Combinaison de connaissances structurelles et d'apprentissage profond pour l'analyse d'images de documents complexes

Direction de thèse : Bertrand COÜASNON, Irisa/Intuidoc, bertrand.couasnon@irisa.fr

Co-direction : Aurélie LEMAITRE, Irisa/Intuidoc, aurelie.lemaitre@irisa.fr
Nathalie GIRARD, Irisa/Intuidoc, nathalie.girard@irisa.fr

Mots-clés : Analyse d'images de documents, reconnaissance de structure de documents, segmentation contextuelle, apprentissage profond, analyse itérative, modélisation de connaissances, grammaire bidimensionnelle, extraction d'information, documents multilingues, documents hétérogènes

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 [1]), conduisant à une reconnaissance de la structure des documents, décrite par l'introduction de grammaires bidimensionnelles permettant d'exprimer la connaissance visuelle sur les documents et des mécanismes perceptifs, qui reposent sur la reconnaissance d'éléments de base par l'intermédiaire de reconnaisseurs de symboles et d'écritures, l'ensemble étant validés sur plus de 700 000 documents et de nombreux types de documents différents. Ces travaux ont été évalués lors de différentes compétions internationales : RIMES 2007, RIMES 2008, ICDAR 2009, ICFHR 2010, ICDAR 2011, ICDAR 2013, HIP 2013, MAURDOR 2013.

Dans le cadre du projet Maurdor, PEA (Plan d'Étude Amont) financé par la DGA, différents laboratoires et entreprises partenaires, pilotés par Cassidian (Airbus), ont construit des chaînes prototypes de traitement de documents numérisés hétérogènes (tout type, contenant du manuscrit et de l'imprimé) multilingues (français, anglais et arabe), pour effectuer de l'extraction automatique d'information dans le cadre du renseignement militaire.

L'équipe Intuidoc a travaillé sur différentes tâches : reconnaissance de la structure physique de documents hétérogènes (Module 1, décomposition en blocs), séparation entre les blocs manuscrits et imprimés (Module 2) et détection de la structure logique (Module 5). Des partenaires (par exemple [2] sur le Module 1) ont également travaillé sur ces modules, mais également sur les Module 3 (reconnaissance de la langue), Module 4 (reconnaissance d'écriture imprimé et manuscrite, en français, anglais et arabe), Module 6 (extraction d'informations). Sur le Module 1, le système

INSA RENNES
20, Avenue des Buttes de Cöesmes
CS 70839 - 35 708 Rennes Cedex 7
Tél.+ 33 (0)2 23 23 82 00 - Fax + 33 (0)2 23 23 83 96
www.insa-rennes.fr

développé s'appuie sur une description générale des différents éléments pouvant se trouver dans un document hétérogène [3]: des blocs de texte imprimés, manuscrits, en langue latine ou langue arabe, des structures tabulaires, des formulaires, des graphiques, des séparateurs… combinée avec des classifieurs permettant de détecter localement si une zone est imprimée, manuscrite ou graphique.

Les différents modules prototypes développés par les partenaires, ont été évalués lors des deux compétitions internationales Maurdor (http://www.maurdor-campaign.org), sur une base de 10 000 documents dont la vérité terrain a été complètement annotée dans le cadre du projet Maurdor. Cette base de données est la plus avancée et la plus complète sur des documents hétérogènes multilingues au niveau mondial. L'évaluation des différents prototypes, sur cette base réaliste et difficile, a pu montrer que :

- la marge de progression était importante car les documents à traiter sont particulièrement complexes et difficiles de par leur hétérogénéité et leur dégradation ;
- le Module 1 de segmentation physique en blocs était particulièrement important puisque sa qualité impacte directement les capacités de reconnaissance des modules qui lui succèdent ;
- il était nécessaire, pour atteindre un niveau de qualité et de précision plus important, de briser l'enchainement séquentiel de modules indépendants tel que cela a été imposé dans le cadre de Maurdor.

En effet, actuellement, la prise de décision sur la segmentation et l'homogénéité des différents blocs de la structure physique, même si elle utilise des classifieurs spécialisés, reste trop locale au vu de la complexité et de la variabilité des documents à traiter. Il est indispensable d'aller vers une segmentation de la structure physique construite sur une véritable compréhension du document s'appuyant sur la reconnaissance du contenu textuel. Ainsi, en exploitant des reconnaisseurs de lignes manuscrites, imprimées ou mixtes, tels que ceux développés dans le projet Maurdor pour le Module 4 (reconnaissance de l'écriture manuscrite et imprimée), il est possible d'envisager d'intégrer un contexte beaucoup plus large lié au contenu textuel du document pour valider des hypothèses de segmentation. Cette compréhension doit apporter une meilleure qualité et une précision dans la segmentation. De même, des éléments de la structure logique, comme ceux par exemple liés aux intitulés de champs et leurs contenus associés dans les formulaires, doivent être intégrés pour améliorer et fiabiliser la segmentation dans un contexte de documents difficiles.

Afin de lever ce verrou scientifique, le sujet de thèse proposé porte sur l'intégration de connaissances multiples liées au contenu textuel et à la structure logique, en brisant le mécanisme classique, séquentiel et linéaire, de traitement de document et en proposant une analyse par itérations successives intégrant de plus en plus de contenu au fur et à mesure de leur fiabilisation. Ce mécanisme original par rapport à l'état de l'art pourra s'appuyer par exemple sur l'analyse itérative proposée dans DMOS-PI [4] et le système d'apprentissage sans vérité terrain EWO [5], tout en combinant les capacités de segmentation contextuelle des systèmes de reconnaissance à base d'apprentissage profond (réseaux de neurones récurrents à convolutions). Il faudra dans cette thèse proposer un mécanisme d'interaction forte entre les systèmes d'apprentissage profond et les systèmes de reconnaissance syntaxiques.

INSA RENNES
20, Avenue des Buttes de Cöesmes
CS 70839 - 35 708 Rennes Cedex 7
Tél.+ 33 (0)2 23 23 82 00 - Fax + 33 (0)2 23 23 83 96
www.insa-rennes.fr

Les mécanismes proposés dans la thèse seront évalués sur la base internationale Maurdor.

Références

[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] Philippine Barlas, Sébastien Adam, Clément Chatelain, **A Typed and Handwritten Text Block Segmentation System for Heterogeneous and Complex Documents.** *Document Analysis Systems 2014.*

[3] Baptiste Poirriez, Aurélie Lemaitre, Bertrand Coüasnon. **Visual perception of unitary elements for layout analysis of unconstrained documents in heterogeneous databases**. In *14th International Conference on Frontiers in Handwriting Recognition (ICFHR-2014)*, 2014.

[4] Joseph Chazalon, Bertrand Coüasnon, Aurélie Lemaitre. **Iterative Analysis of Pages in Document Collections for Efficient User Interaction**. In *International Conference on Document Analysis and Recognition (ICDAR'2011)*, Pages 503-507, Septembre 2011.

[5] Cérès Carton, Aurélie Lemaitre, Bertrand Coüasnon. **Eyes Wide Open: an interactive learning method for the design of rule-based systems**. In International Journal on Document Analysis and Recognition, Springer Verlag, 2017, 20 (2), pp.91-103.

INSA RENNES
20, Avenue des Buttes de Cöesmes
CS 70839 - 35 708 Rennes Cedex 7
Tél.+ 33 (0)2 23 23 82 00 - Fax + 33 (0)2 23 23 83 96
www.insa-rennes.fr