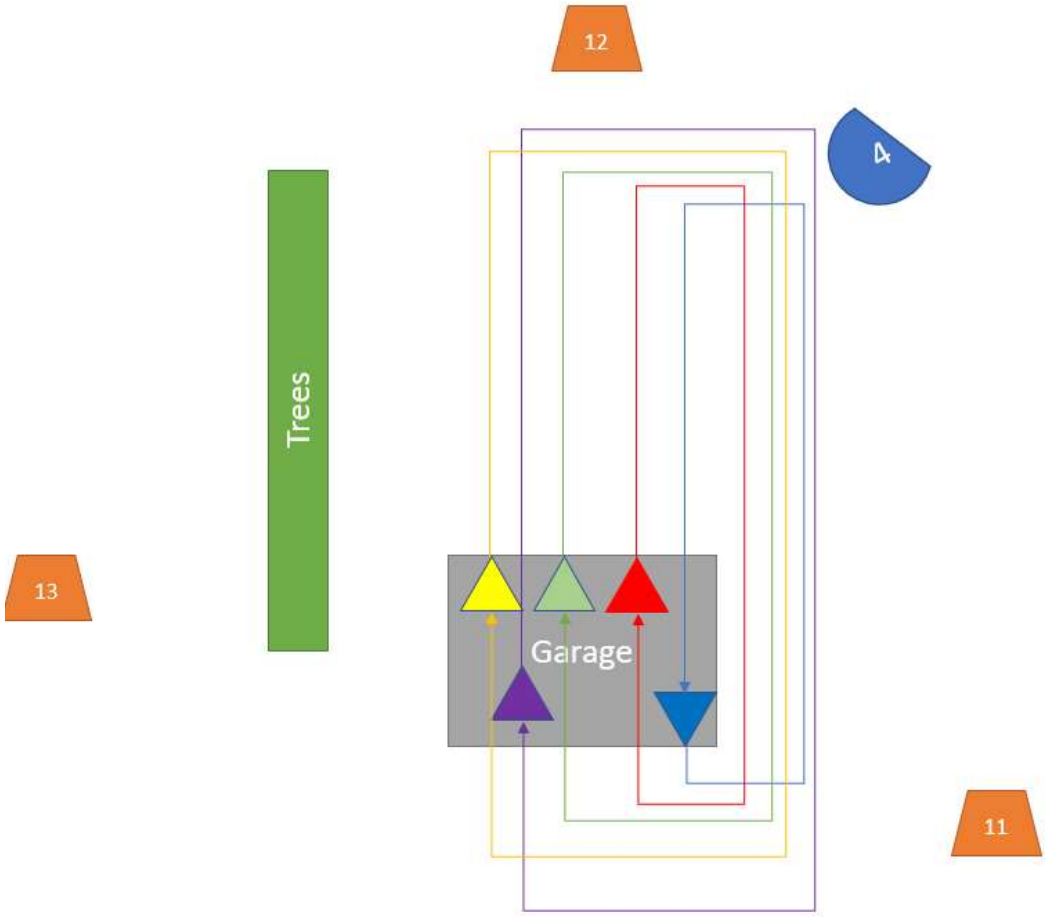

Figure 3. Scenario 2D Overview

For scenario 2D, the nature of the scenario is considerably more chaotic. Four vehicles seemingly emerge from the front of the garage, a pickup truck, a deisel motor Gator utility vehicle, a gas motor Gator utility vehicle, and a stake rack truck and then from behind the garage is a panel van. The deisel motor Gator utility vehicle will then pass the pickup truck in the first fifteen seconds of the scenario, changing the order, and then after thirty seconds the panel van will pass by the four vehicles. After forty five seconds the gas motor Gator utility vehicle will then pass the pickup truck, and five seconds later pass the diesel motor Gator utility vehicle as well.

While it might seem that the order continues in this way based on only the EO data, from the ground truth what in fact happens is that the gas motor Gator utility vehicle pauses and allows its diesel motor Gator counterpart to in fact enter first. While it might seem that the scenario is simply five vehicles going around in a loop, the passing and change in traffic order provides challenges for most systems when tracked with respect to the ground truth.

3.2 Preprocessing and CCA Labeling

In order to better exploit the combined view that the P-RF and EO sensors provide, certain steps had to be taken in order to implement the fusion. From previous research with this dataset, the input of the raw passive radiofrequency data by itself is not sufficient for fusion, even via neural network. There had been other attempts using supervised learning with other classifiers, but the performances of those classifiers were also insufficient for even classifying scenarios, let alone handling the task of identifying specific targets. While the ESCAPE dataset does provide the sensory information required for radar, this research is primarily focused on the exploitation of using the

passive RF information available in concert with the EO data. Because the raw in-phase and quadrature components (I/Q data) was insufficient for even classification purposes, the data is transformed into a series of histograms that correspond to the same points in time as each of their respective frames.

This approach comes with the added benefit that the data to be synchronized with respect to the time domain, which drastically helps simplify the fusion architecture required. In previous work with a related dataset from MTRI, a Digital Imaging and Remote Sensing Image Generation (DIRSIG) simulation, the P-RF histograms actually provided a visually obvious difference for the same application. The differences in histograms were strong enough that on their own the signals were sufficient for distinguishing between scenarios at a rate of 100% accuracy. For the ESCAPE dataset, the results were nowhere near as clean, owing to the fact that the dataset is not a simulation. That being said there were some instances in which the data did have some visually discernable changes that potentially correlated to certain events.

From Scenario 1's data using the input of P-RF sources 11 and 12, there was some noticeable changes in the P-RF histograms that correspond with the vehicle entering the garage. As seen above in Figure 4 and below in Figure 5 (11,12,4), the changes are visually discernable throughout all three sources of P-RF data. While initial results with the raw I/Q data as a P-RF input were less than successful, upon switching to the histogram preprocessing approach the performance increased considerably.

As for the EO data, in previous work the implementation of canny edge detection had been used in an attempt to improve accuracy, even for just a standalone EO based classifier. Applying the Gaussian filter to smooth out the image and then finding the

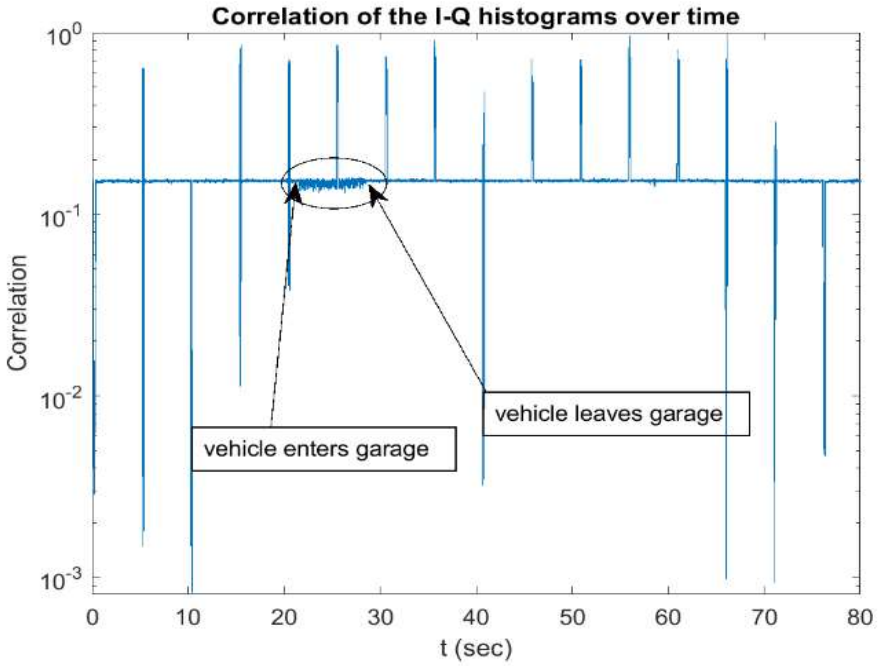

Figure 4. Correlation of P-RF Histograms of P-RF sources 11 and 12

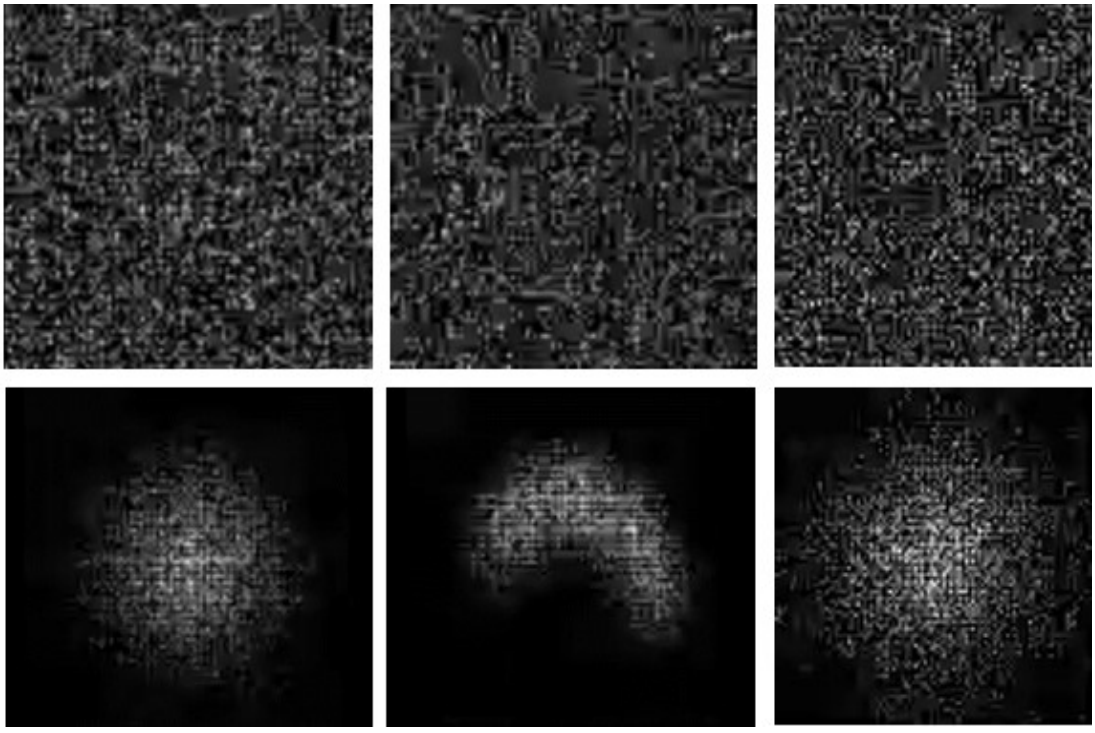

Figure 5. Timeframe 0:01(above) and timeframe 0:25(below).

intensity gradients of the image produce the optimal conditions to apply non-maximum suppression. Following the double threshold to determine potential edges the hysteresis tracking should suffice for finding the outlines of vehicles. Unfortunately for this application the results showed that the application of canny edge detection was almost always detrimental for most of the image-based classifiers used, with the exception of using Gradient Boosting and Nearest Centroid classifiers. This unexpected result lead to a few questions regarding experiment design and a reevaluation of prior assumptions about the current experimental design.

Table 2. EO vs CED Results

Traditional Classifier Method	EO P-RF Histogram	CED-EO P-RF Histogram
Decision Tree	0.64	0.63
SVM	0.65	0.59
Naïve Bayes	0.41	0.38
Gradient Boosting	0.39	0.44
KNN	0.72	0.69
Nearest Centroid	0.94	0.94
Standalone EO	0.52	0.49

At this point it became clear that the current method of processing the EO data was far from sufficient for even classification purposes. While running only the EO data by itself through a number of traditional classifiers, the results were less than satisfactory for even simple matters of classification of input image's scenario. Something that should

have in theory been easy enough to accomplish given the completely different number of vehicles the EO input can detect, especially after applying canny edge detection.

As seen above in Table 2, there is a major disparity between the classification for Scenario 2C and Scenario 2D, one in which the KNN model, which performed the strongest of the standalone EO models, still manages to largely confuse the 2C with 2D and vice versa. While the P-RF data by itself was far from being capable of independently classifying the data, as from the results of Table 3 its clear that it was biased towards Scenario 2C, the standalone EO model was facing more difficulty than it should have.

Table 3. Standalone EO KNN Classifier Normalized CM

True Label\Predicted Label	Scenario 1	Scenario 2C	Scenario 2D
Scenario 1	1	0	0
Scenario 2C	0	0.21	0.79
Scenario 2D	0	0.73	0.27

Table 4. Standalone RF KNN Classifier Normalized CM

True Label\Predicted Label	Scenario 1	Scenario 2C	Scenario 2D
Scenario 1	0.16	0.64	0.2
Scenario 2C	0.12	0.64	0.24
Scenario 2D	0.12	0.64	0.23

After another look at the experimental result's confusion matrices as well as the EO input, it became clear that there was background "noise" in the Scenario 2C and Scenario 2D data. This "noise" takes the form of multiple stationary and inactive vehicles on the sidelines in those two scenarios, an example of which can be seen in Figure 4. In the confusion matrixes for these classifiers it became clear that for Scenarios 2C and 2D, there was an unbalanced confusion between the scenarios, but Scenario 1 was always correctly identified, as it did not have the additional vehicles. While it was not clear what the addition of more vehicles had done in order to make the standalone RF based model to continuously identify Scenario 2C over scenarios 2D and 1, the reason for the classification issues in the EO data became very clear in comparison once the data was visually inspected again.

The application of canny edge detection would in fact make the vehicles in the background more distinguishable, thereby incorrectly making the classifier determine that there were more targets. Up until that point, all prior assumptions in the experiment design was being reviewed, making sure labeling and even base code was all operating under the correct mindset for the experiment. However, the fact that something this minor could massively impact the results became a driving factor for determining the impact that the data inputs have on classification results.

With that information in mind, it became clear that a different approach was needed for image preprocessing. Enhancing the image itself to make the vehicles more visible would end up confusing the classifier on the number of targets, making even just classifying the number of targets more difficult. There are many methods to enhance image inputs, especially for tracking, but the largest concern is falling back onto the same

problem that made even differentiating between Scenario 2C and 2D difficult using the EO input. In order to ensure that the preprocessing method did not repeat the same error, the obvious solution was to use a method that focuses on a target's "movement".

To more specific, estimating pixel level motion, between consecutive images in the three scenarios using MTRI EO 4. In order to handle this crucial issue of computer vision, the implementation of Dense Optical Flow (DOF) [46] was applied on the EO images. Optical flow itself is a fairly commonly used motion estimation technique that is split up between the sparse and dense varieties, and while sparse optical flow is a considerably less costly tracking method, the results with DOF provided a much more desirable change in the EO input. Some of the transformed sparse optical flow results also were lacking in terms of tracking, which also made the choice to switch to DOF considerably easier.

As seen above in Figure 6, the application of Dense Optical Flow removes the "noise" that is the inactive and stationary vehicles while also enhancing the five moving targets in Scenario 2D. Unlike the application of canny edge detection, the inactive vehicles are not detected and erroneously treated as targets for the DOF-EO input. While the transformed images lack the visual input that might help differentiate between

Figure 6. DOF (left) vs EO (right)

different vehicles (some of the targets have different colors even if they appear visually similar, coming in black, red, and green), the movement based changes provided a greater enhancement to the fusion model's performance.

$$\sum_{\Delta x \, \epsilon \, \mathrm{I}} w(\Delta \mathrm{x}) |A(x + \Delta \mathrm{x}) d(x) - \Delta \mathrm{b}(\mathrm{x} + \Delta \mathrm{x})||^2 \tag{3.1}$$

$$d(x) = (\Sigma \, w \, A^T A)^{-1} \Sigma \, w A^T \Delta \mathrm{b} \tag{3.2}$$

For the purposes of preprocessing, the approach taken was the Farneback method [47]. The Farneback method uses a two-frame motion estimation algorithm that utilizes polynomial expansion in order to estimate pixel displacement. The neighborhood of each image pixel is approximated by the polynomial, creating a quadratic polynomial that is used to produce a local signal model represented on a local coordinate system. Changes with respect to weight and displacement are covered above in Equations 3.1 and 3.2 respectively, where A is a symmetric 2x2 matrix of unknowns used to capture the signal vector b, x is the image intensity, and w is the weight function for the points in the

neighborhood equation in Equation 3.1. Using this coordinate system on the two subsequent images, the Farneback method makes it possible to directly derive the displacement between the two, thereby measuring the "flow" of each image pixel. The two-channel array of flow vectors are computed via OpenCV, producing and generated the enhanced DOF-EO input that is used for experimentation.

Lastly, the application of Canonical Correlation Analysis and the input of the variates between the P-RF and EO data is a necessary step for the creation of the current fusion model. In previous research, the application of CCA variates had drastically improved the performance of the classifiers for discriminating between different targets. In the case of the P-RF and EO fusion for the ESCAPE dataset, the results of the explainable AI indicate that the variates provide an insight that can be almost as valuable as the DOF-EO input. As the previous results with the canonical variates had proven to increase performance with respect to F1 score and the Tracking Detection Rate, the use of CCA variates in training for the fusion model was a clear choice for the explainable AI research.

The use of the variates done by converting the P-RF histograms and the DOF-EO input into two separate column vectors and then calculating the variates using sklearn's CCA function in Python. The column vectors of the two modalities are then used as the random variable input for defining the cross-covariance matrix between the EO and P-RF inputs. The covariate input along with the original modality input are provided as a frame of reference for the fusion model to use. Based on the results of previous work, the variates played a large role in improving the model's performance, but until experimentation with the explainable AI, how much of a role was unclear.

$$\rho = \frac{a^T \sum_{XY} b}{\sqrt{a^T \sum_{XX} a}\sqrt{b^T \sum_{YY} b}} \tag{3.3}$$

$$c = \sum_{XX}^{1/2} a \tag{3.4}$$

$$d = \sum_{YY}^{1/2} b \tag{3.5}$$

In order to find the cross-covariance matrix of the EO and P-RF data, the parameter (ρ) to maximize is in equation 3.3. Using the change of basis by Equation 3.4 and 3.5, and by applying the Cauchy-Schwarz inequality, the variates can be computed using the eigenvectors of the orthogonal correlation matrixes. The values of Equations 3.4 and 3.5 are the left and right singular vectors of the correlation between the two, corresponding to the highest singular value. Once these variates are obtained, the variates used as an input in the labeling system, elaborated further in Experiment Design.

For the purposes of comparison research for fusion level models, the fusion model comparison will be done using LSTM-CCA. The model uses the same inputs as the data frame fusion model for Explainable AI will, and therefore can provide a fair comparison to see if the greedy algorithm prototype can perform competitively with a less transparent machine learning model. The LSTM-CCA model uses a CCA layer derived from the work of Deep CCA [48]. The Deep CCA creates a layer that computes the representations of the two views (P-RF histograms and DOF-EO frames for corresponding points in time) that are connected from two deep networks. While the neural network type is different, using the same architecture to process the two views separately with CCA was a desirable approach for this dataset, based on the target detection and classification results. The outer layers are trained to be maximally correlated with each view.

$$(\theta_1^*, \theta_2^*) = argmax\ corr_{[\theta_1,\theta_2]}\big(f_1(X_1;\theta_1), f_2(X_2,\theta_2)\big) \tag{3.6}$$

$$corr(H_1, H_2) = \big||T|\big|_{tr} = tr(T'T)^{\frac{1}{2}} \tag{3.7}$$

$$\frac{\partial \text{cor}\ (H_1,H_2)}{\partial H_1} = \frac{1}{m-1}\big(2\nabla_{11}\overline{H}_1 + \nabla_{12}\overline{H_2}\big) \tag{3.8}$$

$$\nabla_{12} = \hat{\Sigma}_{11}^{-1/2} U\ DU' \hat{\Sigma}_{22}^{-\frac{1}{2}} \tag{3.9}$$

$$\nabla_{11} = -\frac{1}{2}\hat{\Sigma}_{11}^{-1/2} U\ DU' \hat{\Sigma}_{22}^{-1/2} \tag{3.10}$$

Equation 3.6 covers the goal of Deep CCA, which is the network jointly learning parameters for both views, training to maximize the correlation. $\theta_1$and θ_2 represent the vectors of all the parameters of the two views, finding (θ_1^*, θ_2^*) by following the gradient of correlation objective estimated on the training data. Equation 3.7 is treated as the matrix norm of T. Following the training of the parameters for the DCCA via gradient based optimization, the gradient is calculated with respect to (H_1, H_2),which is used for backpropagation. The singular value decomposition is used to derive equation 3.8, and these correlations are then treated as a function of the training dataset. From there, that data is used to calculate the top-level representations, (H_1, H_2), as seen above in equation 3.9 and then used further in equation 3.10.

LSTM-CCA relies on the CCA Layer approach of Deep CCA, but with some minor changes. The correlations of the two outermost views are used and the correlations and two views are then fed into a LSTM. Once the training is completed, evaluation begins. For the purposes of implementation in the equations remains the same, with a change in optimization function to RMSProp, and then using a standard sigmoid activation function.

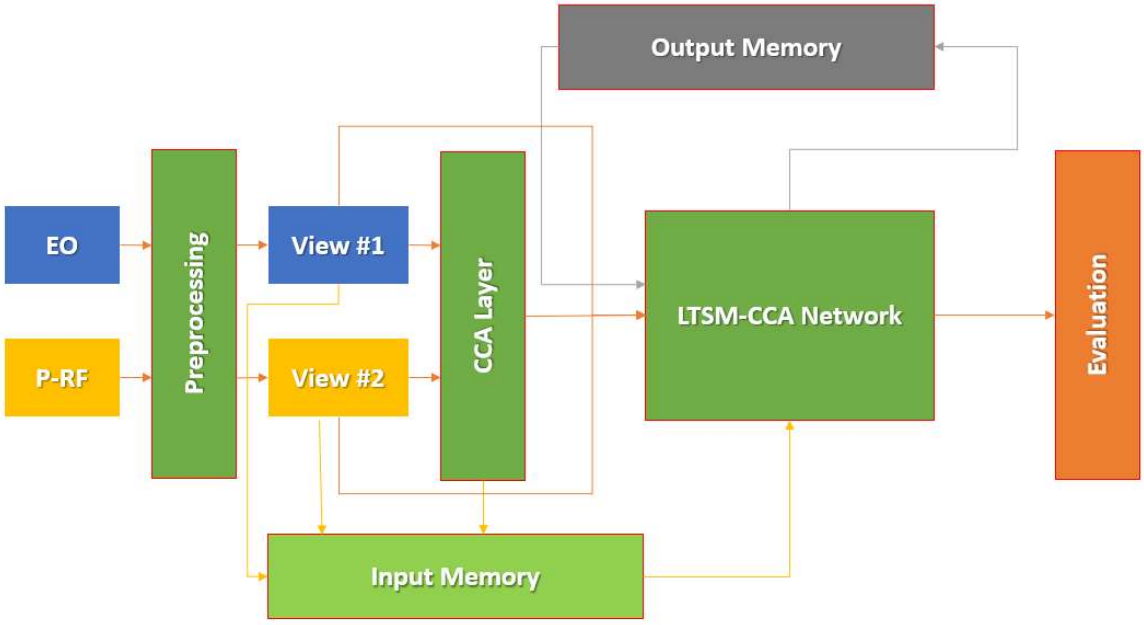

Figure 7. LSTM-CCA Architecture Overview

Rather than implementing the system with just deep neural networks, the resulting views are saved with respect to time, being loaded as a sequential vector for the LSTM portion of the network. While the neural network architecture and implementation of the original Deep CCA layer differs, the purpose of the CCA layer's implementation in the LSTM remains the same.

3.3 Experiment Design

In order to better exploit the combined view that the P-RF and EO sensors provide, certain measures needed to be taken in order to use the Explainable AI input and as a basis of the framework for the fusion model. The reasons for this design are twofold,

one is that by creating a system whose inputs are processed with respect to time makes the fusion process more streamlined. In previous research, the implementation of reducing the input to labeled frameworks using an LSTM-CCA fusion model provided an improvement in performance. Additionally, the framework system performed well for a number of different inputs, such as prior research using confusion matrixes from Decision Level Naïve Bayes fusion. And the other reason for the use of this framework system is in order to implement Explainable AI. While the current system only weighs and compares the modality and its accompanying covariate input, in the future expanding this system to potentially include other factors makes the basis for the framework considerably easier.

The implementation of the framework system is relatively simple. Once preprocessing is completed the information is then processed and placed inside of a pandas data frame. Inside of the data frame is not only the enhanced modality input (EO-DOF and P-RF histograms) for each corresponding point in time but also the canonical variates between both of the enhanced modality inputs. Using this data frame and the ground truth, the model is then evaluated using an F1 Score, confusion matrix, tracker detection rate, and through explainX.ai, an end to end explainable AI framework that provides insights into machine learning and the results of the model tested, the results of which are elaborated further in Chapter 4.

With regards to evaluation for research leading up to the implementation of the data frame and explainable AI, the F1 score is the harmonic mean of the precision and recall, the measurements of positive predictive value and sensitivity for machine learning. Precision is the measurement of type I error, false positives, while recall is the

measurement of type II error, false negatives. Besides the use of an F-1 score for evaluation, the use of a normalized confusion matrix was used in order to provide a better understanding of the portions of the dataset that gave the classifiers difficulty.

$$F_1 \text{ Score} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision+Rec}} \tag{3.11}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}} \tag{3.12}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+Fals Negative}} \tag{3.13}$$

For the evaluation of tracking, in order to objectively categorize Target Object Tracking classification, every category and label was evaluated by the tracker detection rate metric (TDR) [49], shown below in equation 3.14.

$$\text{Tracker Detection Rate} = \frac{Total\ True\ Positives}{Total\ number\ of\ ground\ truth\ points} \tag{3.14}$$

With respect to the design of the experimentation, the primary focus on this research was in order to implement the explainable AI and attempt to discover potential features or discriminative aspects of the fusion process. From previous research, the experimentation had divided up the targets in each of their respective scenarios and used the CCA input along with the DOF-EO and P-RF histogram input in order to attempt to track the targets with respect to the ground truth. As the DOF-EO input would be incapable of accurately assessing the status of the target, the model would therefore need to rely on the covariate input and the P-RF histograms in order to produce an accurate decision.

The other major objective was to see if, for the purposes of the limited detection and tracking the model was conducting, if the explainable AI model would be capable of matching the performance of the LSTM-CCA model. While the LSTM-CCA model is a

blackbox for the purposes of detecting and tracking the up to five unique targets from Scenarios 1, 2C, and 2D, the tracking detection rate as well as the F1 score was 1, which is to say a perfect score. It is reasonably acceptable if the performance for the explainable AI to not reach a perfect score, so long as the model performed competitively with respect to the less transparent fusion model. However, considering the relative simplicity of the application and experimental objective, the goal for this experiment was to achieve a perfect score.

With that in mind, the breakdown of the scenarios means that between Scenario 1, Scenario 2C, and Scenario 2D, there are a total of five independent targets, with some appearing in more than one scenario. In order to prevent unevenly distributing the data, each of these targets is treated as independent entities with respect to each of the scenarios experimented with. The ground truth labels are used in the implementation of the Tracking Detection Rate, and each of the scenarios is relabeled with respect to each of those vehicles. Scenario 1 only has two vehicles, Scenario 2C only has three vehicles, and Scenario 2D has all five vehicles in its ground truth.

For the purposes of determining the impact of the new DOF-EO input and to see how it fares in comparison to the P-RF data, the use of four different classifiers for each of the modalities is conducted. This provides a new frame of reference as the previous standalone modality classifiers were not built with respect to tracking but with the ability to discern between different scenarios unlike the results from tables 2 and 3. Naïve Bayes, Decision Tree, KNN, and Nearest Centroid were selected of the four and then ran separately for both the EO and RF input results. These results of the comparison research are elaborated more in Chapter 4.

Once this comparison research was completed, the next step was to implement the explainable AI. For the purposes of experimentation, the analysis of the CCA-Fusion occurs after preprocessing and compares the local and global weights of the fusion model for each of the scenarios with respect to their performance with individual targets. In order to facilitate this, explainX.ai is used. ExplainX.ai uses a streamlined and optimized version of ProtoDash [50], a versatile algorithm that works with any blackbox machine learning algorithm to identify similar prototypes. The creation of these prototypes, representatives that optimally describe the blackbox algorithm, allow for the coherent framework to find and determine non-negative weights by importance. Using this framework combined with the current data frame and CCA inputs is what composes the Explainable AI model being used in experimentation.

ProtoDash uses a greedy algorithm in order to achieve optimization, seeking to assess the importance of the generated prototype and using the nonnegative weights in order to produce a more natural and easier to interpret comparison. The prototype framework focuses primarily on deriving the theoretical bounds for the selection methods, generating approximation guarantees at a higher level of efficiency than its predecessor, ProtoGreedy. The 2019 algorithm showcases its actionability, utility, and insight when summarizing a variety of different datasets and applications (MNINST, Retail, CDC questionnaires case study). ExplainX.ai is based off of this algorithm, though the number of available features it provides is considerably increased, capable of providing insights based on a single prediction point to providing a global overview of the interactions of different features.

With respect to the features that explainX.ai provides, the four major categories of transparency the software provides is global explanation, feature interaction, distribution, and local interaction. For the purposes of this thesis, the primary focus is on global explanation, and the comparison of weights in decision making. Certain aspects of local interaction will be brought up in Chapter 4, but the primary results and interpretation are focused on the global explanation due to the limited number of features being used in the current data frame fusion model. In future work, other aspects might be integrated once the data frame fusion model is expanded, which will be expanded in Chapter 5.

Global feature impact identifies which features in a dataset have the greatest positive or negative effect on the outcomes of a machine learning model. The impact value is the weight by which the input provided is used to produce a decision. For the purposes of this thesis, the impact of different variables on the decision making of the explainable AI fusion model is the main focus. Due to the fact that the calculations of the feature importance and feature impact remain the same for the four sources of input data, the discussion of weight on decision making is focused only feature importance. In future work, depending on the results of more than four types of information placed into the model, other aspects of the global explanations might be implemented.

Feature interaction contains a number of visual representations of the model, specifically a partial dependence plot and a summary plot. It does this by decomposing the predictions into different terms, a constant term, a term for the first feature, a term for the second feature, a term for the interaction between two features, etc. The interaction between the two features is the change in the prediction that occurs by varying the features after considering individual feature effects. The partial dependence plot shows

the marginal effect one or two features have on the outcome of a machine learning model, while the summary plot provides the first indications of the relationship between a value of a feature and the impact on the prediction, using different colors to represent the value of the feature from low to high.

Distribution provides the option to view the impact of different variables via histogram or violin plot and the option to implement a multi-level Exploratory Data Analysis (EDA) based on chosen variables. Histograms of the different features can be individually produced. For the joint violet plots, the statistics summary provides mean, median, model, interquartile values, etc. The distribution of the predicting variable can be found on top of the other input variables in order to find the join distribution.

Finally, local interaction provides options to view local feature impact and the option to view similar profiles to the data. Local feature impact narrows down the global feature impact graph, showing the decision plot and how much each feature contributes to the overall model prediction for a specific point. The similar profiles generate similar profiles from within the dataset, based on how similar the attributes are with respect to model prediction and ground truth values.

CHAPTER FOUR

RESULTS

4.1 Comparison Research

As a reference for the impact of the dense optical flow enhanced EO and the effectiveness of the histograms alone, there was a need to gauge the accuracy of the modalities without using fusion. Four conventional classifiers were used, Naïve Bayes, Decision Tree, KNN, and Nearest Centroid for each respective vehicle's input of P-RF and EO data. The results of the experimentation for each of the three scenarios for standalone modality input were largely to be expected. In previous work with the dataset, even with the histogram preprocessing the P-RF data for the ESCAPE dataset is not capable of providing enough discriminating features for the purposes of target detection and classification. Between the enhancements via dense optical flow and histogram input, the EO data performed considerably better.

Table 5. Standalone Classifiers Performance for Scenario 1.

Modality/Input	Naïve Bayes	Decision Tree	KNN	Nearest Centroid
Vehicle 1 EO	0.94914	0.63874	0.96214	0.59335
Vehicle 1 RF	0.64027	0.45116	0.62184	0.41289
Vehicle 2 EO	0.95864	0.63874	0.96215	0.63874
Vehicle 2 RF	0.58418	0.45119	0.78328	0.41289

As seen above in Table 5, Decision Tree and Nearest Centroid classifiers performed the poorest of the four standalone classifiers, regardless of modality. The scores for Naïve Bayes and the KNN classifier with respect to the EO input for vehicles 1 and 2 are to be expected, as the removal of the background vehicles via dense optical flow drastically reduced the classification issues. The performance for the KNN and Naïve Bayes are very close to the CCA-Label Fusion results on its own for Scenario 1.

Unlike the results from Scenario 1, the results for the Decision Tree and Nearest Centroid classifiers are more or less on par with KNN and Naïve Bayes. Scenario 2C was by design more difficult than that of Scenario 1. Vehicle 3 only appears briefly, but an interesting anomaly is the relatively higher performance with the Vehicle 3 EO and Decision Tree performance.

Table 6. Standalone Classifiers Performance for Scenario 2C.

Modality Input	Naïve Bayes	Decision Tree	KNN	Nearest Centroid
Vehicle 1 EO	0.79968	0.75587	0.78039	0.7994
Vehicle 1 RF	0.50536	0.56209	0.57213	0.62871
Vehicle 2 EO	0.78681	0.75588	0.78103	0.811136
Vehicle 2 RF	0.51801	0.56209	0.57212	0.0634488
Vehicle 3 EO	0.62696	0.846473	0.50931	0.6131
Vehicle 3 RF	0.54327	0.61196	0.57992	0.62127

Vehicles 1 and 2 however, having a more balanced timeframe within the scenario and also benefiting from the similar make and model of vehicle have a consistent performance. Vehicle 2, which spends the most time on the FMV EO input performs slightly better with Nearest Centroid, but ultimately from the rest of the standalone results on Scenario 2C the results are hardly unexpected considering the classifiers performance in previous experiments.

Table 7. Standalone Classifiers Performance for Scenario 2D.

Modality Input	Naïve Bayes	Decision Tree	KNN	Nearest Centroid
Vehicle 1 EO	0.80022	0.90696	0.9034	0.91158
Vehicle 1 RF	0.71424	0.71854	0.90543	0.75004
Vehicle 2 EO	0.8937	0.89909	0.88101	0.91674
Vehicle 2 RF	0.66409	0.61028	0.79358	0.6815
Vehicle 3 EO	0.86384	0.91316	0.84978	0.92828
Vehicle 3 RF	0.64302	0.57736	0.73593	0.60355
Vehicle 4 EO	0.81769	0.91651	0.91324	0.78472
Vehicle 4 RF	0.69407	0.81593	0.82586	0.64322
Vehicle 5 EO	0.9165	0.87561	0.87376	0.86224
Vehicle 5 RF	0.8202	0.64199	0.85406	0.61775

Finally, with respect to Scenario 2D, the most chaotic of the three scenarios, there were a few notable anomalies in the standalone experiments. The most pleasant of which

was the clustering performance, as KNN just manages to break past 0.9 on accuracy for vehicle 1 in Scenario 2D. For this Scenario and in future work revisiting the performance for KNN and vehicle 1 definitely warrants the use of activation maximization or any form of visualization in the future to determine the impact the P-RF data had on vehicle 1. The overall performance of the P-RF data is much stronger in Scenario 2D than in scenarios 1 and 2c. While there are not any other standalone instances of P-RF scoring above a 0.9, there are still many that are above 0.8, in comparison to the other two scenarios.

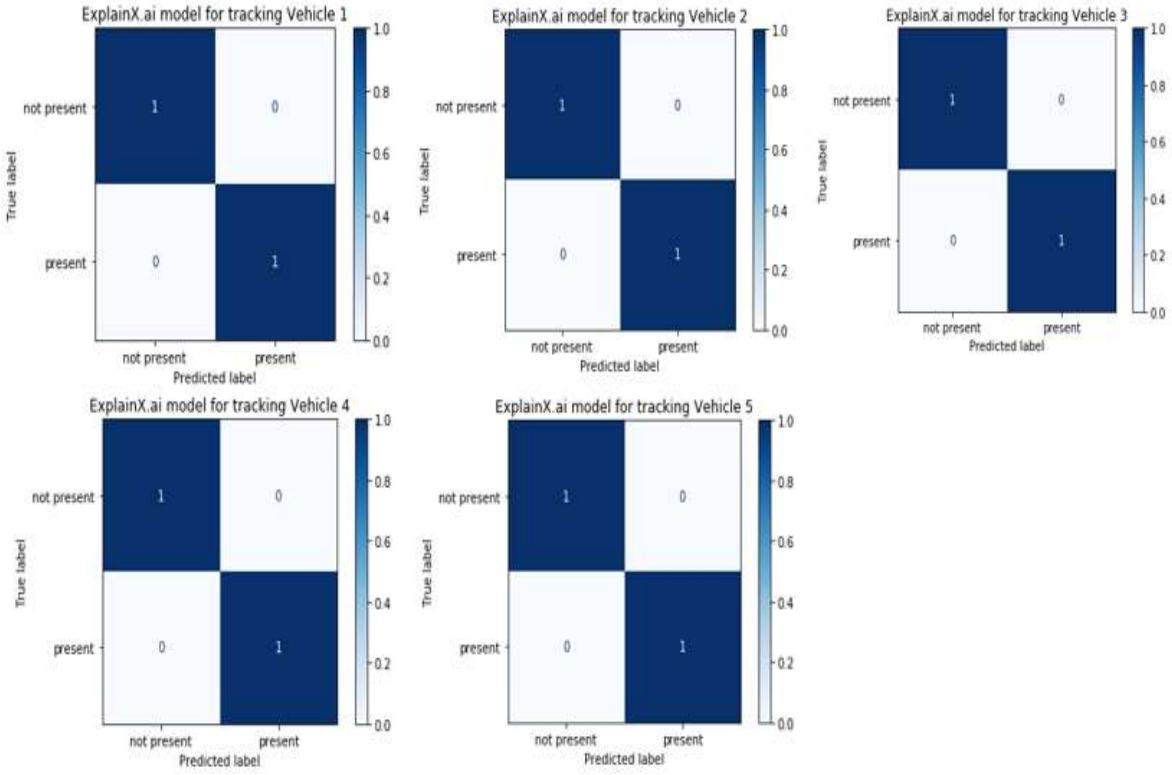

Figure 8. Comparison of Normalized CM for Explainable AI model

With respect to the performance of the tracking detection rate metric and the F1 score, the results for the Explainable AI fusion model were able to meet the expectations

and also achieve a perfect score in both metrics for each respective target. These results would indicate that for the application and dataset in question, the CCA values and the data frame fusion model together are able to sufficiently track and detect the vehicles with the EO and P-RF data. While there was a slight change in distribution between all of them based on the target being tracked, the results from the Explainable AI model would indicate that the CCA input drastically helps improve the heterogenous EO/P-RF sensor fusion.

4.2 Explainable AI

As the results from the LSTM-CCA and the current explainable AI data both were able to achieve a perfect F1 score and perfect score for the tracking detection rate metric, the primary focus for this research is on the results for the explainable AI model. More specifically the research would be more focused on the weights and transparency of the current fusion model. The use of the greedy algorithm in order to derive a prototype approximation for the fusion model provides benefits for determining the impact of different inputs for different scenarios. For the experimentation, the four inputs for the data frame were the DOF-EO frames, the EO-CCA input, the P-RF histogram input, and the RF-CCA input. For Scenarios 1, 2C, and 2D, tables 7-9 below provide the explainX.ai results with respect to the weights of each of the four inputs.

The overall results of the weights experiment with explainX.ai demonstrate a reasonably well spread and balanced result from the prototype. From the weights it is clear that the P-RF data still plays something of a role within the fusion model, even if the histogram input is generally the lowest input weight relative to the other three data inputs. It is less than surprising that the EO input consistently receives the highest weight,

but an interesting scenario in which the P-RF CCA variates will occasionally be almost as important to the decision-making process as the EO. The fused view performs better given the focus on vehicle 1, which in the context of screen time on Scenario 1 makes sense with respect to weight distribution.

For most part Scenario 2C's results were within expectations. The EO input predominantly leads in terms of weight for decision making, followed closely by the P-

Table 8. explainX.ai Weights results for Scenario 1

Vehicle	Input	Weight
Vehicle 1	EO Input	0.212249
	EO CCA Input	0.1844515
	RF Input	0.1246046
	RF CCA Input	0.2122194
Vehicle 2	EO Input	0.1269909
	EO CCA Input	0.09609033
	RF Input	0.08085641
	RF CCA Input	0.08058309

RF CCA values. That being said there are brief but unexpected instances of EO-CCA covariate input dominating the results of the weights over the P-RF CCA covariate

input. Given the importance of the EO input with regards to detecting the targets, it should not be a surprise that the covariate input of that data should also be important. But as seen in Table 8 the EO-CCA covariate input being used more in the decision making for certain vehicles is clearly not a fluke, and in some cases is a major discriminating feature for decision making.

In the case of table 9, Vehicle 1 and Vehicle 3 both rely more on the EO-CCA input, but interestingly enough Vehicle 5 practically does not use the RF histogram or EO-CCA input. Between the five targets, Vehicle 5's classification relies almost entirely on the P-RF CCA and DOF-EO inputs. Considering Vehicle 5 spends the least amount of time on the EO source, this indicates that the P-RF CCA input provides enough insight into the activities of Vehicle 5 to be given the same weight as the DOF-EO input.

4.3 Data Analysis

As both the Explainable AI and the LSTM-CCA models were both capable of achieving a perfect F1 score and tracking score of 1, the value in this experiment comes from the ability to see how the Explainable AI used the data frame fed information for decision making. For each of the scenarios and each of their respective targets, there are corresponding values for the weights each of the four inputs, the DOF-EO, EO-CCA variates, P-RF histograms, and the RF-CCA variates held on decision making. The tables however do not address the bigger picture, instead focusing on what impact each of the individual aspects of the data frame had on the classification of the vehicles.

Table 9. explainX.ai Weights results for Scenario 2C

Vehicle	Input	Weight
Vehicle 1	EO Input	0.1313921
	EO CCA Input	0.1134093
	RF Input	0.09182034
	RF CCA Input	0.100984
Vehicle 2	EO Input	0.1841983
	EO CCA Input	0.07001276
	RF Input	0.06620991
	RF CCA Input	0.1185514
Vehicle 3	EO Input	0.1304092
	EO CCA Input	0.1116768
	RF Input	0.0984916
	RF CCA Input	0.1072683

Table 10. explainX.ai Weights results for Scenario 2D

Vehicle	Input	Weight
Vehicle 1	EO Input	0.09108965
	EO CCA Input	0.07937257
	RF Input	0.05895651
	RF CCA Input	0.0654112
Vehicle 2	EO Input	0.1057887
	EO CCA Input	0.09031994
	RF Input	0.06189732
	RF CCA Input	0.1057738
Vehicle 3	EO Input	0.1332544
	EO CCA Input	0.1116768
	RF Input	0.0984916
	RF CCA Input	0.1072683
Vehicle 4	EO Input	0.1841983
	EO CCA Input	0.0825
	RF Input	0.08589286
	RF CCA Input	0.09032434
Vehicle 5	EO Input	0.1343745
	EO CCA Input	0.04249655
	RF Input	0.007944699
	RF CCA Input	0.134089

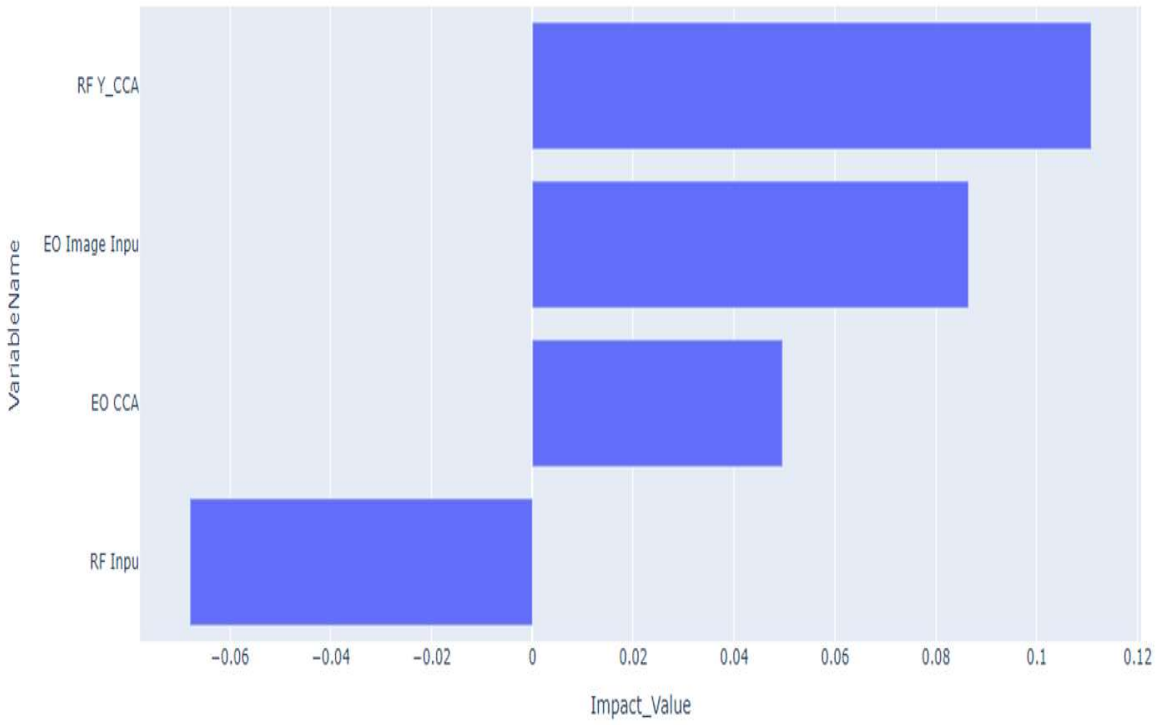

Figure 9. Local impact of data frame attributes

To that end, in order to gain further insight on the Explainable AI model, the first thing to do was to explore the local feature impact. The P-RF histograms appear to have a negative impact on the outcome on the prediction process. In comparison however, the RF CCA variates appeared to have the highest positive impact, surpassing the DOF-EO image input and the EO CCA variate inputs. The results of this are consistent with the experiments in this thesis, as the P-RF histograms have never consistently performed anywhere near the DOF-EO performance. The CCA input drastically improved the results of both the LSTM-CCA model and the Explainable AI model, and it would make

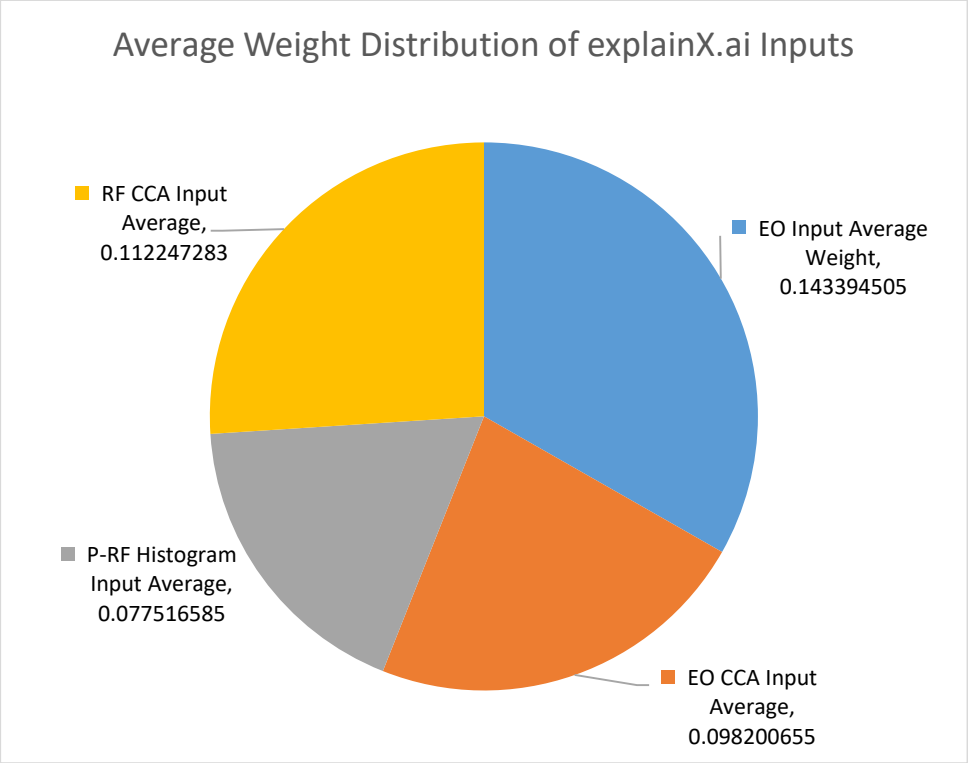

Figure 10. Comparison of Average Weights for all five targets

sense based off the results that the RF CCA variates bridge the gap between the P-RF histograms and accurately detecting and tracking the individual targets.

The next thing to consider is determining if overall the RF CCA variates is as major of an impact as the local feature impact graph claims. By averaging out the weight values for each vehicle and each scenario, Figure 10 shows that for the ESCAPE dataset the RF CCA input holds the second highest value, after the DOF-EO input. The average weight value is slightly higher than that of the EO CCA variates, and naturally the P-RF histogram inputs are the lowest on average. The average weight of the P-RF data is easily expected, as the P-RF weight is almost always the lowest, with the sole exceptions of

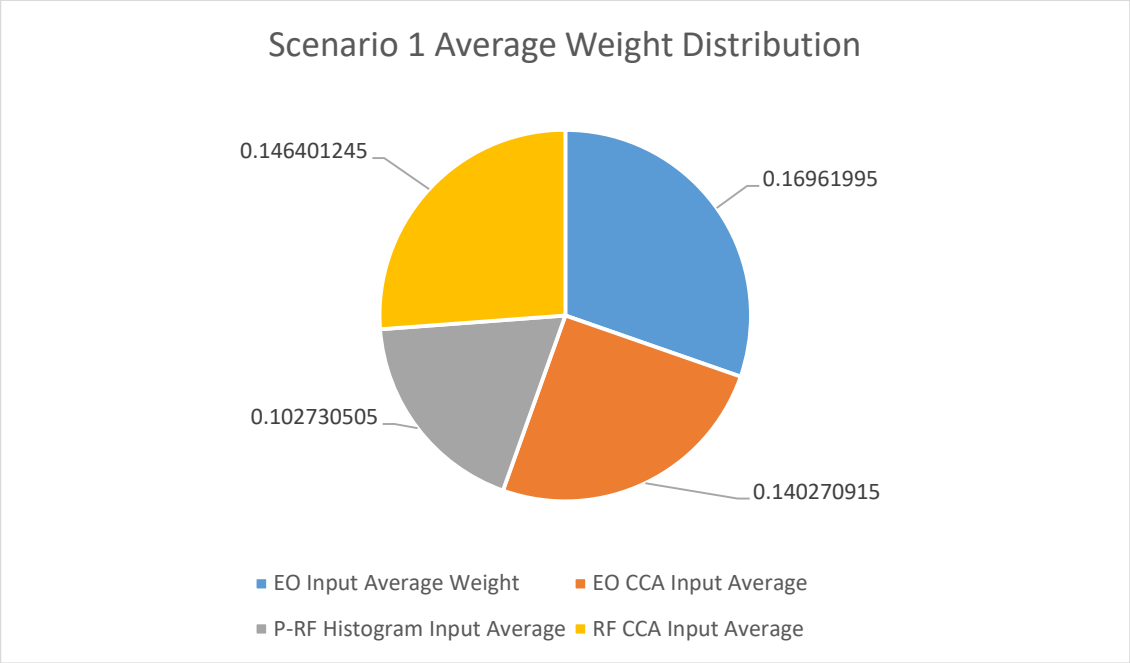

Figure 11. Comparison of Average Weights for Scenario 1

Scenario 1, Vehicle 2 and Scenario 2D, vehicle 4, in which the RF-CCA input and the EO-CCA values are lower in weight.

Taking the results on a scenario by scenario basis, and averaging out the values for each of them, Figure 11 displays the results for Scenario 1. As seen above, the difference in average weight between the DOF-EO input and the RF CCA input is considerably close. The RF CCA data has a little more than a 0.025 difference in weight for decision making compared to the DOF-EO input. It should be noted however that in Scenario 1 the EO CCA input's average weight has a little more than a 0.006 difference than that of the RF CCA input, compared to the average which would suggest a less than 0.02 difference between the two inputs. For Scenario 1 the CCA input is almost on par with that of the DOF-EO input on average.

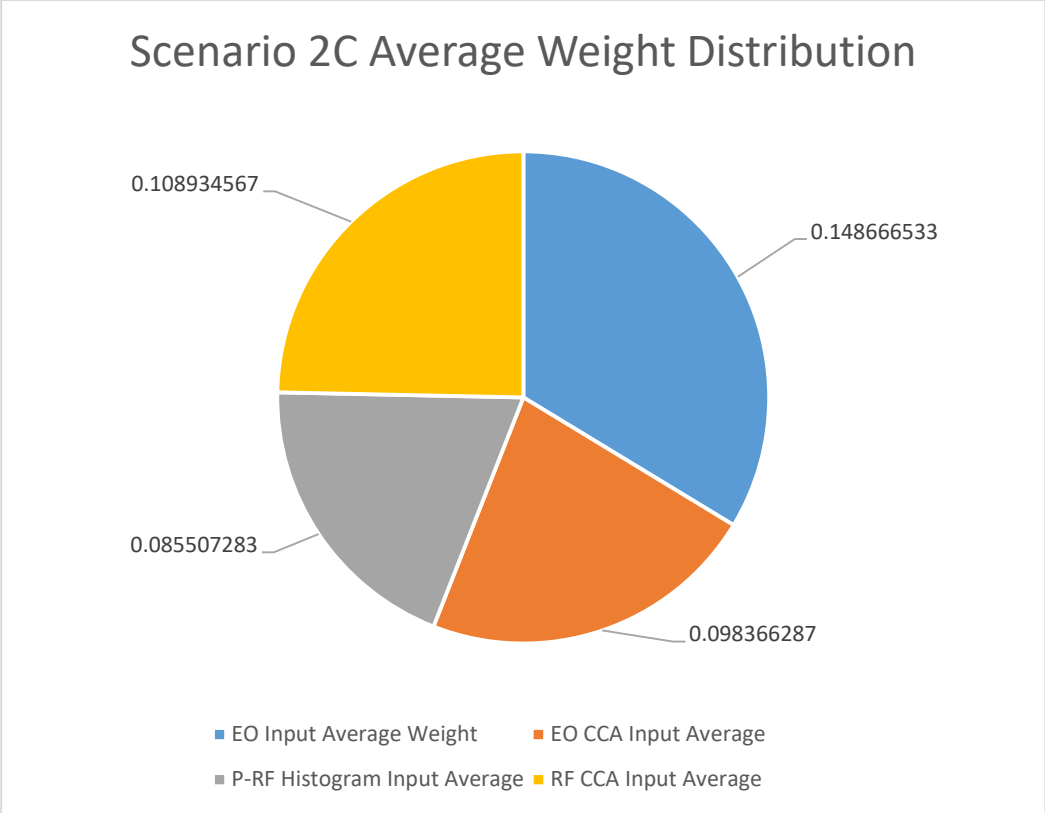

Figure 12. Comparison of Average Weights for Scenario 2C

As for Scenario 2C, the difference is nowhere as close to that of the averaged results in Scenario 1. As seen above, there is an over 0.4 difference in average weight input from the DOF-EO input to the RF CCA variate input. The EO CCA variate input is not far behind from the RF CCA's average weight, having an even closer gulf between the two than in Scenario 1. And as is a reoccurring theme in this experiment, the P-RF histogram input remains at the lowest average weight. The focus on inputs for decision making in the Explainable AI model remains on the DOF-EO input followed by the RF and EO CCA variate inputs.

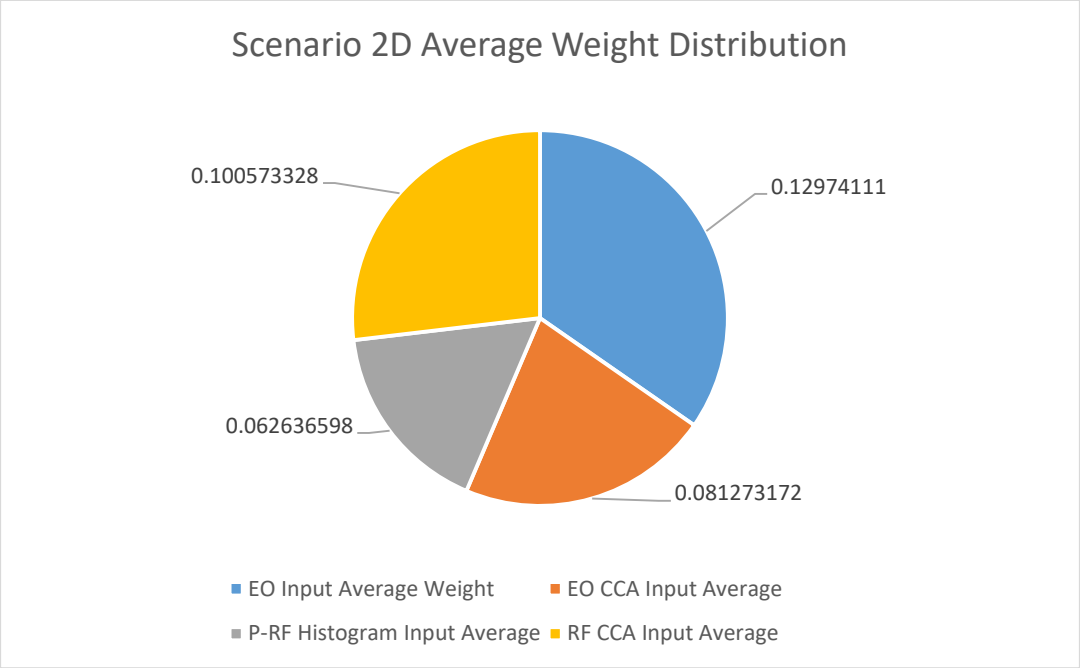

Figure 13. Comparison of Average Weights for Scenario 2D

Finally, for Scenario 2D, which holds the highest number of potential targets between the three, the results show that in terms of average weight distribution the DOF-EO input remains the largest impact on the decision-making process. RF CCA variates have the second highest average weight, followed by a bigger gulf between RF CCA variates and the EO CCA variates than in Scenario 2C. The results in figure 13 indicate that as always, the P-RF histogram input average weight remains the lowest of the four main features provided by the data frame input. From these results, it appears that while the EO CCA variates input does have moments in which its weight is higher than that of the RF CCA variates, globally and on average the decision making will focus on the DOF-EO image input, then the P-RF CCA variates, followed by the EO CCA variate inputs and the P-RF histograms.

CHAPTER FIVE

CONCLUSION

5.1 Overview of Thesis

In this thesis I have analyzed and implemented the impact of CCA on the heterogenous multimodal fusion of EO and passive RF information. This study is focused on the effectiveness and more importantly the interpretability of the results. While the target tracking application of a smaller dataset is relatively simpler to work with, having only a maximum of five targets at once, the larger question is how to best measure the impact of the passive RF data with respect to detection and tracking.

Our research group has had success in using passive RF data for human detection, as well as a few other applications, and as the initial steps taken in this paper have shown, the passive RF data and its canonical variates do provide a notable weight in the decision making process for almost all of the scenarios tested. Only in Scenario 2D was there a noticeable drop in global impact weight for the P-RF data, but even in such situations such as vehicle 5, the importance of the RF canonical variates still played a major role in the prediction process of the fusion model.

Additionally, while on average not as relevant of a weight, the EO canonical variates still largely made stronger contributions to the decision making of the Explainable AI model than that of the P-RF histograms. The results in terms of the P-RF histograms average weight in prediction and decision making would indicate that

research into the further enhancement of P-RF data is sorely needed in the future. The use of this information will be useful in future research.

5.2 Relevance

There are a number of methods for measuring the impact of image-based inputs, which the P-RF histograms and the DOF-EO images are. That being said, based on the research done with the global impact and weights, it becomes clear that the P-RF histograms do not necessarily provide as much of an impact image wise as expected, and only due to time constraints is why the implementation of activation maximization or saliency maps applied. Given that the method of enhancing the EO input was a method of tracking movement between frames, it seemed reasonably redundant to implement an image based method, and as two of the other sources of data in the data frame are not images, the prioritization of using explainX.ai is arguably more prudent.

While not the focus of the paper, the LSTM-CCA applies a branch of Deep CCA, using the time series nature of the data in order to achieve a perfect performance based on the individual tracking of different vehicles. The application of CCA in the case of the Explainable AI is nowhere near as sophisticated. But given the relationship between the P-RF data and the EO data the results the simple application of CCA in the data frame network provided an equally impressive performance as the LSTM-CCA model.

In terms of choices for other comparison research, the four classifier methods were chosen based on prior research done with P-RF/EO fusion. Decision Tree, Naïve Bayes, KNN and Nearest Centroid had performed better with this dataset than methods such as SVM or Gradient Boosting. And as the use of such classifiers without the CCA input was only to provide a comparison for how the classifiers fared in comparison, the

choice of which classifiers used was simply a means to pick which classifiers would likely perform well without the CCA input.

5.3 Future Direction

In future research with both EO/P-RF fusion and the ESCAPE dataset, or tracking applications in general, further research into different tracking metrics and objectives will be necessary. This is to say that while the addition of providing the CCA variates between the modalities has boosted the results considerably, the tracking detection rate and F1 score remain insufficient to test the differences between the three scenarios. It is possible to increase the number of image samples for the EO data in a number of ways, but for the P-RF data this is not the case, and is in fact the largest limiter for this research in terms of modality input.

Of course, there are larger aspects of Explainable AI and CCA to explore with respect to heterogenous sensor fusion. As mentioned earlier, the use of image based Explainable AI methods such as activation maximization or saliency maps would have preferably been added into this thesis if not for time constraints. While the P-RF histograms had an almost universally lowest weight to decision making and prediction, they're still used as an input. For this reason, seeing if there is any image-based input in the noise of the P-RF histograms would be extremely desirable. The results for the DOF-EO input might provide an unexpected surprise, but at the moment is something I would personally find doubtful.

More importantly, as the data frame input used in the Explainable AI only has four inputs at the moment, the two modality inputs (P-RF histograms and DOF-EO) as well as canonical variates between the two, there is a clear need to expand on the number

of inputs the data frame input provides. For the purposes of the LSTM-CCA fusion model, the data is read through the time domain. However, in the future, determining the impact of a timestamp with respect to the Explainable AI is worth considering.

Also, regarding the data, the improvement of the raw I/Q data is highly desirable, and a direction I intend on perusing in the future. The average weight that the P-RF histogram input in the data frame had was the lowest, and it might be worth implementing other interpretations of the raw I/Q data, such as Density estimation or Entropy estimation. Even changing what domain in which the information is read for each of these modalities is a direction worth investigating as well. Currently the information is synchronized with respect to time, and as such, the fusion is primarily accomplished via image inputs and canonical correlation analysis. Researching what other opportunities might be available is another direction to potentially look into.

References

[1] N. M. Correa, T. Adali, Y. Li and V. D. Calhoun, "Canonical Correlation Analysis for Data Fusion and Group Inferences," *IEEE Signal Processing Magazine,* vol. 27, no. 4, pp. 39-50, 2010.

[2] Ö. M. Polat and Y. Özkazanç, "Image enhancement via Multiple Canonical Correlation Analysis," in *21st Signal Processing and Communications Applications Conference (SIU)*, Haspolat, 2013.

[3] L. Du, C. Liu, M. Laghate and D. Cabric, "Sequential detection of number of primary users in cognitive radio networks," in *2015 49th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, 2015.

[4] A. KIRILENKO, A. S. KYLE, M. SAMADI and T. TUZUN, "The Flash Crash: High-Frequency Trading in an Electronic Market," *The Journal of Finance,* vol. 72, no. 3, p. 967–998, 2017.

[5] D. Bau, B. Zhou, A. Khosla, A. Oliva and A. Torralba, "Network Dissection: Quantifying Interpretability of Deep Visual Representations," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017.

[6] C. Barabas, M. Virza, K. Dinakar, J. Ito and J. Zittrain, "Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment," *Proceedings of Machine Learning Research,* vol. 81, pp. 62-76, 2018.

[7] T. Germa, F. Lerasle, N. Ouadah and V. Cadenat, "Vision and RFID data fusion for tracking people in crowds by a mobile robot," *Computer Vision and Image Understanding,* vol. 114, pp. 641-651, 2010.

[8] D. Wu, D. Zhang, C. Xu, H. Wang and X. Li, "Device-Free WiFi Human Sensing: From Pattern-Based to Model-Based Approaches," *IEEE Communications Magazine,* vol. 55, pp. 91-97, 2017.

[9] J. Liu, A. Vakil, R. Ewing, X. Shen and J. Li, "Human Presence Detection via Deep Learning of Passive Radio Frequency Data," in *NAECON*, Dayton, OH, USA, 2019.

[10] A. Vakil, J. Liu, P. Zulch, E. Blasch, R. Ewing and J. Li, "Feature Level Sensor Fusion for Passive RF and EO Information Integration," in *2020 IEEE Aerospace Conference*, Big Sky, MT, USA, 2020.

[11] N. Mallinar and C. Rosset, "Deep Canonically Correlated LSTMs," *ArXiv,* vol. abs/1801.05407, 2018.

[12] H. Hotelling, "Relations Between Two Sets of Variates," *Biometrika,* vol. 28, no. 3-4, p. 321–377, 1936.

[13] X. Yang, L. Weifeng, W. Liu and D. Tao, "A Survey on Canonical Correlation Analysis," *IEEE Transactions on Knowledge and Data Engineering,* no. Early Access, pp. 1-1, 2019.

[14] J. S.-T. Jan Rupnik, "Multi-View Canonical Correlation Analysis," *Proc. Conference on Data Mining, Data Warehouses,* pp. 1-4, 2010.

[15] J. D. Caroll, "Generalization of canonical correlation analysis to three or more sets of variables," *Proceedings of the 76th annual convention of the American Psychological Association,* vol. 3, pp. 227-228, 1968.

[16] Y. Luo, D. Tao, K. Ramamohanarao, C. Xu and Y. Wen, "Tensor Canonical Correlation Analysis for Multi-View Dimension Reduction," *IEEE Transactions on Knowledge and Data Engineering,* vol. 27, no. 11, pp. 3111-3124, 2015.

[17] M. J. Francis Bach, "A probabilistic interpretation of canonical correlation analysis," 2005.

[18] C. Wang, "Variational bayesian approach to canonical correlation," *IEEE Transactions on Neural Networks,* vol. 18, no. 3, pp. 905-910, 2007.

[19] X. Yang, W. Liu, D. Tao and J. Cheng, "Canonical correlation analysis networks for two-view image recognition," *Information Sciences,* Vols. 385-386, pp. 338-352, 2017.

[20] D. R. Hardoon, S. Szedmak and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *IEEE Transaction on Neural Networks,* vol. 16, no. 12, pp. 2639-2644, 2004.

[21] T. Sun, S. Chen, J. Yang, X. Hu and P. Shi, "Discriminative Canonical Correlation Analysis with Missing Samples," *WRI World Congress on Computer Science and Information Engineering,* vol. 6, pp. 95-99, 2009.

[22] D. MWitten, R. Tibshirani and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics,* vol. 10, no. 3, pp. 515-534, 2009.

[23] T. Sun and S. Chen, "Locality preserving CCA with applications to data visualization and pose estimation," *Image and Vision Computing,* vol. 25, no. 5, pp. 531-543, 2007.

[24] D. K. Seo, "Fusion of SAR and Multispectral Images Using Random Forest Regression for Change Detection," *ISPRS Int. J. Geo-Information,* vol. 7, no. 10, p. 401, 2018.

[25] S. Kim, W.-J. Song and S.-H. Kim, "Double Weight-Based SAR and Infrared Sensor Fusion for Automatic Ground Target Recognition with Deep Learning," *Remote Sensing,* vol. 10, no. 1, p. 72, 2018.

[26] D. L. Hall and J. Llinas, "An Introduction to multisensory data fusion,," *Proceedings of the IEEE,* vol. 85, no. 1, pp. 6-23, 1997.

[27] W. C. Barott, E. Coyle, T. Dabrowski, C. Hockley and R. S. Stansbury, "Passive multispectral sensor architecture for radar-EOIR sensor fusion for low SWAP UAS sense and avoid," in *IEEE*, Monterey, CA, 2014.

[28] N. T. B. Bui, D. C. Pham, B. Q. Nguyen and S. T. Le, "Tracking a 3D target with fusion of 2D radar and bearing-only sensor," in *2018 IEEE International Conference on Industrial Technology (ICIT)*, Lyon, 2018.

[29] Q. Zhang, Y. Liu, R. S. Blum, J. Han and D. Tao, "Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review.," *Information Fusion,* vol. 40, pp. 57-75, 2018.

[30] D. Garagic, G. Von Pless, R. Hagan, F. Liu, J. Peskoe, P. Zulch and B. J. & Rhodes, "Unsupervised Upstream Fusion of Multiple Sensing Modalities Using Dynamic Deep Directional-Unit Networks for Event Behavior Characterization," in *2019 IEEE Aerospace Conference*, Big Sky, MT, USA, 2019.

[31] D. Shen, E. Blasch, P. Zulch, M. Distasio and J. L. R. Niu, "A Joint Manifold Leaning-Based Framework for Heterogeneous Upstream Data Fusion," *Journal of Algorithms and Computational Technology (JACT),* vol. 12, no. 4, pp. 311-332, 2018.

[32] M. Robinson, J. Henrich, C. Capraro and P. Zulch, "Dynamic sensor fusion using local topology," in *2018 IEEE Aerospace Conference*, Big Sky, MT, 2018.

[33] D. Gunning, "Explainable Artifical Intelligence (XAI)," in *DARPA*, 2017.

[34] S.-M. Moosavi-Dezfooli, A. Fawzi and P. Frossard, "DeepFool: a simple and accurate method to fool deep neural networks.," *Seyed-Mohsen ,* pp. 2574-2582, 2016.

[35] Z. Lipton, "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery," *Association for Computing Machinery,* vol. 16, no. 3, pp. 1542-7730, 2018.

[36] B. Letham, C. Rudin, T. H. McCormick and D. Madigan, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," *The Annals of Applied Statistics,* vol. 9, no. 3, pp. 1350-1371, 2015.

[37] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-cam: Visual explanations from Deep Networks via Gradient-based Localization," *International Journal of Computer Vision,* vol. 128, pp. 336-359, 2019.

[38] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs and H. Lipson, "Understanding Neural Networks Through Deep Visualization," *ArXiv,* vol. abs/1506.06579, 2015.

[39] A. Dosovitskiy and T. Brox, "Inverting Visual Representations with Convolutional Networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* vol. 6, pp. 4829-4837, 2016.

[40] J. T. Springenberg, A. Dosovitskiy, T. Brox and M. Riedmiller, "Striving for Simplicity: The All Convolutional Net," *ICLR (workshop track),* 2015.

[41] M. D. Zeiler and R. Fergus, Visualizing and Understanding Convolutional Networks, Springer, 2014.

[42] G. Montavon, S. Bach, A. Binder, W. Samek and K.-R. Müller, "Explaining NonLinear Classification Decisions with Deep Taylor Decomposition," *Pattern Recongition ,* vol. 65, pp. 211-222, 2017.

[43] M. Du, N. Liu, Q. Song and X. Hu, "Towards Explanation of DNN-based Prediction with Guided Feature Inversion," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London, 2018.

[44] A. Henelius, K. Puolamäki and A. Ukkonen, "Interpreting Classifiers through Attribute Interactions in Datasets," in *arXiv:1707.07576*, 2017.

[45] P. Zulch, M. Distasio, T. Cushman, B. Wilson, B. Hart and E. Blasch, "ESCAPE Data Collection for Multi-Modal Data Fusion Research," *2019 IEEE Aerospace Conference,* pp. 1-10, 2019.

[46] V. Vaquero, A. Sanfeliu and F. Moreno-Noguer, "Hallucinating Dense Optical Flow from Sparse Lidar for Autonomous Vehicles," in *International Conference on Pattern Recognition (ICPR)*, Beijing, China, 2018.

[47] T. G. p. i. G. Josef Bigun, "Two-Frame Motion Estimation Based on Polynomial Expansion," in *SCIA'03: Proceedings of the 13th Scandinavian conference on Image analysis*, Berlin, Heidelberg, 2003.

[48] G. Andrew, R. Arora, J. Bilmes and K. Livescu, "Deep Canonical Correlation Analysis," in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, Atlanta, GA, USA, 2013.

[49] J. Black, T. Ellis and P. Rosin, "A novel method for video tracking performance evaluation," *Joint IEEE Intl. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance,* pp. 125-32, 2003.

[50] K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi and C. Aggarwal, "Efficient Data Representation by Selecting Prototypes with Importance Weights," *International Conference on Data Mining,* pp. 260-269, 2019.

[51] W. Yang, Y. Gao, Y. Shi and L. Cao, "MRM-lasso: A sparse multiview feature selection method via low-rank analysis," *IEEE Transactions on neural networks and learning systems,* vol. 26, no. 11, pp. 2801-2815, 2015.

[52] A. Orynbaikyzy, U. Gessner and C. Conrad, "Crop type classification using a combination of optical and radar remote sensing data: a review," *International Journal of Remote Sensing,* vol. 40, no. 17, pp. 1-43, 2019.

[53] J. J. Zhang, A. Papandreou-Suppappola and M. Rangaswamy, "Multi-target tracking using multi-modal sensing with waveform configuration," in *EEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, 2010.

[54] G. Fasano, D. Accardo, A. C. C. Moccia, U. Ciniglio, F. Corraro and S. Luongo, "Multi-Sensor-Based Fully Autonomous Non-Cooperative Collision Avoidance System for Unmanned Air Vehicles," *Journal of Aerospace Computing, Information,* vol. 5, no. 10, pp. 338-360, 2008.

[55] S. Kemkemian and M. Nouvel, "Sense-and-Avoid System Based on Radar and Cooperative Sensors," in *Encyclopedia of Aerospace Engineering*, John Wiley & Sons, Ltd., 2015, pp. 1-