STAGE DE MASTER

Extraction automatique de textes incrustés dans des vidéos à l'aide d'un OCR
– Développement d'un système et étude de robustesse sur des vidéos de cours magistraux –

La Reconnaissance Optique de Caractères (Optical Character Recognition ou OCR en Anglais) se définit comme l'ensemble des techniques informatiques et électroniques permettant d'extraire les textes incrustés dans un document numérique (photo, vidéo, document scanné etc.). L'OCR est un champ de recherche du domaine de l'apprentissage automatique, de l'intelligence artificielle et plus particulière de la vision par ordinateur. Dans un grand nombre d'application, l'OCR est combiné à d'autres technologies, comme la synthèse de la parole, dans le but par exemple de permettre aux personnes non-voyantes d'accéder aux informations écrites incrustées dans des images ou dans une vidéo.

Le but de ce stage est de développer une solution complète prenant en entrée une vidéo ou une séquence d'images et produisant en sortie un fichier contenant le texte détecté extrait. Si plusieurs technologies open-source comme OCR Tesseract (https://github.com/tesseract-ocr) ou GNU OCR (https://www.gnu.org/software/ocrad/) sont actuellement disponibles, leur robustesse sur des enregistrements vidéo réels reste assez limitée et leurs performances très modestes. La partie analyse de ce stage consistera à mettre en évidence les limitations de ces logiciels pour tenter de les dépasser.

Le stage se compose de 4 étapes :

- Une étude bibliographique et technologique afin d'identifier les logiciels et les méthodes existants et leurs cadres applicatifs courants ;
- Le développement et le déploiement d'une solution complète à l'aide de technologies open-source existantes ;
- L'évaluation de ces technologies sur une variété de vidéos contenant des présentations de cours magistraux dans des domaines variés tels que ceux accessibles sur les sites https://openclassrooms.com/ et https://ocw.mit.edu/index.htm ;
- L'analyse des résultats, la proposition et la mise en œuvre d'améliorations pour élever les performances et ouvrir ce travail à de nouvelles pistes de recherche.

Les livrables du stage consisteront en un rapport (rédigé en Français ou en Anglais comprenant l'étude bibliographique, la justification et la description de la solution implémentée, une étude expérimentale et la discussion des résultats obtenus), et d'un dépôt du logiciel documenté sur un système de gestion de version.

Les compétences attendues sont une connaissance de l'environnement Linux, des langages de programmations python ou java, d'un niveau d'anglais permettant la lecture d'articles scientifiques et de tutoriaux, et globalement d'une bonne communication et capacité à transmettre ses résultats à l'équipe encadrante.

Encadrant Laboratoire	Dr. Christine Sénac (christine.senac@irit.fr) Maître de Conférence, Université Toulouse III Institut de Recherche en Informatique de Toulouse
Encadrant Entreprise	Dr. Benjamin Bigot (bbigot@authot.com) Directeur Recherche et Développement AUTHOT – authot.com