Multimodal audio-video person recognition using Deep Neural Networks

Thesis advisor:

Conf. Dr. Ing. Horia Cucu

Student:

Sandu Marian Gabriel

Table of contents

Introduction

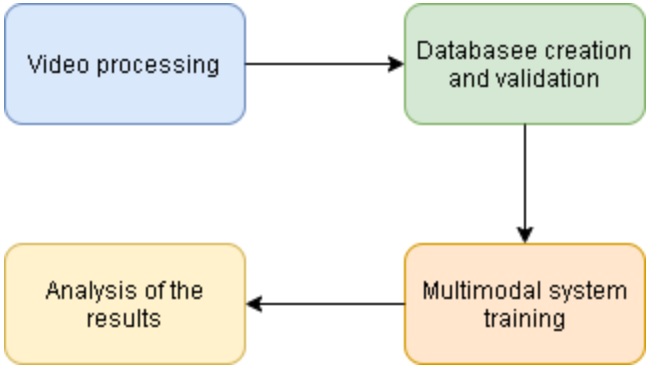
- *Motivation*
- Objectives
- Implementation steps
- Initial data specifications

Introduction

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

Introduction

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

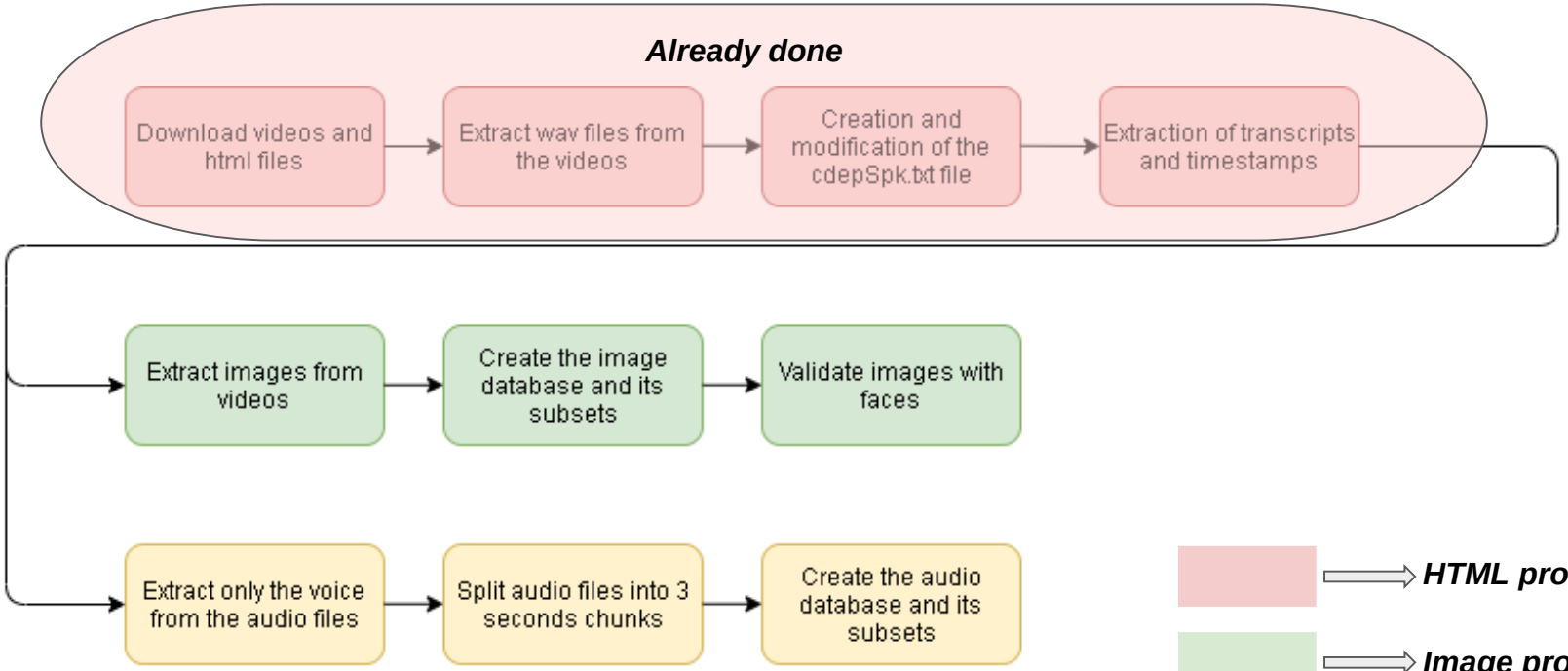


Introduction

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

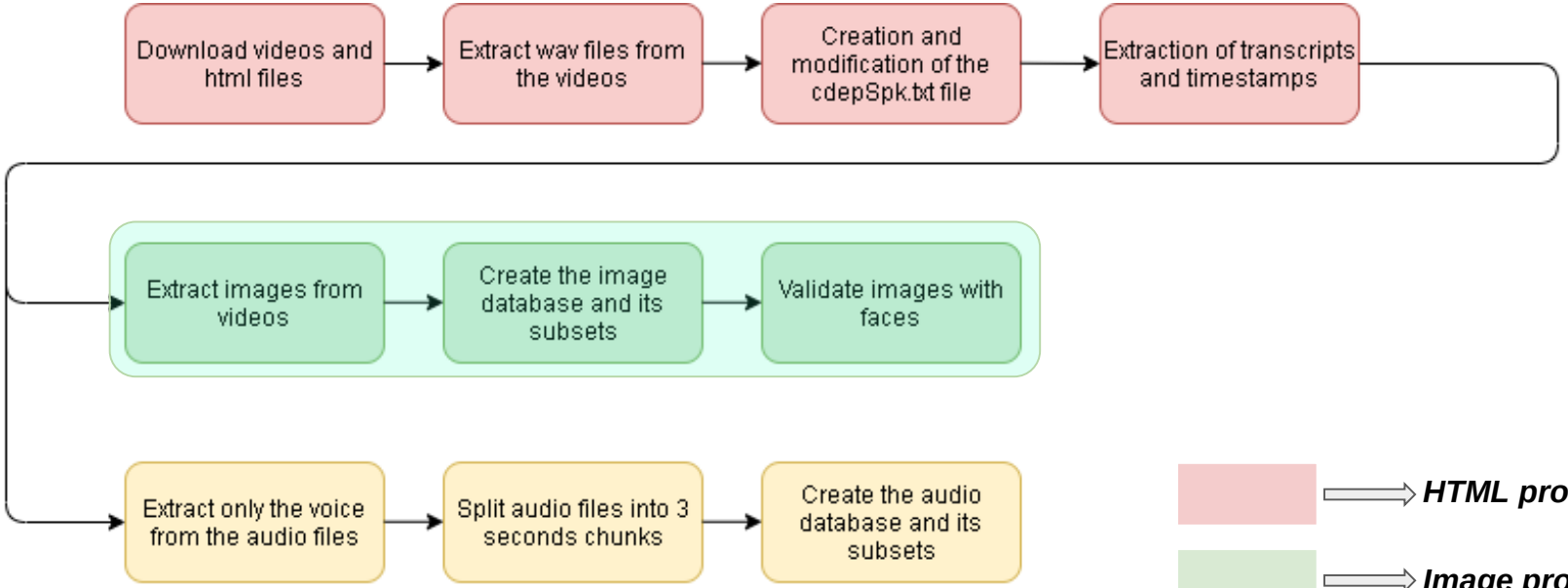
- *Videos*
- *HTML files* corresponding to each video
- *Audio files*, each corresponding to a speech
- Text file which contains every *speaker* in the database with an associated unique *ID*

CDEP* Dataset Development



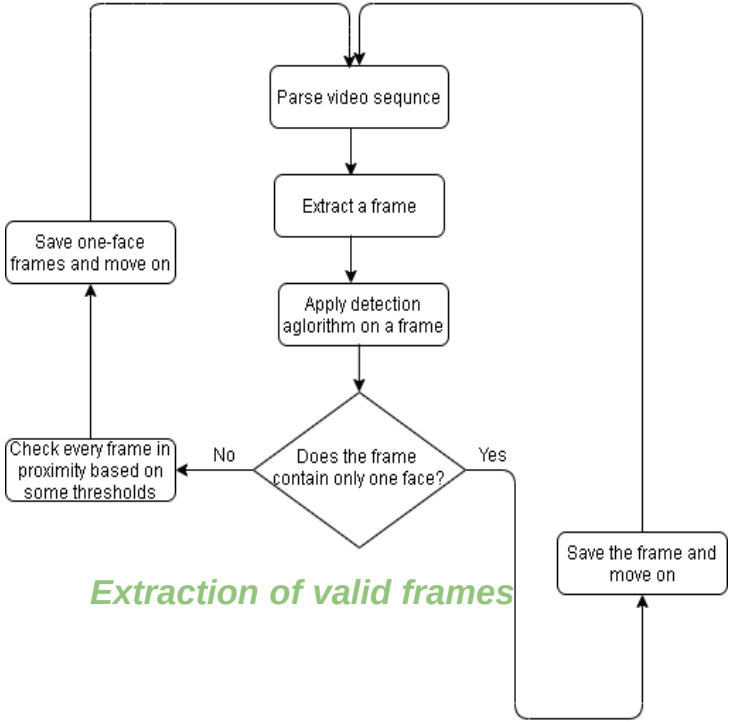
*CDEP = Chamber of Deputies

CDEP Dataset Development



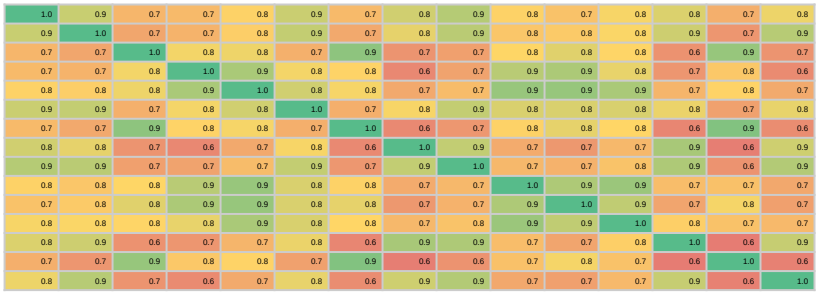
CDEP Dataset Development

Examples of faces

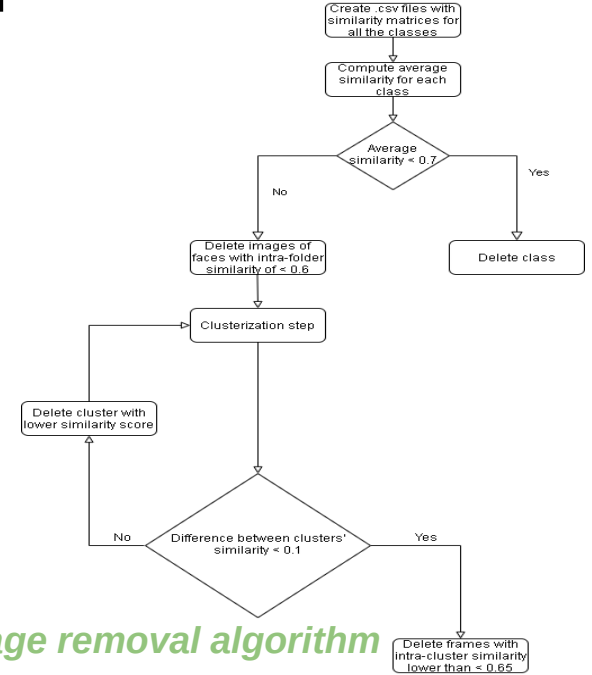


Extraction of valid frames

CDEP Dataset Development



Similarity matrix

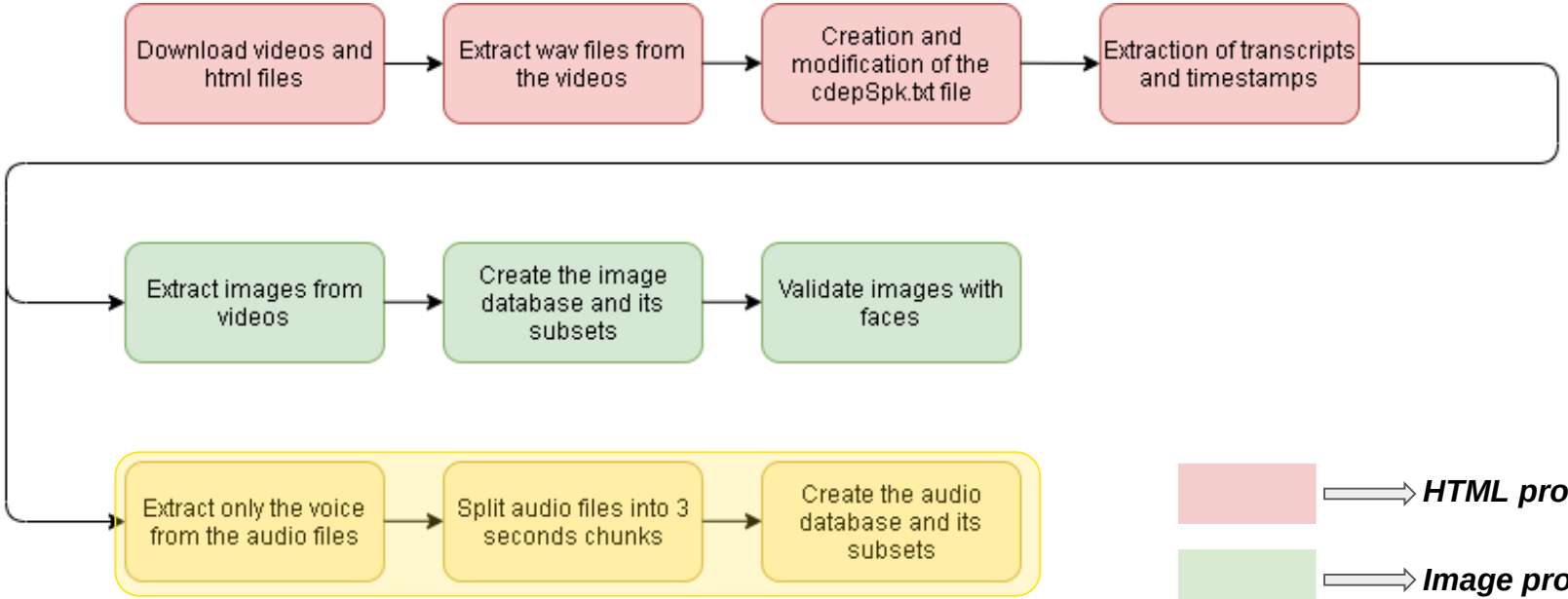


Faulty image removal algorithm

Bad cases

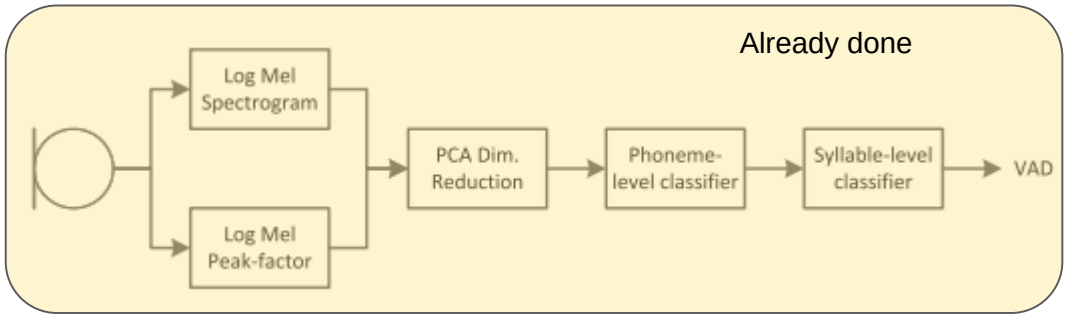
- ***Bad images***
- ***More persons in a folders***

CDEP Dataset Development

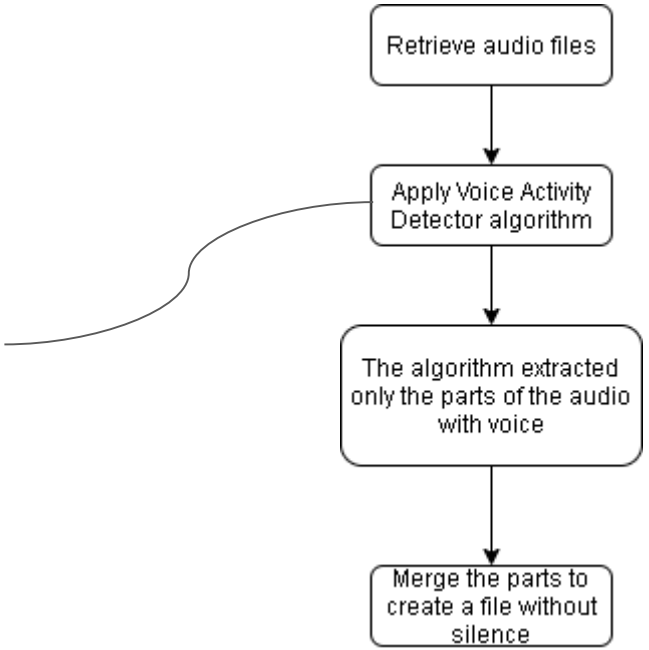


CDEP Dataset Development

Audio processing algorithms



Voice Activity Detector [1]



Audio pre-processing algorithm

CDEP Dataset Development

Database Development conclusions

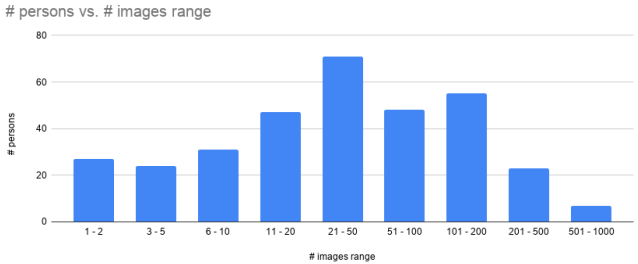


Image histogram

# audio files range	# persons
3 - 5	2
6 - 10	23
11 - 20	15
21 - 50	39
51 - 100	250

Audio histogram

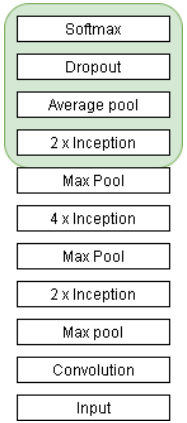
CDEP Dataset Development

Database Development conclusions

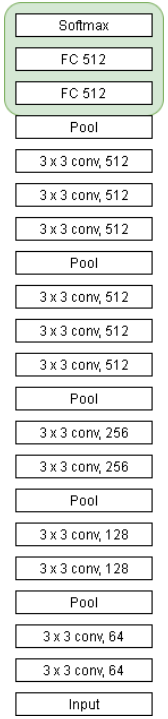
Dataset	No. of samples / class	No. of classes	No. of training samples	No. of evaluation samples	No. of test samples
Image10	10	257	1542	514	514
Image50	50	132	3960	1320	1320
Image100	100	84	5040	1680	1680
Audio_3s_10	10	257	1542	514	514
Audio_3s_50	50	132	1542	514	514

Derived datasets configuration

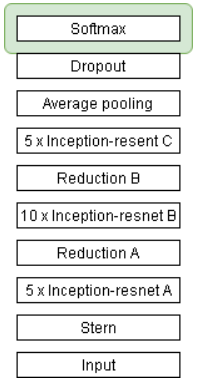
Neural network architectures - Face recognition



GoogLeNet

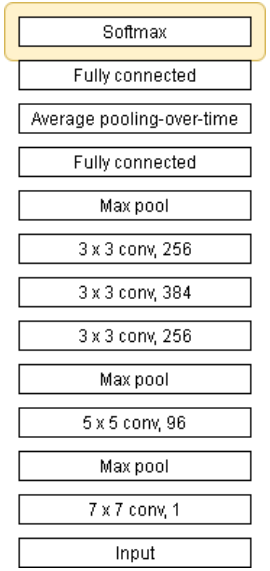


VGG16



FaceNet

Neural network architectures - Speaker recognition



VGGVOX [5]

Monomodal results

Face recognition results

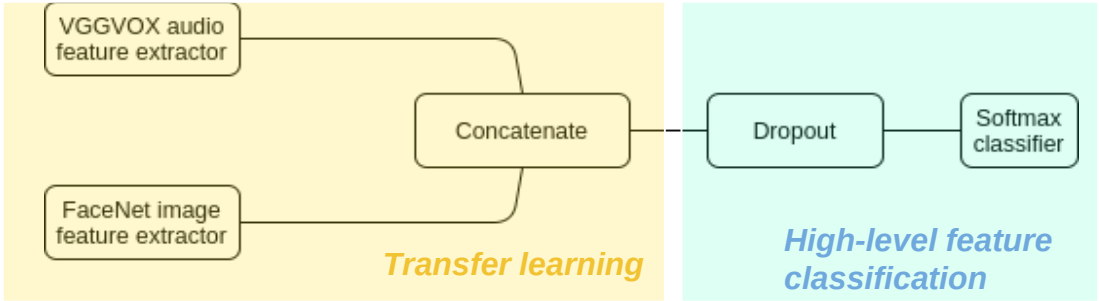
Database	No. of classes	Architecture	Test accuracy
Image10	257	FaceNet	93.2%
Image10	257	VGG16	81.3%
Image10	257	GoogLeNet	67%
Image50	132	FaceNet	98.3%
Image50	132	VGG16	95%
Image50	132	GoogLeNet	95%
Image100	84	FaceNet	99.2%
Image100	84	VGG16	97%
Image100	84	GoogLeNet	97%

Speaker recognition results

Database	No. of classes	Test accuracy
Audio_3s_10	257	98.24%
Audio_3s_50	132	99.23%

Multimodal recognition. Results

Multimodal architecture



Results

Database	Number of classes	Batch size	Optimizer	Test accuracy
Image10 Audio_3s_10	257	32	SGD	99.82%
Image50 Audio_3s_50	132	32	SGD	99.92%

Conclusions

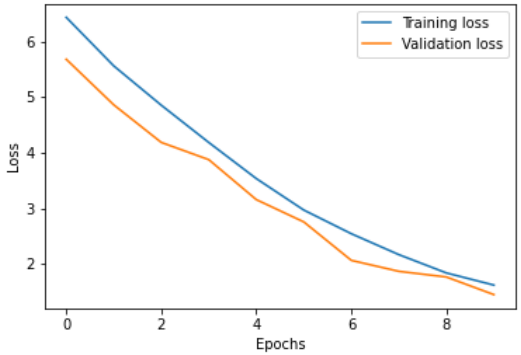
- **Final results**
- Further validation of the model
- Personal contributions
- Further development

	Test accuracy	
	Face recognition vs **Multimodal recognition**	**Speaker recognition** vs **Multimodal recognition**
10 samples / class	+ 6%	+ 1.58 %
50 samples / class	+ 1.6 %	+ 0.7 %

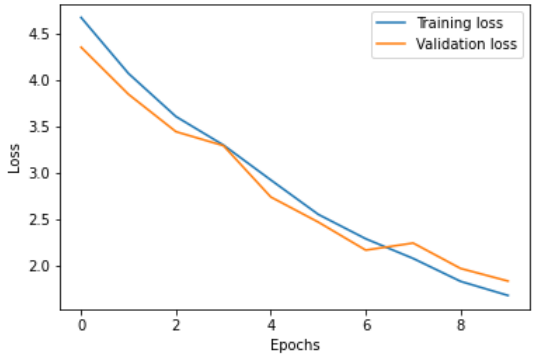
Conclusions

- 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

- [1] Salishev, Sergey & Barabanov, Andrey & Kocharov, Daniil & Skrelin, Pavel & Moiseev, Mikhail. (2016). Voice Activity Detector (VAD) Based on Long-Term Mel Frequency Band Features. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 9924. 352-358. 10.1007/978-3-319-45510-5_40.
- [2] *Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich*; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1-9
- [3] Karen Simonyan, Andrew Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition
- [4] Florian Schroff, Dmitry Kalenichenko, James Philbin, FaceNet: A Unified Embedding for Face Recognition and Clustering
- A. Nagrani*, J. S. Chung*, A. Zisserman, VoxCeleb: a large-scale speaker identification dataset, INTERSPEECH, 2017


SpeeD
Speech & Dialogue
Research Laboratory


POLITEHNICA
UNIVERSITATEA
BUCUREŞTI
200
ani
aniversare
1818

Thank you !

Source code

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