Détection de tableaux dans des documents complexes

Thotreingam Kasar* — **Philippine Barlas*** — **Sébastien Adam*** — **Clément Chatelain**** — **Thierry Paquet***

* *Laboratoire LITIS - EA 4108, Universite de Rouen, FRANCE 76800*
** *Laboratoire LITIS - EA 4108, INSA Rouen, FRANCE 76800*

RÉSUMÉ. *Dans cet article, nous présentons les résultats obtenus par un détecteur de tableau dans le cadre des campagnes MAURDOR, pour lesquelles le corpus présente la particularité de contenir des document fortement hétérogènes dans leur mise en page, leurs scripts et les langues utilisées.*

ABSTRACT. *This paper presents the results obtained by a table detector during the MAURDOR campaign, the corpus of which contain heterogeneous documents in French, English and Arabic with various types of table structures.*

MOTS-CLÉS : *Détection de lignes, Détection de tableaux, Maurdor*

KEYWORDS: *Line detection, Table Detection Maurdor*

CIFED 2014, pp. 363–366, Nancy, 18-21 mars 2014

1. Introduction

Les tableaux sont des représentations qui permettent de synthétiser efficacement des informations ainsi que les relations qu'elles entretiennent. Ils sont fréquemment utilisés dans de nombreuses classes de document telles que les articles de presse, les articles scientifiques, les formulaires, les factures ou encore les documents financiers. Leur reconnaissance est donc un maillon important dans le cadre d'une chaîne d'analyse d'images de documents. Dans [1], nous avons proposé un système de détection de tableau dans des images de documents. L'originalité de ce système est qu'il repose sur un processus d'apprentissage, ce qui permet au système de s'adapter à la variabilité des configurations de tableaux. Dans cet article court, nous présentons les résultats obtenus par ce système de détection lors des campagnes MAURDOR [2]. L'article est structuré de la façon suivante. La section 2 rappelle les principaux aspects de ce système. Puis, dans la section 3, les résultats expérimentaux obtenus lors des deux campagnes d'évaluation MAURDOR sont décrits.

2. Description du système

La méthode proposée dans [1] repose sur un processus d'apprentissage permettant d'apprendre à détecter les tableaux. Un classifieur, de type SVM, est appris à partir d'un ensemble de documents annotés. Ainsi, la méthode est capable de s'adapter à la variabilité des structures possibles de tableaux, sans avoir à définir de modèle (sous forme de règles, par exemple) des tableaux. Par ailleurs, la méthode ne repose pas sur le texte contenu dans le tableau, ce qui permet qu'elle soit indépendante du type d'écriture (manuscrit, imprimé) et de la mise en page du document. Elle peut ainsi traiter des documents multi-colonnes, sans avoir à adapter quelque seuil que ce soit. Enfin, la méthode ne nécessite pas la reconnaissance du contenu textuel, qui est souvent source d'erreur, et très consommateur de ressources. La méthode s'appuie uniquement sur la présence de lignes horizontales et verticales et c'est un classifieur qui décide si les lignes extraites appartiennent ou non à un tableau, par une analyse de leur structure. La figure 1 donne un aperçu schématique de l'approche.

Dans une première étape, pour renforcer la continuité des lignes fines, l'image d'entrée est d'abord lissée avec un filtre Gaussien, avant d'appliquer une opération morphologique originale de type "black-hat". Après l'application de ces opérations, l'image est ensuite binarisée, avec un seuil adaptatif dont la valeur est déduite du niveau de gris moyen de l'image. La méthode ne requiert pas une binarisation très précise, puisqu'elle repose uniquement sur l'information "ligne" et non sur les éléments textuels. L'image résultant de ce seuillage est alors analysée par *run-length*, horizontalement et verticalement, afin d'aboutir une image contenant l'ensemble des lignes horizontales et verticales de l'image.

L'étape suivante de notre détecteur de tableau commence par une détection des composantes connexes dans l'image de lignes préalablement extraite. Un filtrage est appliqué pour ne conserver que les composantes connexes composées d'un mimum de

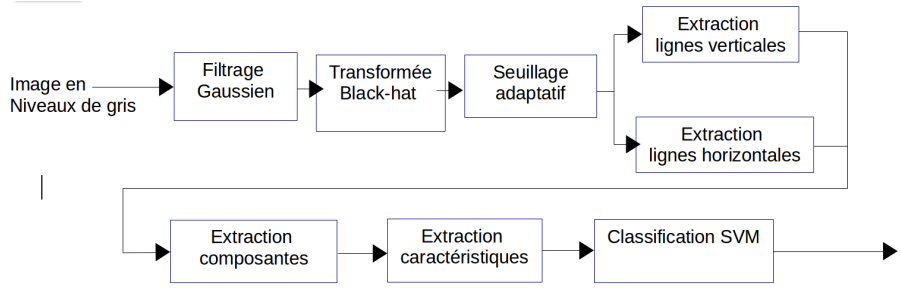

Figure 1. *Approche proposée dans [1] pour la détection de tableau*

trois lignes, susceptibles d'être des tableaux. Pour chacune des composantes connexes (groupes de lignes horizontales et verticales connexes), 26 caractéristiques numériques (décrites dans [1]) sont alors extraites.

A partir des caractéristiques décrites ci-dessus extraites sur un corpus d'apprentissage, il est ensuite possible de réaliser l'apprentissage d'un classifieur. Pour nos travaux, nous avons utilisé un classifieur de type SVM, appris avec un noyau RBF, dont les paramètres ont été réglés par validation croisée sur 5 sous-ensembles.

3. Résutats expérimentaux

Dans cette section, nous présentons les résultats de notre système lors des deux campagnes MAURDOR [2], et les comparons avec ceux des autres participants. Pour chaque tour de campagne, les résultats ont été évalués sur un corpus de 1000 documents inconnus, à l'aide de la métrique principale *ZoneMap* (proposée par le laboratoire national de métrologie et d'essais afin dévaluer la tâche de segmentation des documents en zones), et d'une métrique secondaire *indice de Jaccard*, plus classique. Le score ZoneMap doit être minimisé, alors que l'indice de Jaccard doit être maximisé. Les résultats obtenus lors de la première compagne sont présentés dans le tableau 1, ceux de la seconde campagne sont présentés dans le tableau 2.

Tableau 1. *Résultats de la première campagne pour la détection et la segmentation des tableaux (meilleure performance en gras).*

Participants	**ZoneMap**	**Indice de Jaccard (%)**
LITIS (this work)	**44.80**	**0.400**
participant_1	80.40	0.101
participant_2	97.12	0.270
participant_3	60.28	0.183

Tableau 2. *Résultats de la seconde campagne pour la détection et la segmentation des tableaux (meilleure performance en gras).*

Participants	**ZoneMap**	**Indice de Jaccard (%)**
LITIS (this work)	**59.13**	**0.363**
participant_1	83.90	0.174
participant_2	71.39	0.307

Comme on peut le constater, le détecteur décrit dans [1] obtient pour chaque campagne les meilleures performances pour la segmentation des tableaux avec les deux métriques. Ces résultats valident ceux présentés dans [1], montrant l'intérêt de la phase d'apprentissage pour s'adapter aux formes variables des tableaux. Notons que la dégradation des performances entre les deux tours de campagne s'explique par une complexité supplémentaire des documents lors de la seconde campagne.

4. Conclusion

Nous avons présenté dans cet article les résultats obtenus par un outil de détection de tableaux dans le cadre des campagnes MAURDOR réalisées en 2013. L'approche repose sur un apprentissage, sans règles heuristiques, et peut donc s'adapter à tout type de tableau, pourvu qu'il comporte des lignes. Ces résultats sont très prometteurs, au vu de la complexité des documents traités. L'essentiel des non détection provient du fait que l'approche ne considère pas les tableaux dans lesquels il n'y a pas de lignes. Nos futurs travaux concerneront donc la prise en compte de tels tableaux, en s'appuyant sur les résultats d'un détecteur de texte décrit dans [3].

5. Bibliographie

[1] T. Kasar, P. Barlas, S. Adam, C. Chatelain, T. Paquet, *Learning to Detect Tables in Scanned Document Images Using Line Information*, International Conference on Document Analysis and Recognition (ICDAR), 1185-1189, 2013.

[2] MAURDOR campaign dataset, http ://www.maurdor-campaign.org/

[3] P. Barlas, S. Adam, C. Chatelain, T. Paquet, *A typed and handwritten text block segmentation system for heterogeneous and complex documents*, Document Analysis Systems (DAS), to appear, 2014.