# SUR: Project Technical Report

**Vladislav Sokolovskii**
xsokol15@stud.fit.vutbr.cz

## 1 Introduction

In this project, we have developed a multi-modal identification system that leverages both facial images and voice recordings to recognize 31 different individuals. The primary goal of this project is to create a robust and accurate classifier that can efficiently distinguish between the given subjects based on the provided training data.

## 2 Method

This section outlines the methodology employed for the person recognition system, focusing on the image and audio components.

### 2.1 Image

#### 2.1.1 Data Collection and Preprocessing

A low-resource image dataset was used, containing facial images in various conditions. The images were preprocessed by converting them to grayscale. Data augmentation was performed on-the-fly using the Albumentations (A. Buslaev and Kalinin, 2018) library to enhance the dataset size and improve the model's generalization capabilities.

#### 2.1.2 Model Training and Hyperparameter Optimization

A CNN was trained for 12,000 epochs, with each epoch presenting different augmented data to the model. Weights & Biases (Biewald, 2020) was used for monitoring the training progress. An initial set of hand-crafted configurations was tested, followed by the application of Optuna (Akiba et al., 2019) for hyperparameter search. To prevent overfitting, a high dropout rate and L2 regularization were used. The best model achieved an average accuracy of 72% on the validation set, and this checkpoint was used for subsequent inference.

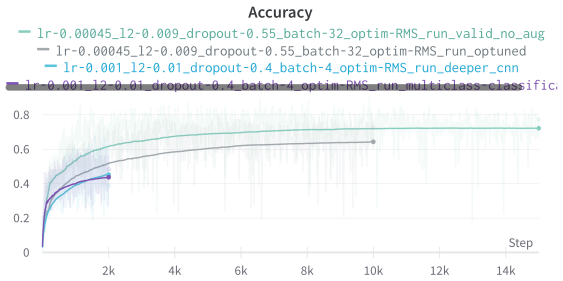


Figure 1: Average accuracy on validation dataset

### 2.2 Audio

#### 2.2.1 Data Collection and Preprocessing

The audio dataset consists of voice recordings of individuals. The raw audio was processed using a Voice Activity Detection (VAD) [1] model, followed by the extraction of Mel Frequency Cepstral Coefficients (MFCC) features using the librosa library (McFee et al., 2015).

#### 2.2.2 Model Training

Due to time constraints, a Support Vector Machine (SVM) model was trained only on the provided data without applying any augmentation or noise filtering. The SVM achieved an accuracy of 64% on the validation dataset. It is believed that with the addition of data augmentation, improved VAD, and noise filtering, the accuracy could be increased to over 80%.

### 2.3 Fusion

The fusion of the image and audio modalities is achieved by combining the probability outputs of the CNN and SVM models, which produce probabilities for 31 classes. A linear interpolation of probabilities is performed using a parameter $\alpha$. The fusion process is described by the following equations:

[1] https://github.com/mathigatti/silence-removal

audio_probs = $[p_1^{aud}, p_2^{aud}, \ldots, p_{31}^{aud}]$

fused_probs = $\alpha \cdot$ image_probs $+ (1 - \alpha) \cdot$ audio_probs

fused_pred = $\arg\max_i(\text{fused\_probs\_norm}_i) + 1$

By employing this simple fusion approach with $\alpha = 0.1$, an accuracy of 83% was achieved on the validation set, demonstrating the effectiveness of combining image and audio features for person recognition.

## 3 Reproduce the Results

To reproduce the results of the person recognition system, follow these steps:

1. Create a Python 3.9 environment and install all dependencies listed in the `requirements.txt` file.
2. Add a `dataset` folder to the project directory. This folder must contain `train`, `dev`, and `eval` subfolders with the respective data.
3. Run the script `./run.sh`. This script will train SVM, load the best CNN checkpoint and perform inference on the data located in the `dataset/eval` directory and create a text file in the required format named `results.txt` in the `src` directory.

Following these steps will allow you to reproduce the results achieved by the person recognition system.

## 4 Possible Improvements

Although the current person recognition system has demonstrated promising results, there are several potential improvements that can be made to enhance its performance and capabilities:

1. **Data Augmentation for Audio:** Similar to the image modality, incorporating data augmentation techniques for the audio data could increase the robustness of the model and improve its generalization capabilities.
2. **Improved Voice Activity Detection:** Using a more advanced VAD model or fine-tuning the existing VAD model on the specific dataset could lead to better identification of speech segments, resulting in improved accuracy for the audio modality.
3. **Noise Filtering:** Implementing noise filtering techniques to reduce the impact of background noise on the audio features can potentially enhance the model's performance.
4. **Deep Learning Models for Audio:** Exploring the use of deep learning models, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), for the audio modality might lead to improved feature representation and classification accuracy.
5. **Optimization of Fusion Parameter:** Investigating different approaches for optimizing the fusion parameter $\alpha$, such as grid search or Bayesian optimization, could result in better performance when combining the image and audio modalities.
6. **Advanced Fusion Techniques:** Instead of simple linear interpolation, more advanced fusion techniques, like feature-level fusion or decision-level fusion using ensemble learning methods, could be explored to better exploit the complementary information provided by the image and audio modalities.

Incorporating these improvements into the current person recognition system has the potential to significantly enhance its performance and applicability to a broader range of scenarios.

## 5 Conclusion

This project developed a multi-modal person recognition system using image and audio data. A Convolutional Neural Network (CNN) with data augmentation was employed for images, while a Support Vector Machine (SVM) with Voice Activity Detection (VAD) and Mel Frequency Cepstral Coefficients (MFCC) feature extraction was used for audio. A simple fusion approach with linear interpolation combined the two modalities, achieving 83% accuracy on the validation set. Future work includes exploring advanced fusion techniques, optimizing model parameters, and incorporating data augmentation for audio to further enhance the system's performance.

## References


E. Khvedchenya V. I. Iglovikov A. Buslaev, A. Parinov and A. A. Kalinin. 2018. Albumentations: fast and flexible image augmentations. *ArXiv e-prints*.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8.