

DATA NOTE

Open Access



ITC-MNP: a diverse dataset for image file fragment classification

Behnam Tavassoli¹, Zhino Naghshbandi¹ and Mehdi Teimouri^{1*}

Abstract

Objectives Image file fragment classification is a critical area of study in digital forensics. However, many publicly available datasets in this field are derived from a single source, often lacking consideration of the diversity in image settings and content. To demonstrate the effectiveness of a given methodology, it is essential to evaluate it using datasets that are sampled from varied data sources. Therefore, providing a sufficiently diverse dataset is crucial to enable a realistic assessment of any proposed method.

Data description The dataset includes image file fragments of 4096 bytes from five formats (JPG, BMP, GIF, PNG, and TIFF), each processed with different conversion settings. The source images are categorized into three content types: Nature, People, and Medical. In total, the dataset contains 501,000 fragments. These fragments consist of file headers and incomplete end-of-file fragments, completed with random bytes to approximate how operating systems handle data when file sizes are not multiples of the sector size. This approach aims to simulate typical scenarios where fragments are recovered from a hard drive, though it may not capture all real-world complexities such as data corruption and complex file structures.

Keywords File fragment classification, File type identification, Image file fragment, Dataset

Objective

Hard drives store data in fragments, with the size of each fragment determined by the file system. Fragments from a single image can be scattered across the hard drive. In digital forensics, for instance, recovering image data from a damaged hard drive, even partially, is often necessary. This challenge is one of the reasons researchers study image file fragment classification [1]. Recent studies in this field frequently use machine learning to classify file fragments [2–4]. A diverse dataset is crucial for effectively training and evaluating these models. Our dataset

aims to include a wide range of image file fragments, varying in both the content and conversion settings of the source images.

Most available datasets in this area tend not to emphasize the source images from which fragments are collected. This raises several questions: Is it beneficial to consider the content of the source files? Could it influence performance if the dataset includes source files with varying settings, such as compression? And how significant might the impact of these factors be on the performance of classifiers?

Our dataset includes fragments from the five most common image file formats, which are also present in the FFT-75 [5] and Fakouri and Teimouri's [6] datasets. This dataset is expected to help partially address the aforementioned questions and offer useful insights to experts in the field. It also facilitates the evaluation and

*Correspondence:

Mehdi Teimouri
mehditeimouri@ut.ac.ir

¹Information Theory and Coding (ITC) Laboratory, University of Tehran, Tehran, Iran



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

comparison of model performance by providing a more diverse and carefully curated dataset. It will also support the evaluation and comparison of model performance by offering a more diverse and carefully curated dataset. The results section of our supporting document presents findings that illustrate the potential effectiveness of this dataset.

Data description

In the first step, we downloaded source images from the internet and categorized them into three content groups: Nature, People, and Health. A total of 500 images were randomly selected from each category. We aimed to obtain high-quality images within each group. Detailed information about the source images is provided in the supporting document. The subsequent steps of our work are as follows:

Choosing settings Each of the five formats (JPEG, GIF, BMP, PNG, and TIFF) required specific settings to be considered. For instance, since TIFF is a container format, multiple encoders could be used to encode images within it. We applied similar considerations to the other formats, focusing on settings that impact the entire file after conversion. For example, varying levels of compression alter the internal bytes of a file, making the resulting fragments distinct. All format-specific settings are detailed in the supporting document.

Data augmentation After examining the settings for each format, we observed an imbalance in our converted data. The JPEG format, in particular, had significantly more settings compared to the other four formats, requir-

ing each file to be converted 78 times to ensure that each source image was processed with a diverse set of 10 different setting combinations. To balance the dataset, we applied horizontal flipping and rotations of 0, 90, 180, and 270 degrees to the remaining four formats for each content category.

Conversion Given the large number of source images and setting combinations, we utilized Python and the Pillow library to convert the images to the desired formats. Due to limitations in converting raw image formats, Nature images were first converted to full-quality JPEGs before being further processed into the desired format/settings using the Rawpy library. To cover a broad range of JPEG settings, each image was converted multiple times with random combinations of settings. Details of these settings are listed in the supporting document, which is available in Data file 16, Table 1.

Taking fragments Unlike many other file fragment datasets, such as those used by Fakouri and Teimouri [6], we retained all fragments from the beginning and end of the files. In cases where files did not contain enough fragments, we employed sampling with replacement. Additionally, when an end-of-file fragment was not long enough, the remaining portion was filled with random bytes to simulate a typical scenario where fragments are recovered from a hard drive. More information on the fragments can be found in the supporting document.

Provided dataset The dataset includes 15 data files in .dat format, with each file representing a unique format/content pair.

Table 1 Overview of data file/files

Labels	Name of data file	File extension	Data repository
Data file 1	Nature_JPEG	Generic Binary Data (.dat)	Mendeley data [7] https://doi.org/10.17632/r9b362tthp.1
Data file 2	Nature_PNG	Generic Binary Data (.dat)	Mendeley data [7] https://doi.org/10.17632/r9b362tthp.1
Data file 3	Nature_GIF	Generic Binary Data (.dat)	Mendeley data [7] https://doi.org/10.17632/r9b362tthp.1
Data file 4	Nature_TIFF	Generic Binary Data (.dat)	Mendeley data [7] https://doi.org/10.17632/r9b362tthp.1
Data file 5	Nature_BMP	Generic Binary Data (.dat)	Mendeley data [7] https://doi.org/10.17632/r9b362tthp.1
Data file 6	Medical_JPEG	Generic Binary Data (.dat)	Mendeley data [7] https://doi.org/10.17632/r9b362tthp.1
Data file 7	Medical_PNG	Generic Binary Data (.dat)	Mendeley data [7] https://doi.org/10.17632/r9b362tthp.1
Data file 8	Medical_GIF	Generic Binary Data (.dat)	Mendeley data [7] https://doi.org/10.17632/r9b362tthp.1
Data file 9	Medical_TIFF	Generic Binary Data (.dat)	Mendeley data [7] https://doi.org/10.17632/r9b362tthp.1
Data file 10	Medical_BMP	Generic Binary Data (.dat)	Mendeley data [7] https://doi.org/10.17632/r9b362tthp.1
Data file 11	People_JPEG	Generic Binary Data (.dat)	Mendeley data [7] https://doi.org/10.17632/r9b362tthp.1
Data file 12	People_PNG	Generic Binary Data (.dat)	Mendeley data [7] https://doi.org/10.17632/r9b362tthp.1
Data file 13	People_GIF	Generic Binary Data (.dat)	Mendeley data [7] https://doi.org/10.17632/r9b362tthp.1
Data file 14	People_TIFF	Generic Binary Data (.dat)	Mendeley data [7] https://doi.org/10.17632/r9b362tthp.1
Data file 15	People_BMP	Generic Binary Data (.dat)	Mendeley data [7] https://doi.org/10.17632/r9b362tthp.1
Data file 16	Supporting_doc	Portable Document Format (.pdf)	Mendeley data [7] https://doi.org/10.17632/r9b362tthp.1
Data file 17	Conversion	Python Script (.py)	Mendeley data [7] https://doi.org/10.17632/r9b362tthp.1
Data file 18	Frag_extractor	MATLAB Script (.m)	Mendeley data [7] https://doi.org/10.17632/r9b362tthp.1

In addition to the data files, the supporting materials include results from machine learning models trained on the dataset (refer to the supporting document). The MATLAB source code for converting the fragment data files to CSV format, as well as the Python source code used in the conversion process, are also provided.

All of the mentioned files are available online as listed in Table 1.

Limitations

Formats Our study focused exclusively on the five most common image formats (JPEG, BMP, GIF, PNG, and TIFF). While these formats are widely used, this selection may not represent the full spectrum of image file formats encountered in real-world scenarios. As a result, the findings may not be generalizable to less common formats that could exhibit different characteristics or behaviors in fragmentation and recovery.

Content We chose only three categories of image content—Nature, People, and Health—based on their distinctiveness. This limited selection may overlook other important content types that could provide further insights into the performance of classifiers. Consequently, our results may not fully capture the diversity of image content encountered in practical applications of file fragment recovery.

Settings The use of the Pillow library for image conversion introduced certain limitations, as not all possible effective settings may have been available for our dataset. This constraint might affect the diversity of fragments generated, potentially limiting the dataset's robustness for training and evaluating machine learning models. Further research could explore additional settings or conversion libraries to enhance the dataset's comprehensiveness.

Abbreviations

JPEG	Joint photographic experts group
BMP	Bitmap
GIF	Graphics interchange format
PNG	Portable network graphic
TIFF	Tagged image file format

Acknowledgements

We would like to thank Alireza Chale Chale for his kind assistance.

Author contributions

MT designed, reviewed the manuscript, and supervised the study; ZN Found and downloaded suitable images; BT wrote the original manuscript, converted images, extracted fragments, and coordinated the teamwork. All authors have read and agreed to the published version of the manuscript.

Funding

The authors declare no source of funding.

Data availability

The data described in this Data note can be freely and openly accessed on Mendeley Data under the name ITC-MNP: A diverse dataset for image file fragment classification at <https://doi.org/10.17632/r9b362tthp.1>.

Declarations

Ethics approval and consent to participate

No human subjects were involved in this study, and therefore, permission was not required according to the Institutional Review Board guidelines of the first author.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 28 September 2024 / Accepted: 9 December 2024

Published online: 19 December 2024

References

1. Beebe NL, Maddox LA, Liu L, Sun M. Scedan: using concatenated N-Gram vectors for Improved file and data type classification. *IEEE Trans Inf Forensics Secur.* 2013;8(9):1519–30.
2. Mittal G, Korus P, Memon N, FiTy. Large-scale file fragment type identification using Convolutional neural networks. *IEEE Trans Inf Forensics Secur.* 2021;16:28–41.
3. ByteRCNN. Enhancing File Fragment Type Identification With Recurrent and Convolutional Neural Networks. [cited 2024 Jul 6]. <https://ieeexplore.ieee.org/document/10347203>
4. Khodadadi A, Molaei S, Teimouri M, Zare H. Classification of audio codecs with variable bit-rates using deep-learning methods. *Digit Signal Process.* 2021;110:102952.
5. Korus P. File Fragment Type (FFT) – 75 Dataset. IEEE; 2019 [cited 2024 Jul 6]. <https://ieee-dataport.org/open-access/file-fragment-type-fft-75-dataset>
6. Fakouri R, Teimouri M. OSF. 2019. <https://doi.org/10.17605/OSF.IO/YH3XP>. Image file fragments dataset and code.
7. Tavassoli B, Naghshbandi Z, Teimouri M. ITC-MNP: A diverse dataset for image file fragment classification. 2024. Mendeley Data. <https://doi.org/10.17632/r9b362tthp.1>

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.