WP4_D1 : T0 + 12 : : Liste de cas d'usages et de descripteurs

Coordinateur : Jean-Yves Ramel

Introduction

Dans une campagne de numérisation où la qualité est prépondérante, la majeure partie des coûts de production sont des coûts de main-d'œuvre, principalement liés au temps de réglage ou de re-réglage du scanner. Il est même parfois nécessaire de recommencer la phase d'acquisition lorsque les images produites ne peuvent être utilisées par les autres maillons de la chaîne. C'est le cas par exemple lorsque le taux de reconnaissance de l'OCR est trop faible. Il est donc primordial d'essayer d'offrir un processus de numérisation permettant aux scanners de s'auto-adapter aux données, de détecter et éventuellement corriger les défauts de prises de vues.

Nous proposons d'intégrer directement dans les scanners les briques logicielles permettant d'extraire des descripteurs caractérisant l'image acquise en même temps que l'ensemble des pixels eux-mêmes. Ainsi, les scanners ne fourniront plus en sortie uniquement un fichier TIF ou JPEG mais également des métadonnées de niveau supérieur, utilisables pour améliorer et simplifier l'exploitation future. Ces caractéristiques des images seront aussi utilisées pour ajuster semi-automatiquement les paramètres d'acquisition et les paramètres de prétraitement par bouclage de prise de vue.

De plus, puisqu'il est complexe d'établir un protocole de numérisation adéquat pour l'ensemble des pages à numériser, nous prenons le parti d'ajouter au scanner un module d'apprentissage qui, sur la base des caractéristiques extraites, sera en mesure de déterminer quel protocole de numérisation doit être appliqué sur chaque page. Plus précisément, ce module d'apprentissage et de classification incrémental aura pour rôle d'associer à différentes classes d'images un protocole d'acquisition adapté

Ainsi, le processus de numérisation pourra s'auto-adapter aux données, pour chaque ouvrage, chaque page d'un ouvrage, voire même pour chaque région dans une page en fonction des objectifs de l'opérateur. Un tel mécanisme que nous appellerons *numérisation cognitive*, nous semble très pertinent compte tenu de la diversité des documents et des conditions d'acquisition. Les collections de documents à numériser se caractérisent en effet par une forte hétérogénéité tant au niveau de leurs contenus (langue, thème,...), que des techniques d'imprimerie utilisées (insertion d'illustration et de texte, technique de mise en page, qualité des techniques d'imprimerie...), que de leur qualité physique après plusieurs siècle d'archivage (apparition de tâches d'encres, détérioration du papier).

IL nous a paru judicieux de faire apparaitre deux types de cas d'usages des scanners :

- Les cas d'usages utilisateur
- Les cas d'usages fonctionnels

Les deux sections suivantes listent, pour chacun de ces deux cas, les principaux cas d'usages qui ont été prévus pour ces scanners cognitifs.

1/ Cas d'usages Utilisateurs

Il s'agit ici de lister les objectifs que peuvent avoir les utilisateurs lorsqu'ils exploitent un scanner pour numériser des documents, l'objectif étant de permettre à terme le choix des paramètres de numérisation optimaux vis-à-vis de cet objectif.

Usage 1 : Préservation, archivage
Il peut s'agir simplement de garder une trace numérique d'un document permettant plus tard d'accéder au message qu'il véhiculait ou au contraire d'en faire une copie fidèle dans tous ces détails.

Usage 2 : Reproduction, Diffusion, impression, web, DVD
Dans ce cas, il s'agit bien souvent de trouver le meilleurs compromis possible entre fidélité à l'original et mémoire de stockage ou débit nécessaire.

Usage 3 : Enrichissement du document : Transcription, OCR sur texte + indexation des images, GED
L'objectif est ici différent puisque la fidélité à l'original n'est plus forcément requise. L'objectif est plutôt d'améliorer la qualité de l'image, d'éliminer certains défauts afin que le l'image produite soit la plus adaptée possible au traitement qui vont suivre.

Usage 4 : Recherche : épigraphe, paléographe, philologie, codicologie, Histoire des textes, …
Il s'agit là de faire des acquisitions très précises ou voire même avec des paramétrages très spécifiques nécessaires à des traitements et analyses très spécifiques qui vont suivre

L »analyse de ces cas d'usages amène à la conclusion qu'il s'agit bien souvent de trouver le meilleur compromis entre fidélité à l'original et espace mémoire ou temps de calcul associé.
Il est aussi possible que l'utilisateur veuille à la fois archiver et enrichir les documents (combinaison de cas d'usage). Il semble donc inopportuns de baser la conception et le fonctionnement futur des scanners sur le choix à priori d'un de ces cas d'usage lors du lancement d'une session de numérisation.

2/ Cas d'usages fonctionnels

Pour concevoir le scanner cognitif, il nous semble préférable de se baser sur la liste des cas d'usage fonctionnels suivants :

Usage 1 : usage de base
Bien évidemment le principal usage d'un scanner concerne la production d'images (pixels, TIFF) de qualité adaptée aux besoins. Les scanners cognitifs produiront bien évidemment, comme leurs prédécesseurs des fichiers images.
Cependant, en plus des images les scanners cognitifs produiront un fichier au format DIGIDOC correspondant à une session complète de numérisation d'un document ou ouvrage (et pouvant donc correspondre à plusieurs pages)

Usage 2 : Auto-calibration ou calibration assistée (réglages des paramètres)
Dans certains cas, avant la phase de numérisation, il est nécessaire de régler le scanner selon les spécifications fournies par le client. Le scanner cognitif devra fournir une aide à ces réglages.

Respect des niveaux de gris, balance des blancs, précision colorimétrie & résolution, profondeur de champ, focale (flou), défaut d'éclairage, …

Sélection des prétraitements (correctifs) à appliquer

Usage 3 Détection des défauts (classes de défauts)

A l'aide de son module d'apprentissage, le scanner cognitif sera capable d'apprendre et donc de détecter certains défauts de numérisation et ce à 3 niveaux différents :

Niveau Documents : Page manquante, pliée, dérive progressive…
Niveau image : Prise de vue incorrecte : orientation, structure, cadrage, bords, …
Niveau EoC : résolution, contraste, couleur, luminosité, bruits vide, floue,...

Usage 4 : Détection de types de contenus (documents, pages, régions : classes d'EoC)

Afin d'utiliser les bons paramètres d'acquisition, le scanner sera capable d'apprendre et ensuite de reconnaitre certains types de contenus spécifiques. Le protocole correspondant au type reconnu sera alors appliqué pour obtenir une image de qualité optimale. Les types de contenu (EoC) qui pourront être appris seront par exemple :

Zones dactylographiées
Zones manuscrites
Dessins, plans, figures
Tableaux
Photos
Dégradations

L'opérateur pourra définir lui-même les types de contenus à reconnaitre

Usage 5 : Détectons de types de documents (classes de documents)

Au niveau document (multi-pages ou page simple), il est aussi envisageable de prévoir la reconnaissance de différents types de documents :

Lettre,
Articles,
Livres,
Formulaire, …

L'opérateur pourra définir lui-même les types de documents à reconnaitre

3/ Descripteurs retenus

La liste des descripteurs retenus sera spécifiée dans les livrables correspondants au WP2 : Format Digidoc