ASMA

Anti-Spoofing-Focused Multi-Biometric Authentication

Biometric Systems - Computer Science 2024/25

SAPIENZA
UNIVERSITÀ DI ROMA

Francesco De Persio
Andrea Donato

Overview

- Introduction and goals
- Project structure
- Dataset
- Implementation
- Results
- Future works

Introduction

In recent years, biometric authentication has become a fundamental tool in both the IT and physical fields. Systems such as Apple FaceID, Windows Hello or Samsung's Iris Scanner are concrete and widespread examples

Goals

We want to develop a multi-biometric system for open-set identification in a controlled environment, based on face and voice recognition, with active anti-spoofing verification

An application context for our idea is a high-security environment (banks, government agencies) where an explicit manifestation of the access intent is required

Project structure

1. Generate and show a random sentence to pronounce aloud
2. A speech-to-text module evaluates what the user said
3. Features of RGB and IR images and audio are extracted
4. Perform a Feature Level Fusion of RGB and IR data to obtain an overall score by a Single Matching Module. The voice is matched alone
5. Perform a Score Level Fusion of voice and face to take a decision

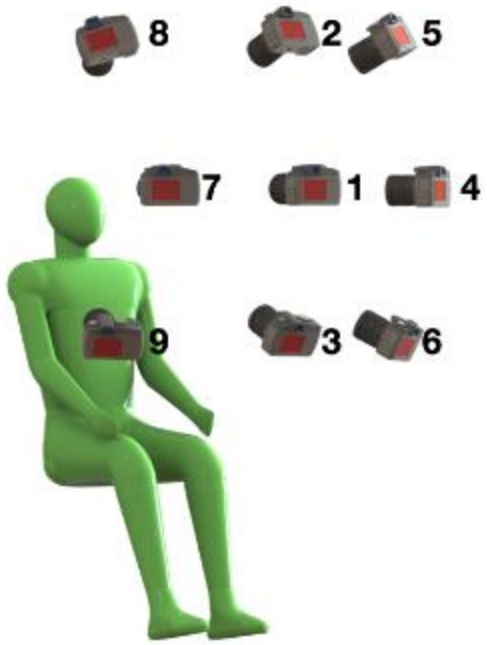
Dataset

SpeakingFaces provided by ISSAI

It consists of aligned high-resolution thermal and visual spectra image streams of fully-framed faces synchronized with audio recordings of each subject speaking 100 imperative phrases. Data were collected from 142 subjects, yielding over 13,000 instances of synchronized data

Dataset

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



Implementation

We split the different tasks on multiple colab notebooks:

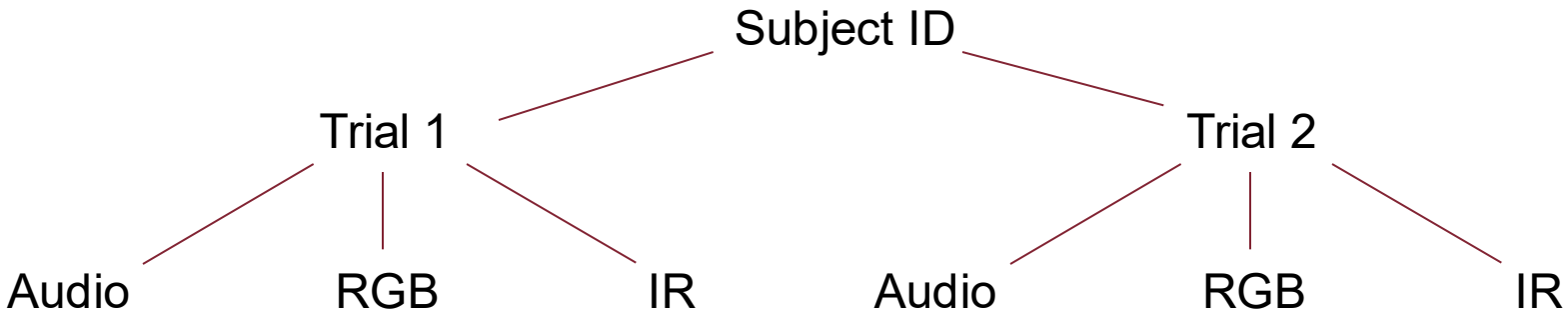
- Dataset Management
- Enrollment
- Feature Extraction Module
- Evaluation
  - Matching module
  - Decision module
  - Evaluation module

Implementation

Dataset Management

Used mainly to scan the dataset to find incomplete subjects

With the list of valid subjects we build a json file containing the paths of images and audio. This allows us not to iterate over the entire dataset everytime



Implementation

Dataset Management

We took:

- 3 images per camera, per subject, for both RGB and IR images
- All the audio

The RGB and IR images are syncronized on the same frame of the same video

Implementation

Feature Extraction Module

The Feature Extraction Module (FEM) notebook loads all the audio and a random sample of RGB and IR pictures

All the loaded data was then passed to the two pre-trained Neural-Network-based models for feature extraction

- **SpeechBrain**
- **InsightFace**

The so-obtained features are again stored in a json file

Implementation

Feature Extraction Module

- **SpeechBrain – ECAPA-TDNN**

ECAPA-TDNN is a deep learning model for speaker recognition, integrated into SpeechBrain. It extracts accurate voice embeddings, ideal for verifying or comparing voice identities

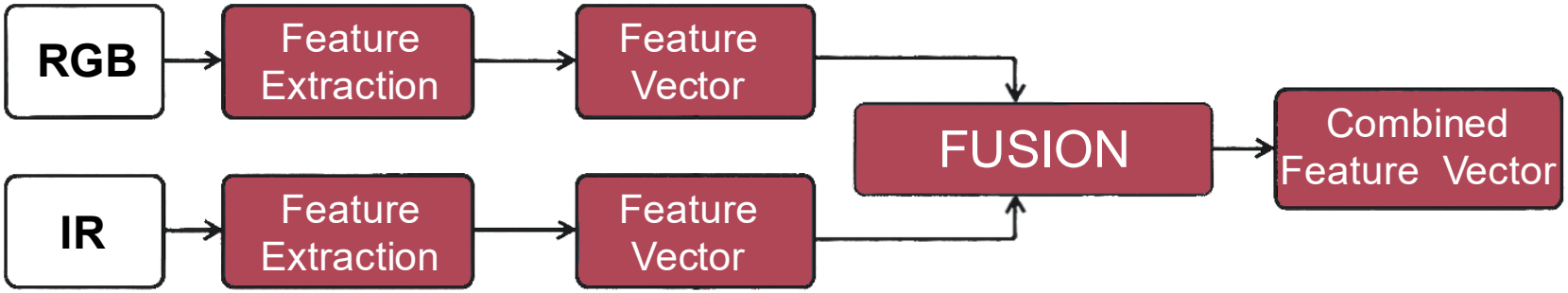
- **InsightFace**

Open-source, state-of-art library for RGB face analysis, based on three modules that provide Face Detection, Face Alignment and Face Recognition

Implementation

Feature Extraction Module

We perform our Feature Level Fusion concatenating both RGB and IR features, both ([512])-shaped, in a single ([1024]) tensor



Implementation

Evaluation

In this notebook are implemented:

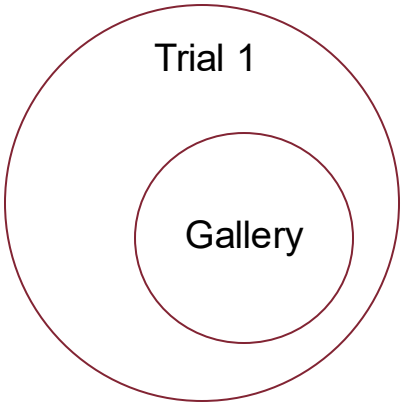
- Matching module
- Decision module
- Evaluation Module

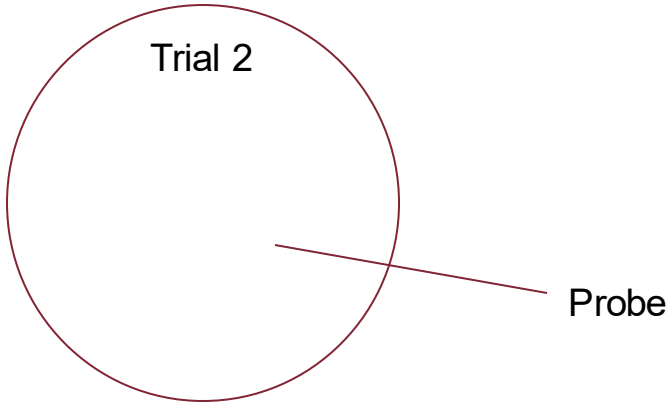
Implementation

Evaluation

The gallery is built by taking extracted features from trial 1 of a random subset of subjects

The probes are chosen from trial 2



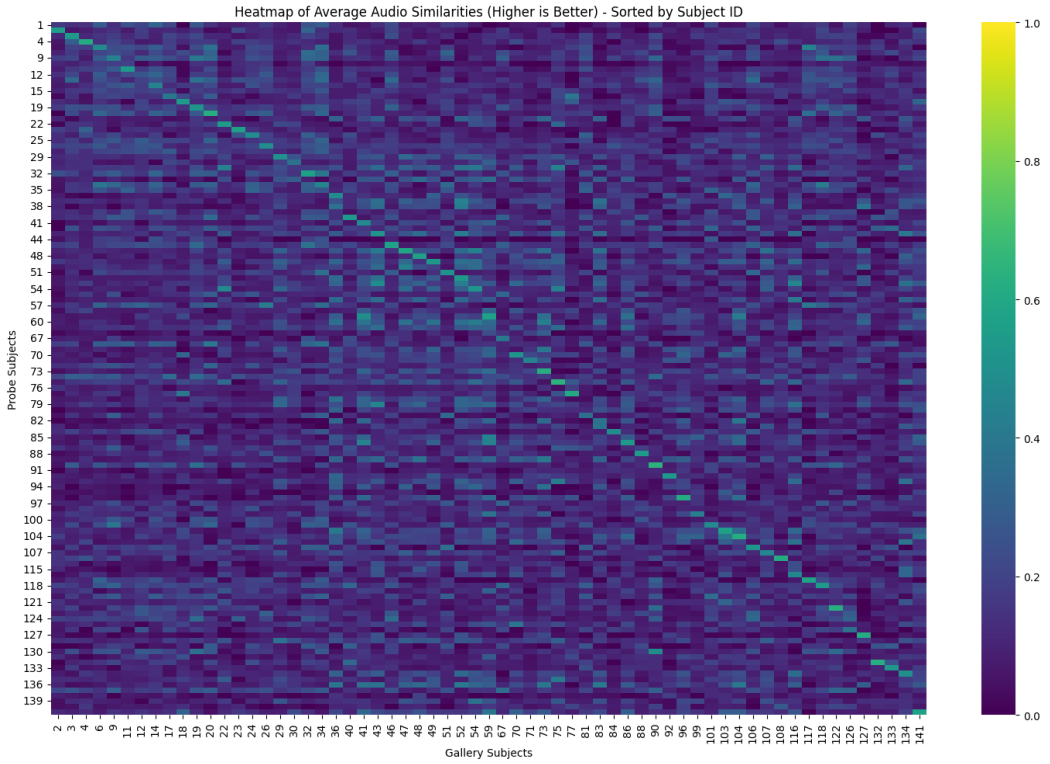


Implementation

Evaluation - Matching module

To evaluate audio matchings we decided to use similarity

The audio of the individual probe is matched with all the audio in the gallery and the average similarity score is taken

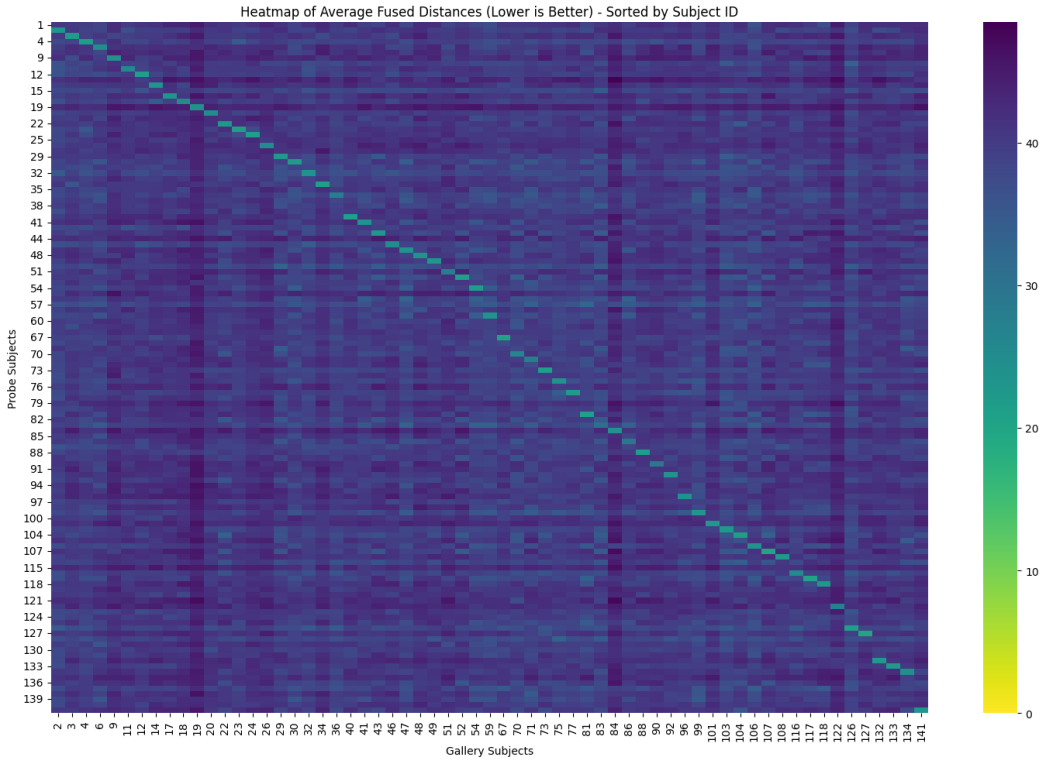


Implementation

Evaluation - Matching module

Since we implemented the concatenation of two different InsightFace embeddings we decided to use the euclidean distance

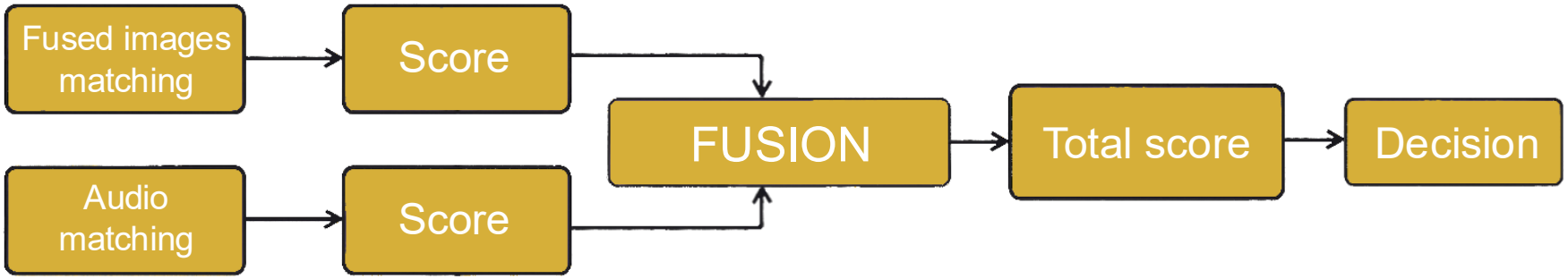
The matching process is analogous to the previous one



Implementation

Evaluation - Decision module

The decision module is a simple linear combination between the two scores followed by a threshold



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

Implementation

Evaluation - Evaluation module

```
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
```

We want to work in very delicate scenarios so DIR at rank 1 is a must

The goal is to avoid any False Acceptance

Results

The model exceeded expectations, achieving zero-FAR and even zero-FRR, without risk of overfitting due to the use of pre-trained models

The excellent results could be influenced by a dataset collected in optimal conditions, but show that similar performances are achievable in controlled environments, by appropriately adjusting the threshold

Future works

- Training an ad-hoc classifier for thermal images or Fine tuning an existing one
- Explore more sophisticated techniques of Feature Level Fusion
- Time-correlate all the data
- Build and use a dataset containing actual spoofing attempts

Thank you for your attention