ASMA
Anti-Spoofing-Focused Multi-Biometric Authentication

Biometrics Systems Fall 2024

Francesco De Persio, Andrea Donato

July 5, 2025

{depersio.1938022; donato.1808606}@studenti.uniroma1.it

Abstract

In this work we present and evaluate a multi-biometric authentication system based on both face and voice recognition. Our focus in building the architecture is to provide multiple but simple anti-spoofing layers and to minimize probability of false acceptance. The system has been evaluated in an open-set identification setting, and we achieved zero-`FAR` and even zero-`FRR`. These results are obtained leveraging state-of-the-art pre-trained models only. Despite the limitations of using a dataset recorded under controlled conditions, our findings suggest that similar results can be achieved in real applications designed for high-security environments.

Contents

1 Introduction

In the last few years, biometric authentication has become central to both cyber and physical security. It is well-known that several big tech companies have implemented such systems to provide a quick and secure access to their devices: just think of Apple's FaceID, Microsoft's Windows Hello or Samsung Galaxy's Iris Scanner.

However, even a well-trained model such as FaceID [1] is error-prone: for instance, identical twins [2] and close relatives [3] were shown to be able to unlock each other's device. These (few) zero-effort attacks suggest that one may obtain better results with "some-effort" attacks. A device might be put in front of an unwilling subject in order to obtain a Face Recognition unlock [4], or forced against the finger of an incapacitated one. This directly leads us to the Spoofing problem: what if we actively pretend to be someone else?

1.1 Spoofing and Anti-spoofing

A Spoofing Attack (also known as Presentation Attack) is the active attempt of being recognized as someone else by a Biometric System, usually trying to leverage this identity to access restricted areas. This can include coercive methods, as seen earlier, but it typically refers to the fabrication of items such as silicone replicas, masks and pictures.

A common solution to hand-crafted features is a Liveness Detection check, which consists in looking for vital signs. For instance, a severed finger won't show vein pulsations caused by the heartbeat, and a mask won't blink. Challenges such as moving the head or showing a hand are included in this kind of verifications, but one must be aware that predictable patterns might be easily overcome by a video[1]. For this reason, challenges should always be "randomized" enough and possibly supported by other Liveness Detection checks: for instance, showing a video will produce Moiré Patterns, wearing a mask will present peculiar light reflections on its surface.

1.2 Multi-Biometrics as Anti-Spoofing Layer

We have a wide variety of biometric traits at our disposal, each with its own unique characteristics in terms of class separation and robustness w.r.t. aging or spoofing itself. Genotypic an randotypic traits are normally more accurate, but they are more prone to spoofing attacks. Behavioral traits, on the other hand, are usually less accurate and less persistent, but they are also more difficult to replicate. Our goal is to take the best of both worlds by exploiting both face and voice.

Multi-Biometric Systems combine the analysis of multiple individual biometric traits in order to achieve better results in terms of robustness. Fusion can occur at different levels, namely

- Feature Level - Feature vectors from multiple Feature Extraction Modules (FEMs) are combined before being processed by the Matching Module;

[1] Always asking "move your head left" is a weak challenge. It's easy, for instance, to craft a video (such as a Deepfake) showing the desired, predictable feature of the challenge.

- Score Level - Multiple models (FEM + Matching) assign their own score, which are then combined in the Decision Module;
- Decision Level - Multiple models (FEM + Matching + Decision) take multiple decisions, then depending on the implementation one may consider a majority or a custom overall decision logic.

In any case, requiring multiple traits forces an attacker to spoof-attack multiple features at the same time. This is an implicit Anti-Spoofing layer, and as such it will be exploited in our work.

1.3 Our Work

In this project we present a Multi-Biometric Authentication System in a controlled environment. The overall idea of the model is the following:

- Generate and show the user a random sentence to pronounce aloud. Reading the sentence from the screen will ideally provide a coarse alignment with a camera, allowing to record three elements: RGB, IR and audio data;
- A speech-to-text module evaluates what the user said, and compares the obtained string with the random generated sentence. If the score is high enough we proceed with the analysis, otherwise the first Liveness Detection check is considered failed and the process stops;
- RGB, IR and audio data are preprocessed and given to the respective FEMs. Thermal data provide a second Liveness Detection layer, avoiding the usage of masks or videos;
- A Feature Level Fusion of RGB and IR data is performed, obtaining by a single Matching Module an overall Score regarding the **face**. The **voice** is matched alone;
- A Score Level Fusion of **voice** and overall **face** is performed, leading to the decision.

Different architectures have been taken into account for the implementation, and will be quickly discussed in section 5.1. This one in particular has been chosen for being a good trade-off between simplicity and performance.
The system works in an identification open-set setup balanced to minimize FAR. The potential applications of such a system include critical infrastructures such as government buildings and banks, and more generally any situation in which a clear, manifest intention of access is required.

2 Dataset

To date, public datasets including RGB, thermal and audio data are not widely available. The SpeakingFaces [7] dataset is in fact the only valuable option we found, consisting in 142 subjects with adequate male/female balance (slightly less balanced in terms of ethnicity distribution).

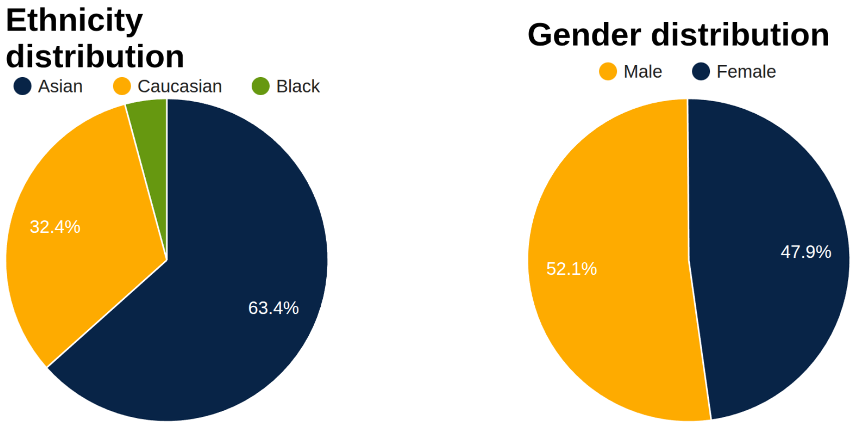


Figure 1: Dataset ethnicity and gender distribution.

For each subject we're provided with full video recordings by 9 cameras, each positioned at a different angle w.r.t. the subject. Each video records a different phrase pronounced aloud by the subject, and the whole procedure was repeated on a different day for a total of 2 trials.

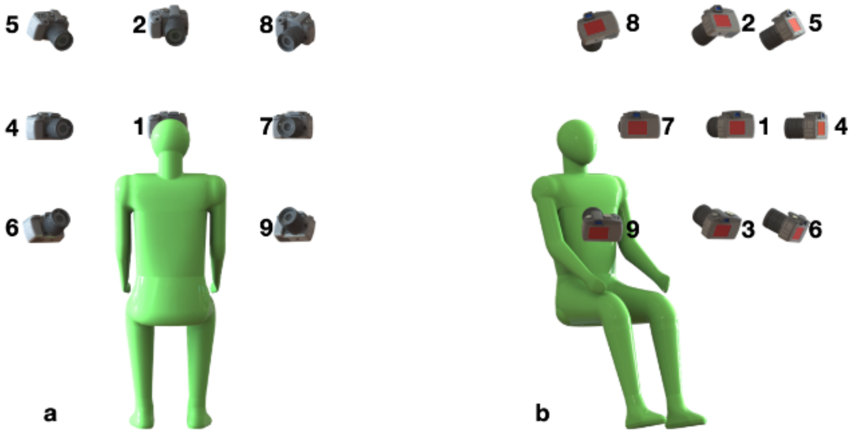


Figure 2: Input devices' positions.

As the paper says [7], file names notation is based on the following scheme:

- subID = {1, ..., 142} denotes the subject number.
- trialID = {1, 2} denotes the trial number.
- sessionID is 1 if the session does not involve utterances and 2 otherwise.[2]
- posID = {1, ..., 9} denotes the camera position.
- commandID = {1, ..., 1298} denotes the command number.[3]
- frameID = {1, ..., 900} denotes the number of an image in a sequence.
- streamID is 1 for thermal images, 2 for visual (RGB) images, and 3 for the aligned version of the visual images.[4]
- micID is 1 for the left microphone and 2 for the right one.[5]

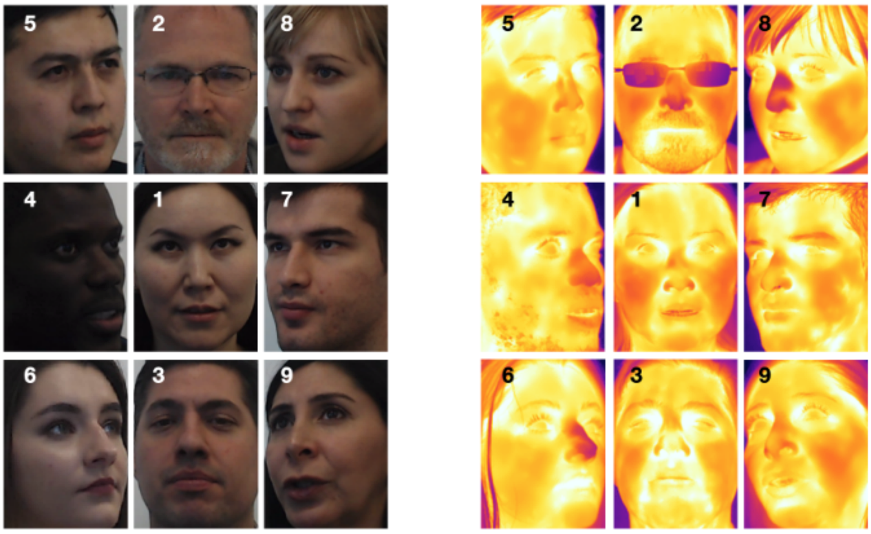


Figure 3: Examples of each camera's recordings for both RGB and IR data.

One of the key features of this dataset is being entirely recorded in a controlled environment. Of course, this is possibly also a flaw, due to the absence of noisy data. Nonetheless, this helped us to figure out a possible application in a controlled setting, which eventually led us to focus on the anti-spoofing layer.

[2] This value is most often = 2.

[3] i.e. the pronounced phrase.

[4] We'll use the aligned, already pre-processed version of the RGB images only.

[5] For simplicity, we decided to use data from the first microphone only, as we concluded that this leads to no significant bias.

2.1 Dataset Analysis and Management

Despite claiming to contain 142 subjects, our preliminary analysis showed that some folders were missing. In addition, we found that some of the remaining subjects had some missing data, e.g. RGB images for trial 2. Since our idea is to use trial 1 for building the Gallery and trial 2 for the probes, this incompleteness led us to directly eliminate incomplete subjects, leaving us with 126 valid identities.

Moreover, as we have seen we have plenty of different phrases. Still, it may happen that the same subject pronounces the same phrases in both trials. If this was a common practice in this dataset it would have led to a very significant bias in the Matching Module[6]. Luckily, we found that phrases were randomized enough so that less than 15% of the subjects pronounced at least one same command in both trials, and even in these cases the percentage of repeated phrases (w.r.t. the total number of audios) is less than 10% (the only exception being subject 129, where 9 phrases over 43 were repeated).

Regarding the visual components, each audio command is associated with RGB and IR recordings taken from each camera. Each video is stored as a frame sequence, for a total of ~ 5000 images per modality (i.e. RGB and IR) per trial per subject. Since this is an enormous amount of pictures we decided to significantly reduce by taking a small number (3 by default) of random images per camera, via the parameter `images_per_camera` to be found in the FEM. This allows us to keep angular and utterance diversity while significantly reduce the time consumption for running the Feature Extraction Module. Like we said, noise may still be a problem: a subject who wears glasses for each and every video take in `trial 1` will have no glass-less pictures in our gallery. We are aware of this kind of problem, still we couldn't do much to solve it.

[6] If, for instance, command 514 is pronounced by subject 14 only, and if this happens in both trials, it is straightforward that we would have a strong matching bias while comparing (subj 14, trial 2, command 514) with all the audios in the gallery.

3 Implementation

We split the different tasks on multiple notebooks, using `json` files to save checkpoints between different phases.

1. First, the Dataset Management notebook scans the dataset to find incomplete subjects, as discussed earlier. Then, builds a `json` file containing the paths for the dataset and defines functions to retrieve such data. Then, this notebook is used to further analyze the dataset (e.g. checking for audio biases as described in the previous section).

2. An Enrollment notebook implements a sample enrollment for a single subject.

3. The FEM notebook takes the whole set of paths and for each valid subject loads all the audios and a random sample of RGB and IR pictures as discussed earlier. Then, it proceeds to extract the features for the selected samples. This is done via pre-trained Neural-Network-based models, which will be discussed in the following sections. The so-obtained features are again stored in a `json` file.

4. The Evaluation notebook builds the Gallery by randomly choosing a percentage of subjects from `trial 1` extracted features and implements both the Matching and the Decision modules by using subjects of `trial 2` as probes. This also includes the Evaluation of the whole process.

3.1 Feature Extraction Module

After considering to train our custom Neural Network for the FEM, we decided for simplicity to rely on pre-trained models. Here we present their main characteristics.

3.1.1 Wav2Vec2

Wav2Vec2 [8] is a powerful self-supervised voice recognition model from Meta AI, whose training consists of over 960 hours of audio. Even without fine-tuning, it's capable of extracting high quality features consisting in `([1, 768])`-shaped tensors for each "voice frame". This means extracting multiple feature tensors per audio, but we decided to take a mean of these fine-grained features in order to extract only one feature tensor per audio. Of course, this kind of approach loses information about time distribution and treats in the same way noisy sections and relevant ones. Considering we are in a controlled environment (so that noise is kind of limited), we thought it to be a good trade-off between simplicity and performances.

Nonetheless, we found out that exploiting this model we obtained low discriminative power, which led us to consider some alternatives. Among the ones we tried to use, the best one was SpeechBrain.

3.1.2 SpeechBrain

SpeechBrain [9] is a deep learning toolkit for speech-related tasks, and its ECAPA-TDNN-based speaker recognition model proved to be an excellent alternative to Wav2Vec2 for our purposes. Unlike Wav2Vec2, which was trained for general speech representation, this model is specifically optimized for speaker verification tasks on large-scale datasets like VoxCeleb1 and 2. It takes raw audio as input and produces a fixed-size embedding of shape (`[1, 192]`), directly optimized to represent speaker identity. Just like in the Wav2Vec2 case, we extracted one embedding per audio sample, relying on the model's ability to internally handle duration variability and attention to relevant voice patterns. Since this model is trained with cosine similarity in mind, its embeddings are much more suited for downstream biometric matching. Indeed, we observed a significant boost in discriminative power using this approach, without the need to fine-tune anything.

3.1.3 InsightFace

An open-source, state-of-art library for RGB face analysis, based on three modules that provide Face Detection, via RetinaFace (localizes keypoints such as eyes, nose and mouth), Face Alignment and Face Recognition, via ArcFace [10] (exploits a deep CNN to give back a (`[1, 512]`) tensor while used in feature extraction mode).

Since this is an RGB tool, it is not intended to be used with thermal data. However, as of today there's shortage of pre-trained state-of-art models for IR data. After considering the option of training our own thermal images Feature Extractor, we decided to rely on InsightFace itself[7]. Despite being a non-optimal choice, it is fairly capable of extracting good quality embeddings. It may sometimes happen a failure in the Face Detection block (as expected, since its training is on a completely different kind of data), but overall the failures are a small percentage of the total number of images.

As discussed before, here is where we perform our **Feature Level Fusion**. We extract both RGB and IR features (both (`[1, 512]`)-shaped) and concatenate them in a single $([1, 1024])$ tensor which we will call **fused** data. As for the failures, couples which presented an error in thermal Feature Extraction are just skipped. We are kind of tolerant w.r.t. this kind of problem, since we've built a fairly redundant database (27 images per subjects from different angles) and the mean number of errors per subject is (2.8 ± 3.8), which is around 10%[8]. Of course, one may further improve the redundancy by setting a higher value of the parameter `images_per_camera`.

[7] We also considered non-Neural-Network techniques such as LBP or Gabor Filters, but we agreed for simplicity to use InsightFace.

[8] Most of the subjects present 0 errors, but some of them are particularly problematic for the Face Detection block, leading to 10/15 errors each (thus only 12/17 valid `fused` tensors). This is the cause of such an high standard deviation. Despite these peaks, we still have a fair amount of samples in the Gallery for each subject.

3.2 Matching Module

Both Speechbrain and ArcFace as embedding extractors are optimized for **cosine similarity**. This would naturally lead us to exploit this metric to evaluate both audio and image matchings, but since we implemented the concatenation of two different InsightFace embeddings we found the **euclidean distance** to better fit image features' characteristics.

The Gallery is built by random-choosing a subset of the trial 1 dataset. The percentage of included subjects is a manually chosen parameter, which defines the behavior of the `random.choice` function. The data for each chosen Gallery subject include all the audios and 27 `fused` images[9]. Probes are loaded by choosing one single random audio file and one single `fused` image.

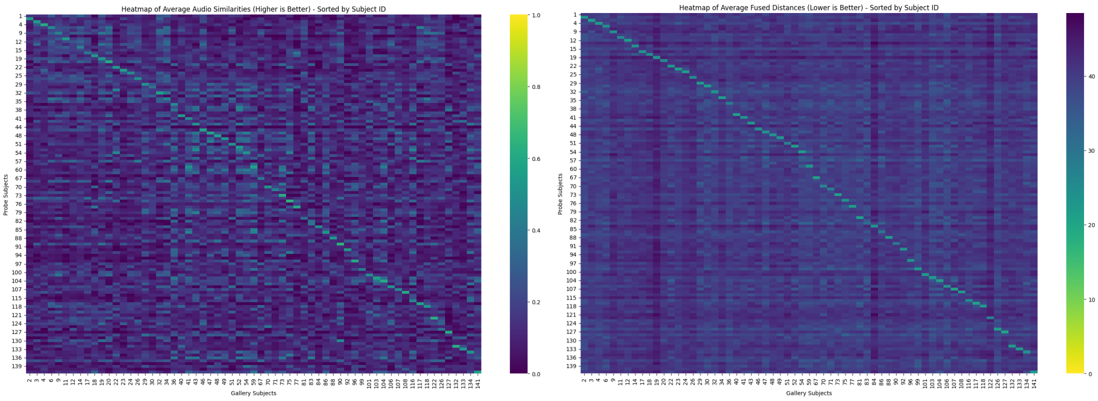


Figure 4: Heatmaps for a Matching Module run, `All vs All_Gallery`. Here, the Gallery consist in 50% of the total valid subjects. Despite being fairly accurate, `audio` (left) is clearly more noisy than `fused` data, confirming its weaker identification power w.r.t face. In the latter we might recognize a couple of Doddington Zoo characters (e.g. a darker vertical line for subjects like `84` may suggest a Dove, while a brighter horizontal line like for subject `32` might indicate the presence of a Wolf). Of course, this is just an intuition: identifying such characters would require further analyses.

By building a random instance of the Gallery and a random choice of the probes, one may obtain a one-vs-all matching for each probe. For instance, the single audio is matched with all the audios in the Gallery, and return a similarity score for each of them. For simplicity we chose to take an average, obtaining the mean match of the probe audio with the best matching "mean audio feature" of each Gallery's subject. Same happens with the `fused` embeddings, leaving us with a list of audio and image matching scores for each probe vs All Gallery.

[9] Slightly less because of the thermal data, as discussed in the InsightFace section.

3.3 Decision Module

Since we chose to use different metrics for audio and image data, we need to re-normalize both range and "growth direction" of the two scores. This was done by re-defining the euclidean score of the `fused` features as

$$\texttt{fused\_score} \leftarrow \frac{\texttt{1}}{\texttt{1 + fused\_score}}$$

That being said, the decision module is a simple linear combination between the two scores followed by a threshold. This can be summarized by the pseudocode

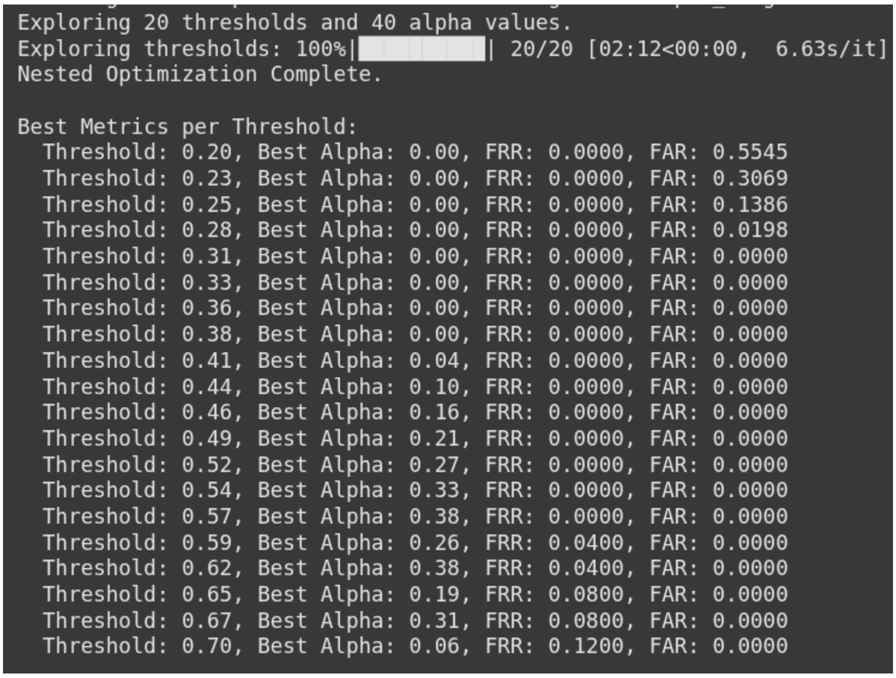
```
**if** $\alpha \times \texttt{audio\_score} \; + \; (1 - \alpha) \times \texttt{fused\_score} \; > \; \texttt{threshold} :$ **accept**
```

The α and the `threshold` parameters are then fine-tuned in the Evaluation Module, which runs multiple instances of the entire model.

4 Evaluation

First of all, we defined the metrics for a classic Identification Open Set task. Since this model is intended to work in very delicate scenarios, we defined the Detect and Identification Rate to be at **rank 1**, obtaining FRR as 1 - DIR. This means that a subject who is correctly accepted but associated with the wrong identity counts as an error.

Then, while fine-tuning the α parameter on different threshold values we found a range in which both FRR and FAR are zero.



```
Exploring 20 thresholds and 40 alpha values.
Exploring thresholds: 100%|██████████| 20/20 [02:12<00:00,  6.63s/it]
Nested Optimization Complete.

Best Metrics per Threshold:
  Threshold: 0.20, Best Alpha: 0.00, FRR: 0.0000, FAR: 0.5545
  Threshold: 0.23, Best Alpha: 0.00, FRR: 0.0000, FAR: 0.3069
  Threshold: 0.25, Best Alpha: 0.00, FRR: 0.0000, FAR: 0.1386
  Threshold: 0.28, Best Alpha: 0.00, FRR: 0.0000, FAR: 0.0198
  Threshold: 0.31, Best Alpha: 0.00, FRR: 0.0000, FAR: 0.0000
  Threshold: 0.33, Best Alpha: 0.00, FRR: 0.0000, FAR: 0.0000
  Threshold: 0.36, Best Alpha: 0.00, FRR: 0.0000, FAR: 0.0000
  Threshold: 0.38, Best Alpha: 0.00, FRR: 0.0000, FAR: 0.0000
  Threshold: 0.41, Best Alpha: 0.04, FRR: 0.0000, FAR: 0.0000
  Threshold: 0.44, Best Alpha: 0.10, FRR: 0.0000, FAR: 0.0000
  Threshold: 0.46, Best Alpha: 0.16, FRR: 0.0000, FAR: 0.0000
  Threshold: 0.49, Best Alpha: 0.21, FRR: 0.0000, FAR: 0.0000
  Threshold: 0.52, Best Alpha: 0.27, FRR: 0.0000, FAR: 0.0000
  Threshold: 0.54, Best Alpha: 0.33, FRR: 0.0000, FAR: 0.0000
  Threshold: 0.57, Best Alpha: 0.38, FRR: 0.0000, FAR: 0.0000
  Threshold: 0.59, Best Alpha: 0.26, FRR: 0.0400, FAR: 0.0000
  Threshold: 0.62, Best Alpha: 0.38, FRR: 0.0400, FAR: 0.0000
  Threshold: 0.65, Best Alpha: 0.19, FRR: 0.0800, FAR: 0.0000
  Threshold: 0.67, Best Alpha: 0.31, FRR: 0.0800, FAR: 0.0000
  Threshold: 0.70, Best Alpha: 0.06, FRR: 0.1200, FAR: 0.0000
```

Figure 5: Placing the threshold somewhere between 0.3 and 0.6 allows both FAR and FRR to be zero. Here we present a single instance on a random Gallery containing 20% of the total trial 1 subjects, but we proved this range to exist by trying on different Gallery instances, with different values of ratio.

We also noticed that optimal α values are sometimes zero too. This is somewhat expected: we knew prior to these observations and we confirmed by looking at the heatmaps that voice has a lower discriminative power w.r.t. face. It is totally normal that in some configurations the model performs better by looking at fused features only.

Considering all this, it is an obvious choice to use a threshold valued within the zero-FAR-zero-FRR range. Also, it is straightforward to take a value which is slightly above the mean of this range: the main goal is to avoid any False Acceptance. It follows that between 0.3 and 0.6 we choose 0.55, and the corresponding optimal α value of 0.35.

As we expected from these values, the ROC curve is somewhat useless, as we need to find no compromise between `FAR` and `FRR` (we are in fact in the middle of a confidence range in which they're both zero).

This of course is a great result from one side, but on the other it might raise some questions. There's no way we can talk about overfitting, since we only used pre-trained models. The Dataset, on the other hand, contains little to no Intra-Class variations (except the angular diversity), which implies poor noise robustness, but as discussed in section 2 there's nothing we could do about it. Nonetheless, if we think about the potential applications (e.g. banks, military infrastructures) we realize that a clean Gallery built in a controlled environment fits much more the task than a noisy one.

5 Results

Overall, we found the model to perform even better than expected. We correctly achieved the zero-`FAR` result that was strictly necessary for such an application, and within a good confidence interval we also achieved zero-`FRR`. These results came with no overfitting risk, as we only exploited state-of-the-art pre-trained models. The only possibly significant bias for such good results might come from the dataset, which is recorded in optimal conditions. However, this also suggests that one may obtain similar results by performing such biometrics in a controlled environment, and in any case to always be capable of reaching at least the zero-`FAR` result with minimum `FRR` by highering the `threshold`.

5.1 Future Works

Even if this simple implementation has proved to work very well, we may still suggest some improvements. First, training an ad-hoc classifier for thermal images (or even fine-tuning an existing one such as InsightFace itself) would allow to better extract these kind of features. One may also explore more sophisticated techniques of Feature Level Fusion to catch an even deeper correlation between RGB and IR spectra. Of course, IR spectrum may significantly change among different trials from the same subject[10]. This may lead to loosen the weight given to the IR spectrum for the identification itself, while taking in strong consideration the correlation with the RGB spectrum.

This directly leads us to one of the originally proposed architecture for this project: **time-correlating all the data**, ensuring that thermal frames movements match the RGB ones and audio is consistent with all these movements. Of course, this would require a much more powerful setup, since we would be obliged to take into account all the frames ($\sim$ 10000 per subject), but it would provide a much stronger Anti-Spoofing structure. Time-correlating RGB, IR and audio and ensuring it is all coherent with the

[10] In fact it's a soft biometric trait. In this model we included it to provide an anti-spoofing layer rather than to improve class separation.

randomly generated sentence to pronounce aloud would really leave little to no chance for an impersonator to perform a spoofing attack.

Finally, the key improvement would be to build a dataset containing **actual spoofing attempts**. In fact, here we have seen performance evaluation for a simple identification task. Multiple choices in order to protect from spoofing were made (i.e. multiple Liveness Detection Layers), still we had no chance to evaluate the system while under a real spoofing attack. As of today, we found no dataset containing RGB, IR and audio data that also includes spoofing attempts. For instance, the CASIA-SURF [11] dataset includes spoofing, but exploits a depth-map instead of voice along RGB and IR data. ASVspoof [12] deals with spoofing, but only for the voice part. We also considered to evaluate this system by separately testing visual and audio components (e.g. by artificially associate audios from ASVspoof to images from CASIA-SURF, building fictitious subjects), but we finally decided it to be beyond the intent of this work.

A Code

All codes are provided as links in section 3.

B Bibliography

References

[1] About FaceID - A brief overview of how FaceID works, including security issues.

[2] The iPhone X can't tell the difference between identical twins, 2017.

[3] This 10-year-old was able to unlock his mom's iPhone using Face ID, 2017.

[4] Feds Force Suspect To Unlock An Apple iPhone X With Their Face, 2018.

[5] Windows Hello Unlocked by a Photo, 2017.

[6] Feds Force Suspect To Unlock An Apple iPhone X With Their Face, 2018.

[7] Abdrakhmanova, M.; Kuzdeuov, A.; Jarju, S.; Khassanov Y.; Lewis M.; Varol H.A, SpeakingFaces: A Large-Scale Multimodal Dataset of Voice Commands with Visual and Thermal Video Streams, Sensors 2021.

[8] A. Baevski, H. Zhou, A. Mohamed, M. Auli: Wave2Vec 2.0 A Framework for Self-Supervised Learning of Speech Representations, 2020.

[9] M. Ravanelli et al., SpeechBrain: A General-Purpose Speech Toolkit.

[10] J. Deng et al., ArcFace: Additive Angular Margin Loss for Deep Face Recognition.

[11] S. Zhang et al., CASIA-SURF: A Large-scale Multi-modal Benchmark for Face Anti-spoofing, 2020.

[12] X. Wang et al., ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech, 2020.