Interprétation de documents pdf natifs : Application à la presse magazine

Encadrement : Bertrand COÜASNON, Irisa/Intuidoc, bertrand.couasnon@irisa.fr
Aurélie LEMAITRE, Irisa/Intuidoc, aurelie.lemaitre@irisa.fr
Lieu du stage : IRISA – Rennes

Mots-clés : Analyse d'images de documents, traitement de pdf natif, reconnaissance de pages de presse magazine, modélisation de connaissances, grammaire bi-dimensionnelle, expression de règles, interprétation de structures

L'équipe de recherche Intuidoc de l'Irisa (http://www.irisa.fr/intuidoc) travaille notamment sur la reconnaissance du contenu et de la structure de documents anciens, manuscrits ou dégradés (partitions musicales, registres d'archives, journaux, courriers manuscrits, schémas électriques...). Dans ce contexte, les travaux de l'équipe ont abouti à des chaînes de traitement complètes (méthode DMOS-P [2]), conduisant à une reconnaissance de la structure des documents, décrite par des grammaires bidimensionnelles permettant d'exprimer la connaissance visuelle du document et des mécanismes perceptifs.

Nous souhaitons pouvoir travailler sur une nouvelle source de documents : des fichiers pdf natifs, c'est-à-dire des fichiers pdf générés électroniquement, ayant servi de base à une impression de documents. Dans ces fichiers pdf, le contenu électronique est composé d'éléments (caractères, symboles) permettant d'expliciter un rendu visuel du document à imprimer (cf. Figure 1). Cependant, la cohérence logique entre ces symboles (organisation en mots, blocs de textes, tableaux…) est souvent inexistante au sein du fichier pdf. De plus, pour un même rendu visuel, les symboles utilisés sont parfois très variables.

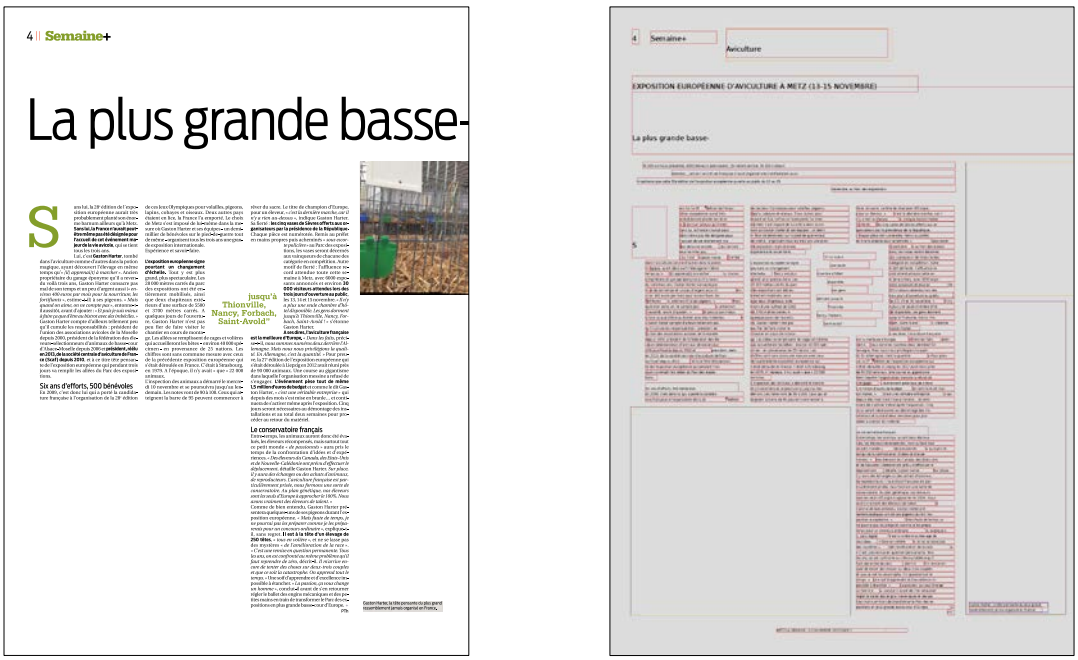

Figure 1 : Exemple de fichier PDF (gauche) et des éléments de bases extraits (droite)

L’objectif du stage de Master est de créer un système de reconnaissance de la structure de documents pdf natifs. Il s’agira de modéliser des mises en page modernes, par l’apprentissage de règles dans un langage de description de structure. Ce travail sera appliqué principalement sur des pdf de presse magazine. En effet, nous avons une forte demande de la part des industriels pour retrouver l’organisation de fichiers pdf. D’un point de vue scientifique, il s’agira de combiner des informations fiables du contenu du fichier avec des connaissances structurelles. On pourra proposer des mécanismes d’interprétation de la structure en s’inspirant de ceux utilisés par la vision humaine.

À la suite de ce stage, un financement de thèse pourra être proposé dans le cadre du projet ANR HBDEX qui démarre en 2018. Une possibilité de thèse CIFRE est également envisagé.

Référence

[1] Aurélie Lemaitre, Jean Camillerapp, Bertrand Coüasnon. **Multiresolution Cooperation Improves Document Structure Recognition**. International Journal on Document Analysis and Recognition (IJDAR), 11(2):97-109, Novembre 2008.

[2] B. Coüasnon. **DMOS: A generic document recognition method, application to an automatic generator of musical scores, mathematical formulae and table structures recognition systems**. In *ICDAR, International Conference on Document Analysis and Recognition*, Seattle, USA, Septembre 2001.

[3] Hu, J. & Liu, Y. **Analysis of Documents Born Digital.** Handbook of Document Image Processing and Recognition, Doermann, D. & Tombre, K. (ed.), Springer London, 2014, 775-804.

[4] Dengel, A. & Shafait, F. **Analysis of the Logical Layout of Documents**. Handbook of Document Image Processing and Recognition, Doermann, D. & Tombre, K. (ed.), Springer London, 2014, 177-222