Multimodal audio-video person recognition using Deep Neural Networks

*Thesis advisor:*
Conf. Dr. Ing. Horia Cucu Gabriel

*Student:*
Sandu Marian
Table of contents

● Introduction
● CDEP Dataset Development
● Neural network architectures
● Face and speaker recognition results
● Multimodal results
● Personal contributions
● Conclusions
Introduction

- Motivation
- Objectives
- Implementation steps
- Initial data specifications
Introduction

- Motivation
- Objectives
- Implementation steps
- Initial data specifications
Introduction

- Motivation
- Objectives
- *Implementation steps*
- Initial data specifications
Introduction

- Motivation
- Objectives
- Implementation steps
- *Initial data specifications*

- Videos
- *HTML files* corresponding to each video
- *Audio files*, each corresponding to a speech
- Text file which contains every *speaker* in the database with an associated unique *ID*
CDEP* Dataset Development

Already done

Download videos and html files → Extract wav files from the videos → Creation and modification of the cdepSpk.txt file → Extraction of transcripts and timestamps

Extract images from videos → Create the image database and its subsets → Validate images with faces

Extract only the voice from the audio files → Split audio files into 3 seconds chunks → Create the audio database and its subsets

*CDEP = Chamber of Deputies
CDEP Dataset Development

1. Download videos and html files
2. Extract wav files from the videos
3. Creation and modification of the cdepSpk.txt file
4. Extraction of transcripts and timestamps

- Extract images from videos
- Create the image database and its subsets
- Validate images with faces

- Extract only the voice from the audio files
- Split audio files into 3 seconds chunks
- Create the audio database and its subsets

HTML processing
Image processing
Audio processing
CDEP Dataset Development

Examples of faces

Extraction of valid frames
CDEP Dataset Development

Bad cases
- Bad images
- More persons in a folders

Faulty image removal algorithm

Similarity matrix

<table>
<thead>
<tr>
<th>1.0</th>
<th>0.9</th>
<th>0.7</th>
<th>0.6</th>
<th>0.4</th>
<th>0.3</th>
<th>0.2</th>
<th>0.1</th>
<th>0.0</th>
<th>-0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>1.0</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
<td>0.4</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
<td>-0.1</td>
</tr>
<tr>
<td>0.8</td>
<td>0.9</td>
<td>1.0</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
<td>0.4</td>
<td>0.3</td>
<td>0.2</td>
<td>-0.1</td>
</tr>
<tr>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>1.0</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
<td>0.4</td>
<td>0.3</td>
<td>-0.1</td>
</tr>
<tr>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>1.0</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
<td>0.4</td>
<td>0.3</td>
</tr>
<tr>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>1.0</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
<td>-0.1</td>
</tr>
<tr>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>1.0</td>
<td>0.8</td>
<td>0.7</td>
<td>-0.1</td>
</tr>
<tr>
<td>0.3</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>1.0</td>
<td>0.8</td>
<td>-0.1</td>
</tr>
<tr>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>1.0</td>
<td>-0.1</td>
</tr>
<tr>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>1.0</td>
</tr>
<tr>
<td>0.0</td>
<td>-0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
</tr>
<tr>
<td>-0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Group: row files with similarity matrices for all the classes
Group average similarity for each class

Average similarity > 0.7
Delete Images of classes with non-valid similarity of < 0.6

Yes
Delete class

No
Clusterization step

Difference between clusters' similarity < 0.1
Delete cluster with lower similarity score

Yes

No
CDEP Dataset Development

1. Download videos and html files
2. Extract wav files from the videos
3. Creation and modification of the cdepSpk.txt file
4. Extraction of transcripts and timestamps
5. Extract images from videos
6. Create the image database and its subsets
7. Validate images with faces
8. Extract only the voice from the audio files
9. Split audio files into 3 seconds chunks
10. Create the audio database and its subsets

HTML processing
Image processing
Audio processing
CDEP Dataset Development

Audio processing algorithms

Voice Activity Detector [1]

Retrieval audio files

Apply Voice Activity Detector algorithm

The algorithm extracted only the parts of the audio with voice

Merge the parts to create a file without silence

Audio pre-processing algorithm
CDEP Dataset Development

Database Development conclusions

![Image histogram](image_url)

![Audio histogram](image_url)

<table>
<thead>
<tr>
<th># audio files range</th>
<th># persons</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 - 5</td>
<td>2</td>
</tr>
<tr>
<td>6 - 10</td>
<td>23</td>
</tr>
<tr>
<td>11 - 20</td>
<td>15</td>
</tr>
<tr>
<td>21 - 50</td>
<td>39</td>
</tr>
<tr>
<td>51 - 100</td>
<td>250</td>
</tr>
</tbody>
</table>
## CDEP Dataset Development

### Database Development conclusions

<table>
<thead>
<tr>
<th>Dataset</th>
<th>No. of samples / class</th>
<th>No. of classes</th>
<th>No. of training samples</th>
<th>No. of evaluation samples</th>
<th>No. of test samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image10</td>
<td>10</td>
<td>257</td>
<td>1542</td>
<td>514</td>
<td>514</td>
</tr>
<tr>
<td>Image50</td>
<td>50</td>
<td>132</td>
<td>3960</td>
<td>1320</td>
<td>1320</td>
</tr>
<tr>
<td>Image100</td>
<td>100</td>
<td>84</td>
<td>5040</td>
<td>1680</td>
<td>1680</td>
</tr>
<tr>
<td>Audio_3s_10</td>
<td>10</td>
<td>257</td>
<td>1542</td>
<td>514</td>
<td>514</td>
</tr>
<tr>
<td>Audio_3s_50</td>
<td>50</td>
<td>132</td>
<td>1542</td>
<td>514</td>
<td>514</td>
</tr>
</tbody>
</table>

### Derived datasets configuration
Neural network architectures - Face recognition

**GoogLeNet**
- Input
- Convolution
- Max pool
- 2 x Inception
- 4 x Inception
- Max pool
- 2 x Inception
- Average pool
- Dropout
- Softmax

**VGG16**
- Input
- 3 x 3 conv, 64
- 3 x 3 conv, 64
- Pool
- 3 x 3 conv, 128
- 3 x 3 conv, 128
- Pool
- 3 x 3 conv, 256
- 3 x 3 conv, 256
- Pool
- 3 x 3 conv, 512
- 3 x 3 conv, 512
- Pool
- 3 x 3 conv, 512
- 3 x 3 conv, 512
- Pool
- FC 512
- FC 512
- Softmax

**FaceNet**
- Input
- Stem
- Reduction A
- Reduction B
- Reduction B
- Inception-resepnet B
- Inception-resepnet C
- Average pool
- Dropout
- Softmax
Neural network architectures - Speaker recognition

VGGVOX [5]
# Monomodal results

## Face recognition results

<table>
<thead>
<tr>
<th>Database</th>
<th>No. of classes</th>
<th>Architecture</th>
<th>Test accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image10</td>
<td>257</td>
<td>FaceNet</td>
<td>93.2%</td>
</tr>
<tr>
<td>Image10</td>
<td>257</td>
<td>VGG16</td>
<td>81.3%</td>
</tr>
<tr>
<td>Image10</td>
<td>257</td>
<td>GoogLeNet</td>
<td>67%</td>
</tr>
<tr>
<td>Image50</td>
<td>132</td>
<td>FaceNet</td>
<td>98.3%</td>
</tr>
<tr>
<td>Image50</td>
<td>132</td>
<td>VGG16</td>
<td>95%</td>
</tr>
<tr>
<td>Image50</td>
<td>132</td>
<td>GoogLeNet</td>
<td>95%</td>
</tr>
<tr>
<td>Image100</td>
<td>84</td>
<td>FaceNet</td>
<td>99.2%</td>
</tr>
<tr>
<td>Image100</td>
<td>84</td>
<td>VGG16</td>
<td>97%</td>
</tr>
<tr>
<td>Image100</td>
<td>84</td>
<td>GoogLeNet</td>
<td>97%</td>
</tr>
</tbody>
</table>

## Speaker recognition results

<table>
<thead>
<tr>
<th>Database</th>
<th>No. of classes</th>
<th>Test accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio_3s_10</td>
<td>257</td>
<td>98.24%</td>
</tr>
<tr>
<td>Audio_3s_50</td>
<td>132</td>
<td>99.23%</td>
</tr>
</tbody>
</table>
Multimodal recognition.

Results

Multimodal architecture

Results

<table>
<thead>
<tr>
<th>Database</th>
<th>Number of classes</th>
<th>Batch size</th>
<th>Optimizer</th>
<th>Test accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image10 Audio_3s_10</td>
<td>257</td>
<td>32</td>
<td>SGD</td>
<td>99.82%</td>
</tr>
<tr>
<td>Image50 Audio_3s_50</td>
<td>132</td>
<td>32</td>
<td>SGD</td>
<td>99.92%</td>
</tr>
</tbody>
</table>
Conclusions

- Final results
- Further validation of the model
- Personal contributions
- Further development

<table>
<thead>
<tr>
<th></th>
<th>Face recognition vs Multimodal recognition</th>
<th>Speaker recognition vs Multimodal recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 samples / class</td>
<td>+ 6%</td>
<td>+ 1.58 %</td>
</tr>
<tr>
<td>50 samples / class</td>
<td>+ 1.6 %</td>
<td>+ 0.7 %</td>
</tr>
</tbody>
</table>
Conclusions

- Final results
- **Further validation of the model**
- Personal contributions
- Further development

**VidTIMIT Database examples**

**CDEP database**

**VidTIMIT database**
Conclusions

- Final results
- Further validation of the model
- Personal contributions
- Further development

- Developing an algorithm for face database creation and validation
- Developing an algorithm for audio database creation and validation
- Fine-tuning three state-of-the-art neural networks for face recognition and choosing the best option, and one state-of-the-art network for audio recognition
- Creating and evaluating a multimodal architecture for person identification
Conclusions

● Final results
● Further validation of the model
● Personal contributions
● Further development

● Model optimization
  ○ Multimodal parameter tuning
● Training on a larger dataset
  ○ Cloud computing for a higher data volume
● Live recognition
  ○ API support for video upload
  ○ Automatic database update
● Website integration
  ○ User-friendly interface for identification
Bibliography


- A. Nagrani*, J. S. Chung*, A. Zisserman, VoxCeleb: a large-scale speaker identification dataset, INTERSPEECH, 2017
Thank you !
Source code

https://git.speed.pub.ro/diploma/multimodal-person-identification