L'analyse de cartes anciennes par l'intelligence artificielle

Aurélie Lemaitre

Palimpseste

sciences · humanités · sociétés
RECHERCHE À L'UNIVERSITÉ RENNES 2

numéro 7 Printemps-été 2022

Penser les relations humains • non - humains

L'analyse de cartes anciennes par l'intelligence artificielle

Aurélie Lemaitre*

Les dernières années, les services d'archives ont entrepris de vastes campagnes de numérisation, dans le but de préserver les fonds documentaires. Ces documents scannés sont alors disponibles sous forme d'images, matrices de pixels. Notre objectif est de reconnaître automatiquement le contenu de ces images pour en extraire de l'information interprétée. C'est ce que l'on appelle *l'analyse automatique d'images de documents*.

Il existe dans le commerce des systèmes nommés OCR (*Optical Character Recognition*), capables de reconnaître le texte contenu dans des images de documents. Ces OCR sont particulièrement performants pour les documents imprimés, récents, de bonne qualité. En revanche, ils obtiennent encore de piètres performances sur des documents anciens, manuscrits ou à structure complexe. L'analyse automatique d'images de documents est donc un champ de recherches en informatique encore ouvert.

Reconnaissance d'images de documents

Dans ce domaine de l'intelligence artificielle, les méthodes qui obtiennent aujourd'hui les meilleures performances sont basées sur des réseaux de neurones. On les appelle méthodes d'apprentissage supervisé. Le fonctionnement des réseaux de neurones consiste à présenter un grand nombre d'images à un algorithme, ainsi que les résultats de reconnaissance attendus. Le système va alors apprendre à inférer les bonnes réponses pour des images qui n'auraient pas été vues précédemment. Par exemple, si on montre à l'ordinateur de nombreuses images de mots manuscrits, et leur transcription, le réseau de neurones peut apprendre suffisamment d'indices pour être capable de proposer une transcription pour une image d'un mot jamais vu avant.

Les réseaux de neurones sont omniprésents dans tous les domaines de l'intelligence artificielle : requêtes sur un moteur de recherche, reconnaissance de visages, prévisions météo, cybersécurité, aéronautique, etc. Cependant, pour que ces algorithmes soient performants, il est nécessaire de disposer d'une très grande quantité de données, pour lesquelles on dispose de la transcription attendue. Ceci est la limite principale de ces approches pour l'analyse de documents anciens. En effet, la transcription manuelle de document anciens est souvent très chronophage, et parfois compliquée selon les dégradations de l'objet. De nombreux travaux de recherche en analyse de documents portent donc sur la manière d'entraîner des réseaux de neurones en disposant de peu de données étiquetées.

Nous proposons un point de vue différent. Avant l'émergence des méthodes à base de réseaux de neurones, le problème de reconnaissance automatique d'images de documents était parfois abordé sous l'angle de systèmes à base de règles : pour un type de document à reconnaître, le programmeur décrit la manière dont le contenu de l'image doit être interprété. L'intérêt majeur de ces méthodes est qu'elles ne nécessitent pas de disposer d'exemples d'images transcrites. De plus, elles se rapprochent du système de réflexion humain, et sont donc facilement explicables en cas d'erreur, contrairement aux réseaux de neurones. Malgré la suprématie des méthodes d'apprentissage supervisé, nous pensons que, dans certains contextes applicatifs précis, les méthodes à bases de règles restent compétitives. C'est ce que nous montrons dans le cadre de l'analyse de cartes anciennes.

Reconnaissance de cartes anciennes avec un système de règles

Nous présentons ici des travaux[1] réalisés lors de la compétition internationale de reconnaissance de structure de documents, MapSeg 2021. Cette compétition a été

* Maîtresse de conférences HDR en informatique, membre de l'UMR 6074 Institut de recherche en informatique et systèmes aléatoires (IRISA).

proposée en partenariat avec le LASTIG, laboratoire en sciences et technologies de l'information géographique pour la ville intelligente et les territoires durables. Les données étudiées sont des images d'atlas de la ville de Paris, de 1894 à 1937 [figure 1]. La numérisation des cartes historiques est un des enjeux forts pour les systèmes d'information géographiques. Elle détient en effet un grand potentiel pour de nombreuses études historiques.

La numérisation des cartes historiques est un des enjeux forts pour les systèmes d'information géographiques. Elle détient un grand potentiel pour de nombreuses études.

La compétition MapSeg 2021 se focalise sur une des premières étapes de la vectorisation des cartes anciennes. L'objectif est ici de localiser automatiquement les contours de la carte, en excluant les cartouches de légendes, ainsi que les intersections de lignes graticules (lignes de coordonnées géographiques). Pour les géographes, le fait de pouvoir replacer les coordonnées des cartes anciennes dans un système de coordonnées actuel ouvre de nombreuses possibilités d'études sur l'évolution des territoires au fil du temps.

Nous utilisons la méthode DMOS , mise au point par l'équipe Intuidoc de l'Institut de recherche en informatique et systèmes aléatoires (IRISA). Le principe est le suivant. Dans un premier temps, le programmeur décrit le contenu attendu dans les images : paragraphes, structures tabulaires, images, ainsi que leur positionnement relatif. Cette description s'appuie sur la présence de blocs de pixels dans l'image, mais aussi sur des éléments construits comme des segments de droite ou des lignes de texte. Dans un second temps, à partir des règles décrivant le contenu, la méthode DMOS génère automatiquement un logiciel capable d'analyser les images. Il est alors possible avec ce logiciel d'interpréter l'intégralité d'un corpus dont on a décrit le contenu.

Voyons maintenant le cas spécifique des cartes de l'atlas parisien. Nous décrivons le contenu du document à reconnaître à l'aide de règles. La première règle de l'algorithme a pour but de localiser la zone de contenu en vert [figure 2]. Cette zone est matérialisée sur le document par un rectangle avec un trait double. À une certaine distance de l'image, ces traits doubles peuvent apparaître comme un seul trait épais. La règle de reconnaissance décrite est donc la suivante : *sélectionner des segments épais pour construire un grand rectangle à proximité des bords de la page*.

La seconde étape de l'algorithme vise à détecter les contours de la carte représentés en jaune [figure 3]. Ces contours sont matérialisés par des traits d'épaisseur moyenne. On ne connaît pas *a priori* le nombre de cartouches de légende dans les cartes, ni leur position. On suppose néanmoins que, dans le corpus étudié, ces cartouches sont situés dans les angles. Nous décrivons donc la règle suivante : *utiliser les segments d'épaisseur moyenne, à l'intérieur du cadre vert, pour construire un rectangle. Dans chaque coin, chercher s'il y a un cartouche de légende (un rectangle).*

Enfin, la troisième étape de l'algorithme a pour but de détecter les lignes graticules qui définissent le système de coordonnées géographiques [figure 4]. Ces lignes sont représentées par des traits fins parallèles, qui forment une grille régulière. La difficulté de reconnaissance vient du fait que les traits fins sont très nombreux dans les documents, ils sont notamment utilisés pour tracer les rues, les bâtiments. De plus, les lignes graticules sont souvent morcelées, et parfois partiellement effacées. Nous avons donc mis au point une stratégie en deux étapes.

1. *En utilisant les traits fins, sélectionner les plus longs segments avec lesquels il est possible de former une croix de deux droites perpendiculaires.* Cette croix (en violet) sert de base à la reconnaissance de la grille.
2. *En utilisant les traits fins, reconstruire la grille des lignes graticules en cherchant des segments parallèles à la croix de base, à intervalles réguliers* (en orange).

Après avoir programmé les règles de reconnaissance du contenu des documents, nous avons utilisé la méthode DMOS, décrite plus haut, pour produire le logiciel capable de reconnaître l'intégralité du corpus. Si les règles sont simples, notons que l'algorithme est confronté à une complexité importante pour sélectionner les segments nécessaires. Par exemple, pour l'étape 3, les lignes graticules sont sélectionnées parmi 6424 segments candidats. Un des points forts de cette approche est que les règles décrites sont suffisamment génériques pour reconnaître des configurations différentes de documents [figure 5, p. 48] : document au format portrait, lignes graticules inclinées.

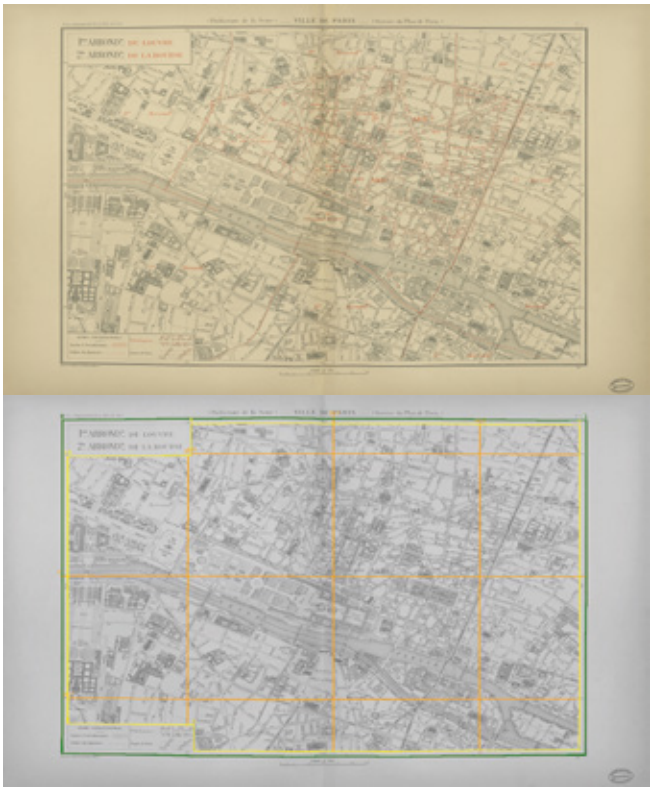

Figure 1. Image d'atlas parisien, et le résultat de l'analyse automatique du contenu : en vert la zone de contenu, en jaune les contours de la carte, et en orange les lignes graticules.

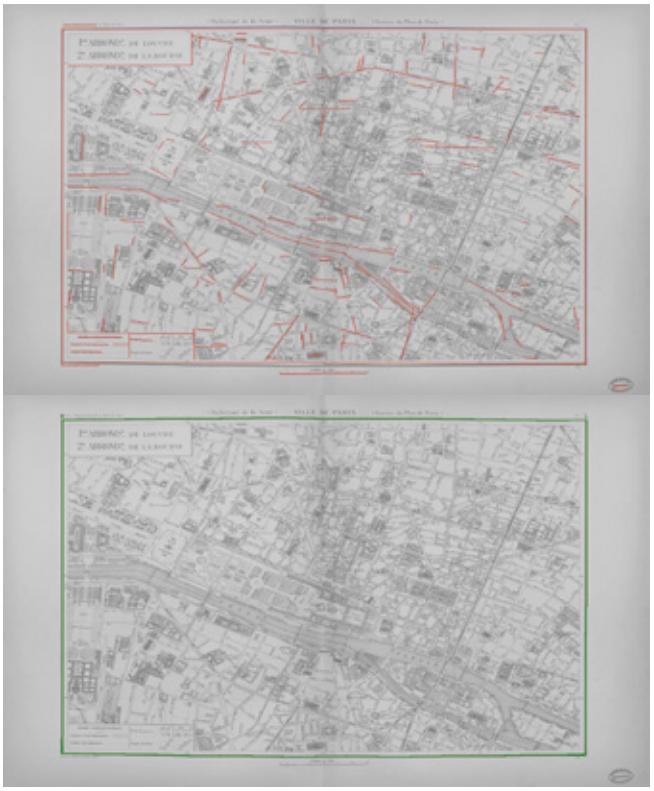

Figure 2. Étape 1 : détection des segments épais (ici 181 segments en rouge) et reconstruction du cadre extérieur en vert en sélectionnant les segments constituant un grand rectangle.

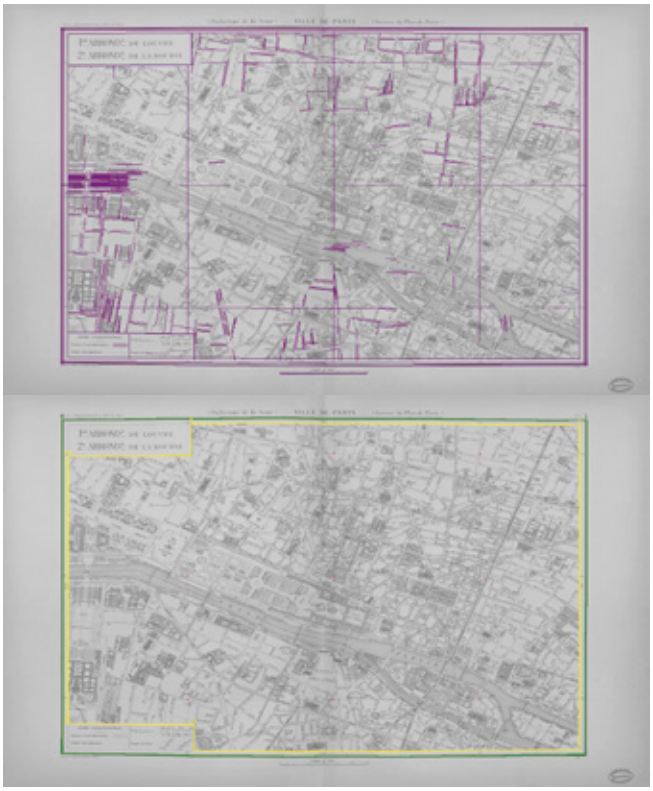

Figure 3. Étape 2 : détection des segments d'épaisseur moyenne (ici 2295 segments en violet), utilisés pour construire les contours de la carte en jaune.

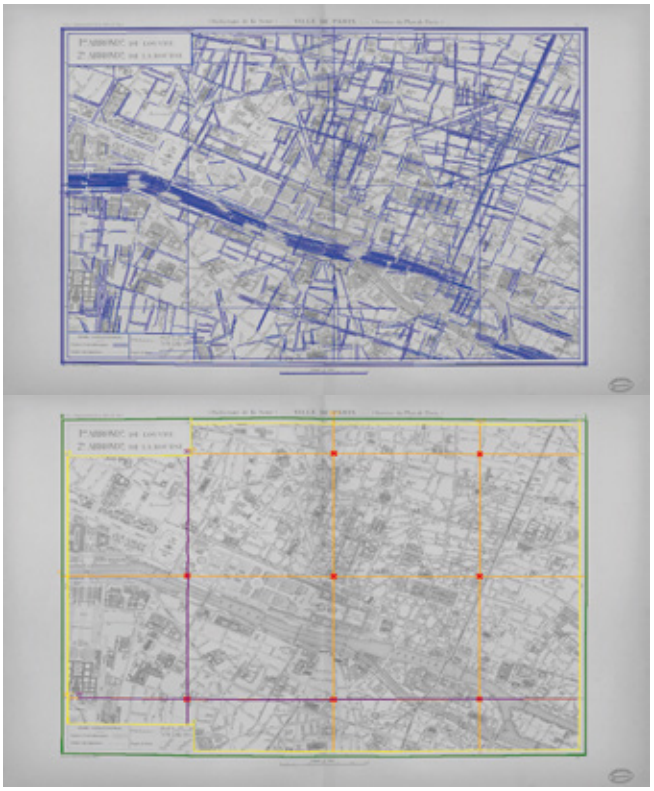

Figure 4. Étape 3 : détection des segments fins (ici 6424 segments bleus), utilisation des plus longs pour construire une croix en violet, et construction de la grille des lignes graticules en orange avec les segments parallèles.

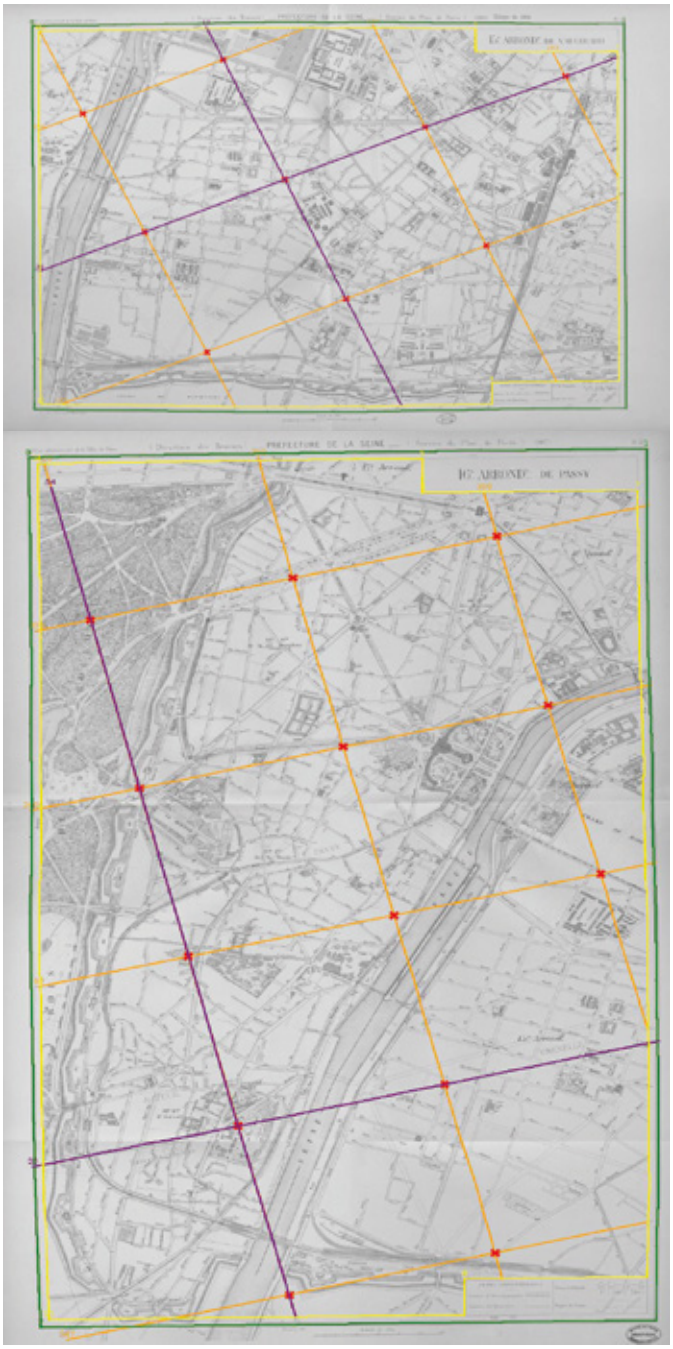

Figure 5. Reconnaissance sur plusieurs pages du corpus ayant des configurations différentes.

Figure 6. Exemple d'erreur : le système ne reconnaît pas la zone de la carte qui sort du contour. Cette erreur est explicable car la règle de description du contenu ne prévoit pas ce cas.

Évaluation et intérêt de l'« explicabilité »

Cette approche a été évaluée lors de la compétition MapSeg 2021[3]. Il s'agissait de détecter automatiquement la structure de 95 images d'atlas de la ville de Paris, datant de 1894 à 1937. Notre système a été capable de détecter les lignes graticules avec un score de 89,2 % de reconnaissance. Il est arrivé en deuxième place du concours, les concurrents obtenant des scores de reconnaissance entre 73,6 % et 92,5 %.

Un des grands intérêts de ces approches est leur « explicabilité » : le résultat dépend directement des règles qui ont été programmées. Par exemple, la figure 6 montre une erreur de prédiction sur une image de la compétition : la carte sort de la zone de contenu, ce que notre règle de détection des bords (en vert et en jaune) ne prévoit pas. Cette erreur est donc tout à fait explicable, et pourrait être corrigée par l'ajout d'une nouvelle règle prévoyant ce cas.

Conclusion

Nous avons présenté ici des travaux de reconnaissance automatique d'image de documents. À l'heure actuelle, les systèmes les plus performants sont ceux basés sur des techniques d'apprentissage supervisé, comme les réseaux de neurones. Ces méthodes ont récemment fait de grands progrès, notamment sur la reconnaissance d'écriture manuscrite. Cependant, ces méthodes nécessitent, pour leur phase d'apprentissage, des données transcrites avec les résultats attendus, très coûteuses à obtenir pour les documents anciens. C'est pourquoi les méthodes à base de règles restent compétitives dans certains contextes, notamment lorsque le contenu à reconnaître peut être décrit de manière simple. Ces approches ont l'intérêt majeur d'être compréhensibles et interprétables. Selon le problème de reconnaissance abordé, il sera donc judicieux de choisir l'approche la plus pertinente, sans se focaliser exclusivement sur les réseaux de neurones. ■

Notes de l'article

1 Aurélie Lemaitre et Jean Camillerapp, « Segmentation of historical maps without annotated data », *6th International Workshop on Historical Document Imaging and Processing (HIP'21)*, Lausanne, septembre 2021, p.19-24.

2 A. Lemaitre, J. Camillerapp et B. Couasnon, « Multiresolution cooperation makes easier document structure recognition », *International Journal on Document Analysis and Recognition*, vol.11, n° 2, 2008, p. 97-109.

3 J. Chazalon, E. Carlinet, Y. Che, J. Perret, B. Duménieu, C. Mallet et T. Géraud, « Competition on Historical Map Segmentation », *International Conference on Document Analysis and Recognition*, 2021.