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

November 29, 2023

Contents

1 Introduction

1.1 Motivation

PRIVACY

RADAR PENETRATE HAIR, CLOTH [1]

3D FACE RECOGNITION SURVEY [2]

Facial Recognition is a crucial area of research for its wide range of applications spanning security surveillance, forensics, human-computer interaction and healthcare. Its most popular application being access control using biometric authentication. This removes the need to remember passwords and provides a non-invasive, hands-free approach to human verification. Facial metrics are naturally more accessible in comparison to other biometrics like fingerprint, iris or palm print.

Facial recognition systems have come a long way since its dawn in the 1960s. The earliest work by Bledsoe [3] distinguished faces by comparing distances of manually annotated landmark features such as the nose, eyes, ears and mouth. In more recent years, the advent of Deep Learning techniques has greatly improved face recognition performance with help of the sheer number of images of faces online. However, these systems primarily rely on 2D imagery from RGB cameras which are vulnerable to lighting changes and pose variations. To compensate for this, depth information of facial attributes are required. Additionally, moving to 3D facial recognition increases the security of biometric authentication systems.

3D face recognition systems are becoming more popular with the likes of many smartphone companies integrating a type of face unlock such as Apple's Face ID [4]. Furthermore, this demand has pushed this depth sensing technology to smaller form factors and requiring little power and computation to work efficiently on mobile devices on the fly.

Most depth cameras used for this purpose use a form of active face acquisition where non-visible light is emitted and reflected back from a person's face which is subsequently captured by sensors and measured to estimate facial features. The most popular approach uses LiDAR which emit waves in the near-infrared spectrum. The main disadvantage to this is that it is usually too weak to penetrate clothing or hair. In contrast, millimetre waves used in Radar can penetrate thin objects to directly reach the dermal layer of the skin meaning that it may perform better against occlusion like obstacles or even rain or fog.

Very little research has been done in the effectiveness of using Radar waves for 3D face recognition but what has been done show positive results [5, 6, 7, 8, 9]. Radar technology is often less expensive in terms of acquisition cost and computational cost since it requires little energy to power compared to LiDAR cameras. However, Radar has its drawbacks since it is less accurate and sparse which may hinder its performance for facial recognition.

A solution is to combine RGB information with the depth captured by the Radar sensors to effectively learn facial features and identify them.

1.2 Aims

This project aims to investigate the effectiveness of using RGB cameras in conjunction with mmWave Radar sensors for 3D facial recognition. Since there are no easily accessible datasets online for this purpose, we will require acquiring this data ourselves. We will be using an Intel RealSense L515 [10] camera to capture the RGB information of a subject's face. The Google Soli Radar sensor [11] will be used to capture depth information from reflected millimetre waves.

This RealSense camera also includes a LiDAR sensor which produces a more accurate dense depth image. As a backup, a separate model can be trained to transform the sparse Soli data into a more dense representation before inputting into the facial recognition model. However, if the Radar data works well this may not be needed. (MAY NOT INCLUDE)

Since data must be collected we aim to collect faces of around 50 participants which will be enough to train and evaluate our proposed model given the tight timeframe. We aim to acquire face data of different poses, lighting conditions and with multiple occlusion scenarios.

Next we aim to produce our very own face recognition model using a Deep Concurrent Neural Network to learn facial features using both the RGB and depth information acquired from the Soli sensor. We will investigate different data fusion techniques in order to find the best approach to combine RGB features with the Radar data.

2 Background Survey

Before any work is conducted, a review of relevant literature in the field is conducted in order to corroborate best practices and gain a deeper understanding of the strengths and weaknesses of using a Radar sensor, specifically for face identification. Firstly, a look into the datasets available for our purpose and the subsequent reasoning behind the requirement to gather our own dataset will be explored. This section also provides recent study of using machine learning techniques to classify radar signatures of human faces.

2.1 mmWave Radar Technology

FMCW has advantage of better range resolution than other modulation techniques due to high pulse compression [12]

Radio Detection and Ranging, or Radar, has been around for decades and is used in many important applications such as space exploration, military and commercial aircraft and ship navigation, as well as, weather forecasting. However, only recently with the miniaturisation of Radar sensors to the millimetre wave band (mmWave) has research expanded to other small scale domains [13]. Most notably, Google integrated their own Soli sensor for face detection and motion gesture recognition into their Pixel 4 smartphone [14]. Furthermore, mmWave sensing is being explored in the domain of autonomous driving for instance, the application of collision warning and adaptive cruise control systems [15]. mmWaves offer an advantage over the traditional near-infrared waves used by Light Detection and Ranging (Lidar) cameras, with its robustness to atmospheric conditions such as dust, smoke, fog and rain as well as extreme lighting conditions [16]. This is counter-balanced by the lower accuracy of mmWaves in comparison.

2.2 Data Acquisition

There are various techniques available in capturing 3D face data, but they can be categorised into two main types: *Active* and *Passive* [2]. Active acquisition systems emit non-visible light on to the target and use the reflected light information to estimate a 3D representation of the object. Contrastingly, passive systems use the available light in the scene to capture facial feature information. For instance, stereoscopic cameras use two or more cameras to capture images of the target from different perspectives enabling depth perception. In addition, 3D facial features can be inferred using shape-from-shading by analysing the luminance values of a 2D grayscale image.

While passive systems offer advantages of not requiring emission of light and being able to work in real-time, it is highly sensitive to the lighting conditions in the environment. Active systems are robust against this since they use their own light to illuminate the target being able to work in extremely dim conditions.

Active systems often combine the use of structured light and triangulation based methods to capture depth information. In the case of Lidar cameras such as the Intel RealSense, the time-of-flight of emitted light is measured to infer the distances of points on the target. The Radar sensor is also a form of active acquisition using 3 receiving antennas to capture reflected mmWaves from the target, measuring the phase difference and Doppler shift to estimate the distance and velocity of the target.

When collecting data of human faces for 3D face recognition, it is often ideal to ensure the data encompasses a wide range of facial poses and expressions, lighting conditions and occlusion scenarios. Unlike 2D face images, the pure geometric information from 3D face scans ensure the models are insensitive to pose and lighting changes. [17] found that the maximum angle that their Local Binary Pattern (LBP) based model is robust against is 60°. A wide range of genders, ages and ethnicities are also vital for the model to be robust enough against real-world scenarios.

In recent years, the number of 3D face databases available has grown using a variety of

different acquisition techniques and devices. Most notable datasets include the BU-3DFE [18] and the FRGC [19] database, widely accepted as the standard reference database for performance evaluation of 3D face recognition systems. The BU-3DFE dataset primarily focusses on expression-invariance containing 6 types of expressions from 100 individuals using stereo photography. While these datasets are unsuitable for this project, it is useful to survey the data collection process used to amass them. From the extensive research done, there is currently only one public database including Radar signatures of 206 human faces available [20]. This dataset was captured with a Qualcomm 60GHz mmWave Radar, however, does not include any RGB face images of the participants of the study. These factors motivate building our own dataset with both RGB face images and mmWave Radar face signatures. This allows us to investigate the model's effectiveness against varying conditions such as lighting, pose variation and occlusion.

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 miniaturisation of Radar sensor technology to the millimetre wave band. One of the earliest paper found to investigate human identification using mmWaves dates back to 2019 [21]. While this paper focusses on simultaneous classification of people by their gait and body shape rather than facial features, it displayed the ability for mmWaves to capture subtle idiosyncrasies between individuals for machine learning models to achieve accurate classification accuracies.

Following this, Hof et al. [5] proposed a Deep Neural Network (DNN) based Autoencoder that is able to distinguish human faces captured by a 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 a reference encoding of a face. The study involved a decently-sized data acquisition procedure capturing mmWave signatures of 206 individuals of varying genders and ages with 5 poses each: frontal, as well as, 15° and 25° left and right. This dataset was subsequently made available through an IEEE Data Port [20]. While this dataset encapsulates faces from a wide range of people, including some with beards and spectacles, it does not feature other common occlusion scenarios that we aim to investigate such as wearing head accessories. Additionally, the chipset used contained a large sensor containing over 1024 transmitting and receiving antenna pairs. This is in contrast to the compact mobile Soli chip with a single transmit and three receiver antennas. The study also simulated the effect of reducing the number of antennas to 10 which made a significant reduction in the distinctiveness of the faces. Promisingly, increasing the number of neurons and an additional hidden layer to their Neural Network was able to maintain high accuracy.

Lim et al. [6] also proposed another Deep Neural Network model, however with a more representative Multi-Layer Perceptron (MLP) architecture where every layer is fully connected to adjacent layers. The study utilised a small-scale 61 GHz FMCW Radar sensor by bitsensing Inc. similar to the Google Soli with a single transmit and 3 receiver antennas.

The model achieved a mean classification accuracy of 92% on 8 subjects. The study also showed that their DNN approach outperformed two other approaches, namely a Support Vector Machine (SVM) and a tree-based Ensemble Learning to learn the mmWave facial features. It is important to note the very small dataset used to train the model indicating a high probability of overfitting since the data is not representative enough. Additionally, the data collection method used is never explained in detail except that the face distances varied from 30 cm to 50 cm. However, it can be assumed that only frontal face poses are captured with zero occlusion on all 8 subjects. The paper also investigates using just a single receiving antenna reducing accuracy to 73.7%. This aligns with the results found by Hof et al. [5] that increasing the number of receiving antennas improves classification accuracy. The paper also suggests that a Concurrent Neural Network (CNN) may be more appropriate if signals were stacked on the time axis rather than the frequency axis.

Around the same time, Kim et al. [7] studied using the same 61 GHz FMCW Radar sensor by bitsensing Inc. with a range resolution of 2.5 cm. This paper proposes a CNN model composing of three convolutional layers and 3 fully connected layers. Heavy preprocessing of the Radar data is done to make it more image-like in format before inputting into the CNN model. With a test, validation and test split of 70/15/15, it achieved an average classification accuracy of 98.7% with an even smaller dataset of 3 people. However, the paper does investigate the affect of wearing a cotton mask with all 3 subjects' faces. The model was found to only drop in average classification performance by 0.9% showing promising for what we aim to investigate in this project. These results are taken with caution due to the very small dataset, it cannot be ascertained that the performance would be constant for a larger number of more diverse faces and occlusions.

Pho et al. [8] takes 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 could be useful when there is 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 face similarity of inputs. The model is trained for *binary classification* by inputting pairs of face signatures of the same or distinct people to learn embeddings that push distinct faces into distinct Euclidean regions of the embedding space. The same bitsensing Inc. BTS60 chipset used by [6, 7] was utilised to capture 500 frames of the faces of 8 participants. An average classification of 97.6% was achieved which is an improvement from the previous DNN approach by Lim et al. [6] testing on the same number of people. t-Stochastic Neighbour Embedding (t-SNE) [22] is used to reduce the dimensionality of output embeddings in order to visualise their distribution. The visualisations presented show that their one-shot Siamese network approach was able to distinctively embed each person's face for an easy classification task. While a small dataset is used only scanning frontal poses with no occlusion, the proposed method is well documented and is most likely robust enough against a larger dataset.

Challa et al. [9] employs two different machine learning models on the data made available at [20], first a CNN-based Autoencoder followed by a Random Forest Ensemble Learning approach. A total of 9 Autoencoders are built and trained for different frame rates each

compressing and learning to rebuild the original data from the compressed, latent representation. The Autoencoders are trained from randomly selected data samples of 186 mmWave face signatures from the data port. The flattened and labelled outputs are then used to train and test 9 discrete Random Forest models using identical hyperparameters recommended by the Sci-kit library []. The resulting model achieved promising results with an average of 99.98% classification prediction accuracy from using all 1400 frames per person. The model is still able to achieve a 97.1% accuracy even with just using 70 frames worth of data for each person. 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 comparable to Hof et al. [5].

From the research conducted here, all papers tackling this problem investigate solely using the information gathered from the Radar sensor, a major motivation being privacy preservation. One problem of this approach is that in order for an accurate scan of the face to be captured, the sensor must be run for several seconds ranging from 7 to 10 seconds. In real world scenarios this would be infeasible as the subject would require to keep their face still for significant period of time. No study, as of yet, has looked at the effect of combining the Radar signatures with the RGB information of the same face and if this could further improve facial recognition performance. With many existing deep learning models performing incredibly well with just 2D RGB information such as InsightFace [23], utilising their power along with mmWave Radar could help speed up data acquisition while still having the robustness to lighting and occlusion as mmWaves offer.

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 development of the InsightFace toolkit implementing algorithms tackling the various intricacies with face analysis and recognition. Key papers include the preliminary ArcFace model, introduced by Deng et al. [23]. ArcFace employs a novel Additive Angular Margin Loss to maximize class separability, further enhancing the discriminative power in mapping 2D facial images to feature embeddings. While this method was found to face challenges with label noise requiring many real-world images found on the web to be "cleaned" beforehand. To counter this, further progress was made with Sub-center ArcFace [24] introducing the idea of sub-classes resulting in increased robustness to 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) [25] and YouTube Faces (YTF) datasets [26].

The integration of pretrained models offered by InsightFace into our system allows a dominant focus on increasing the performance of solely our CNN embedding mmWave face signatures. By fusing the depth and contour detection capabilities of mmWave radar with the rich textural information gathered by ArcFace from RGB images, the system could achieve improved accuracy and robustness, showing promise in environments where conventional optical methods falter.

2.5 Multimodal Data Fusion Techniques

Multimodality, as defined by Lahat et al. [27], refers to the use and analysis of multiple types of data, that may be sourced from multiple sensors, with the aim of extracting and mixing the important information gathered by each sensor. The integration of this diverse data leads to outputs with richer representation and performance than what could be achieved by 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 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 lower-dimensional common space. Techniques like Principle Component Analysis (PCA) and Canonical Correlation Analysis (CCA) are used for this. This technique would come with a variety of challenges since it must be ensured that the RGB frames are synchronised with the Radar frames which is difficult since both sensors use wildly different sampling rates. Additionally, the continuous mmWaves signals would need to be discretised appropriately to match the discrete RGB values in each frame. A major disadvantage of early fusion is the fact that combining the different modalities often squashes salient information within each individual modality, impacting the training efficacy.

Late Fusion or **Decision Level Fusion** operates by feeding data sources independently into separate models and then fusing them at the decision-making stage. Common approaches of this is taking a weighted average providing a way to minimising or maximising the effect of certain modalities. Late Fusion is often simpler and flexible, and can be very effective when dealing with extremely dissimilar data sources either in terms of sampling rate, dimensionality or unit of measurement. Additionally, late fusion often provides 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 representation 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 [28]. This approach offers more flexibility than early and late fusions in being able to fuse features at different depths within the neural network. However, it can lead to challenges like overfitting or failure in learning inter-modal relationships.

Each data fusion technique comes with its own set of challenges and considerations, so identifying the best approach for combining RGB and mmWave signatures requires experimentation of each. Late and intermediate fusion are feasible, however, it would be challenging to integrate early fusion due to the huge variation in the two modalities. Heavy preprocessing of the Radar data is required, possibly transforming it into 3D point cloud.

2.6 Deep Learning for Face Recognition

The most popular approaches observed for face recognition models involve deep learning, specifically using Concurrent Neural Networks (CNN).

3 Proposed Approach

The following section delineates the planned approach to take on the Radar-based facial recognition problem. Each decision and step will be justified with a premonition of upcoming problems and intricacies that may occur during the project lifecycle.

3.1 Data Acquisition and Experiments

3.2 CNN Model: mmFace

3.3 Progress

4 Work Plan

show how you plan to organize your work, identifying intermediate deliverables and dates.

References

[1] 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.

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

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

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

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

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

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

[24] 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.

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

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

[27] 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.

[28] 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.