Funded in part by the German Federal Ministry of Education and Research (BMBF) under grant number (FKZ): 13N15736

Joining of Data-driven Forensics and Multimedia Forensics for Deepfake Detection on the Example of Image and Video Data

Dennis Siegel, Christian Kraetzer, Jana Dittmann
Otto-von-Guericke University, Magdeburg, Germany
dennis.siegel@ovgu.de

The Seventeenth International Conference on Emerging Security Information, Systems and Technologies
SECURWARE 2023

About the presenter

Dennis Siegel
dennis.siegel@ovgu.de
Otto-von-Guericke University, Magdeburg, Germany

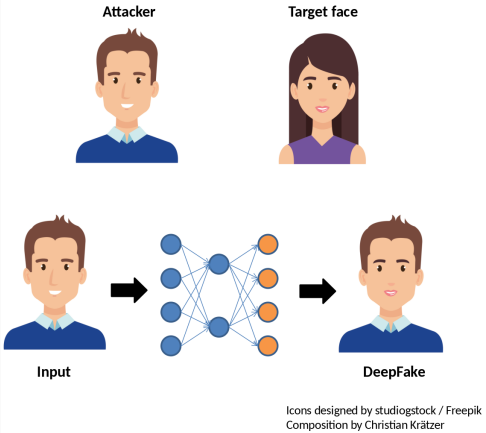
- PhD student at Advanced Multimedia and Security Lab (AMSL) at Otto-von-Guericke University Magdeburg (OvGU)
- 2021: received his masters degree in Computer Science at OvGU
- currently working on the research project “FAKE-ID”
- research field: media forensic

Content

- Introduction and Motivation
- Brief summary of the state of the art
- Conceptional joining of IT and media forensic methodologies
- Application of the concept
- Summary, Conclusions and Future Work

Introduction and Motivation

- DeepFakes pose a new challenge to digital medias integrity and authenticity
- with further advances it will be near impossible to spot them with the bare eyes
- Current state of the art to mitigate the threat of fake identities is based on detectors utilize machine learning (mostly deep learning) to detect DeepFake.
- But with the intended usage in the forensic context, further requirements have to be met (i.e., aspects of human oversight and control).



Introduction and Motivation

- Many **process models** for forensic processes exist for 'traditional' forensic sub-disciplines (e.g. dactyloskopy), the purpose of these is to **make the corresponding investigations fit for courtroom usage** (i.e. define standards for application of methods, certification of practitioners, etc.)
- Most media forensic approaches today still lack maturity in this regard
- Academic focus here lies mostly only on **proposing detectors** for specific forensic tasks like image manipulation detection or DeepFake detection
- This scientific domain is in need of **modeling work, aiming at creating corresponding domain specific process models**

Introduction and Motivation

- Our contributions in this paper:
  - conceptional **joining of IT and media forensic methodologies** on the selected example of the existing *Data-Centric Examination Approach (DCEA) Kiltz* [2020]; *Siegel et al.* [2022] and the *Best Practice Manual for Digital Image Authentication (BPM-DI)* from the *European Network of Forensic Science Institute (ENFSI) European Network of Forensic Science Institutes* [2021].
  - illustration of **applicability and benefits** of our concept on the example of **three existing applications** ExifTool (*Phil Harvey* [2016]), the hand-crafted DeepFake detector DF_{mouth} (*Siegel et al.* [2021]) as well as the deep learning based DeepFake detector LipForensics (*Haliassos et al.* [2021]).

State-of-the-art: Domain specific process modeling for media forensics

- European SOTA on media forensic process models: ENFSI best practice manuals, e.g. on Digital Image Authentication (*European Network of Forensic Science Institutes* [2021]), Forensic Video and Image Enhancement (*European Network of Forensic Science Institutes* [2018a]) and Facial Image Comparison (*European Network of Forensic Science Institutes* [2018b])

- German situation regarding IT forensics in general (including media forensics): BSI "Leitfaden IT-Forensik" *BSI* [2011]
- Accepted models / standardization / certification of media forensic methods are rare in general (see e.g. discussion in *Krätzer* [2013])
- For novel media forensics tasks like DeepFake detection no dedicated process models are currently existing (to our knowledge)

State-of-the-art: Domain specific process modeling for media forensics

- **Starting point of our own modeling work**:
  - ENFSI "Best Practice Manual for Digital Image Authentication" *European Network of Forensic Science Institutes* [2021]
  - BSI "Leitfaden IT-Forensik" *BSI* [2011]
  - and existing publications extending this process model, especially *Kiltz* [2020]
  - initial steps towards taylor-made model for media forensics (including DeepFake detection): *Siegel et al.* [2022]

  - regulatory documents, i.e., Artificial Intelligence Act (AIA) *European Commission* [2021]; *European Parliament* [2023]

State-of-the-art: Domain specific process modeling for media forensics

ENFSI "Best Practice Manual for Digital Image Authentication"

- "*aims to provide a framework for procedures, quality principles, training processes and approaches to the forensic examination*"
- describes investigation steps for image authenticity validation
- investigation steps are categorized in four aspects:
  - Auxiliary data analysis
  - Image content analysis
  - Strategies
  - Peer review

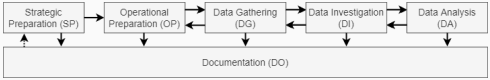


Figure 1: Phase model (based on *BSI* [2011])

State-of-the-art: Deriving a domain adapted context model for media forensics

Forensic data type	Description
MFDT1 "digital input data"	The initial media data considered for the investigation.
MFDT2 "processed media data"	Results of transformations to media data (e.g. grayscale conversion, cropping)
MFDT3 "contextual data"	Case specific information (e.g. for fairness evaluation)
MFDT4 "parameter data"	Contain settings and other parameter used for acquisition, investigation and analysis
MFDT5 "examination data"	including the traces, patterns, anomalies, etc that lead to an examination result
MFDT6 "model data"	Describe trained model data (e.g. face detection and model classification data)
MFDT7 "log data"	Data, which is relevant for the administration of the system (e.g. system logs)
MFDT8 "chain of custody & report data"	Describe data used to ensure integrity and authenticity (e.g. hashes and time stamps) as well as the accompanying documentation for the final report.

Table 1: **M**edia **F**orensic **D**ata **T**ypes (MFDT) proposed in *Siegel et al.* [2022]

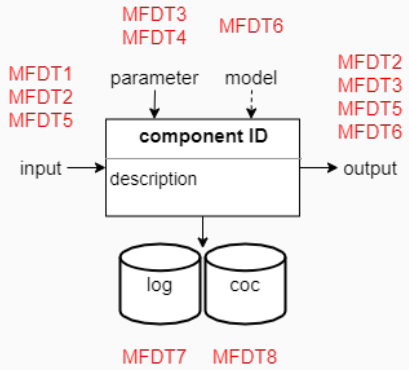


Figure 2: Template structure for a single component introduced in *Siegel et al.* [2022]

Conceptional joining of IT and media forensic methodologies (1/2)

- Aim: identify investigation steps to validate integrity and authenticity of digital media (here: DeepFakes)
- on the basis of the existing Best Practice Manual for Digital Image Authentication and DCEA
- according to the phase model: SP is excluded

Conceptional joining of IT and media forensic methodologies

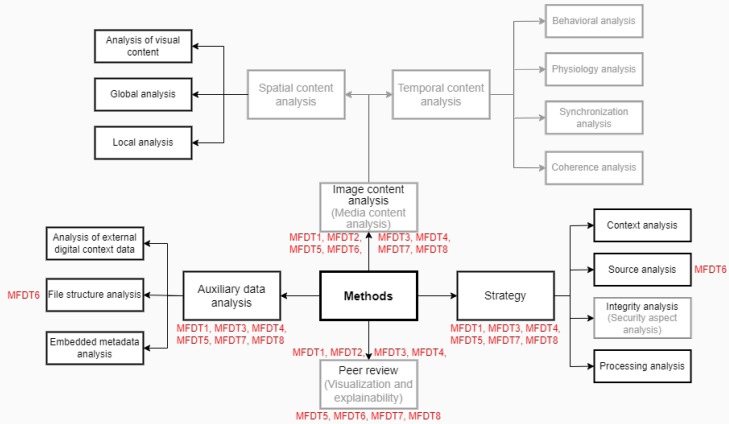


Figure 3: Extension of the forensic methodology proposed in *European Network of Forensic Science Institutes* [2021]. Extensions are marked in gray. Application of data types can be found in red.

Tools: ExifTool

- open source tool by *Phil Harvey* [2016] to read, write and edit metadata
- applicable to a wide range of image and video formats
- in total, eight features are extracted, including both required and optional metadata fields

Ext. BPM-DI		feature	description	value	processing step	analysis	strategy	data type
Auxiliary data analysis	Analysis of external digital context data	ID-exif$_1$	MACtime	timestamp	PS-exif	File system metadata	Processing analysis	MFDT3
		ID-exif$_2$	file size	string				
		ID-exif$_3$	system feature flags					
	File structure analysis	ID-exif$_4$	file format	string		File structures	Source & Processing analysis	
		ID-exif$_5$	file format version	version number				
		ID-exif$_6$	video codec	string				
	Embedded meta-data analysis	ID-exif$_7$	file resolution	int [0, ∞]		Additional metadata	Context analysis	
		ID-exif$_8$	file frame rate	real [0, ∞]				

Table 2: Categorization of ExifTool according to the proposed methodology.

Tools: DF_{mouth}

- DeepFake detector based on the mouth region, proposed in *Siegel et al.* [2021]
- usage of traditional machine learning (i.e., hand-crafted feature spaces and classification)
- *Krätzer et al.* [2023]: benchmarking of the detector based on datasets FaceForensics++ (*Rössler et al.* [2019]), DFD (*Rössler et al.* [2019]), Celeb-DF (*Li et al.* [2020]) and HiFiFace (*Wang et al.* [2021])
- detection performance: 69.9% accuracy

→ suitable only for certain DeepFake synthesis methods
→ integration for decision support

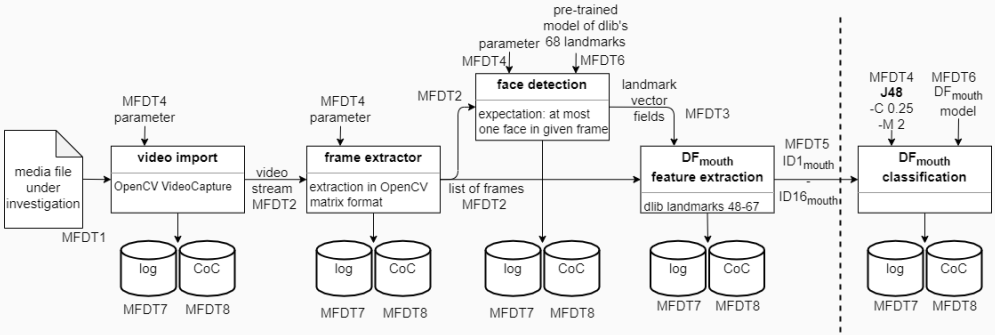


Figure 4: Process pipeline of DF_{mouth} as proposed in *Siegel et al.* [2021] and *Krätzer et al.* [2023].

Tools: DF_{mouth}

Ext. BPM-DI		feature	description	value	processing step	analysis	strategy	data type
Media content analysis	Temporal content analysis	ID-mouth$_1$	abs max change Y	*real [0, ∞]*	PS-mouth$_4$	Physiology analysis	Context analysis	MFDT5
		ID-mouth$_2$	max change Y	*real [0, ∞]*				
		ID-mouth$_3$	min change Y	**real [-∞, 0]**				
		ID-mouth$_4$	abs max change X	*real [0, ∞]*				
		ID-mouth$_5$	max change X	*real [0, ∞]*				
		ID-mouth$_6$	min change X	**real [-∞, 0]**				
		ID-mouth$_7$	percentage time state 1	**real [0, 1]**				
		ID-mouth$_{12}$	percentage time state 2	*real [0, 1]*				
	Spatial content analysis	ID-mouth$_8$	max regions state 1	*real [0, ∞]*	PS-mouth$_5$	Local analysis	Processing analysis	
		ID-mouth$_9$	max FAST keypoints state 1	*real [0, ∞]*				
		ID-mouth$_{10}$	max SIFT keypoints state 1	*real [0, ∞]*				
		ID-mouth$_{11}$	max sobel pixel state 1	*real [0, ∞]*				
		ID-mouth$_{13}$	min regions state 2	**real [0, ∞]**				
		ID-mouth$_{14}$	min FAST keypoints state 2	**real [0, ∞]**				
		ID-mouth$_{15}$	min SIFT keypoints state 2	**real [0, ∞]**				
		ID-mouth$_{16}$	max sobel pixel state 2	**real [0, ∞]**				

Table 3: Categorization of DF_{mouth} according to the proposed methodology.

Tools: LipForensics

- DeepFake detector based on the mouth region and its movement, proposed in *Haliassos et al.* [2021]
- theoretical inclusion to further validate the suitability for deep learning based approaches
- The detection process can be separated in three processing steps:
  1. preprocessing
  2. lip reading for feature extraction using pre-trained ResNet-18
  3. DeepFake classification using multiscale temporal convolutional network (MS-TCN)

Tools: LipForensics

Ext. BPM-DI		feature	description	value	processing step	analysis	strategy	data type
Media content analysis	Spatial content analysis	ID-LF_1	extraction of 25 frames, grayscale, crop and align	int [0, 255]	PS-LF_1	Local analysis	Context analysis	MFDT2
		ID-LF_2	feature extraction utilizing ResNet-18	feature vector of size 512	PS-LF_2	Local analysis	Context analysis	MFDT3
	Temporal content analysis	ID-LF_3	classification of mouth movement based on MS-TCN	label: {real, fake} probability: real [0, 1]	PS-LF_3	Physiology analysis	Processing analysis	MFDT5

Table 4: Categorization of LipForensics according to the proposed methodology.

Integrating the human operator in DeepFake detection

- each individual step of the underlying process have to be clear
- instead of providing the decision the features resulting in that decision have to be provided
- first conceptual example consists of four segments:
  1. filter for forensic categorization
  2. media player
  3. feature visualization
  4. decision-making by the human operator

Integrating the human operator in DeepFake detection

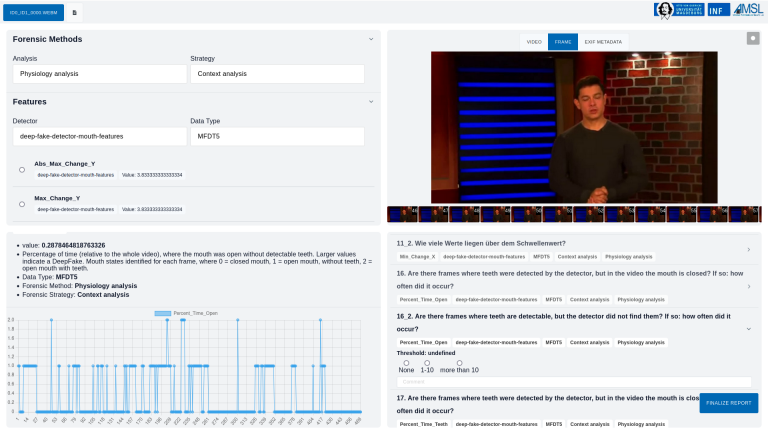


Figure 5: Demonstration of the extended Methods, exemplified on DF$_{mouth}$ for video id0_id1_0000 of the Celeb-DF dataset *Li et al.* [2020] - from *Siegel and Dittmann* [2023]

Summary, Conclusions and Future Work

- suitability of the proposed extensions can be validated for DeepFake detection
- in this exemplary approach not all methods can be addressed by using three tools
- only case specific investigation steps are addressed, the detectors suitability have to be validated beforehand, including steps of benchmarking and certification

- further research to enable a more specific integration of audio domain *European Network of Forensic Science Institutes* [2022]
- 'explainable AI', especially in the context of 'human in the loop' and 'human in control' have to explored in more detail to minimize the potential of error and uncertainty

Thank you for your attention!

References

BSI, *Leitfaden IT-Forensik*, German Federal Office for Information Security, 2011.

European Commission, Proposal for a Regulation of the european parliament and of the council Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, *COM(2021) 206 final*, [Online]. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 [Last retrieved: 14.09.2021]., 2021.

European Network of Forensic Science Institutes, Best practice manual for forensic image and video enhancement, *ENFSI-BPM-DI-02*, [Online]. Available at: https://enfsi.eu/wp-content/uploads/2017/06/Best-Practice-Manual-for-Forensic-Image-and-Video-Enhancement.pdf [Last retrieved: 12.09.2023]., 2018a.

European Network of Forensic Science Institutes, Best practice manual for facial image comparison, *ENFSI-BPM-DI-01*, [Online]. Available at: https://enfsi.eu/wp-content/uploads/2017/06/ENFSI-BPM-DI-01.pdf [Last retrieved: 12.09.2023]., 2018b.

European Network of Forensic Science Institutes, Best practice manual for digital image authentication, *ENFSI-BPM-DI-03*, [Online]. Available at: https://enfsi.eu/wp-content/uploads/2022/12/1.-BPM_Image-Authentication_ENFSI-BPM-DI-03-1.pdf [Last retrieved: 12.01.2023]., 2021.

European Network of Forensic Science Institutes, Best practice manual for digital audio authenticity analysis, *ENFSI-FSA-BPM-002*, [Online]. Available at: https://enfsi.eu/wp-content/uploads/2022/12/FSA-BPM-002_BPM-for-Digital-Audio-Authenticity-Analysis.pdf [Last retrieved: 12.09.2023]., 2022.

European Parliament, Amendments adopted by the european parliament on 14 june 2023 on the proposal for a regulation of the european parliament and of the council on laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, *COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)*, [Online]. Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html [Last retrieved: 12.09.2023]., 2023.

Haliassos, A., K. Vougioukas, S. Petridis, and M. Pantic, Lips don't lie: A generalisable and robust approach to face forgery detection, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 5039–5049, Computer Vision Foundation / IEEE, doi: 10.1109/CVPR46437.2021.00500, 2021.

Kiltz, S., Data-centric examination approach (DCEA) for a qualitative determination of error, loss and uncertainty in digital and digitised forensics, Ph.D. thesis, Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, 2020.

Krätzer, C., D. Siegel, S. Seidlitz, and J. Dittmann, Human-in-control and quality assurance aspects for a benchmarking framework for DeepFake detection models, *Electronic Imaging*, *35*(4), 379–1–379–6, doi: 10.2352/ei.2023.35.4.mwsf-379, 2023.

Krätzer, C., Statistical pattern recognition for audio-forensics - empirical investigations on the application scenarios audio steganalysis and microphone forensics, Ph.D. thesis, Otto-von-Guericke-Universität Magdeburg, 2013.

Li, Y., X. Yang, P. Sun, H. Qi, and S. Lyu, Celeb-df: A large-scale challenging dataset for deepfake forensics, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 3204–3213, IEEE, doi: 10.1109/CVPR42600.2020.00327, 2020.

Phil Harvey, Exiftool, [Online]. Available at: https://exiftool.org/ [Last retrieved: 12.09.2023], 2016.

Rössler, A., D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, Faceforensics++: Learning to detect manipulated facial images, in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 1–11, IEEE, doi: 10.1109/ICCV.2019.00009, 2019.

Siegel, D., and J. Dittmann, Tracemap, student project within the lecture of Multimedia and Security [MMSEC], Otto-von-Guericke-University Magdeburg, unpublished, 2023.

Siegel, D., C. Krätzer, S. Seidlitz, and J. Dittmann, Media forensics considerations on deepfake detection with hand-crafted features, *Journal of Imaging*, *7*(7), doi: 10.3390/jimaging7070108, 2021.

Siegel, D., C. Krätzer, S. Seidlitz, and J. Dittmann, Forensic data model for artificial intelligence based media forensics - illustrated on the example of DeepFake detection, *Electronic Imaging*, *34*(4), 324–1–324–6, doi: 10.2352/ei.2022.34.4.mwsf-324, 2022.

Wang, Y., et al., Hififace: 3d shape and semantic prior guided high fidelity face swapping, in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, edited by Z.-H. Zhou, pp. 1136–1142, International Joint Conferences on Artificial Intelligence Organization, doi: 10.24963/ijcai.2021/157, main Track, 2021.