BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF INTELLIGENT SYSTEMS
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

DEEPFAKE DETECTION FRAMEWORK

FRAMEWORK PRO DETEKCI DEEPFAKES

MASTER'S THESIS
DIPLOMOVÁ PRÁCE

AUTHOR **Bc. JAN BERNARD**
AUTOR PRÁCE

SUPERVISOR **Mgr. KAMIL MALINKA, Ph.D.**
VEDOUCÍ PRÁCE

BRNO 2022

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Master's Thesis Assignment

140642

Institut:	Department of Intelligent Systems (UITS)
Student:	**Bernard Jan, Bc.**
Programme:	Information Technology and Artificial Intelligence
Specialization:	Cybersecurity
Title:	**Deepfake Detection Framework**
Category:	Security
Academic year:	2022/23

Assignment:

1. Learn about deepfakes (voice and video). Explore the current state of deepfakes detection methods (voice and video).
2. Learn about the technologies needed to create web extensions and technologies for creating scalable server applications.
3. Learn about existing deepfake detection solutions (e.g. other commercial web browser plug-ins)
4. Design an extensible framework (server-client or client-only) for deepfakes detection (support for at least 3 detection methods (voice and video)). Design a web extension for deepfakes detection that will use this framework. The solution should support multiple browsers and allow the detection of displayed content and uploaded files.
5. Implement the tool according to the design.
6. Test the functionality and reliability of the resulting implementation. Perform testing on at least two independent publicly available deepfakes datasets.
7. Discuss usability, detection efficiency and possible extensions.

Literature:
Puspita Majumdar, Akshay Agarwal, Mayank Vatsa, and Richa
Singh, “Facial retouching and alteration detection,” in Handbook of Digital Face Manipulation and Detection, pp. 367–
387. Springer, 2022
FIRC Anton a MALINKA Kamil. The dawn of a text-dependent society: deepfakes as a threat to speech verification systems. In: Brno: Association for Computing Machinery, 2022

Requirements for the semestral defence:
Items 1 to 4.

Detailed formal requirements can be found at https://www.fit.vut.cz/study/theses/

Supervisor:	**Malinka Kamil, Mgr., Ph.D.**
Consultant:	Ing. Anton Firc
Head of Department:	Hanáček Petr, doc. Dr. Ing.
Beginning of work:	1.11.2022
Submission deadline:	17.5.2023
Approval date:	3.11.2022

Abstract

Do tohoto odstavce bude zapsán výtah (abstrakt) práce v anglickém jazyce.

Abstrakt

Do tohoto odstavce bude zapsán výtah (abstrakt) práce v českém (slovenském) jazyce.

Keywords

Sem budou zapsána jednotlivá klíčová slova v anglickém jazyce, oddělená čárkami.

Klíčová slova

Sem budou zapsána jednotlivá klíčová slova v českém (slovenském) jazyce, oddělená čárkami.

Reference

BERNARD, Jan. *Deepfake Detection Framework*. Brno, 2022. Master's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Mgr. Kamil Malinka, Ph.D.

Deepfake Detection Framework

Declaration

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana X... Další informace mi poskytli... Uvedl jsem všechny literární prameny, publikace a další zdroje, ze kterých jsem čerpal.

.....
Jan Bernard
December 17, 2022

Acknowledgements

V této sekci je možno uvést poděkování vedoucímu práce a těm, kteří poskytli odbornou pomoc (externí zadavatel, konzultant apod.).

Contents

Chapter 1

Introduction

- deepfake is buzzword (no agreed-upon technical definition) - ...

Chapter 2

Deepfakes

The creation of fake media and their detection have been a problem since photography was invented. Digital photography or video with tools such as GIMP, Adobe Photoshop or Adobe After Effects allows more people to create fakes than before, still some experience in this area is needed. Media that have been modified or otherwise manipulated are called synthetic media, and they do not depend on whether it is an analogue or digital medium. Deepfakes also fall under this category [2]. Tools powered by deep learning allow unexperienced users to easily create trusted fakes.

The quality of deepfakes reached a level when a trained person or even an experienced researcher in this field has a problem of spotting them. This fast development allows creating realistically looking assets to art photography or movie production, unfortunately, it can be used for malicious purposes like creating fake porn videos to blackmail people or manipulate public via fake news. There are many use cases where deepfakes can be applied.

It is putting huge pressure on researchers to develop new forensics tools or any technology which will prevent malicious usage of deepfakes. As mentioned before, creating fakes is not new, and a whole field of study engaged in spotting fakes and developing techniques over 15 years. Despite continuous research efforts in the past, the advent of deep learning changed the rules of the game. [5]

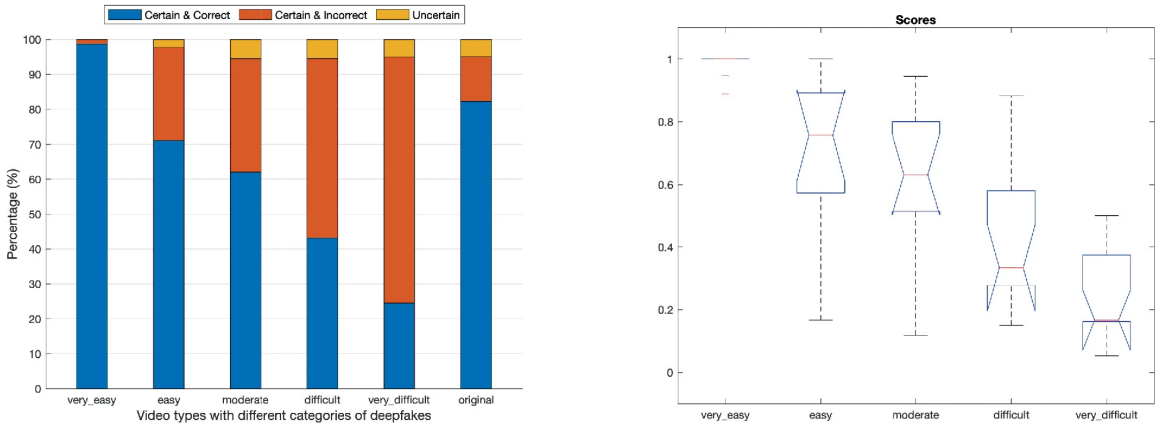
2.1 Human capabilities of deepfake detection

The human ability to recognize fake materials from the originals is in contradiction to their quality. Korshunov and Marcel confirmed this in their research. They created a questionnaire containing several videos and the subject (interviewee) had to answer after watching the video whether the person in the video was genuine, fake or they are uncertain. The videos were manually divided into five categories (very easy, easy, moderate, difficult, and very difficult, original).

Videos were split into several categories manually by researchers probably without usage of any metrics but based on their personal feelings, and ANOVA test shows there is an overlap in several categories so several videos could be moved to different category. However, the categories are still significantly different.

The results of test in figure 2.1 certainly demonstrate that people's recognition ability decreases significantly as the quality of deepfakes increases. Also, audience of this test knows they are looking for fakes, otherwise we can expect worse results if there will be unsuspected audience (for example: deepfakes on social media). It is quite alarming that

the correct answers in the category of "very easy" reach only 71,1 %. The quality of deepfake increases over time, thus it can be expected that human recognition ability will continue to decrease. [3]



(a) Subjective answers

(b) ANOVA test

Figure 2.1: Subjective answers and median values with error bars from ANOVA test for different deepfake categories [3].

Another research tested only recognition of audio tracks and they were comparing humans versus computer programs. Attendees had correct classification between fakes and origins 67 % after the first several rounds. Their accuracy increased while listening and answering to more tracks, but the value stabilizes on 80 %. On average trained AI performs about 10 % better than human, but this result highly depends on difference of learning and test dataset. Still, it shows that the computer can outperform humans in spotting deepfakes. [4]

2.2 Potentional risks

[1] [2]

2.3 Types of deepfakes and their generation

Chapter 3

Analysis of existing tools for detecting deepfakes

3.1 A

3.2 B

3.3 C

Chapter 4

Deepfake detection

4.1 Voice deepfakce detection

4.2 Image/Video deepfakce detection

4.3 A

4.4 B

4.5 C

Chapter 5

Architecture analysis

Chapter 6

Framework architecture

6.1 High level architecture

6.2 Containerazation and scaling

6.3 Input layer

6.4 Data preparation layer

6.5 Inividual detection containers

Chapter 7

Client architecture

7.1 Web plugin

Chapter 8

Framework implementation

Chapter 9

Client implementation

Chapter 10

Test experiment and results

Chapter 11

Conclusion

Bibliography

[1] Firc, A. and Malinka, K. The dawn of a text-dependent society: deepfakes as a threat to speech verification systems. In: *SAC '22: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. Association for Computing Machinery, 2022, p. 1646–1655. DOI: 10.1145/3477314.3507013. Available at: https://www.fit.vut.cz/research/publication/12595.

[2] Homeland Security. *Increasing Threat of Deepfake Identites* [https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf]. 2021 [cit. 2022-12-16].

[3] Korshunov, P. and Marcel, S. The Threat of Deepfakes to Computer and Human Visions. In: Rathgeb, C., Tolosana, R., Vera Rodriguez, R. and Busch, C., ed. *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*. Cham: Springer International Publishing, 2022, p. 97–115. DOI: 10.1007/978-3-030-87664-7_5. ISBN 978-3-030-87664-7. Available at: https://doi.org/10.1007/978-3-030-87664-7_5.

[4] Müller, N. M., Pizzi, K. and Williams, J. Human perception of audio deepfakes. In: *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. Association for Computing Machinery, 2022, p. 85–91. DOI: 10.1145/3552466.3556531. ISBN 9781450394963. Available at: https://doi.org/10.1145/3552466.3556531.

[5] Verdoliva, L. Media Forensics and DeepFakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing*. 2020, vol. 14, no. 5, p. 910–932. DOI: 10.1109/JSTSP.2020.3002101.