University of Glasgow | School of Computing Science

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 30, 2023

Contents

1 Introduction

1.1 Motivation

Facial recognition technology is a crucial field of research due to its broad applications in areas such as security surveillance, forensic analysis, human-computer interaction, and healthcare. Its most notable use being in biometric authentication for access control. This enables a non-invasive, hands-free approach to identity verification, removing the need to remember 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 improved the performance of human face classification, benefiting from the vast online database of face images. However, these systems primarily rely on 2D images from RGB cameras making them susceptible to variations in lighting and pose. To compensate for this, the depth information of facial attributes are essential. 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's Face ID [2]. 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 face. However, its weakness in penetrating materials like clothing and hair is a notable limitation. In contrast, millimetre Radar waves 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 existing studies show positive results [4, 5, 6, 7, 8]. Radar technology is generally more cost-effective, both in terms of acquisition and computationally since it consumes less power compared to Lidar cameras. However, the lower accuracy and sparsity of mmWave 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. The Google Soli Radar sensor [10] will be employed to gather depth data 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 different poses, lighting environments and involving 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 individually.

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

2 Background Survey

Before any work is conducted, a review of relevant literature in the field is undertaken. This is essential to corroborating best practices and gaining a deeper insight of the strengths and limitations of mmWave Radar sensors, in the context of face recognition. Next, a look into existing datasets compatible with our research objectives is performed, followed by a detailed justification for the need to compile our own dataset. Furthermore, an analysis of the recent work with Radar technology specifically for the purpose of human face classification is provided.

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 smartphone 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 being 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 methods exist for capturing 3D face data which can be broadly categorised into two main types: *Active* and *Passive* techniques [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 in 2D 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 another form of active acquisition, using three receiving antennas to capture reflected mmWaves, measuring the phase difference and Doppler shift to estimate the distance and radial velocity of the target.

When collecting data of 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 2D 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°. Additionally, 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, focussing 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. This also enables flexibility to tailor our experiment design to investigate the model's effectiveness under specific conditions such as lighting and pose variations, 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 [22]. 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. [4] 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 [21]. 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. [5] 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. [4] that increasing the number of receiving antennas improves classification accuracy. The paper also suggests that a Convolutional 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. [6] 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. [7] 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 [5, 6] 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. [5] testing on the same number of people. t-Stochastic Neighbour Embedding (t-SNE) [23] 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. [8] employs two different machine learning models on the data made available at [21], 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. [4].

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 [24], 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. [24], as well as their robust Face Alignment model Gho et al. [25]. 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 [26] 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) [27] and YouTube Faces (YTF) datasets [28].

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. [29], 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 [30]. 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 Convolutional Neural Networks (CNN). Before proceeding to build our own model to learn mmWave face signatures it is vital to understand the underlying concepts involved to identify the best strategy by observing recent work on facial recognition.

CNNs are a class of deep neural networks, most commonly applied to the analysis of image data. They were inspired by the organisation of the animal visual cortex and designed to automatically and adaptively learn spatial hierarchies of features from input images. A typical CNN architecture comprises a series of layers, including convolutional layers, pooling layers, and fully connected layers.

Unlike traditional algorithms, CNNs learn to identify the necessary features for classification directly from the input data, minimizing the need for manual feature extraction. This is particularly advantageous in face recognition, where 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 layer structures. 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 used across different parts of the input, reducing the number of parameters and thereby the complexity of the model. This makes CNNs particularly memory efficient data with high dimensionality such as the mmWave signatures.

CNNs are ideal for face recognition due to their ability to handle variability in images, such as changes in angle, lighting, or facial expressions. Notable examples include the DeepFace by Facebook and FaceNet by Google. DeepFace uses a nine-layer CNN to identify faces with an accuracy of 97.35% on the Labeled Faces in the Wild (LFW) dataset (Taigman et al., 2014). FaceNet, on the other hand, achieved a record-breaking accuracy of 99.63% on the same dataset by directly learning a mapping from face images to a compact Euclidean space (Schroff et al., 2015).

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

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

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

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

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.