Fusion des connaissances en analyse de documents

Exemples sur des documents d'archives

Bertrand Coüasnon – *IRISA / INSA Rennes*

RÉSUMÉ. *La reconnaissance de collections de documents structurés numérisés et notamment de documents d'archives est difficile non seulement par la complexité de l'organisation des documents, mais aussi par la dégradation des documents (tâches, déchirures, encre traversant le papier, courbures produites à la numérisation...). Afin d'améliorer la qualité de la reconnaissance tout en gérant le bruit induit par ces dégradations, il est nécessaire d'exploiter le maximum de connaissances dans le processus d'analyse.*

Or, les sources de connaissances pour l'analyse de documents sont multiples. En se centrant sur la notion de page, nous pouvons les décomposer selon trois types : les connaissances a priori sur la page (liées à un type de document), les connaissances internes à la page (présentes dans l'image et qu'il est nécessaire d'extraire), et les connaissances externes à la page (provenant d'autres pages d'une collection de documents ou bien d'utilisateurs interrogés sur le contenu dans un processus interactif).

Nous montrerons comment il est possible de fusionner et d'exploiter ces différents types de connaissances en s'appuyant sur un langage de description de documents, des calques perceptifs, une mémoire visuelle et une analyse itérative. Ces éléments peuvent être ajoutés à un système existant pour lui fournir de nouvelles capacités. Nous avons ainsi pu construire un système générique multirésolution de traitement de collections de documents intégrant des mécanismes de vision perceptive tout en proposant une interaction asynchrone, capable d'amener au niveau de la page, des connaissances externes à la page, provenant d'un utilisateur, d'autres pages ou d'autres traitements. Ce système perceptif engendre des mécanismes d'analyse plus complexes, tout en étant plus simple à définir, et ayant une combinatoire plus faible.

Ces principes ont été validés sur plus de 600 000 documents de types différents, allant de partitions musicales, de formules mathématiques à des documents d'archives. Nous présenterons des résultats sur des registres matricules militaires, des décrets de naturalisation, de la presse ancienne ou des registres de ventes de la Révolution Française. Sur ces derniers documents nous montrerons, à titre d'illustration, comment une analyse interactive asynchrone combinée à des reconnaisseurs d'écriture manuscrite, des mécanismes de word spotting et des utilisateurs, permet de mettre en place une transcription assistée de patronymes manuscrits, dans laquelle l'utilisateur est pratiquement deux fois moins sollicité.

CORIA 2012, pp. 5–5, Bordeaux, 21-23 mars 2012