

mmFace: 3D Face Recognition with RGB and Millimetre Wave Radar

Stergious Aji (2546916A)

March 26, 2024

ABSTRACT

TODO

1. INTRODUCTION

Facial recognition is an evolving research domain within the field of computer vision, finding extensive use across areas including human-computer interaction, security surveillance, and forensic analysis. Its primary application revolves around biometric authentication, granting individuals access to their devices or restricted zones. This enables a non-intrusive, hands-free means of verifying identity, eliminating the need to memorise passwords. Additionally, facial biometrics are naturally more attainable than other modalities such as fingerprints, palm prints, or iris scans [1].

Since its inception in the 1960s, facial recognition technology has undergone significant growth. Initially pioneered by Bledsoe [2], early systems distinguished faces by comparing manually annotated landmark features such as the nose, eyes, and mouth. In recent years, the emergence of deep learning has reshaped human face classification, leveraging extensive online repositories of face images for improved performance and efficiency. However, these systems predominantly rely on images captured by RGB cameras, leaving them vulnerable to variations in lighting and facial pose [3]. By incorporating depth data and drawing attention to the geometric details of the face, the impact of such environmental factors can be regulated. Furthermore, the transition to three-dimensional facial recognition not only increases accuracy but also bolsters the security of biometric systems against spoofing attacks [4].

1.1 Motivation

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 [5] technology. This growing demand has pushed the commercialisation of depth-sensing technology to smaller form factors, facilitating its efficient real-time operation on mobile devices [6]. Face ID, in particular, has garnered a level of security that enables payment authentication within services such as Apple Pay. However, Apple's use of costly proprietary hardware and restrictive patents make it harder for smaller companies to adopt an equally compact and secure face recognition system.

Depth cameras, used in this context, typically employ an active form of acquisition. This involves projecting non-visible light onto the face, which is then reflected back, allowing sensors to gauge and delineate facial features. Lidar cameras, emitting waves in the near-infrared (NIR) spec-

trum, are the most prevalent choice given their capacity to acquire a dense 3D map of the subject's face [7]. However, they are limited by their inability to penetrate thin materials such as clothing and hair. In contrast, millimetre wave radar (mmWaves) can penetrate such materials and directly reach the skin's dermal layer [8]. This could enable greater performance in scenarios involving facial hair or adverse environmental conditions such as rain or fog.

Research into the efficacy of radar waves for 3D face recognition remains relatively limited, although recent studies indicate promising outcomes [9, 10, 11, 12, 13]. Radar sensors typically offer greater cost efficiency in terms of both acquisition and computation, as they consume less power compared to sensors NIR-based systems. Nevertheless, it is crucial to acknowledge the trade-off, as mmWaves often result in a sparser representation. This could impact recognition accuracy, where precision in detecting and mapping of facial features is paramount. Thus, we aim to address this limitation by integrating information from colour images, potentially paving the way for more resilient and versatile systems.

1.2 Research Contributions

TODO CHANGE

Our work explores the efficacy 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 have collated this data ourselves. We use the Intel RealSense L515 RGB-D camera [14] for photographing individuals' faces. Meanwhile, the Google Soli 60 GHz radar sensor [15] is employed to gather depth information by transmitting and measuring millimetre waves reflected from the target.

We initially planned to gather face data from approximately 50 participants, however, only 21 participants came forward within the limited timeframe. We obtained face data under various conditions including diverse poses, lighting environments, and common occlusion scenarios. This approach yielded a system that is invariant to environmental conditions in comparison to RGB-only systems.

We developed a novel face recognition model using a deep convolutional neural network. This model was trained on the captured data in order to learn facial features from both the RGB and depth characteristics acquired from the Soli sensor, simultaneously. We investigated different techniques in fusing these two modalities, aiming to pinpoint the most effective strategy that provides rich and distinctive representations, for clean identity separation. The model's effectiveness is benchmarked against prior radar-based facial

recognition systems, as well as, a comparison to using each modality by themselves.

The main contributions of this paper are listed below:

1. We have collated a face dataset of colour images and mmWave signatures from a diverse array of 21 participants. This data includes five diverse poses, two lighting environments, and two occlusion scenarios.
2. We present *mmFace*, a novel face recognition model that leverages both modalities yielding a system that is robust, invariant to occlusion and spoofing. This model is able to classify a given face and identify if it is live or fake with strong abilities to generalise to unseen faces.
3. We present an empirical analysis of seven feature-level fusion methods and determine the most optimal approach for mixing the RGB and mmWave facial features.
4. Our models and their evaluations are open sourced¹ to facilitate further research in small-scale, 3D face identification using mmWave technology.

2. BACKGROUND

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, as well as maritime navigation. Recently, the miniaturisation of radar sensors to the millimetre wave band has brought its application to more small-scale domains [6]. mmWave sensing has been very successful in the domain of autonomous vehicles for object detection, specifically in systems such as collision warnings and adaptive cruise control [16]. This is primarily due to its edge over traditional near-infrared waves employed by lidar cameras, particularly in its resilience to atmospheric conditions such as dust, smoke, fog, and rain [17]. This penetrative power of mmWaves make it a promising candidate for reliable facial recognition in uncertain, real-world scenarios.

Another notable example is Google’s integration of their Soli sensor into the Pixel 4 smartphones for motion detection and gesture recognition [18]. However, its application to face recognition is unexplored offering a unique research area. Consequently, this is the sensor we used to capture mmWave face signatures during our data collection phase. A key driving factor for this choice is the Soli’s miniature form factor of just $6.5\text{ mm} \times 5.0\text{ mm}$, portability, and its 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 [19], a vital aspect for extracting accurate facial features. The Soli chip has a relatively low power consumption due to the fact that it sends 16 chirps every burst at a pulse-repetition frequency of 2 kHz, then stops transmitting until the next burst of chirps [20]. Each burst is transmitted at 25 Hz giving an overall transmission duty cycle of 2% meaning the radar chip is turned off during the

majority of its operation saving a lot of power for mobile applications.

2.2 Related Work

The use of millimetre waves for face identification is a relatively new research field, spurred by the commercialisation of radar sensor technology. One of the earliest studies found to investigate human identification using mmWaves can be traced back to 2019, conducted by Zhao et al. [21]. Although this paper focuses on classifying subjects by their gait and body shape rather than facial features, it demonstrates the ability of mmWaves to encapsulate the subtle idiosyncrasies among individuals. These nuanced differences are vital for learning models to effectively differentiate between unique subjects, leading to high classification accuracies.

Following this, Hof et al. [9] proposed a Deep Neural Network (DNN) based Autoencoder that can 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 between positive and negative instances by measuring their Mean Squared Error (MSE) against reference facial embeddings. The study conducted an extensive data collection process, capturing face scans of 206 participants comprising various genders and ages, in five different poses: frontal, as well as, 15° and 25° head rotations to the left and right. This collection was subsequently made available through an IEEE Data Port [22]. 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 larger sensor containing a total of 1024 transmit and receiver antenna pairs, found to capture redundant information. This is in contrast to the compact Soli chip with a single transmit and three receiver antennas, intended to work within a smartphone. 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. [10] proposes another Deep Neural Network model, however with a more traditional 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, surpassing the performance of both, a Support Vector Machine (SVM), and a tree-based Ensemble Learning approach trained on the same 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 then 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 [9] observation that an increased num-

¹<https://github.com/StergiousAji/mmFace-3D-Face-Recognition-using-RGB-and-mmWave-Radar>

ber of receiving antennas can enhance classification accuracy by the ability to capture 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. [11] conducted research using an identical 61 GHz FMCW radar sensor from bitsensing Inc., featuring a range resolution of 2.5 cm. Their study introduces a CNN model comprising 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. [12] adopts a One-Shot Learning approach to the problem. This is where the model is trained with a single or only a few labelled instances, 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 a latent space. A distance metric between the outputs of both CNNs is used during the training and testing phases to measure the similarity between face inputs. The model is specifically trained for *binary classification* by inputting pairs of face signatures from either the same or different people. This process resulted in the model learning embeddings that push different faces into distinct Euclidean regions of the embedding space. The same bitsensing Inc. BTS60 chipset, used by Lim et al. and Kim et al. [10, 11], is employed to capture 500 frames of the faces of eight participants. An average classification of 97.6% was achieved, an improvement over the previous DNN model by Lim et al. involving the same number of people. t-Stochastic Neighbour Embedding (t-SNE) [24] is then applied for dimensionality reduction. The resulting visualisations demonstrated that the one-shot Siamese network effectively separated each individual’s face into exclusive regions, simplifying the classification task. Although a small dataset is used, only encompassing frontal poses with no occlusion settings, the proposed method is well documented and is likely robust against larger datasets.

Challa et al. [13] employs two different machine learning models on the dataset made available through the IEEE port [22]. Their approach began with CNN-based Autoencoders followed by a Random Forest Ensemble Learning approach. A total of nine Autoencoders were built, each tailored to different frame rates, focusing on compressing and learning to reconstruct the original data from its compressed, latent form. The Autoencoders were trained using randomly selected data samples from a subset of 186 mmWave face signatures. The flattened and labelled outputs were then used to train and test nine discrete Random Forest models using identical hyperparameters, as recommended by the Scikit 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 was able to maintain 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 an efficient 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 RGB images alone, such as InsightFace [25], integrating these models with mmWave radar data presents a promising avenue. This combination could accelerate face acquisition time, while leveraging the advantages of mmWaves in terms of their robustness to lighting variations and occlusions.

2.3 InsightFace

In the evolving field of face recognition, deep CNNs have emerged as a dominant approach due to their ability to extract 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 maximise 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.

2.4 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 lead to outputs with richer representations than what could be achieved by the indi-

vidual 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. 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 more flexible, and it 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 involves the idea of combining different modalities within 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 as complex as using 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, feature-level 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.

3. METHODOLOGY

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 ?? of 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 induces pose-invariance into the system, accommodating real-world use cases where individuals may not always present an exact frontal pose to the face recognition system. Most studies concentrate 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 left and right rotations of the face. For this reason, we will similarly focus on head rotations around the yaw axis. We plan to record facial poses at 0° , as well as, 30° and 45° to the left and right of the sensors, denoted 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 be incorporated in our data collection experiments. Namely, regular and dim lighting scenarios. We hypothesise that the mmWave face signatures would be unaffected by environmental lighting due to the sensor using its own active illumination of the target face, unlike the RGB camera. Therefore, if the system is able to demonstrate higher accuracy utilising both modalities as opposed to relying solely on RGB data, it would conclusively show that mmWaves offer robustness against varying 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 for facial recognition systems to be robust against typical obstructions such as glasses, hats, masks and so on. Currently, users would be required to remove such accessories for systems to accurately identify and grant them access to particular devices or areas. With mmWaves, we hypothesise that this may not be needed since facial features could be captured regardless. This could greatly benefit security surveillance where individuals deliberately obscure their faces in order to hide their identities. In our experiment, we aim to capture scenarios both with and without occlusion. Since cotton masks have already been explored by Kim et al. [11], other common items like hats, sunglasses, and scarves will be used to mirror day-to-day scenarios.

On top of investigating environmental invariances, our research seeks to determine the least amount of radar data necessary to still yield easily separable facial embeddings. Previous studies were observed to record between 200 and 2,000 total frames which, depending on the sampling rate, requires running the sensor for eight to a maximum of 80 seconds for each scenario. This is due to the lower accuracy of mmWaves demanding a longer exposure to provide a dense enough representation. Nevertheless, requiring an individual to keep their face still for over three seconds is impractical in real-world applications. Our approach, which integrates RGB data, may reduce the need for prolonged radar frame capture. Consequently, in our experiments, we will collect 10 RGB frames and a maximum of 250 frames worth of mmWave bursts, equivalent to operating the Google Soli sensor for 10 seconds at a 25 Hz sampling rate. This approach will provide ample data to assess the minimum number of mmWave frames needed for effective identification.

To ensure a diverse range of facial data, we aim to recruit 50 participants, given the tight timeframe of this project. Adhering to ethical standards regarding sensitive personal

information, our participant pool will consist of adults, predominantly university students. While this results in an over-representation of individuals aged 20–25 years, it should not impact our study as age variance is not something that is being explored. 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.

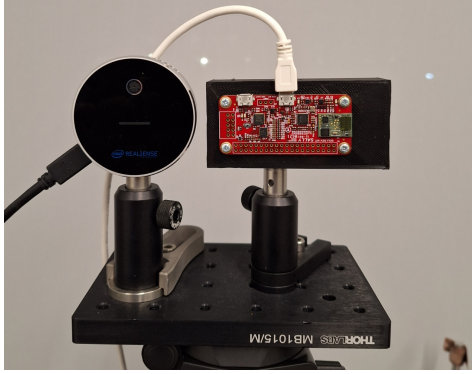


Figure 1: Equipment: Intel Realsense L515 RGB-D Camera (Left) and the Google’s Soli 60 GHz radar sensor (right)

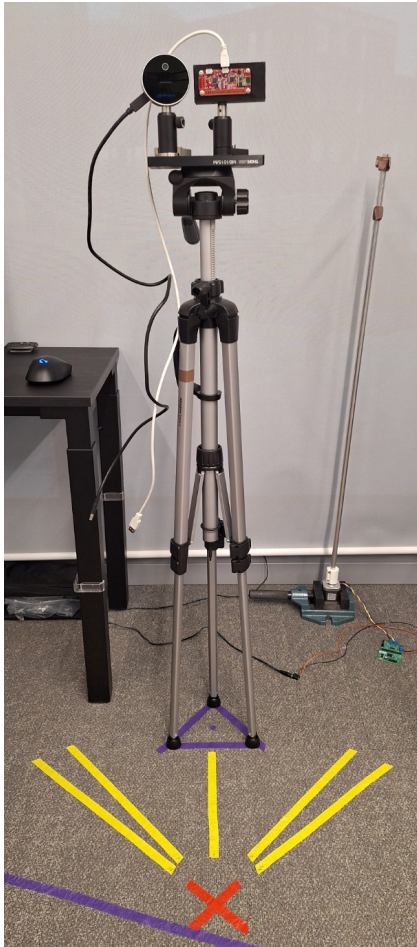


Figure 2: Experiment Setup

3.2 mmFace

Building on the intuition from Section 2.3 of the Background chapter, it is clear that the ArcFace model from the InsightFace toolkit emerges as the best choice for our project. It attains state-of-the-art classification results on accepted benchmark sets, outperforming the previous bests such as Facebook’s DeepFace [34] and Google’s FaceNet [35]. This selection allows us to treat the RGB data processing as a *black-box* framework, enabling us to concentrate efforts on perfecting the radar-based model we are naming, *mmFace*. Furthermore, this facilitates exploration into the various methods in fusing the two modalities as discussed in Section 2.4. Figure 4 shows a high-level diagram of the model workflow described here.

Our proposed model, mmFace, will employ a CNN-based architecture, which is particularly effective for processing image-like data. The radar bursts obtained during the data collection phase will be transformed into Complex Range-Doppler (CRD) maps [15, 20] prior to feeding it into the model. Face scans will be collected using the Google Soli’s short configuration which operates at a centre frequency of 60 GHz, with a maximum bandwidth B of 5.5 GHz, and bursts sampled at 25 Hz. This gives a range resolution of $\frac{c}{2B} = 2.7$ cm, where c denotes the speed of light. Given that the Intel RealSense captures RGB-D frames at a different sampling rate of 30 frames per second (FPS), timestamp information is also recorded for the possibility of synchronising the two modalities for early data fusion.

TODO: The Soli chip as single transmit antenna and three receiver antennas capturing a superposition of scattered reflections from the target

As explained before, the data fusion techniques we plan to investigate include late, feature-level fusion and late, decision-level fusion. Pure intermediate fusion is not feasible due to the black-box treatment of the InsightFace model making it difficult to integrate information from both modalities within its hidden layers. Nonetheless, late, feature-level fusion remains viable, combining the outputs of the final layers of each model to form an embedding containing both the RGB and radar features. Similarly, decision-level fusion will be explored since this entails mixing the predictions from the individual models. Early fusion presents significant challenges due to the dissimilarities in sampling rates and data formats of the two modalities. The CRD maps must be synchronised and transformed into a depth image-like format before merging with the RGB images. Furthermore, this would require a whole new training cycle with the mmFace model which may be infeasible within the project’s timeframe. However, should time permit, we will consider investigating this approach.

We plan to adopt a ResNet-based architecture for mmFace due to its refinements over its predecessors like AlexNet [36] and VGGNet [37]. The ResNet framework [38] incorporates “skip connections” and residual blocks to resolve the vanishing gradient problem encountered in VGGNet, allowing scaling of the network beyond the 19-layer limitation. This support for deeper networks provides a strong foundation for the mmFace model for learning the complex radar face signatures.

The dataset, comprising 50 participants, will be divided by subjects into an 80%/10%/10% split for training, validation, and testing. Given the dataset’s small size, a larger proportion is allocated for training to ensure the model can

learn effectively. Following training, the testing phase will evaluate the distinctiveness of the output face embeddings. For accurate classifications, each person's face must be spatially separable in the high-dimensional embedding space allowing for unambiguous identification. t-SNE visualisations [24] will be employed to visually inspect and confirm this is the case by comparing against the original data. In addition, standard classification accuracies will be calculated to verify the model's identity recognition capabilities against randomly selected ground truths. This also allows benchmarking our results against previous studies on radar-based 3D facial recognition.

3.3 Feature-Level Fusion

4. EVALUATION

4.1 Results

5. CONCLUSIONS

5.1 Future Work

Acknowledgments. This is optional; it is a location for you to thank people, most probably your family and your supervisor.

6. REFERENCES

- [1] Song Zhou and Sheng Xiao. 3d face recognition: a survey. *Human-centric Computing and Information Sciences*, 8(1):1–27, 2018.
- [2] Woodrow Wilson Bledsoe. The model method in facial recognition. *Panoramic Research Inc., Palo Alto, CA, Rep. PR1*, 15(47):2, 1966.
- [3] Chenghua Xu, Yunhong Wang, Tieniu Tan, and Long Quan. Depth vs. intensity: Which is more important for face recognition? In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 1, pages 342–345. IEEE, 2004.
- [4] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015.
- [5] Apple Inc. About Face ID advanced technology, 2023. Accessed: 2023-11-19 <https://support.apple.com/en-gb/102381>.
- [6] 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.
- [7] Xin Wang, HuaZhi Pan, Kai Guo, Xinli Yang, and Sheng Luo. The evolution of lidar and its application in high precision measurement. In *IOP Conference Series: Earth and Environmental Science*, volume 502, page 012008. IOP Publishing, 2020.
- [8] 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.
- [9] 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 (ICAIIIC)*, pages 320–324, 2020.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] Intel Corporation. Intel RealSense LiDAR Camera L515, 2023. Accessed: 2023-11-19 <https://www.intelrealsense.com/lidar-camera-l515/>.
- [15] Jaime Lien, Nicholas Gillian, M Emre Karagozler, Patrick Amihoud, 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.
- [16] DF Robot. Eight Practical Applications of mmWave Radar Technology, 2023. Accessed: 2023-11-19 <https://www.dfrobot.com/blog-1650.html>.
- [17] 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>.
- [18] 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>.
- [19] Bassem R Mahafza. *Radar systems analysis and design using MATLAB*. Chapman and Hall/CRC, 2005.
- [20] 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.
- [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] Evyatar Hemo, Amichai Sanderovich, and Eran Hof. mmwave radar face signatures, 2018. <https://dx.doi.org/10.21227/wr67-kx23>.
- [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*

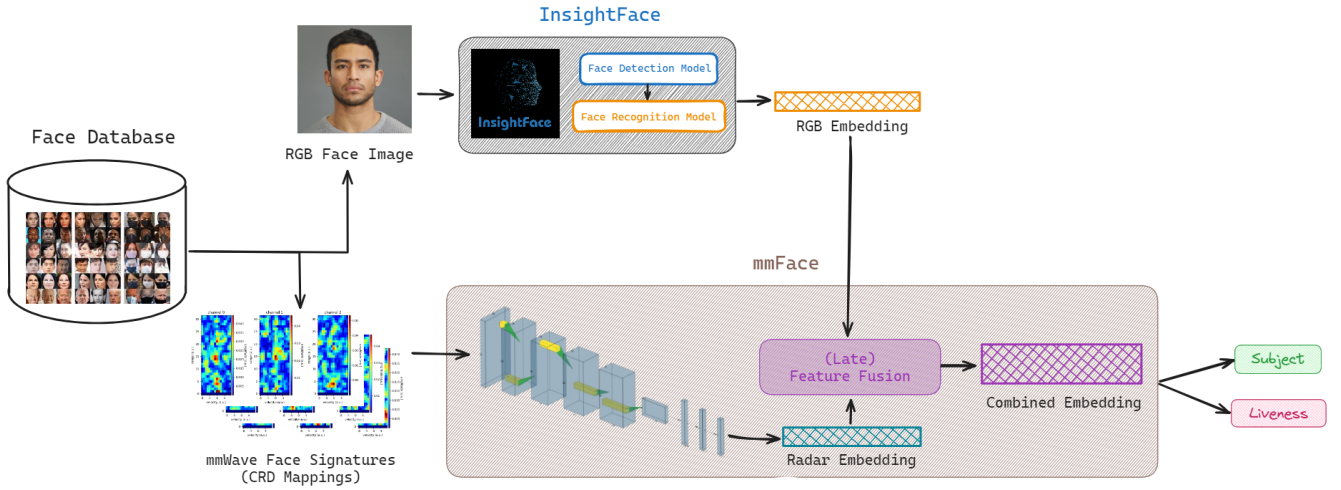


Figure 3: Workflow Diagram

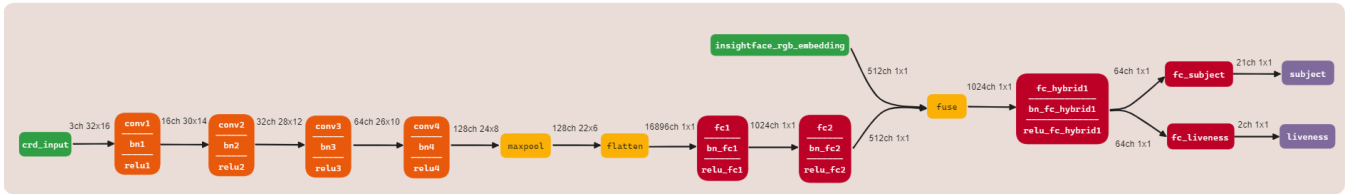


Figure 4: Architecture Diagram

- 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 Adalı, 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] 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.
 - [35] 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.
 - [36] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
 - [37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
 - [38] 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.