ABSTRACT

The increasing accessibility of digital tools makes it easier than ever to forge documents. This thesis presents the development and implementation of a modular analysis pipeline designed to support investigators in the examination and assessment of potentially manipulated documents. The aim is to provide a flexible and explainable foundation that combines various analysis methods and presents the results in a structured and interpretable way.

The pipeline consists of several modules that incorporate both classical computer vision techniques and machine learning methods, including neural networks. Among the implemented methods are feature vector-based pre-sorting to group similar documents, binarization to highlight visual artifacts, visual anomaly detection, and frequency-based analyses. In addition, OCR-based alignment analysis is used to detect irregularities in horizontal text alignment.

For the design and evaluation of the pipeline, several custom datasets were created. Different types of document manipulations were defined and reproduced by copying, overlaying, and replacing content to simulate realistic forgery scenarios.

Evaluation results show that certain types of manipulations, such as imitation forgeries, where new characters are digitally inserted, can be reliably detected using the proposed methods. Other manipulations, such as the repositioning of characters within a document, proved more difficult to identify. Feature vector-based pre-sorting using deep learning models effectively grouped visually similar documents and enabled a level of structured analysis that was previously not feasible.

Future work could focus on further automation and the integration of additional analysis techniques, such as content-based or metadata-based document examination.

ZUSAMMENFASSUNG

Die Digitalisierung ermöglicht immer einfacheren Zugang zu Software, mit der Dokumentenfälschungen einfacher werden. Diese Arbeit beschäftigt ich mit der Entwicklung und Umsetzung einer modular aufgebauten Analysepipeline, die Ermittler bei der Sichtung und Bewertung potenziell gefälschter Dokumente unterstützen soll. Das Ziel hierbei soll eine flexible und erklärbare Grundlage zu schaffen, die verschiedene Analysemethoden kombiniert und die Ergebnisse aufbereitet.

Die Pipeline ist in mehrere Module unterteilt, die sowohl klassische Computer Vision Verfahren als auch Methoden des maschinellen Lernen wie neuronale Netze integrieren. Zu den eingesetzten Verfahren gehören unter anderem die Feature-Vektor basierte Vorsortierung zur Gruppierung ähnlicher Dokumente, die Binarisierung zur Hervorhebung von Artefakten, die visuelle Anomalieerkennung sowie frequenzbasierte Analysen. Darüber hinaus wird eine OCR-gestützte Alignment-Analyse zur Erkennung von Unregelmäßigkeiten in der horizontalen Ausrichtung verwendet.

Zur Konzeption und Evaluierung der Pipeline wurden mehrere selbst erstellte Datensätze verwendet. Es wurden Arten von Fälschungen klassifiziert und diese durch Kopieren, Überlagern und Ersetzen von Inhalten erzeugt.

Bei der Evaluierung der Pipeline zeigte sich, dass bestimmte Manipulationsarten wie sogenannte Imitationsfälschungen, bei denen digital neue Zeichen eingefügt werden, zuverlässig erkennbar sind. Andere Manipulationen wie das Verschieben von Zeichen innerhalb eines Dokuments erwiesen sich hingegen als schwer zu erkennen. Die Vorsortierung durch Feature-Vektoren auf Basis von Deep-Learning-Modellen zur Gruppierung ähnlicher Dokumente war sehr erfolgreich und erlaubt bereits ein strukturiertes Arbeiten, was vorher nicht möglich war. Zukünftige Arbeiten können eine stärkere Automatisierung sowie die Integration weiterer Analyseverfahren beinhalten.