MSc Project Report

AuthDoc: Content based fingerprinting of physical documents for tamper detection

H. K. Hashani Sashiprabha Atapattu

2022

A report submitted as part of the requirements for the degree of

MSc Big Data Analytics at Robert Gordon University,
Aberdeen, Scotland

Abstract

Most of the important documents on a life of a person such as birth certificates, passports, license, school or academic certificates, cheques, bills, marriage certificates and many kinds of other certificates and assertions are paper based physical documents. Those are typically kept in both digital and physical forms in order to promote the availability regardless of the availability of a digital infrastructure. Unlike the digital documents validating the authenticity of a physical document is not straight forward and cost effective. The study mentioned in this thesis proposes several possible methods to validate the authenticity of a physical document through a digitally obtained content-based and robust fingerprints. In particular, the study first proposes an autoencoder to digest an image of a document to a fixed length code which can be used as a fingerprint. After evaluating and identifying the limitations of this method the second method is proposed. The second method utilizes the features learned by the previous autoencoder followed by an image processing pipeline. The two methods have been evaluated in terms of robustness and reliability using recall, precision, and F1-score by comparing with a state-of-the-art technique using a self-made dataset and an existing benchmark dataset. The benchmark technique obtained 70.10% F1-score while autoencoder-based method and hybrid method gained 72.14% and 75.33% respectively. Thus, the experimental results prove that the two proposed methods outperformed the existing method. Furthermore, out of the two proposed methods the image processing and machine learning based hybrid method performs better than autoencoder-based method.