Proposition de stage

Sujet de stage

Apprentissage profond pour l'extraction de caractéristiques de documents – Application à la vérification d'intégrité

Mots clés

Apprentissage profond, extracteurs flous, vérification d'intégrité, traitement de documents

Contexte de l'étude

La situation sanitaire actuelle force la population et les administrations à utiliser des copies numérisées des documents papier. Néanmoins, la large disponibilité d'outils professionnels d'édition d'images, de dispositifs simples de numérisation et l'accessibilité des outils d'impression de haute qualité augmentent le nombre de contrefaçons de documents. Les copies numérisées peuvent être facilement falsifiées à l'aide de certains outils d'édition d'images (comme Photoshop ou Gimp) ou de nouvelles approches basées sur l'utilisation de l'apprentissage profond. De ce fait, il existe un besoin important de solutions efficaces et robustes pour la vérification de l'intégrité des documents imprimés et numérisés par la suite. Il s'agit alors d'extraire la signature d'un document électronique pouvant être utilisée pour la vérification d'intégrité de documents numérisés.

Description du sujet

Lorsqu'un document électronique est imprimé et numérisé plusieurs fois, une image du document légèrement différente – en raison des caractéristiques optiques des dispositifs de capture – est obtenue à chaque fois. Un problème similaire se pose en biométrie. Par exemple, lorsque nous réalisons plusieurs fois la capture d'une même empreinte digitale, il ne sera jamais possible d'obtenir des images parfaitement identiques, même si celles-ci seront proches. La difficulté de la mise au point d'une méthode de vérification d'intégrité de documents imprimés puis numérisés est semblable aux difficultés rencontrées en biométrie : on souhaite relever puis comparer des caractéristiques afin de déduire avec une assez grande certitude que deux jeux de données représentent bien la même chose, malgré la présence de bruits.

Pour prendre en compte la propriété d'unicité des données biométriques, des extracteurs flous robustes aux bruits de capteurs sont utilisés [1]. Contrairement à des données biométriques falsifiées, un document falsifié ne diffère pas significativement de sa version authentique, ce qui rend la vérification d'intégrité plus complexe.

Dans de précédents travaux réalisés dans le cadre de ce projet, un premier système de vérification d'intégrité des documents a été mis en place [1,2]. Les caractéristiques extraites sont basées sur l'analyse des intersections et des bifurcations au sein des caractères alpha-numériques.

Nous sommes à présent intéressés par l'exploration de l'utilisation de l'apprentissage automatique pour l'extraction de caractéristiques floues, en nous basant sur une méthode développée précédemment en biométrie [4].

Les objectifs de ce stage sont :

1. Analyse des bases de données des documents falsifiés existantes.

2. Adaptation des extracteurs flous [3, 5] pour la détection de falsifications minimes – par exemple, la suppression ou l'ajout d'un caractère, et des modifications dues aux processus d'impression et de numérisation.
3. Exploration de l'utilisation des réseaux de neurones pour l'extraction de caractéristiques floues [4].
4. Adaptation de la méthode [4] à la vérification d'intégrité de documents.
5. Comparaison des méthodes existantes [1,2] avec les méthodes développées, basées sur l'apprentissage profond.

Profil recherché

- Le candidat doit suivre actuellement une formation de Master 2 ou dernière année d'école d'ingénieur (Bac+5) en informatique
- Langage : Python
- Outils de programmation pour l'analyse et le traitement d'images : OpenCV et/ou scikit-image (Python)
- Connaissances scientifiques : analyse et traitement d'images, apprentissage automatique (en particulier apprentissage profond), extracteurs flous/hachage perceptuel seront un plus
- Langues : français ou anglais

Durée et lieu de stage

Ce stage de recherche en laboratoire s'inscrit dans le cadre du projet exploratoire FuzzyDoc financé par le GdR ISIS (Information, Signal, Image et ViSion). Il sera encadré par Iuliia Tkachenko (Université Lyon 2, LIRIS, Lyon) et Pauline Puteaux (CNRS, CRIStAL, Lille).

Le financement couvre 5-6 mois de stage, le début souhaité est février-mars 2023. Le stagiaire sera rattaché au LIRIS (Laboratoire d'Informatique en Image et Systèmes d'information) sur le campus de l'Université Lyon 2 à Bron. Des missions ponctuelles au CRIStAL (Centre de Recherche en Informatique, Signal et Automatique de Lille) sont prévues.

Références

[1] P. Puteaux, I. Tkachenko, "Crossing number features: from biometrics to printed character matching", IWCDF@ICDAR 2021, September 2021, Lausanne, Switzerland.

[2] F. Yriarte, P. Puteaux, I. Tkachenko, "A Two-Step Method for Ensuring Printed Document Integrity using Crossing Number Distances", IEEE WIFS 2022, December 2022, online.

[3] K. Nandakumar, A. K. Jain, and S. Pankanti. "Fingerprint-based fuzzy vault : Implementation and performance." IEEE Transactions on Information Forensics and Security, 2(4) :744–757, 2007.

[4] C. Rathgeb, J. Merkle, J. Scholz, B. Tams, V. Nesterowicz, "Deep face fuzzy vault: Implementation and performance", Computers & Security, Volume 113, 2022.

[5] Python implementation of fuzzy extractor https://pypi.org/project/fuzzy-extractor/

Contact

Courriel : iuliia.tkachenko@liris.cnrs.fr et pauline.puteaux@cnrs.fr

Merci de fournir un CV, une lettre de motivation, les relevés de notes des deux années de Master.