3D Face Recognition from RGB Camera and Radar Sensor

Stergious Aji

School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
G12 8QQ

MSci Interim Report

December 12, 2023

Contents

1 Introduction

1.1 Motivation

Facial recognition technology is a crucial field of research in computer vision, due to its broad applications in areas such as security surveillance, forensic analysis, human-computer interaction, and healthcare. Its most notable use case being the biometric authentication of individuals, in order to allow access to personal devices or restricted areas. This enables a non-invasive, hands-free approach to identity verification, removing the need to recall passwords. Furthermore, facial biometrics are naturally more accessible than other forms such as fingerprints, iris, or palm prints.

Since its inception in the 1960s, face recognition systems have evolved significantly. The pioneering work by Bledsoe [1] distinguished faces by comparing distances of manually annotated landmark features such as the nose, eyes, and mouth. In more recent years, the advent of Deep Learning has greatly enhanced the performance and efficiency of human face classification, benefiting from the vast online databases of face images. Nevertheless, these systems primarily rely on images captured by RGB cameras, making them susceptible to variations in lighting and pose. By integrating the depth of facial attributes, which draws attention to the geometric details of the face alone, the effect of such environmental factors can be mitigated. Additionally, transitioning to 3D facial recognition not only improves accuracy but also enhances the security of biometric systems.

The popularity of 3D face recognition is on the rise, evidenced by its adoption in smartphones with the likes of Apple and their Face ID [2] technology. This growing demand has pushed the miniaturisation of depth-sensing technology to smaller form factors, enabling it to operate efficiently in real-time on mobile devices with minimal computational power. Depth cameras used in this context typically employ an active face acquisition method, where non-visible light is projected onto the face and reflected back, allowing sensors to measure and map facial features. The most common approach uses Lidar, emitting waves in the Near-Infrared spectrum, due to its ability to capture a dense 3D map of the subject's face. However, its weakness in penetrating materials like clothing and hair is a notable limitation. In contrast, millimetre Radar waves (mmWaves) can penetrate such obstacles to directly reach the dermal layer of the skin [3], potentially offering better performance in occlusion scenarios or even within the presence of rain or fog.

Research into the efficacy of Radar waves for 3D face recognition is relatively sparse but recent studies show positive results [4, 5, 6, 7, 8], discussed in Section 2.3. Radar technology is generally more cost-effective, both in terms of acquisition and computationally-speaking as it consumes less power in comparison to the more widely used Lidar cameras. However, the lower accuracy and sparsity of millimetre wave signatures may hinder its effectiveness in facial recognition. A solution is to integrate RGB information with the depth data obtained by Radar sensors to enhance the system's ability to accurately learn and identify facial characteristics.

1.2 Aims

This project aims to explore the effectiveness of using RGB cameras in conjunction with mmWave Radar sensors for 3D facial recognition. Since there are no appropriate datasets available for this purpose, we will be required to collate this data ourselves. We plan to use the Intel RealSense L515 camera [9] for capturing RGB images of an individual's face. Meanwhile, the Google Soli Radar sensor [10] will be employed to gather depth information through reflected millimetre waves.

Given the necessity of data collection, our goal is to gather facial data from approximately 50 participants. This number is expected to be sufficient for both training and evaluating our proposed model within the limited timeframe. We aim to obtain face data under various conditions including diverse poses, lighting environments, and common occlusion scenarios. The objective is to empirically validate the benefits of utilising mmWave technology in this context. We hypothesise that this approach would yield a system that is invariant to pose, lighting, and occlusion compared to using RGB or depth information alone.

Next, we plan to develop a novel face recognition model using a Deep Convolutional Neural Network. This model will be trained on the captured data in order to learn facial features from both the RGB and depth characteristics acquired from the Soli sensor. We also intend to investigate different techniques in fusing these two modalities in order to identify the most effective strategy that provides rich and distinctive representations for accurate face classification performance.

2 Background Survey

Before any implementation, a review of relevant literature in the field is undertaken and summarised in the following chapter. This is essential in corroborating best practices and gaining a deeper insight into the strengths and limitations of mmWave Radar in the context of face recognition. A look into existing databases compatible with our research objectives is then performed, followed by a detailed justification for the need to compile our own dataset. Furthermore, a critical analysis of the recent work studying Radar-based human face classification is provided. Finally, an outline of multimodal data fusion techniques and deep learning architectures for face recognition will be laid out.

2.1 mmWave Radar Technology

Radio Detection and Ranging, or Radar, has been around for decades and is instrumental in fields such as space exploration, military and commercial aviation, maritime navigation, as well as, meteorology. Recently, the miniaturisation of Radar sensors to the millimetre wave (mmWave) band has brought its application to more small-scale domains [11]. A notable example is Google's integration of the Soli sensor into their Pixel 4 smartphones

for facial detection and motion gesture recognition [12]. This is the exact sensor we plan to utilise during this project's data collection phase. A key driving factor for this choice is the Soli's use of Frequency Modulated Continuous Wave (FMCW) technology. This is proven to offer superior range resolution in comparison to other modulation techniques due to its high pulse compression [13], a vital aspect for generating accurate facial embeddings.

mmWave sensing is also being explored in the domain of autonomous vehicles, specifically in systems such as collision warnings and adaptive cruise control [14]. This is primarily due to its edge over traditional Near-Infrared waves employed by Light Detection and Ranging (Lidar) cameras, particularly in its resilience against atmospheric conditions such as dust, smoke, fog, and rain [15]. This penetrative power of mmWaves makes it a promising candidate for reliable facial recognition in diverse, real-world scenarios. However, it is important to note the trade-off as mmWaves tend to have lower accuracy in comparison, which could impact 3D facial recognition performance where precision in detecting and mapping facial features is paramount. This project will therefore explore counter-balancing this limitation with the information gained from RGB images, potentially paving the way for more resilient and versatile systems.

2.2 Data Acquisition

Various techniques exist for capturing 3D face data which can be broadly categorised into two main types: *Active* and *Passive* systems [16]. Active systems emit non-visible light onto the target and use the reflected light to construct a 3D point cloud of the subject. Contrastingly, passive systems rely on the available light in the scene to capture facial features. For instance, stereoscopic cameras use two or more cameras to capture images of the subject from different perspectives to enable depth perception. In addition, 3D facial features can be inferred using shape-from-shading techniques which analyse the luminance values contained within grayscale images [17].

Passive systems, advantageous for their ability to operate in real-time without light emission, are highly influenced by environmental lighting conditions. Active systems, in contrast, are more robust as they use their own light to illuminate the subject, permitting functionality even in dimly lit settings. Active acquisition often involves structured light and triangulation methods to gather depth information. In the case of Lidar cameras like the Intel RealSense, the time-of-flight of emitted light is measured to gauge the distances of points on the target. The Soli sensor is a form of active acquisition, incorporating a single transmitting antenna and three receiving antennas. The system measures the phase difference and Doppler shift of reflected mmWaves to estimate the distance and radial velocity of the target.

When collecting data on human faces for 3D face recognition, it is crucial to ensure the data comprises a wide range of facial poses, expressions, lighting conditions and occlusion scenarios. Unlike standard images, the pure geometric information from 3D face scans ensure that models are insensitive to pose and lighting changes during training. Research by [18] found that the maximum pose angle that their Local Binary Pattern-based model is

robust against, is 60°. For this reason, the poses we capture will not exceed this since our goal is not to investigate extreme pose invariance, instead the capabilities of Radar waves. Furthermore, a diverse range of genders, ages, and ethnicities is essential for real-world applicability.

In recent years, the number of 3D face databases available has grown, encompassing various acquisition techniques and devices. Noteworthy datasets include the BU-3DFE [19] and the FRGC [20] database, widely accepted as standard references for evaluating the performance of 3D face recognition systems. The BU-3DFE dataset, focusing on expression variance, contains six types of expressions from 100 individuals, captured using stereo photography. While these datasets are unsuitable for our project, the data collection procedure used to amass them provide valuable insights. Presently, there is only one public database featuring Radar signatures of 206 human faces [21]. This dataset was captured with a Qualcomm 60 GHz mmWave Radar, however, lacks any RGB face images of the participants in the study. These factors motivate building our own dataset including both RGB images and mmWave Radar face signatures of subjects. Such an approach affords us the flexibility in tailoring our experiment design to investigate the model's effectiveness under specific conditions, all while adhering to established data collection protocols.

2.3 Prior Work on Radar-based Face Recognition

The use of mmWaves for human face identification is a relatively new research field driven by the advancement of Radar sensor technology to the millimetre wave band. One of the earliest papers found to investigate human identification using mmWaves dates back to 2019 [22]. While this paper focuses on the simultaneous classification of people by their gait and body shape rather than facial features, it displayed the ability of mmWaves to encapsulate the subtle idiosyncrasies among individuals. These nuanced differences are crucial for learning models to effectively distinguish between different people, leading to high classification accuracies.

Following this, Hof et al. [4] proposed a Deep Neural Network (DNN) based Autoencoder that is able to distinguish human faces captured by an 802.11ad/y networking chipset operating at a centre frequency of 60 GHz. The Autoencoder is able to encode mmWave face signatures of over 200 individuals with enough separation to distinguish positive and negative instances by measuring their Mean Squared Error (MSE) against reference facial embeddings. The study conducted an extensive data collection process, capturing mmWave face signatures of 206 individuals of varying genders and ages, in five different poses: frontal, as well as, 15° and 25° rotations to the left and right. This dataset was subsequently made available through an IEEE Data Port [21]. While this dataset encapsulates faces from a wide range of people, including some with beards and spectacles, it lacks representation of other common occlusion scenarios like head accessories, that our project aims to explore. Moreover, the study utilised a large sensor containing 1024 transmitting and receiving antenna pairs. This is in contrast to the compact Soli chip with a single transmit and three receiver antennas, intended to work in a smartphone. Notably, the study simulated the effect of reducing the antenna count to 10, markedly decreasing the distinctiveness of facial

signatures. Promisingly, increasing the number of neurons in their Neural Network and an additional hidden layer could compensate for this reduction, maintaining high accuracy.

Lim et al. [5] also proposes another Deep Neural Network model, however with a more representative Multi-Layer Perceptron (MLP) architecture where every layer is fully connected to adjacent ones. The study utilised a small-scale 61 GHz FMCW Radar sensor developed by bitsensing Inc. [23], comparable to the Google Soli with a single transmit and three receiver antennas. The model attained a mean classification accuracy of 92% across eight subjects. The study also showed that their DNN approach surpassed the performance of both, a Support Vector Machine (SVM), and tree-based Ensemble Learning approaches, trained on the same mmWave face signatures. It is important to note the relatively small-sized dataset used to train the model raising concerns about potential overfitting as the data is not representative enough. The paper provides limited details on the data collection methodology used, only mentioning that the facial distances ranged from 30 cm to 50 cm. It can be assumed that the study likely focussed on frontal poses without any occlusions for all eight subjects. The research also explored the impact of using a single receiving antenna, which resulted in a reduced accuracy of 73.7%. This finding is in line with Hof et al.'s [4] observation that an increased number of receiving antennas can enhance classification accuracy by capturing more nuanced facial features. The paper also suggests that a CNN may be more appropriate if signals were stacked on the time axis rather than the frequency axis.

During the same period, Kim et al. [6] conducted research using the same 61 GHz FMCW Radar sensor from bitsensing Inc., featuring a range resolution of 2.5 cm. Their study introduced a CNN model composed of three convolutional layers and three fully connected layers. The radar data underwent heavy preprocessing to transform it into a more image-like format suitable for the CNN model. With a data split of 70%/15%/15% for training, validation, and testing, the model achieved an average classification accuracy of 98.7% on a limited dataset of only three individuals. Interestingly, the study also examined the impact of wearing cotton masks. The results showed a minimal drop in average classification accuracy by 0.9%, which is encouraging for the objectives of our project. However, these findings are to be taken with caution due to the small size of the dataset. It remains unclear whether this level of performance would hold consistently across a larger and more diverse group of subjects, with more varied occlusions.

Pho et al. [7] adopts a One-Shot Learning approach to the problem. This is where a model is trained with a single or only a few labelled instances. This is beneficial when there is a lack of training samples available. The proposed method constitutes a Siamese structure of two identical CNNs with shared parameters that map the input Radar signals into the embedding space. A distance metric between the outputs of both CNNs are used in training and testing to measure the similarity of facial inputs. The model is trained for *binary classification* by inputting pairs of face signatures from either the same or different people. This process aims to learn embeddings that push different faces into distinct Euclidean regions of the embedding space. The same bitsensing Inc. BTS60 chipset, used by [5, 6], was employed to capture 500 frames of the faces of eight participants. An average classification of 97.6% was achieved, an improvement from the previous DNN model by Lim

et al. [5] with the same number of people. t-Stochastic Neighbour Embedding (t-SNE) [24] was applied for dimensionality reduction. The resulting visualisations demonstrated that the one-shot Siamese network effectively distinguished each individual's face, simplifying the classification task. While a small dataset is used only encompassing frontal poses with no occlusion, the proposed method is well documented and is likely robust against larger datasets.

Challa et al. [8] employs two different machine learning models on the dataset made available at [21]. Their approach began with a CNN-based Autoencoder followed by a Random Forest Ensemble Learning approach. A total of nine Autoencoders are built, each tailored to different frame rates, focusing on compressing and learning to reconstruct the original data from its compressed, latent form. The Autoencoders are trained using randomly selected data samples from 186 mmWave face signatures from the data port. The flattened and labelled outputs are then used to train and test nine discrete Random Forest models using identical hyperparameters, as recommended by the Sci-kit library. This methodology yielded impressive results, achieving an average classification accuracy of 99.98% using all 1400 frames per individual. Even reducing the number of frames to 70 per person, the model maintained a high accuracy of 97.1%. The paper presents an approach that is unique in comparison to the rest of the research papers tackling this subject, showcasing a model that is able to be deployed on mobile chips.

The research in this area exclusively focuses on utilising data from Radar sensors, largely driven by concerns of privacy preservation. However, a significant limitation of this approach is the required duration for capturing an accurate facial scan. The sensor needs to operate for several seconds, typically in the range between 10 and 15 seconds, in order to obtain a detailed scan. Such a time frame is impractical in real-world situations, as it necessitates the subject to remain motionless for a prolonged period. Up to this point, no study was found to explore the potential benefits of combining Radar signatures with corresponding RGB data to enhance facial recognition capabilities. Given the high performance of existing deep learning models using 2D RGB data alone, such as InsightFace [25], integrating these models with mmWave Radar data presents a promising avenue. This combination could accelerate data acquisition while leveraging the advantages of mmWaves in terms of their robustness to lighting variations and occlusions.

2.4 InsightFace

In the evolving field of face recognition, deep CNNs have emerged as a dominant approach due to their ability to extract rich, discriminative facial features from images. One significant advancement in this area is the InsightFace toolkit, implementing algorithms designed to address the intricacies of face analysis and recognition. Key works include the preliminary ArcFace model, introduced by Deng et al. [25], alongside the robust Face Alignment model by Gho et al. [26]. ArcFace employs a novel Additive Angular Margin Loss to maximize class separability, further enhancing the discriminative power in mapping face images to feature embeddings. However, this method was found to face challenges with label noise, requiring the "cleaning" of many real-world images sourced from the web. To address this,

further progress was made with the Sub-center ArcFace model [27], introducing the concept of sub-classes to boost resilience against intra-class variations and label noise. It achieved state-of-the-art performance on many widely used benchmark datasets such as the Labeled Faces in the Wild (LFW) [28] and the YouTube Faces (YTF) datasets [29].

The integration of pretrained models offered by InsightFace into our system enables us to concentrate primarily on enhancing the performance of our CNN learning mmWave face signatures. By fusing the depth and contour detection capabilities of mmWave Radar with the rich textural features gathered by ArcFace from RGB images, the system has the potential to attain improved accuracy and robustness. This approach is particularly promising in environments where conventional optical methods falter.

2.5 Multimodal Data Fusion Techniques

Multimodality, as defined by Lahat et al. [30], refers to the use and analysis of multiple types of data, potentially arriving from multiple sensors. The aim is to extract and blend salient information gathered by each sensor. The integration of this diverse data leads to outputs with richer representations than what could be achieved by the individual modalities alone. We hypothesise that coupling the colour information from face images with the depth gathered by the Radar sensor could greatly improve class separation, and subsequently, face recognition performance.

A common technique involves fusing the multiple data modalities before feeding them into a learning model, referred to as **Early Fusion**, or **Data-level Fusion**. It includes combining data by removing correlations between sensors or fusing data in a common, lower-dimensional space [31]. Techniques such as Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA) are commonly employed for this purpose. One key issue with applying early fusion is ensuring synchronisation between the RGB and Radar frames, which is difficult due to their significantly different sampling rates. Furthermore, the continuous mmWave signals must be effectively discretised to match the form of the RGB data. Moreover, a major disadvantage of early fusion is the potential to squash critical information present within each individual modality, impacting the training efficacy.

Late Fusion, or **Decision-level Fusion**, operates by independently processing different data sources through separate models and then fusing them at the decision-making stage. A standard approach involves taking a weighted average of the separate predictions, providing a way to minimise or maximise the influence of specific modalities [32]. Late fusion is often simpler and flexible, and can be effective when dealing with extremely dissimilar data sources either in terms of sampling rate, dimensionality, or unit of measurement. Additionally, late fusion often yields better performance since errors from multiple models are dealt with independently.

Intermediate Fusion or **Feature-level Fusion** is based on DNN architectures and is the idea of combining different modalities in the feature space where there is a higher level of abstraction of the raw data. This can be as straightforward as a simple concatenation

of the individual latent embeddings or utilising Autoencoders for non-linear feature fusion [33]. This approach offers greater versatility than early and late fusions, as it allows for the integration of features at various depths within the neural network. However, it can lead to challenges such as a risk of overfitting or a failure in learning relationships between the different modalities.

Each data fusion technique comes with its own set of challenges and considerations, necessitating experimentation to determine the most effective way to merge the RGB and mmWave signatures. A variant of late-intermediate fusion is the most feasible where the embeddings from the last layers of each model are combined. It would be challenging to attempt early fusion due to the substantial differences between the two modalities. Such integration would likely require heavy preprocessing of the Radar data, potentially involving its conversion into a depth image-like format.

2.6 Deep Learning for Face Recognition

The most popular methodology observed for creating face recognition models involve deep learning, specifically through the use of Convolutional Neural Networks (CNN). Before proceeding to develop our own model to learn mmWave face signatures, it is vital to grasp the underlying concepts involved and strategies employed in recent research. This understanding will guide us in identifying the most effective strategy for our project.

CNNs are a class of deep neural networks, most commonly applied to the analysis of image data. Inspired by the structure of the animal visual cortex, they are designed to automatically and adaptively learn spatial hierarchies of features from input images. A typical CNN architecture comprises a sequence of layers, including convolutional layers for feature detection, pooling layers for dimensionality reduction, and fully connected layers for final output processing [34].

Unlike traditional algorithms, CNNs learn to identify the necessary features for classification directly from the input data, minimising the need for manual feature extraction. This is particularly advantageous in face recognition where nuanced features like edges, textures, and specific parts of the human face need to be identified. CNNs are able to do this using their deep hierarchical layers of abstraction. Lower layers may identify contours and simple textures, while deeper layers can detect more complex features like the shape of the nose, eyes and mouth. The same weights are shared across different parts of the input, reducing the number of parameters and thereby the complexity of the model. This makes CNNs particularly memory efficient than classical neural networks in dealing with data with high dimensionality such as mmWave signatures.

In conclusion, the use of CNNs in face recognition is justified by their ability to automatically learn and generalize from data, their efficiency in handling high-dimensional data, and their robustness against variations in input data. These characteristics make CNNs a powerful tool for developing accurate and efficient face recognition systems.

3 Proposed Approach

The following section delineates our proposed methodology in tackling the Radar-based facial recognition task. Each decision and step taken thus far will be justified, including an outline of potential workaround plans for problems and intricacies that may arise during the project lifecycle.

3.1 Data Acquisition and Experiments

Following a thorough research of the field, the next steps involve designing and conducting the data acquisition process necessary to train our proposed model with. These experiments require careful planning since the data collected here directly determines the effectiveness of the resulting model. As found in Section 2.2 in the Background, it is vital to compile multiple poses in order for the model to learn a comprehensive 3D scan of the individual's face. Furthermore, it imposes pose-invariance into the system as people would not always provide an exact frontal pose to a face recognition system in real-world use-cases. Most studies only focus on variations in the yaw axis since a person is less likely to tilt or pitch their head by a significant angle in comparison to the left and right rotation of the face. For this reason, we will also focus primarily on head rotation in the yaw axis, specifically capturing face poses of 0°, and 30° and 45° to the left and right of the sensors indicated by positive and negative angles.

Since this experiment aims to explore the benefits of mmWave sensors in the context of face recognition, two different lighting conditions will also be incorporated in the data collection experiments. Namely, regular and dim lighting scenarios. We hypothesise that the mmWave face signatures would be invariant to the environmental lighting due to it using its own active illumination of the target face compared to the RGB camera. Therefore, if the system is able to achieve a higher accuracy using both modalities compared to solely using the RGB information, it would conclusively show that mmWaves offer robustness against lighting conditions.

Finally, we aim to investigate the penetrative power of mmWaves to directly reach the skin through cloth and hair by injecting common occlusion scenarios into our experiments. It would be beneficial if facial recognition systems are robust against common occluding head accessories such as hats, glasses, masks and so on. Currently, individuals would have to take these accessories off in order for the system to thoroughly identify and allow them access to the particular device or area. With mmWaves, we hypothesise that this may not be required since facial features could be captured regardless. This could greatly benefit security surveillance where individuals deliberately obscure their faces to hide their identities. In our experiment, we aim to capture scenarios both with and without occlusion. Since cotton masks have already been investigated by Kim et al. [6], other common items like hats, sunglasses and scarves would be used to mirror day-to-day scenarios.

On top of investigating environmental invariances, our study also aims to identify the

minimal amount of Radar data required to still yield easily separable . Previous work are observed to capture 200 to 2,000 total frames which, depending on the sampling rate, requires running the sensor for eight seconds and a maximum of 80 seconds for each scenario. This is due to the lower accuracy of mmWaves requiring a longer exposure to provide a dense enough representation. Nevertheless, anything over three seconds is an impractical amount of time for an individual to keep their face still for. However, since our model will utilise the extra RGB information, this may mean that only a few seconds of Radar frames are needed to still identify the individual. For this reason, all scenarios in our experiment will involve capturing 10 RGB frames and a maximum of 250 frames worth of mmWave bursts which maps to running the Google Soli for 10 seconds at a sampling rate of 25 Hz. This allows a sufficient amount of data to investigate the minimal number of mmWave frames required.

In order to gather a wide variety of faces, we aim to get 50 participants given the tight timeframe of this project. Following ethical guidelines of sensitive personal information, the participants will comprise of adults mainly university students. However, this over-representation of adults aged 20–25 should not impact our study since age invariance is not something that is being investigated. A total of 15 scenarios will be captured for each participant at a distance of 20 cm from the sensors. Each time the sensors will be run for 10 seconds, totalling 150 RGB frames and 3,750 mmWave frames per person.

3.2 Proposed Model and *mmFace*

Following the research done in Section 2.4 in the Background chapter, it is clear that the ArcFace model available in the InsightFace toolkit is the best choice for this project. It achieves state-of-the-art classification results, outperforming the previous bests such as Facebook's DeepFace [35] and Google's FaceNet [36]. This allows treating the handling of the RGB data with a black-box model and a focus on perfecting the Radar-based model we are naming *mmFace*. Furthermore, this enables exploration into the various ways in fusing the two modalities as discussed in Section 2.5. Figure 1 shows a high-level plan of the model architecture described here.

Our proposed model will be built using a CNN-based architecture since it works best with image-like data. The Radar bursts from the data collection phase will be transformed into a Complex Range Doppler (CRD) map [10, 37] prior to feeding it into the model. Using the short-range configuration of the Google Soli at a centre frequency of 60 GHz with a maximum bandwidth of 5.5 GHz. Bursts are sampled at 25 Hz. The Intel RealSense is set to capture RGB-D frames at a sampling rate of 30 frames per second (FPS). Due to the difference in sampling rates between the two sensors, timestamps information is stored for the possibility of synchronising the two modalities for early data fusion.

As explained before, the data fusion techniques that will be explored are late feature-level fusion and late decision-level fusion. Purely intermediate feature-level fusion is not feasible due to the black-box treatment of the InsightFace model making it difficult to integrate features from both modalities within hidden layers. However, a late feature-level fusion

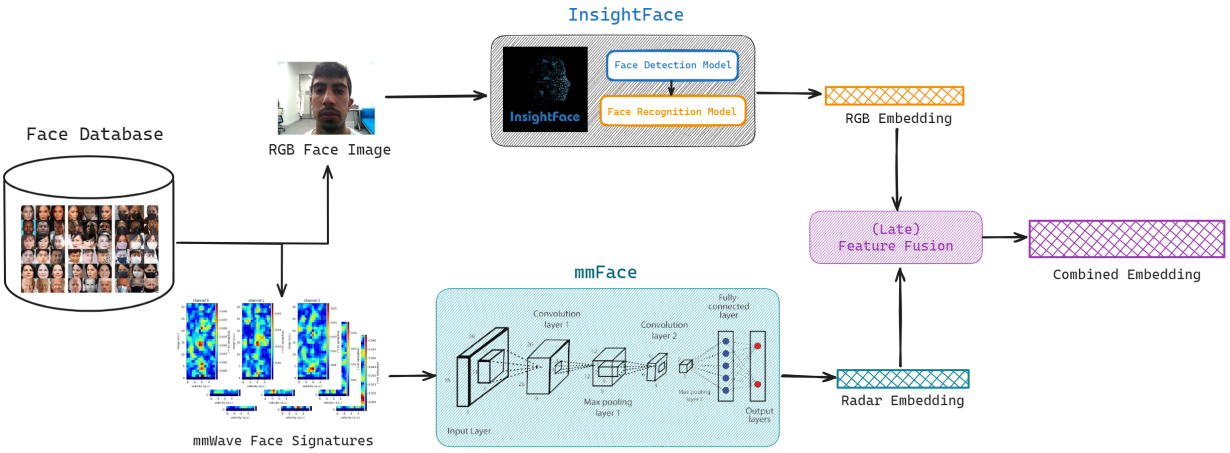

Figure 1: High-level Model Architecture

is possible combining the outputs of the last layers of each model to form an embedding combining both the RGB and mmFace features. Similarly, decision-level fusion will be explored since this involves mixing the predictions from the individual models. Early data-level fusion would be the hardest to perform due to the dissimilarities between the two modalities in sampling rate and data format. The CRD maps must be synchronised and transformed into a depth image-like format before combining with the RGB images. Furthermore, this would require a whole new training cycle with the mmFace model which may be infeasible within the timeframe of the project. If time is available, then this will be looked at.

We plan to adopt a ResNet-based architecture for mmFace due to its refinements over previous architectures such as AlexNet [38] and VGGNet [39]. ResNet [40] provides "skip connections" and residual blocks to resolve the vanishing gradient problem present in the VGGNet allowing the scaling of the network beyond the maximum of 19 layers.

The dataset will be split by subjects into 80%/10%/10% for training, validating and testing. This is due to the small size of the dataset of 50 subjects requiring a higher number of training samples for the mmFace model to learn effectively from. Following training, the testing phase will evaluate the distinctiveness of the output embeddings for distinct individuals. For accurate classifications, the individual faces must be spatially separable in the embedding space in order to discretely identify a person's face without ambiguity. t-SNE visualisations [24] of the higher dimensional embeddings will also be employed to ensure this is the case to compare and contrast the attempt to classify the original data compared to with the output mappings. In addition, standard classification accuracies will be calculated to verify the identity recognition capabilities using the model against randomly selected ground truths. This will also allow comparisons with the results obtained from previous work with Radar-based 3D facial recognition.

3.3 Progress

Thus far, work has been conducted providing confidence that completion is feasible. The data acquisition has begun successfully with the faces of 18 participants having been captured with the plan laid out in Section 3.1. We plan to accelerate this process in January in order to reach the required 50 participants with advertisements. The feasibility of InsightFace has also been evaluated with a short experiment in recognising the faces of famous NBA basketball players. The model achieved high accuracies, trained with a variety of images sourced from the web. All work is being implemented in Python due to its wide array of powerful libraries useful for machine learning such as NumPy [41] and PyTorch [42].

To illustrate the data acquisition process used thus far, data samples of a single subject is provided in the left half of Figure 2. This grid shows all 15 scenarios laid out by the three conditions along the rows and the five pose variations along the columns. For brevity, the conditions are abbreviated as outlined in Table 1.

Abbreviation	Expanded Forms
NO	No Occlusion
O	Occlusion
RLC	Regular Lighting Condition
DLC	Dim Lighting Condition

Table 1: Table displaying full forms for abbreviations describing experiment conditions.

Work has been conducted in transforming the data collected from the acquisition experiments into suitable formats for input into the proposed models. We are able to successfully transform the Radar bursts into a CRD map which can be plotted showing intensities of received waves into discrete Doppler bins along the x-axis and Range bins along the y-axis. A data sample of a single CRD frame can be observed in the right half of Figure 2.

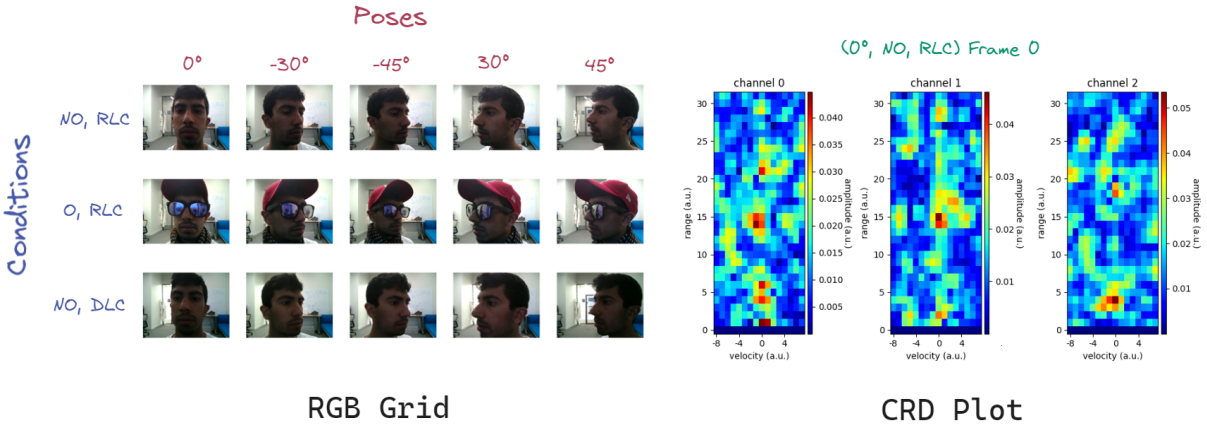

Figure 2: Data samples collected for Subject 0. The left figure shows the RGB frames of all 15 scenarios organised by pose and condition. The right figure plots a single CRD frame showing amplitudes of reflected waves detected by the three receiving channels of the Soli categorised into discrete Range-Doppler bins

4 Work Plan

As the project involves a recently emerging research field of mmWave-based applications, a large amount of prerequisite knowledge is required before implementing a solution to the identified research problem of 3D face recognition. This research process will continue throughout the remaining project lifecycle since there are still hurdles to overcome.

In order to alleviate the effects of these hurdles, it is important to look ahead and postulate the main problems that can occur by delineating a thorough plan of the next steps. As stated in the Progress Section 3.3, the data acquisition process is planned to be accelerated in January in order to gather the remaining 32 participants needed to reach our goal. Simultaneously, it is possible to implement and experiment with different Convolutional configurations of our proposed mmFace CNN that is most optimal with the CRD data format. This allows a sufficient amount of time to train, validate and test our model with the decently sized dataset we will collect. In addition, various different models will be trained with varyingly missing amounts of the dataset in order to evaluate the robustness of utilising both modalities over individually.

This will be followed by an evaluation stage involving t-SNE visualisations and classification accuracy experiments in order to empirically gauge the final model's performances against models trained with different amounts of the data. The performance will also be compared with the previous work with mmWave-based face recognition models. The work plan delineated here can be visualised in Figure 3 in the form of a Gantt chart.

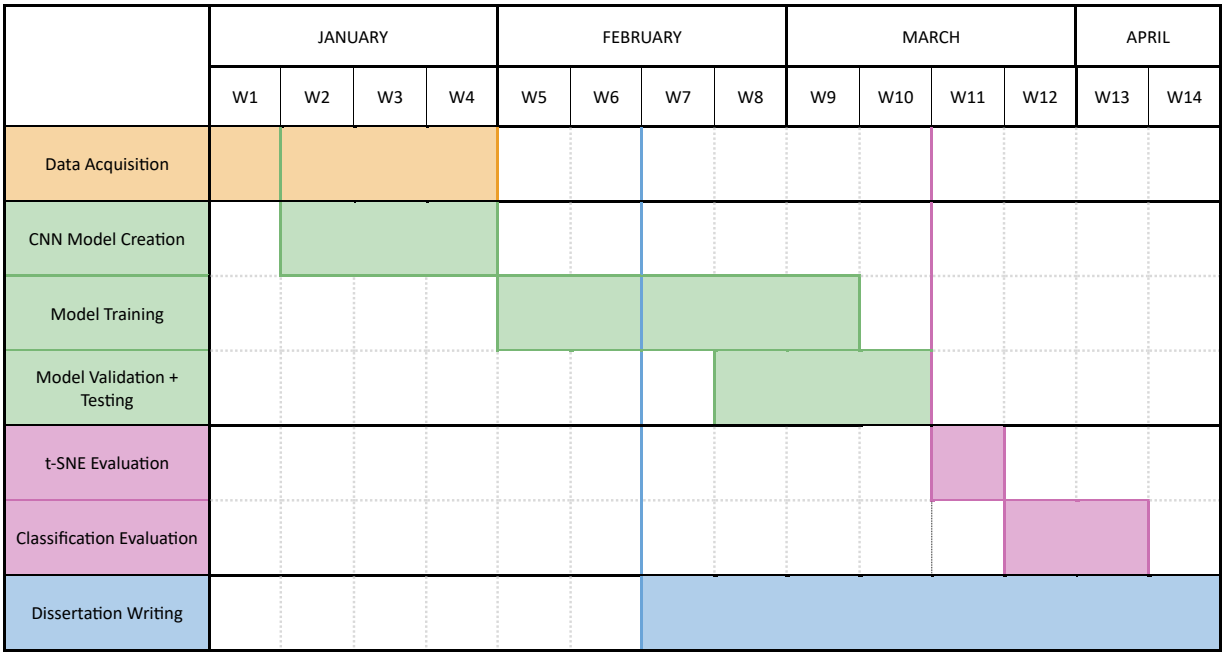

Figure 3: Gantt Chart showing work plan for next semester

References

[1] Woodrow Wilson Bledsoe. The model method in facial recognition. *Panoramic Research Inc., Palo Alto, CA, Rep. PR1*, 15(47):2, 1966.

[2] Apple Inc. About Face ID advanced technology, 2023. Accessed: 2023-11-19 https://support.apple.com/en-gb/102381.

[3] David R Vizard and R Doyle. Advances in millimeter wave imaging and radar systems for civil applications. In *2006 IEEE MTT-S International Microwave Symposium Digest*, pages 94–97. IEEE, 2006.

[4] Eran Hof, Amichai Sanderovich, Mohammad Salama, and Evyatar Hemo. Face verification using mmwave radar sensor. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 320–324, 2020.

[5] Hae-Seung Lim, Jaehoon Jung, Jae-Eun Lee, Hyung-Min Park, and Seongwook Lee. Dnn-based human face classification using 61 ghz fmcw radar sensor. *IEEE Sensors Journal*, 20(20):12217–12224, 2020.

[6] J Kim, J-E Lee, H-S Lim, and S Lee. Face identification using millimetre-wave radar sensor data. *Electronics Letters*, 56(20):1077–1079, 2020.

[7] Ha-Anh Pho, Seongwook Lee, Vo-Nguyen Tuyet-Doan, and Yong-Hwa Kim. Radar-based face recognition: One-shot learning approach. *IEEE Sensors Journal*, 21(5):6335–6341, 2021.

[8] Muralidhar Reddy Challa, Abhinav Kumar, and Linga Reddy Cenkeramaddi. Face recognition using mmwave radar imaging. In *2021 IEEE International Symposium on Smart Electronic Systems (iSES)*, pages 319–322, 2021.

[9] Intel Corporation. Intel RealSense LiDAR Camera L515, 2023. Accessed: 2023-11-19 https://www.intelrealsense.com/lidar-camera-l515/.

[10] Jaime Lien, Nicholas Gillian, M Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. Soli: Ubiquitous gesture sensing with millimeter wave radar. *ACM Transactions on Graphics (TOG)*, 35(4):1–19, 2016.

[11] A Soumya, C Krishna Mohan, and Linga Reddy Cenkeramaddi. Recent advances in mmwave-radar-based sensing, its applications, and machine learning techniques: A review. *Sensors*, 23(21):8901, 2023.

[12] Nicholas Gillian Jaime Lien. Soli: Radar-based perception and interaction, 2020. Accessed: 2023-11-25 https://blog.research.google/2020/03/soli-radar-based-perception-and.html.

[13] Bassem R Mahafza. *Radar systems analysis and design using MATLAB*. Chapman and Hall/CRC, 2005.

[14] DF Robot. Eight Practical Applications of mmWave Radar Technology, 2023. Accessed: 2023-11-19 https://www.dfrobot.com/blog-1650.html.

[15] Cadence Design Systems. mmwave radar applications and advantages, 2022. Accessed: 2023-11-25 https://resources.system-analysis.cadence.com/blog/msa2022-mmwave-radar-applications-and-advantages.

[16] Song Zhou and Sheng Xiao. 3d face recognition: a survey. *Human-centric Computing and Information Sciences*, 8(1):1–27, 2018.

[17] Berthold KP Horn. Understanding image intensities. *Artificial intelligence*, 8(2):201–231, 1977.

[18] Utsav Prabhu, Jingu Heo, and Marios Savvides. Unconstrained pose-invariant face recognition using 3d generic elastic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1952–1961, 2011.

[19] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE, 2006.

[20] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 947–954. IEEE, 2005.

[21] Evyatar Hemo, Amichai Sanderovich, and Eran Hof. mmwave radar face signatures, 2018. https://dx.doi.org/10.21227/wr67-kx23.

[22] Peijun Zhao, Chris Xiaoxuan Lu, Jianan Wang, Changhao Chen, Wei Wang, Niki Trigoni, and Andrew Markham. mid: Tracking and identifying people with millimeter wave radar. In *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 33–40. IEEE, 2019.

[23] Bitsensing. BTS60 Technical Specification, May 2020. Accessed: 2023-12-01 http://bitsensing.com/pdf/Technical_Specification_InCabinRadar_miniV.pdf.

[24] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[25] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

[26] Jia Guo, Jiankang Deng, Niannan Xue, and Stefanos Zafeiriou. Stacked dense u-nets with dual transformers for robust face alignment. In *BMVC*, 2018.

[27] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *Proceedings of the IEEE Conference on European Conference on Computer Vision*, 2020.

[28] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[29] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534. IEEE, 2011.

[30] Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

[31] Bahador Khaleghi, Alaa Khamis, Fakhreddine O Karray, and Saiedeh N Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information fusion*, 14(1):28–44, 2013.

[32] Maciej Pawłowski, Anna Wróblewska, and Sylwia Sysko-Romańczuk. Effective techniques for multimodal data fusion: A comparative analysis. *Sensors*, 23(5):2381, 2023.

[33] Francisco Charte, David Charte, Salvador García, María J del Jesus, and Francisco Herrera. A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software, and guidelines. *arXiv preprint arXiv:1801.01586*, 2018.

[34] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[35] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[37] Eiji Hayashi, Jaime Lien, Nicholas Gillian, Leonardo Giusti, Dave Weber, Jin Yamanaka, Lauren Bedal, and Ivan Poupyrev. Radarnet: Efficient gesture recognition technique utilizing a miniature radar sensor. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[41] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.