Research Institute
Cyber Defence
Universität der Bundeswehr München

MEDIA SOURCE SIMILARITY HASHING (MSSH)

A Practical Method for Large-Scale Media Investigations

Samantha Klier and Harald Baier

Research Institute CODE

University of the Bundeswehr Munich

INTRODUCTION | MOTIVATION & CONTEXT

- Digital forensics faces massive data overload, e.g. 36M CSAM reports in 2023
- Cryptographic & perceptual hashes detect duplicates or similar content

  → Flooding Digital Forensics lab worldwide
- Gap: No fast method to identify files from the same media source…
- For example, to distinguish downloaded vs. self-produced content
- …besides Exif (easily thwarted, error prone and time consuming)

INTRODUCTION | RESEARCH GOAL & CONTRIBUTIONS

- Develop a lightweight similarity hash to group files by media source
- Operate efficiently at scale and robustly, also without metadata

→ Media Source Similarity Hash (MSSH)

- First syntactic approximate matching scheme
- Public Python implementation (open source)
- Evaluation across 7 datasets / 87k+ images

BACKGROUND | SIMILARITY HASHING

- Goal:
  - > Find degree of similarity between digital artifacts, in the range $[0 - 1]$,
- Based on extracted features which allow a binary decision
- Operate at byte, syntactic, or semantic levels
- For example: TLSH, ssdeep, sdhash operate on byte-level:
  - > independent of file format, but sought-after similarity must be reflected at byte-level ❌
- For example: PhotoDNA operates on semantic-level:
  - > Resemble human perception, find similar looking scenes ❌
- No similarity hash known on syntactic-level, considered cheap pre-processing by Breitinger et al.

BACKGROUND | SOURCE CAMERA IDENTIFICATION

- SPN methods verify image source via sensor noise patterns
  - > Similar, but different (used in practice when Exif approach exhausted)
  - > High accuracy but computationally expensive and storage-heavy
- Proposed metadata-based methods for images (Mullan et al.)
  - > are lightweight, but rely on Exif metadata
  - > need training data
- Proposed structure-based methods (Iuliani et al., López et al.)
  - > only for MP4
  - > need training data

CONCEPT | OVERVIEW

- Source: last producer, e.g. a configured device, social media network

- Extract structure
- Generate feature set
- Convert to Similarity Digest (SD)
- Compare SDs, get similarity in the range $[0 - 1]$,

CONCEPT | STRUCTURE EXTRACTION (JPEG CASE STUDY)

- MSSH is format dependent, here: JPEG
- Top-level JPEG structure:
  - > composed of multiple segments
  - > specification allows variation
  - > differences in segment order and amount
  - > capturing modes can influence structure
  - > RST markers cycle
  - > application segments are prone to complete deletion

Huawei		Samsung			Apple	
P20 pro	P20 lite	S9		S9+	iPhone 11	
std. & bokeh	*std. & bokeh*	*std.*	*bokeh*	*std. & bokeh*	*std.*	*bokeh*
SOI	SOI	SOI	SOI	SOI	SOI	SOI
APP1	APP1	APP1	APP1	APP1	APP1	APP0
APP0	APP0	APP4	APP5	APP4	APP2	APP1
DQT	DQT	APP5	APP4	APP5	APP10	APP1
DQT	DQT	APP11	SOF0	APP11	DQT	APP2
SOF0	SOF0	DHT	DQT	DHT	DRI	APP2
DHT	DHT	DQT	DHT	DQT	SOF0	APP10
DHT	DHT	SOF0	DRI	SOF0	DHT	DQT
DHT	DHT	SOS	SOS	SOS	SOS	DRI
DHT	DHT	EOI	RST0	EOI	RST0	SOF0
DRI	SOS	*slack*	RST1	*slack*	RST1	DHT
SOS	EOI		RST2		RST2	SOS
EOI	*slack*		RST3		RST3	RST0
slack			EOI		RST4	RST1
			slack		RST5	RST2
					RST6	RST3
					RST7	RST4
					[repeat...]	RST5
					EOI	RST6
					slack	RST7
						[repeat...]
						EOI
						slack

CONCEPT | APPx STRUCTURE EXTRACTION

- APP1 stores Exif metadata:
  - > known to be dependent on the source
  - > known to be easily changed and deleted by users

    → Consider additionally and regardfully
- Similar, APPx segments, but…
  - > likely more telling (and undeleted), especially exotic ones
  - > extraction needs high effort, b/c not publicly specified

    → Future Work

Huawei		Samsung		Apple
P20 pro	P20 lite	S9	S9+	iPhone 11
0100	0100	0100	0112	010F
0101	0101	0101	0213	0110
0102	0102	0112	011A	0112
010F	010F	0213	011B	011A
0110	0110	011A	0128	011B
0112	0112	011B	010F	0128
011A	011A	0128	0110	0131
011B	011B	010F	0131	0132
0128	0128	0110	0132	0213
0131	0131	0131	8769	8769
0132	0132	0132		0006
0213	0213	8769		
8769	8769			
8825	8825			
A40B	A40B			
0180	0180			

CONCEPT | FEATURE SET GENERATION

- Similarity Hashing: features must allow a binary decision -

- Order & amount of structural elements must be preserved
- Problems: RST markers, deletion of APPx, Exif tampering, capturing modes
- Similar problems in natural language processing → n-grams

```
F = {D8E1, E1E0, E0DB, DBDB, DBC0, C0C4, C4C4,
     C4DD, DDDA, DAD9, D900, 01000101, 01010102,
     0102010F, 010F0110, 01100112, 0112011A,
     011A011B, 011B0128, 01280131, 01310132,
     01320213, 02138769, 87698825, 8825A40B,
     A40B0180}
```

- (Optional) source SD: aggregate features across diverse image base, adapted to source of interest (e.g. Social Media, Brand)

CONCEPT | DIGEST & SIMILARITY CALCULATION

- Similarity Digest: concatenate n-grams
- Comparison based on Feature Sets:
  - > set comparison with symmetric Jaccard
  - > but source SD represents „all“ possible features → asymmetric Tversky? (=weighted Jaccard)
  - > Hypothesis: Tversky Index performs better for source-level comparison
- Similarity of 1.0 does not mean identical sources! For Tversky not even identical structure.

EVALUATION | SETUP & N-GRAM SIZE

- 7 public forensic datasets, 87k+ images, 189 models, 287 devices, 5 social media platforms
- Image types: flat, natural, bokeh, HDR
- Pre-evaluation for n-gram size:
  - > 2-grams sufficient; 3-grams longer but not better at all
  - > Unique 2-gram SDs: Devices 49.5%, Models 69.3%, Brands & Social media 100%
- AUC-ROC evaluation at device, model, brand, and social-media levels

EVALUATION | DEVICE & MODEL

- Jaccard S_J vs. Tversky S_T
- Both yield AUC > 0.9 across datasets, differences are mostly small
- Robust even when metadata removed

JPEG & APP1			JPEG	
S_J	S_T		S_J	S_T
DEVICE		**Data Set**	**DEVICE**	
0.9906	0.9918	**ALL**	0.9745	0.9684
0.9846	0.9849	**FloreView**	0.9590	0.9546
0.9737	0.9719	**IMAGINE**	0.9457	0.9386
0.9846	0.9838	**FODB**	0.9827	0.9729
0.9173	0.9426	**PrnuMD**	0.8976	0.9263
0.9762	0.9837	**HDR**	0.9693	0.9688
0.9682	0.9642	**VISION**	0.9246	0.9187
0.9687	0.9914	**DIDB**	0.9237	0.9037
MODEL		**Data Set**	**MODEL**	
0.9889	0.9906	**ALL**	0.9739	0.9689
0.9726	0.9720	**FloreView**	0.9576	0.9516
0.9766	0.9753	**IMAGINE**	0.9482	0.9421
0.9873	0.9864	**FODB**	0.9859	0.9755
0.9365	0.9562	**PrnuMD**	0.9239	0.9420
0.9781	0.9862	**HDR**	0.9723	0.9689
0.9736	0.9694	**VISION**	0.9410	0.9327
0.9860	0.9859	**DIDB**	0.9381	0.9236

EVALUATION | BRAND & SM

- Jaccard S_J vs. Tversky S_T
- Here: Tversky clearly superior, results support hypothesis
- Robust even when metadata removed

JPEG & APP1			JPEG	
S_J	S_T		S_J	S_T
BRAND		Data Set	BRAND	
0.7595	0.9893	ALL	0.7593	0.9306
0.9652	0.9935	FloreView	0.9023	0.9447
0.9386	0.9986	IMAGINE	0.8482	0.9607
0.8998	0.9868	FODB	0.8347	0.9615
0.9920	0.9943	PrnuMD	0.9543	0.9773
0.9616	0.9983	HDR	0.8783	0.9486
0.9365	0.9956	VISION	0.8485	0.9457
0.9656	0.9866	DIDB	0.9255	0.9073
SOCIAL MEDIA		Data Set	SOCIAL	MEDIA
		FODB	0.7816	0.8505
		VISION	1.0000	1.0000

DISCUSSION | COMPARISON WITH SPN METHODS

- Comparing apples with oranges (but...)
  - > similar use case in practice for performance oriented SCI methods
- MSSH performance similar (but...)
  - > ALL datasets not suited to test device level discrimination, also applies to SPN approaches
  - > datasets are designed for SPN approach & unrealistic

	IMAGINE	VISION	DIDB subset	DIDB
Reported by Bernacki [4]				
Bernacki [4]	0.94	-	0.93	-
Valsesia et al. [39]	0.78	-	0.81	-
Li et al. [25]	0.84	-	0.84	-
Lukas et al. [27]	0.89	-	0.90	-
Bondi et al. [6]	0.89	-	0.88	-
Tuama et al. [37]	0.88	-	0.86	-
Mandelli et al. [28]	0.83	-	0.87	-
Kirchner and Johnson [22]	0.74	-	0.75	-
Reported by Shullani et al. [35]				
Goljan et al. [13]	-	0.99	-	-
Own Experiments				
proposed	0.97	0.97	-	0.99

DISCUSSION & FUTURE WORK

- syntactical approximate hashing is pre-processing:
  > TLSH, ssdeep, sdhash report AUC-ROC in the range of $0.65 - 0.98$!
- Low n/o unique SDs for devices & models questions very good results:

  → *Metrics Matter - Source Camera Forensics for Large-Scale Investigations. Digital Threats: Research and Practice (2025).*
- Real world performance unclear → test under (more) realistic conditions
- include more structural cues, to improve uniqueness of SD
- extend to HEIC and MP4 formats

CONCLUSION

- First syntactic similarity hash:
  - > easily integretable in common workflows
  - > no training dataset needed
- Closes the gap between Exif analysis and Source Camera Identification
- Lightweight and robust to common metadata deletion
- Real world performance (still) unclear

THANK YOU FOR YOUR ATTENTION!

QUESTIONS, PLEASE.

Samantha Klier
Research Institute CODE
University of the Bundeswehr Munich

Samantha.Klier@unibw.de
https://www.unibw.de/digfor

Research Institute
Cyber Defence
Universität der Bundeswehr München