Année 2016-2017

Réseau Thématique de Recherche « Image »
Stage de Master 2

Titre : Identification et comparaison automatique des pages de différents exemplaires de livres anciens numérisés

1 – Noms des équipes proposant le stage

- Laboratoire d'informattique de Tours (EA 6300), équipe RFAI (Jean-Yves Ramel)
- Centre d'études supérieures de la Renaissance, programme « Bibliothèques Virtuelles Humanistes » (Chiara Lastraioli, Rémi Jimenes, Toshinori Uetani)

2 – Noms et adresses courriels des responsables du stage

- Jean-Yves Ramel <jean-yves.ramel@univ-tours.fr>
- Rémi Jimenes <remi.jimenes@univ-tours.fr> / Toshinori Uetani <toshinori.uetani@univ-tours.fr> / Chiara Lastraioli <chiara.lastraioli@univ-tours.fr>

3 – Coordonnées du lieu du stage

Laboratoire d'informatique, Polytech-Tours, 64 avenue Jean Portalis, 37200 Tours

4 – Dates / Durée du stage

5 ou 6 mois entre Février et Septembre 2017

5 – Résumé du contexte et des objectifs du stage

Objectif applicatif: concevoir une «machine à collationner» numérique destinée au livre ancien

L'objectif de ce stage est la production d'un outil informatique destiné à l'étude du livre et des textes anciens, susceptible de rencontrer des applications dans le champ des humanités numériques.

On conserve généralement, dans les bibliothèques, différents exemplaires d'une même édition ancienne (15^{e}-18^{e} siècles). Leur texte est souvent réputé identique, puisqu'ils ont tous été imprimés en même temps, sous une même presse. Or, la comparaison minutieuse des différents exemplaires conservés d'une même édition fait souvent apparaître des variations d'état : des corrections typographiques ont pu être apportées sous presse en cours d'impression, des passages ont pu être censurés, des annotations manuscrites ajoutées, etc. Ainsi deux exemplaires réputés identiques présentent-ils des variantes souvent importantes pour l'histoire du texte et de la réception du livre.

Pour étudier ces variantes, le bibliographe Charlton Hinman avait développé, au milieu du 20^{e} siècle, une « machine à collationner » permettant, par un jeu de miroirs et de lentilles optiques, de projeter sur un même écran les pages de deux exemplaires différents, afin de mieux en faire ressortir, visuellement, les variantes. De tels machines sont rares, fragiles, complexes à mettre en œuvre, et nécessitent, surtout, de réunir dans une même salle deux exemplaires d'une même édition.

La numérisation de corpus massifs de livres anciens dans des bibliothèques du monde entier permet aujourd'hui au chercheur de disposer depuis son domicile des versions numériques de plusieurs exemplaires différents. Il devient donc possible d'envisager la réalisation d'une « machine à collationner » numérique, capable de réaliser automatiquement les tâches suivantes :

- Rapprochement et alignement des images de pages issues d'exemplaires différents afin de pouvoir ensuite comparer plus finement leur contenu
- Suppression du bruit et recalage des images par application de transformations géométriques
- Comparaison page à page et signalement des variantes les plus importantes au travers d'IHMs conviviales

Définition et description des missions en termes scientifiques et informatiques

Ces dernières années, de nouvelles techniques d'analyse et de recherche d'images très performantes ont vu le jour notamment grâce, d'une part à un couplage avec des techniques de détection de points d'intérêt (SIFT, VLAD, …) et de *template matching*, et d'autre part grâce à leur couplage avec des techniques d'apprentissage automatique [2].

L'objectif de ce stage réside dans la mise en place de ce nouveau type d'approches d'apprentissage dans le cadre du recalage et de la comparaison de contenu d'images de documents anciens. Les méthodes proposées seront alors plus adaptatives et performantes car exploitant des capacités de généralisation à partir d'exemples fournis par l'expert. Plus précisément, il s'agira ici de mettre en place une méthode (type CBIR[1]) capable d'apprendre les caractéristiques et métriques adaptées au corpus à traiter.

De plus, des méthodes de recalage habituellement utilisées pour le recalage d'images médicales (IRM) seront également adaptées et exploitées sur des images de types différents.

Une fois les images recalées, des méthodes de *region proposal* [3] et de *template matching* robustes au bruit [1] seront mises en place pour la mise en évidence (détection) des variations entre exemplaires.

Références

[1] S. En, C. Petitjean, S. Nicolas, L. Heutte, and F. Jurie. "Pattern localization in historical document images via template matching". International Conference on Pattern Recognition, 2016, Cancun, Mexico.

[2] ImageNet Classification with Deep Convolutional Neural Networks, Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton, NIPS 2012.

[3] S. En, C. Petitjean, S. Nicolas, F. Jurie, and L. Heutte. "Region proposal for pattern spotting in historical document images". International Conference on Frontiers in Handwriting Recognition, 2016, Shenzhen, China.

6 – Observations

Ce stage s'effectuera au sein du Laboratoire d'informatique de l'Université de Tours afin de réactiver les collaborations fructueuses passées avec le programme « Bibliothèque Virtuelles Humanistes ». Le stagiaire pourra ainsi s'appuyer sur les outils développés dans le cadre d'anciennes collaborations, notamment les logiciels Agora et Rétro. Il sera encadré par une équipe d'informaticiens spécialistes du traitement d'images et suivi étroitement par l'équipe du Centre d'études supérieures de la Renaissance.

Merci de retourner un formulaire par sujet de stage
avant le lundi 18 novembre 2016, à :
bureau.rtrimage@univ-orleans.fr

[1] Content Based Image Retrieval