PROPOSITION DE STAGE

Année 2016

Laboratoire L3i

Sujet de stage :

Titre : Layout analysis (text and non-text segmentation) of born-digital document images

Résumé du travail proposé :

Within the large multimedia data collections available on network media, there are many images that are characterized as having weakly structured content, where text and graphics appear in those images not in the usual paper document layout. Examples of those images are advertisements, micro-blog images and images embedded in web pages. Such images are called born-digital images, and they are difficult to analyze due to their complex layout, low resolution, variations of font type, size and color, mixed graphics and text, multi-languages. Analyzing the contents of such images will help in the development of the next generation of search engines, cyber security applications and commercial data mining.

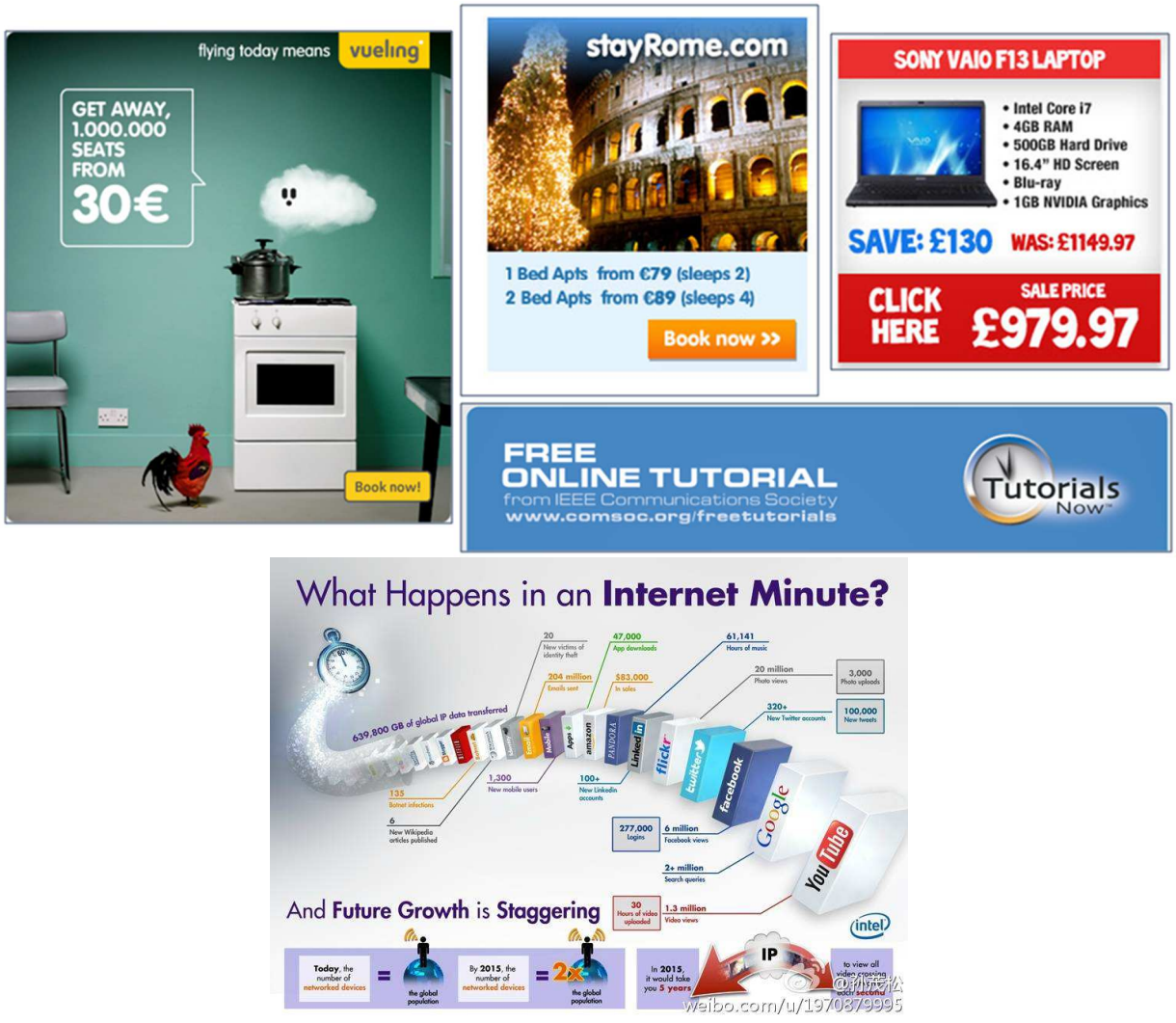

Figure 1: Advertisement images (top row) and images from Weibo micro-blog (bottom row)

This master project focuses on segmenting regions of different content type in born-digital images using image processing and computer vision techniques. In particular, segmenting text regions from non-text (graphical) regions. Once the text regions are localized, the text can be easily recognized and used in interpreting image content. This project deals only with the task of text and non-text region segmentation.

Mots clés :

Born-digital images - text and non-text segmentation – graphical layer separation – complex layout analysis – visual feature extraction – clustering.

Informations complémentaires :

Encadrant(s) : Jean-Marc Ogier, Jean-Christophe Burie, Nibal Nayef

Thématiques :

- ☐ **Ingénierie des connaissances**
- **x Analyse et gestions de contenus**
- ☐**Interactivité et dynamique des systèmes**

Domaine d'application :

- **x Pertinence – contenu – interactions**
- ☐ **Environnement**

Cadre de coopération : Projet ANR AUDINM Franco-Chinois (Académie des Sciences de Pekin)

Date de début du stage: 4 janvier 2015

Durée du stage : 6 months

Financement : 450 Euros per month (for 6 months) + travel and living costs in China if the student wishes to benefit from this.

Contexte de l'étude:

The task of layout analysis of born-digital images is within the AUDINM project (Analysis and Understanding of Document Images in Network Media) [Analyse et Interprétation d'images de documents sur les réseaux sociaux]. AUDINM is funded by ANR. In this project, the laboratory L3i collaborates with the NLPR laboratory in Beijing – China.

The main goal of the project is to analyze scenes images and born-digital images which can be found on the internet. This master project focuses on born-digital images embedded in web-pages and in social media applications, where segmenting text from non-text in those images helps us to search and retrieve web content.

This master project includes the opportunity to spend a fully funded 3-month research visit to NLPR lab in Beijing (3 months at L3i and 3 months at NLPR).

Description du sujet :

The main problem in this project is to develop a method for segmenting born-digital images into regions of homogeneous content, and classify them according to their content type: text or non-text (graphical symbols and shapes, pictures of objects, tables, etc.). The main challenges in this problem is that the images have complex layout, low resolution, variations of font type, size and color, mixed graphics and text and different text alignments.

The work plan for this master project includes the following main steps: study of state-of-the-art in text / graphics layer segmentation and layout analysis, text layer segmentation based on local color uniformity, texture, edge properties and other visual features, segmenting an image into layers using clustering techniques. The master student will implement a system that takes an image as input, and outputs the detected text regions inside the image. See Figure 2 as an example output of the system.

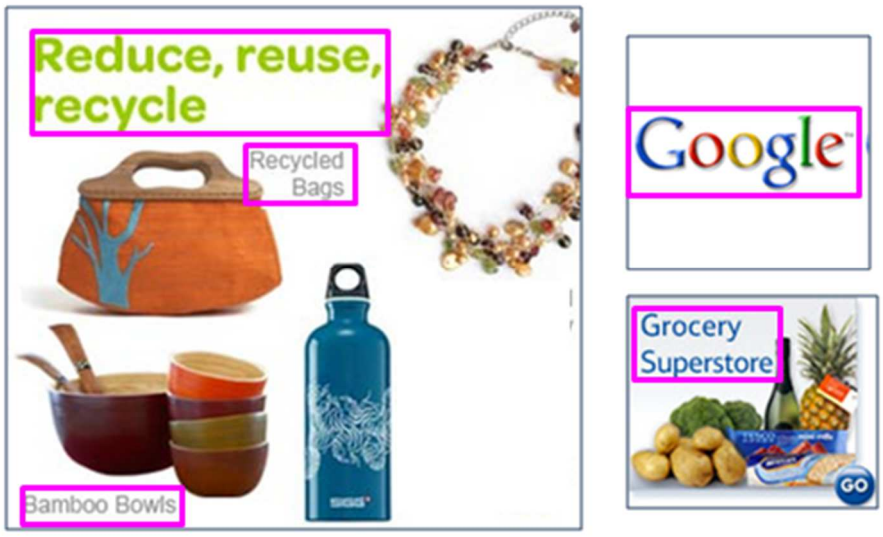

Figure 2: Segmented text regions (text regions marked in pink bounding boxes)

Prérequis et contraintes particulières :

- Basic image processing knowledge (master courses)
- Good programming skills (python or C++ or matlab)
- Good English language skills (specially if you want to spend the research visit in China)

Références bibliographiques :

[1] A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher, «ICDAR2015 Competition on Recognition of Documents with Complex Layouts – RDCL2015», Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR2015), 2015, pp. 1151-1155

[2] MinhHieu Nguyen, Soo-Hyung Kim, and Gueesang Lee, «Recognizing Text in Low Resolution Born-Digital Images» , In Ubiquitous Information Technologies and Applications, volume 280, 2014, pp. 85-92

[3] Viet Phuong Le, Nibal Nayef, Muriel Visani, Jean-Marc Ogier, Cao De Tran, «Text and Non-text Segmentation based on Connected Component Features », ICDAR 2015.

Contacts – liens :

Email :

jean-christophe.burie@univ-lr.fr

nibal.nayef@univ-lr.fr