SUR Projekt 2024/2025

Trénovnání modelů pro identifikaci osob z obrázku obličeje a hlasové nahrávky

Pavel Osinek (xosine00)

5. května 2025

1 Úvod

Cílem projektu bylo vytvořit systém pro rozpoznávání 31 různých osob na základě dvou typů vstupních dat: snímků obličejů a hlasových nahrávek. Pro každý typ vstupu byla navržena samostatná trénovací a vyhodnocovací pipeline.

Výsledná struktura systému:

- Audio klasifikátor využívající MFCC příznaky a Gaussovské směsi (GMM)
- Obrazový klasifikátor založený na extrakci HOG příznaků, následné redukci dimenze pomocí PCA a klasifikaci pomocí SVM
- Evaluační skripty a hlavní řídicí soubor `main.py`, který zajišťuje spuštění všech kroků.

2 Návod ke spuštění

Zdrojové kódy jsou umístěny v adresáři `SRC/`. Výsledné soubory (`voice_eval_result.txt`, `image_eval_result.txt`) je možné vytvořit v kořenovém adresáři projektu po úspěšném spuštění evaluace (viz níže).

Závislosti a instalace: Systém je implementován v jazyce Python 3.12.10 a využívá následující externí knihovny:

- `numpy`
- `scikit-learn`
- `librosa`
- `scikit-image`
- `opencv-python`

Instalaci lze provést:

```
pip install numpy scikit-learn librosa scikit-image opencv-python
```

Data: Před spuštěním je nutné mít data ve složkách `data/train`, `data/dev`, `data/eval` ve formátu dle zadání (31 podadresářů tříd).

Spuštění všech kroků projektu jedním příkazem:

```
python main.py --all
```

Jednotlivé kroky lze spustit volitelně:

```
python main.py --voice-train     # trénink hlasového modelu
python main.py --voice-eval      # evaluace modelu pro audio a uložení výsledků
python main.py --image-train     # trénink obrazového modelu
python main.py --image-eval      # evaluace modelu pro obrázky a uložení výsledků
```

3 Popis implementace

Celá implementace vznikala postupným experimentováním a validací výsledků na předpřipravené validační (`dev`) sadě. V úvodu byly stanoveny dvě nezávislé pipeline – pro hlasová data a pro obrázky obličejů.

voice_train.py

Návrh hlasového klasifikátoru původně vycházel ze standardního SVM aplikovaného na příznaky získané z MFCC (mel-frekvenčních cepstrálních koeficientů). Tato základní metoda však dosahovala na validační sadě nízké přesnosti (typicky pod 30 %).

Vzhledem k těmto výsledkům byl přístup změněn na trénování samostatného GMM (modelu Gaussovských směsí) pro každou osobu na základě akustických příznaků. Experimenty probíhaly s různými parametry: počtem MFCC koeficientů (vyšší počet lépe vystihoval spektrum řeči), rozšířením o různé MFCC koeficienty (pro zvýšení robustnosti), počtem komponent GMM (nejlépe se osvědčilo 32, vyšší počet již nepřinášel znatelné zlepšení) a typem kovariančí matice (nejlepší výsledky přinesla volba "tied"). Ve výsledné implementaci se používá 64 MFCC příznaků rozšířených o první a druhé derivace (delta a delta-delta).

Výrazného zlepšení robustnosti na validační sadě bylo dosaženo použitím jednoduchých augmentací zvukového signálu – zejména přidáním bílého šumu a náhodnými změnami hlasitosti, což pomohlo snížit rozdíly v kvalitě nahrávek a zvýšily schopnost modelu generalizovat. Každá zvuková ukázka prochází také cepstrální normalizací (CMVN), aby se snížil vliv rozdílů v hlasitosti nahrávek.

Výsledný model je uložen do souboru `voice.gmms` a připraven pro evaluaci. Na validační sadě (`dev`) dosáhl hlasový systém přibližně 76 % přesnosti.

voice_eval.py

Evaluační skript načítá uložené GMM modely a aplikuje na evaluační zvuková data stejný postup extrakce příznaků (MFCC + delta + delta delta) jako při trénování. Pro každý vstupní vzorek vypočítá skóre (logaritmickou pravděpodobnost) pro všechny osoby a do výsledného souboru zapíše jak predikovanou třídu (tu s nejvyšším skóre), tak skóre všech tříd dle zadané specifikace. Tohoto výsledku bylo docíleno zejména experimentování s nastavením parametrů GMM modelů a použitím datové augmentace.

image_train.py

Pro rozpoznání osob na základě snímků obličejů bylo vyzkoušeno několik přístupů. Přímé použití SVM klasifikátoru na zmenšených obrázcích vedlo k nízké přesnosti. Výrazného zlepšení bylo dosaženo při využití příznaků HOG (Histogram of Oriented Gradients). Tyto příznaky reprezentují hranovou strukturu objektu a slouží jako vstup pro klasifikaci. Vzhledem k vysoké dimenzi vektoru HOG byla provedena redukce dimenze pomocí PCA se zachováním 98 %

rozptylu, čímž se zlepšila jak výpočetní efektivita, tak i přesnost klasifikace díky potlačení šumu. Klasifikace byla realizována pomocí SVM s RBF jádrem.

Při ladění parametrů se jako nejvhodnější ukázala hodnota $C = 2$, zatímco lineární jádro vedlo ke znatelně nižší přesnosti. Trénink proběhl jak s augmentacemi (např. rotace, zrcadlení, změna jasu), tak bez nich.

Výsledný model (HOG + PCA + SVM s RBF jádrem a augmentacemi) dosáhl na validační sadě (`dev`) přesnosti přibližně 74.2 %. Natrénovaný model je uložen jako jeden objekt pomocí `pickle` do souboru `image.svm`.

image_eval.py

Evaluační pipeline pro obrazová data načte každý PNG soubor, provede extrakci HOG příznaků a následnou transformaci pomocí PCA. Na takto připravených vektorech model SVM vypočítá skóre (logaritmus posteriorní pravděpodobnosti) pro všech 31 tříd. Výstupní soubor obsahuje jak predikovanou třídu, tak log pravděpodobnosti pro všechny třídy. Formát výsledku odpovídá zadání.

main.py

Skript `main.py` umožňuje jednotlivé kroky řetězit v požadovaném pořadí, případně rychle spustit všechny pipeline jedním příkazem (`-all`).

4 Závěr

- Přesnost modelu pro obrazová data: **76.0 %**
- Přesnost modelu pro zvuková data: **74.2 %**

Nejlepší výsledky pro úlohu identifikace osob přineslo použití GMM klasifikátoru a MFCC pro zvuková data a kombinace HOG, PCA a SVM pro obrazová data. U hlasových nahrávek model každé osoby využívá MFCC příznaky rozšířené o delta a delta-delta koeficienty, přičemž byla provedena jednoduchá augmentace (změna hlasitosti a přidání šumu). Pro rozpoznávání podle obličejů byly po extrakci HOG příznaků použity PCA a SVM s RBF jádrem. Experimenty ukázaly, že augmentace (rotace, převrácení, změna jasu) vedla ke zvýšení přesnosti.

Při nastavování parametrů byly u SVM i GMM testovány různé hodnoty počtu komponent, typy kovariance, rozsahy PCA a parametry jádra SVM. Výsledné modely byly ověřovány na validační sadě, kde se potvrdilo, že přechod od jednodušších metod (např. SVM přímo na MFCC nebo snímky) ke kombinovanému přístupu s vhodnou extrakcí příznaků a augmentacemi zásadně zvýšil úspěšnost rozpoznávání. Výsledky obou systémů splňují zadání a připravují vstup pro případné další zpracování či kombinaci modalit.