A Concept for a Comprehensive Understanding of Communication in Mobile Forensics

Jian Xi*[†], Michael Spranger[†] and Dirk Labudde[†][‡]

[†]University of Applied Sciences Mittweida
Forensic Science Investigation Lab (FoSIL), Germany
Email: {xi, spranger}@hs-mittweida.de
[‡]Fraunhofer
Cyber Security
Darmstadt, Germany
Email: labudde@hs-mittweida.de

***Abstract*—Nowadays, mobile devices play a crucial role in our daily life. In practice, criminals also use mobile devices to communicate. Therefore, they have been becoming an important resource for evidence for law enforcement agencies. Especially, the communication between criminals may provide information that could be important for a criminal investigation. Furthermore, the extensive use of mobile devices every day leads to a huge amount of data. Often, it would take too much time to analyze and sort through all the data manually and in some cases it is not even possible. Additionally, investigators are faced with a heterogeneity in the data. Not only are different messengers used to communicate, yet communication is no longer restricted to textual communication and might also include videos or pictures. For this reason, this paper proposes a novel concept that takes different types of media and communication channels in a joint semantic analysis of the content into account. Additionally, a communication network can be derived in terms of topics discussed between users that communicated via smartphone.**

***Index Terms*—Semantic Analysis; Mobile Forensics; Multimodal Machine Learning.**

I. Introduction

In recent years, the emergence of mobile devices changed our life completely. It became the most essential communication medium for acquiring and exchanging information in our daily life. However, it also enables criminals to commit crimes in a very effective manner. Especially, in a well-organized criminal offense, various mobile devices are used for different purposes like locating the targeting places and the victims, organizing the actions or even taking photos for confirming the activities after a crime was committed. Usually, mobile communication is neither limited to one specific medium or communication channel, e.g., email, social networks like Facebook, Telegram, WhatsApp etc. nor to a single data modality like text, image, audio and video etc. Consequently, it inevitably leads to not only isolated and segmented information but also to heterogeneity in the data. In order to support the investigators to understand such communication data in an investigation process, we propose a concept for a joint semantic analysis, which provides comprehensive understanding of communication data for reconstructing an overall view of the data.

The paper is organized as follows: in Section II, we conduct a brief review of related work of semantic analysis in the forensic field. Then we explain the proposed concept of the joint semantic analysis in Section III. Finally, a short discussion is given and some outlooks of future work are discussed in Section IV.

II. Related Work

Most of the existing work in the area of forensic analysis handles the single modalities separately for each case. As shown in [1], Machine Learning-based approaches are reported to detect sexual predatory chats in online text conversations. Focusing on crime scene investigation, [2] presented an Convolution Network-based approach that utilizes feature engineering to improve the image retrieval performance. Similarly, based on feature engineering, video data is examined for detecting illicit content, e.g., pornographic material [3], where periodic patterns and salient regions are respectively analyzed at first in audio-frames and visual-frames. Subsequently, the multi-modal co-occurrence semantics is described by a multi-model fusion approach. Recently, deep learning approaches have also been considered in the forensic filed e.g., detecting drug dealing via social media [4] and detecting video manipulation [5].

As shown the existing methods are not able to jointly process the data in multi-modality. Yet, the semantic context of a mobile conversation is embodied coherently by the data in diverse modalities [6] [7]. In addition, all the data can be transferred via different channels, which in fact inevitably lead to segmented information in data understanding. Furthermore, the amount of mobile messages grows tremendously. Analyzing such a big amount of heterogeneous data manually is overwhelming. As so far, no system reported analyzes all the data modalities in a joint manner in the forensic field. Yet a critical question needs to be answered in such case:

how to understand data consistently whose semantic content is represented by different modalities?

III. Concept of Joint Semantic Analysis

In order to address the aforementioned issues, a feasible and stable solution for a joint semantic analysis in mobile forensics is proposed, as shown conceptually in Figure 1. By means of this joint semantic analysis, the communication data can be analyzed jointly and consistently in an investigation process.

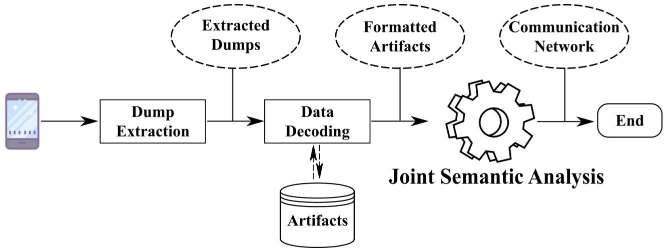

Figure 1. Illustration of the concept for the joint semantic analysis of communication on mobile devices.

The different components are as follows:
Dump Extraction: Extracting data from targeted suspicious mobile devices is a critical step in forensic analysis since there are various different operating systems, hardware, and software. Furthermore, criminals often delete files, which contain information that may be used against them in a criminal investigation. The dump should be extracted in compliance with the chain of custody.
Data Decoding: The foundation of jointly analyzing communication covering all communication channels is a common understanding of the data. For this purpose, extracted dumps need to be restructured in a pre-defined artifact format that includes the necessary information like communication channel, whether a message was send or received and deleted as well, the information about sender, receiver as well caller. The storage information is also necessary for multimedia data. Note that the artifacts can be extracted from multiple dumps (devices).
Joint Semantic Analysis: Aiming at explaining the coherent semantic content and hidden connections in a mobile communication consistently, we formally formulate the joint semantic analysis as follows:

$$\tilde{e} = \text{argmax}_{\boldsymbol{\theta}}\, \tilde{P}(e|d_{cm};\boldsymbol{\theta}) \qquad (1)$$

where e is the semantic context in the conversation data D, which is mostly presented by a topic and possibly connected to a concrete crime, $d_{cm} \in D$ stands for a single artifact message spread via the communication channel $c \in D_c$ $\{$WhatsApp, Telegram, Facebook Messenge, email etc.$\}$ and represented in the modality $m \in D_m$ $\{$Text, Image, Audio, Video etc.$\}$. D is time- and semantically-coherent and organized chronologically. $\boldsymbol{\theta}$ is the parameter set that captures the latent semantic topics in the data and it could be inferred in the topic modeling.

The critical work at this step focuses on finding an intermodal relation that implies a semantic concept between different modalities and channels. For this reason, the individual elements of the respective modalities need to be derived first. Meanwhile, these semantic elements need to be searchable in investigation. Subsequently, the semantic connection (intermodal correspondence) can be determined by considering the whole context in communication. For this purpose, we need at first to map the content of all multimedia data into a textual semantic space to extract semantic topics. For image data, the traditional classification approach [8] or image captioning [9] can be used, where the former delivers only discrete labels like *people* or *car*, etc., while the latter describes the coherent information of image as a whole scene with a natural sentence, e.g., *a man is holding a gun in a bank*. The performance of semantic image captioning can be evaluated by standard evaluation approaches [10]. Instead of merely focusing on describing semantic content of an image, the semantic interpretations and the relations between image and text can be determined as shown in [11]. Meanwhile, a scene graph is planed to be extracted in order to determine how a scene graph contributes to understanding a conversation [12]. Similar, a video can also be translated to a textual representation, i.e., a natural sentence with respect to the content [13]. The audio data can be transcribed into text form by means of Automatic Speech Recognition (ASR) [14]. Note that based on the proposed approaches, the multimedia data is semantically represented in textual form, which can be used for retrieving the forensic information, as well as extracting the coherent semantic topics of the data by using Latent Dirichlet Allocation (LDA) [15]. After topic modeling, each artifact will get a label that has the highest probability with respect to extracted topics. The semantic meaning of this label can be explained by the most important features selected according to the posterior probability of the extracted topics. Finally, a communication in given suspicious data can be represented by these extracted semantic topics. This semantic representation can be used as evidentiary information for clarifying the forensic facts and avoiding misinterpretations of the communication.

IV. Conclusion and Future Work

In this paper, we proposed a concept for a joint semantic analysis in mobile forensics that aims to support investigators when examining the content of an entire communication by taking the multi-modality, as well as multi-channels into account simultaneously. As a result, the investigators are able to capture the overall information of the data in terms of the semantic concepts, which could be related to specific cases. This semantic connection in data is a key information that helps investigators to completely reconstruct the whole criminal scenario. Meanwhile, multiple devices can also be analyzed jointly in this process. In future work, we need to integrate some case-related words, in other words *prior knowledge* from investigators in this pipeline. Furthermore, an alias matching strategy needs to be developed for matching the people who have different names in different communication channels as well as devices.

ACKNOWLEDGMENT

This paper has been funded by the EU-project FORMOBILE. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 832800. Any dissemination of results must indicate that it reflects only the author's view and that the Agency is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] C. Ngejane, J. Eloff, T. Sefara, and V. Marivate, "Digital forensics supported by machine learning for the detection of online sexual predatory chats," *Forensic Science International: Digital Investigation*, vol. 36, p. 301109, 2021.

[2] Y. Liu, Y. Peng, D. Hu, D. Li, K.-P. Lim, and N. Ling, "Image retrieval using cnn and low-level feature fusion for crime scene investigation image database," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1208–1214.

[3] Y. Liu, X. Gu, L. Huang, J. Ouyang, M. Liao, and L. Wu, "Analyzing periodicity and saliency for adult video detection," *Multimedia Tools and Applications*, vol. 79, 02 2020.

[4] J. Li, Q. Xu, N. Shah, and T. Mackey, "A machine learning approach for the detection and characterization of illicit drug dealers on instagram: Model evaluation study," *Journal of Medical Internet Research*, vol. 21, p. e13803, 06 2019.

[5] A. L. Sandoval Orozco, C. Quinto Huamàn, D. Povedano Àlvarez, and L. J. Garcìa Villalba, "A machine learning forensics technique to detect post-processing in digital videos," *Future Generation Computer Systems*, vol. 111, pp. 199–212, 2020.

[6] H.-J. Bucher, "Multimodal understanding or reception as interaction theoretical and empirical foundations of a systematic analysis of multimodality," *Visual linguistics. Theory-method case studies (Bildlinguistik. Theorien-Methoden-Fallbeispiele)*, pp. 123–156, January 2011.

[7] J. R. Hobbs, "Why is discourse coherent?" *SRI International*, November 1978.

[8] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," *CoRR*, 2021.

[9] A. Karpathy and F.-F. Li, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, June 2015, pp. 3128–3137.

[10] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *ECCV*, 2016.

[11] C. Otto, M. Springstein, A. Anand, and R. Ewerth, "Understanding, categorizing and predicting semantic image-text relations," in *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, 2019, pp. 168–176.

[12] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3097–3106.

[13] V. Iashin and E. Rahtu, "Multi-modal dense video captioning," *arXiv e-prints*, Mar. 2020.

[14] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, pp. 993–1022, March 2003.