

Extraction des photos publiées dans le Petit Parisien (1940)

Auteur: Luka LAVAL

Encadrants: Isabelle BLOCH, Marie-Eve BOUILLON, Daniel FOLIARD, Julien SCHUH
PRAT / Sorbonne Université / 2024-2025

1 Introduction.....	2
2 Rapport bibliographique.....	3
2.1 Construction et dégradations des documents.....	3
2.1.1 Impression.....	3
2.1.2 Conservation.....	3
2.1.3 Numérisation.....	4
2.2 Méthodes de segmentation.....	4
2.2.1 Étude de la mise en page et pré-traitement.....	4
2.2.2 Méthodes modernes.....	4
2.2.3 Texture et morphologie.....	7
2.3 Conclusion.....	9
2.4 Notes.....	10
3 Méthodologie.....	11
3.1 Archives du Petit Parisien.....	11
3.2 Expérimentations initiales.....	12
3.2.1 Demi-teinte et numérisation.....	12
3.2.2 Demi-teinte dans le domaine fréquentiel.....	13
3.3 Approche locale.....	16
3.3.1 Description générale.....	16
3.3.2 Choix des paramètres.....	16
3.3.3 Exemple.....	17
3.4 Approche globale.....	18
3.4.1 Description générale.....	18
3.4.2 Choix des paramètres.....	18
3.4.3 Exemple.....	19
3.5 Évaluation des méthodes.....	20
3.5.1 Annotation des données.....	20
3.5.2 Critères d'évaluation.....	21
3.5.3 Performances.....	22
3.6 Résultats de l'extraction.....	26
4 Discussion: limites et améliorations.....	27
5 Conclusion.....	29
6 Annexe.....	29
7 Bibliographie.....	30

1 Introduction

L'objectif de ce projet est d'extraire les photographies publiées dans le journal *Le Petit Parisien* en 1940. La base du travail repose sur les archives scannées du journal accessibles sur le site de la Bibliothèque Nationale de France: gallica.fr. Deux méthodes (locale et globale) analytiques de détection des zones de demi-teinte, méthode d'impression utilisée pour les photos à l'époque, sont décrites. Ces méthodes se basent sur une analyse dans le domaine fréquentiel pour repérer les zones d'intérêt de la page. Les résultats montrent de bonnes performances pour les deux approches sur un jeu de données annoté manuellement.



Figure 1. Première du Petit Parisien, lundi 1er janvier 1940.

2 Rapport bibliographique

2.1 Construction et dégradations des documents

L'étude des méthodes de production, conservation et numérisation des journaux sont des indices pour adapter au mieux la méthode d'extraction. L'objet exploité (page de journal numérisée) est construit en suivant un processus au cours duquel chaque étape laisse une empreinte (impression, conservation, numérisation). Ces indices qui apparaissent sous forme d'artefacts ou de dégradation peuvent être pris en compte pour une segmentation efficace des éléments de la page.

2.1.1 Impression

L'impression des journaux au XXe siècle suit un processus complexe. Le texte est composé à l'aide de caractères en plomb (linotypes ou monotypes) qui offrent une grande clarté (détails relativement précis) et efficacité pour l'impression en masse. Les illustrations sont reproduites grâce à des gravures sur blocs métalliques (souvent en zinc ou cuivre). Les photographies sont retrancrites avec le procédé de tramage (ou impression demi-teinte), qui convertit les images en une matrice de points pour créer l'illusion de nuances de gris. Ce procédé est plus coûteux et demande une préparation plus complexe, avec des résultats souvent affectés par la qualité du papier journal (Stulik & Kaplan, 2013). L'image tramée peut être gravée sur les cylindres de la rotogravure. Celle-ci permet de faire de grands tirages rapidement. La technique repose sur des cylindres gravés en creux, qui permettent une reproduction fine et rapide des détails, bien qu'elle nécessite un investissement initial important pour la gravure des cylindres. Alors même que textes, illustrations, et photographies étaient imprimés simultanément, chacun exigeait des méthodes et des préparations distinctes. Chaque méthode d'impression a sa propre signature (imperfections et dégradation), elles apportent donc des rendus différents dont les détails visuels peuvent être utilisés pour classer la nature des éléments de la page. Une branche de ce domaine consiste à classer des blocs de la page par type (titre, texte, illustration, photographie, etc.) (Lee et al., 2020).

2.1.2 Conservation

Les journaux subissent des dégradations avant et pendant leur conservation. Lors de leur utilisation initiale, ils peuvent présenter des déchirures, des plis, des froissements, ou des taches. Pendant la conservation, des facteurs tels que le jaunissement et la fragilisation du papier, les cassures aux pliures, l'humidité, et les dégâts dus à la lumière peuvent survenir. Ces altérations rendent les journaux cassants et parfois difficiles à manipuler pour la numérisation. On peut supposer que des journaux conservés dans les mêmes conditions auront des signes de dégradation similaires.

2.1.3 Numérisation

Le processus de numérisation fixe la dégradation du temps en conservant le journal sous forme de données numériques. La nature même de l'objet est alors changée, d'une entité continue matérielle nous passons à sa discrétisation (numérique). Les méthodes de numérisation, elles aussi, diffèrent tout comme leur résolution (en dpi). La numérisation peut aussi mener à des dégradations, par exemple, un effet Moiré peut apparaître sur une impression en demi-teintes. Cet effet peut être simplement supprimé pour améliorer l'aspect de l'image (Chingyu Yang & Tsai, 1998).

2.2 Méthodes de segmentation

La segmentation des documents consiste à déterminer la forme et la nature des éléments de la mise en page (blocs de texte, titre, légende, photographies, décoration). Cette section donne une idée des méthodes modernes utilisées et met aussi en lumière des approches alternatives. Un pré-traitement des documents numérisés et une étude de la mise en page facilitent généralement la segmentation. Les marques laissées par la construction et la dégradation des documents, discuté dans la section précédente, sont parfois vues comme des contraintes rendant la segmentation plus difficile que pour des documents plus récents. Ces marques sont aussi parfois utilisées comme indices pour la segmentation.

2.2.1 Étude de la mise en page et pré-traitement

La disposition des éléments dans les journaux anciens diffère de celle des journaux modernes, ce qui complique la segmentation automatique. Les zones de texte, images, titres, et lignes de séparation suivent une logique d'agencement spécifique à chaque époque, rendant la segmentation de mise en page essentielle pour éviter la confusion entre les types de contenu. Ainsi, avant de se lancer dans la construction d'une méthode, il est préférable d'étudier la mise en page pour délimiter des zones de recherches ou trouver des indices permettant l'extraction des photographies (Wang & Srihari, 1989). Par exemple, on peut observer que la police d'un texte diffère de celle d'une légende, et certaines zones de la page ne pourront jamais contenir d'images.

Il sera aussi intéressant de tenter d'améliorer la qualité de documents exploités pour de meilleurs résultats de segmentation. Ainsi, l'application de filtres numériques (gaussien, median) ou la binarisation (potentiellement avec un seuillage adaptatif) par exemple, permettent dans certains cas de débruiter l'image. L'amélioration du document peut se faire avec des filtres morphologiques (ouverture, fermeture) ou en augmentant les contrastes (Schade & Bloch, 2024).

2.2.2 Méthodes modernes

La plupart des méthodes utilisées aujourd'hui pour l'extraction des images sont basées sur de l'apprentissage profond. Ces méthodes permettent d'automatiser l'identification de contenus visuels dans les journaux, un processus essentiel pour des corpus numérisés comme *Le Petit Parisien* de 1940. Il faut aussi noter que ces approches

abordent rarement le problème d'anciens journaux. Dans cette section, trois travaux sont présentés.

PRAT Project Report (Schade & Bloch, 2024)

Dans son rapport, Oskar Schade propose l'utilisation de réseaux de neurones Fully Convolutional Networks (FCN) pour segmenter et identifier des éléments visuels dans des documents historiques. En particulier, le projet explore l'efficacité du modèle dans la détection de zones photographiques et de légendes textuelles dans des journaux anciens.

Le modèle U-Net, populaire pour les tâches de segmentation sémantique, permet de générer une segmentation de pixels et d'isoler les régions d'intérêt avec une précision élevée, même dans des documents comportant des motifs complexes. Le modèle repose sur des couches de convolution pour extraire des caractéristiques détaillées tout en préservant la résolution spatiale, essentielle pour les images historiques souvent dégradées ou de faible résolution. Par ailleurs, le modèle VGG16, utilisé pour classifier et segmenter les images, se montre également performant pour séparer le texte et les images.

Pour compenser le manque de données annotées, les données ont été artificiellement augmentées avec trois types de bruits adaptés aux données historiques: le bruit Gaussien, le Salt-and-Pepper, et l'introduction de petites taches. La nécessité d'augmenter artificiellement les données souligne aussi le manque de données pour l'entraînement des modèles sur des documents historiques.

Le rapport souligne que l'approche globale donne de meilleurs résultats en comparaison avec l'approche en patchs, et utilise des filtres morphologiques en post-traitement pour améliorer la segmentation. Finalement, cela montre que les FCN sont capables d'extraire efficacement les éléments visuels dans des documents historiques en surmontant les défis posés par les artefacts et la faible résolution des images d'archives.

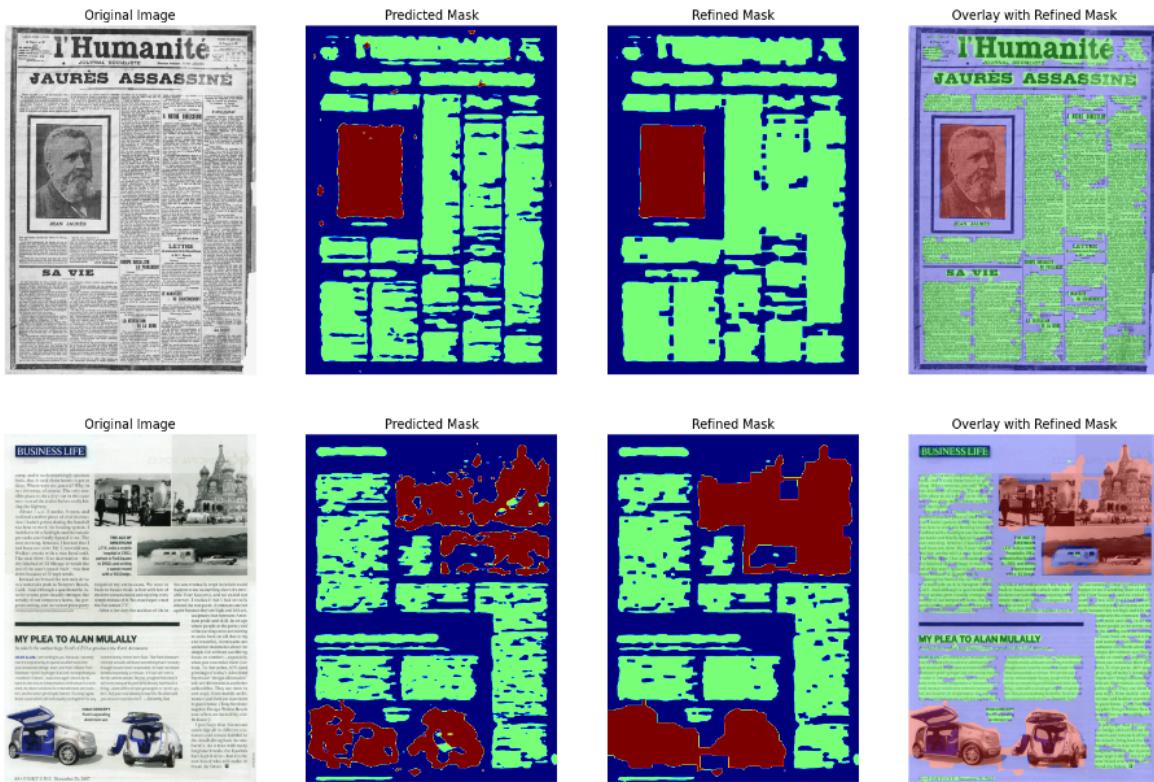


Figure 2. Évolution du masque pour l'extraction des images et du texte.

LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis (Shen et al., 2021)

LayoutParser est un outil open-source développé pour simplifier et standardiser l'analyse des documents. Conçu pour être modulable, cet outil offre des modèles pré entraînés permettant la détection et la segmentation d'éléments de mise en page, tels que les photographies, les titres, et les blocs de texte, rendant son application adaptée à divers types de documents historiques.

Le principal atout de LayoutParser réside dans sa capacité à reconnaître différentes structures de mise en page grâce à un "Zoo" de modèles, notamment ceux basés sur les réseaux neurones convolutifs (CNN). La possibilité de personnaliser les modèles permet d'adapter l'outil aux spécificités des journaux anciens, notamment ceux imprimés en demi-teinte, où l'identification précise des zones photographiques est cruciale. L'adaptabilité de LayoutParser est donc particulièrement pertinente pour des projets d'extraction de photographies, car elle permet de segmenter des images complexes en tenant compte des caractéristiques des documents anciens.

Toutefois, LayoutParser présente certaines limites. Bien que sa conception modulaire facilite son adaptation, l'outil n'est pas spécifiquement conçu pour les documents dégradés. Ainsi, des prétraitements supplémentaires peuvent être nécessaires pour optimiser l'extraction d'images dans des journaux historiques fortement dégradés, et l'adaptation du modèle aux particularités des documents d'archives. Néanmoins,

LayoutParser offre une base solide et adaptable pour l'extraction des éléments visuels dans des documents historiques numérisés, il a notamment été utilisé pour extraire des photographies de journaux historiques dans *Identification de la circulation d'images historiques* (Halimi, 2024).

The Newspaper Navigator Dataset - Extracting Headlines and Visual Content from 16 Million Historic Newspaper Pages in Chronicling (Lee et al., 2020)

Le *Newspaper Navigator Dataset*, est un projet ambitieux en matière d'extraction d'éléments visuels à partir de journaux historiques. Ce projet repose sur l'utilisation du modèle Faster-RCNN pour détecter sept catégories d'éléments visuels, y compris les photographies et les illustrations, sur plus de 16 millions de pages de journaux américains.

Faster-RCNN, un modèle basé sur l'apprentissage profond pour la détection d'objets, utilise un réseau de propositions de régions (Region Proposal Network, RPN) pour identifier rapidement les zones pertinentes dans une image, ce qui le rend idéal pour traiter des corpus volumineux comme celui de *Chronicling America*. Les annotations ont été générées par des volontaires, puis utilisées pour raffiner les prédictions.

Le *Newspaper Navigator Dataset* représente une avancée significative en démontrant comment l'apprentissage profond peut être utilisé pour des applications de grande envergure dans l'analyse de journaux anciens.

Plus largement, (Lombardi & Marinai, 2020) proposent une base théorique solide pour les projets de recherche visant à extraire des éléments visuels dans les journaux historiques numérisés dans *Deep Learning for Historical Document Analysis and Recognition—A Survey*. L'article met en lumière l'impact croissant des techniques de deep learning pour résoudre les problèmes de reconnaissance des documents historiques. Cependant, pour pleinement exploiter ces avancées, des efforts supplémentaires sont nécessaires dans la collecte de données, la réduction des biais et le développement d'outils accessibles pour les chercheurs en sciences humaines.

2.2.3 Texture et morphologie

Classification of Newspaper Image Blocks Using Texture Analysis (Wang & Srihari, 1989)

Wang et Srihari proposent une méthode innovante pour la classification des blocs d'image dans les journaux numérisés en utilisant l'analyse des textures. Leur approche repose sur l'extraction de caractéristiques spécifiques à partir de matrices décrivant la répartition des motifs de texture. Ces matrices, noir-blanc (BW Matrix) et noir-blanc-noir (BWB Matrix), permettent de capturer les propriétés fondamentales des blocs homogènes dans les journaux. Les caractéristiques ainsi dérivées permettent de distinguer des blocs contenant des photographies, du texte en petits ou grands caractères, et des dessins au trait.

La méthode proposée est particulièrement utile pour les documents historiques, où la qualité de l'impression et les artefacts visuels posent des défis considérables. Contrairement aux méthodes traditionnelles basées sur des dimensions géométriques, cette approche s'appuie sur des indices texturaux pour classifier les blocs, offrant ainsi une solution robuste face aux variations de mise en page et aux dégradations des journaux anciens. Les auteurs rapportent une précision de classification atteignant 94% à 100ppi et 100% à 200ppi, ce qui illustre l'efficacité de la méthode pour des résolutions typiques des documents numérisés.

Cette technique présente toutefois certaines limites. Par exemple, elle peut être moins efficace pour des motifs extrêmement complexes ou des documents présentant des artefacts visuels inhabituels.

Document Image Segmentation Using Wavelet Scale-Space Features (Acharyya & Kundu, 2002)

Acharyya et Kundu proposent une méthode basée sur l'analyse multi-échelle utilisant des ondelettes pour segmenter les images de documents en régions textuelles et non textuelles. L'approche repose sur l'exploitation des propriétés énergétiques des textures en décomposant les images en différentes sous-bandes fréquentielles. Cette méthode vise à transformer les frontières entre les textures en discontinuités détectables, facilitant ainsi la segmentation.

L'algorithme fonctionne en trois étapes principales : une décomposition multi-échelle à l'aide de filtres ondelettes multi-directionnels, une estimation locale de l'énergie à partir des sous-bandes décomposées, et une classification non supervisée utilisant l'algorithme k-means. Cette approche est entièrement indépendante des caractéristiques spécifiques des documents, telles que la taille des caractères, la disposition des colonnes ou l'orientation. De ce fait, elle est particulièrement adaptée pour les documents complexes ou dégradés.

Les résultats expérimentaux montrent que cette méthode est efficace pour traiter une large gamme de documents, qu'ils soient structurés ou non structurés. Par exemple, l'algorithme est capable de gérer des textes mal alignés, des documents inclinés, et des mises en page irrégulières. Cependant, cette approche exige un coût computationnel relativement élevé en raison de la complexité de la décomposition multi-bandes, ce qui peut limiter son utilisation pour des projets à grande échelle.

Malgré ces défis, l'approche par ondelettes multi-échelle représente une alternative puissante aux méthodes traditionnelles pour l'analyse de documents historiques.

Improved Document Image Segmentation Algorithm Using Multiresolution Morphology (Bukhari et al., 2011)

Bukhari et al. introduisent des améliorations à l'algorithme de segmentation morphologique multi-résolution de (Bloomberg, 1991), en étendant son application pour inclure des composants non-textuels variés tels que les dessins, les graphiques,

et les cartes. Cette méthode repose sur des opérations de morphologie mathématique pour séparer les composants textuels des non-textuels, en s'appuyant sur des seuils multi-échelles pour filtrer les détails indésirables.

Les améliorations incluent l'ajout d'une étape de "hole-filling" pour préserver les contours creux des dessins et graphiques, ainsi qu'une méthode de reconstruction des lignes brisées, souvent causées par des erreurs de numérisation ou une faible résolution. Ces étapes permettent de garantir que même les composants non-textuels fragmentés soient correctement classifiés. Les résultats expérimentaux montrent une amélioration significative de la précision de segmentation, avec des taux dépassant 90% dans des cas complexes.

Malgré son efficacité, cette approche est plus coûteuse en termes de calcul que l'algorithme original de Bloomberg, en raison des étapes supplémentaires introduites. Malgré tout, sa capacité à traiter une variété de composants non-textuels et sa robustesse face à des artefacts complexes en font une méthode précieuse pour l'analyse de documents historiques.

2.3 Conclusion

L'extraction de photographies à partir d'archives de journaux, représente un défi technique et méthodologique. Ce rapport bibliographique a permis de mettre en lumière les enjeux liés à la construction, à la dégradation et à la numérisation des documents, ainsi que les stratégies employées pour leur segmentation et leur analyse.

Le rapport souligne l'importance des empreintes laissées par les processus d'impression et de conservation, qui influencent directement les méthodes d'extraction. Ces marques peuvent être exploitées comme des indices pour améliorer la précision des modèles d'apprentissage ou, à l'inverse, être considérées comme des contraintes complexes nécessitant des techniques de prétraitement spécifiques.

Les méthodes modernes basées sur l'apprentissage profond ont montré leur efficacité pour l'extraction d'éléments visuels dans des contextes similaires. Cependant, elles exigent souvent un grand volume de données annotées et une adaptation rigoureuse aux particularités des journaux historiques. Parallèlement, des approches alternatives, comme l'analyse des textures ou l'utilisation de transformations multi-échelles, offrent des solutions intéressantes en termes de robustesse face à la dégradation des documents, bien qu'elles présentent des limitations en termes de coût computationnel ou d'adaptabilité.

En somme, l'extraction des photographies de journaux anciens reste un domaine de recherche riche et évolutif. Les progrès réalisés dans le domaine de l'apprentissage profond, combinés à des stratégies de prétraitement innovantes et à une exploitation judicieuse des artefacts, offrent des perspectives pour développer des outils efficaces et adaptés.

2.4 Notes

Certains travaux n'ont pas pu être consultés mais semblent particulièrement pertinents. *Image extraction in digital documents* (Won, 2008) discute de l'identification et de l'extraction d'images dans des documents de manière générale (pas seulement pour des journaux historiques). *A multiresolutional algorithm for halftone detection* (Hidayati et al., 2015) et *Low complexity pixel-based halftone detection* (Ok et al., 2011) proposent deux approches pour détecter des zones d'impressions en demi-tons dans un document. Néanmoins, ces approches sont particulièrement intéressantes dans notre cas pour la segmentation des photographies publiées dans les journaux de l'époque.

3 Méthodologie

Dans cette section je détaille deux approches pour extraire les photos des pages de journaux. Les deux méthodes se basent sur la détection de régions d'impression en demi-teinte dans le domaine fréquentiel. Ces deux méthodes ont été évaluées sur un jeu de données annoté manuellement. La dernière partie décrit le jeu de données final contenant toutes les photos extraites des archives.

3.1 Archives du Petit Parisien

L'étude porte sur les archives du Petit Parisien publiées au cours de l'année 1940. Ces archives sont disponibles sur le site de la BNF Gallica.fr. Plusieurs scripts sont disponibles en ligne pour utiliser l'API du site (Alexander, n.d.; ModOAP, n.d.-a, n.d.-b). J'ai rédigé un script pour télécharger toutes les pages publiées en 1940. Le jeu de données contient 1216 images (336 numéros) au format .jpg d'environ 3400×4800 pixels.

Les pages sont construites en blocs de différentes natures (photographie, illustration, texte, ornementation, publicité...). On remarque aussi une certaine diversité dans la construction des pages, par exemple certaines pages contiennent presque uniquement des illustrations et d'autres que du texte. Néanmoins, ces architectures ne représentent pas la majorité des pages.



Figure 3. Page avec différents blocs; Page entière de texte; Page entière de publicité.

Ces archives présentent aussi parfois des imperfections. Celles-ci n'ont pas été considérées dans l'élaboration des méthodes d'extraction mais sont importantes à souligner puisqu'elles peuvent impacter leur efficacité. Certaines d'entre elles sont illustrées ci-dessous.



Figure 4. Imperfections des archives.

3.2 Expérimentations initiales

3.2.1 Demi-teinte et numérisation

L'impression en demi-teinte, utilisée dans les années 1940, est une technique permettant de reproduire des images en noir et blanc avec différentes nuances de gris. Elle fonctionne en décomposant une image en petits points : les zones sombres sont représentées par des points plus gros et rapprochés, tandis que les zones claires utilisent des points plus petits et espacés. Bien qu'une seule couleur d'encre soit utilisée, cette variation de taille et de densité des points crée l'illusion de dégradés continus. À une certaine distance, l'œil humain perçoit ces points comme une image fluide en niveaux de gris.

Alors même que ces archives ont été numérisées à une résolution relativement haute, on ne peut pas distinguer les points de trame composant les photos. L'image ci-dessous montre que les points de trames (visibles à gauche) ont une apparence très différente après la numérisation. On devine néanmoins une sorte de damier (alternance de pixels clairs et foncés) qui rappelle le tramage pour représenter le gris à l'arrière-plan.

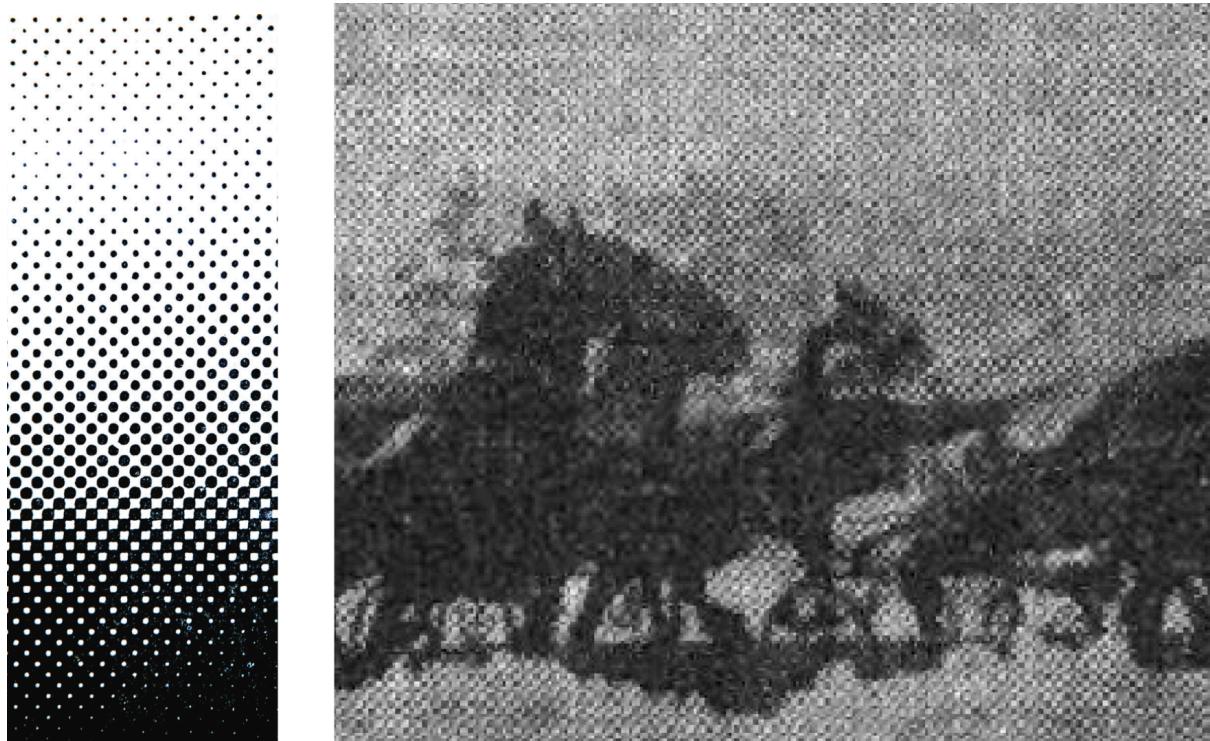


Figure 5. Points de trame théoriques (à gauche) et numérisés (à droite).

3.2.2 Demi-teinte dans le domaine fréquentiel

Dans le domaine fréquentiel, l'analyse d'une image, notamment via la transformée de Fourier (FFT, Fast Fourier Transform), permet de décomposer l'image en ses différentes fréquences spatiales. Les basses fréquences représentent les variations lentes, comme les grandes zones uniformes, tandis que les hautes fréquences capturent les détails fins et les bords nets. Pour une image en demi-teinte, les points réguliers introduisent un motif périodique qui apparaît comme des pics dans le spectre fréquentiel. Cette propriété est utile pour analyser ou filtrer les artefacts spécifiques à ce type d'impression, comme les motifs répétitifs, et pour optimiser le rendu ou la compression numérique des images imprimées. Par exemple, dans *Frequency domain filtering techniques of halftone images* (Tholeti et al., 2015), les auteurs suggèrent d'appliquer un masque dans le domaine fréquentiel pour masquer le bruit lié à la demi-teinte.

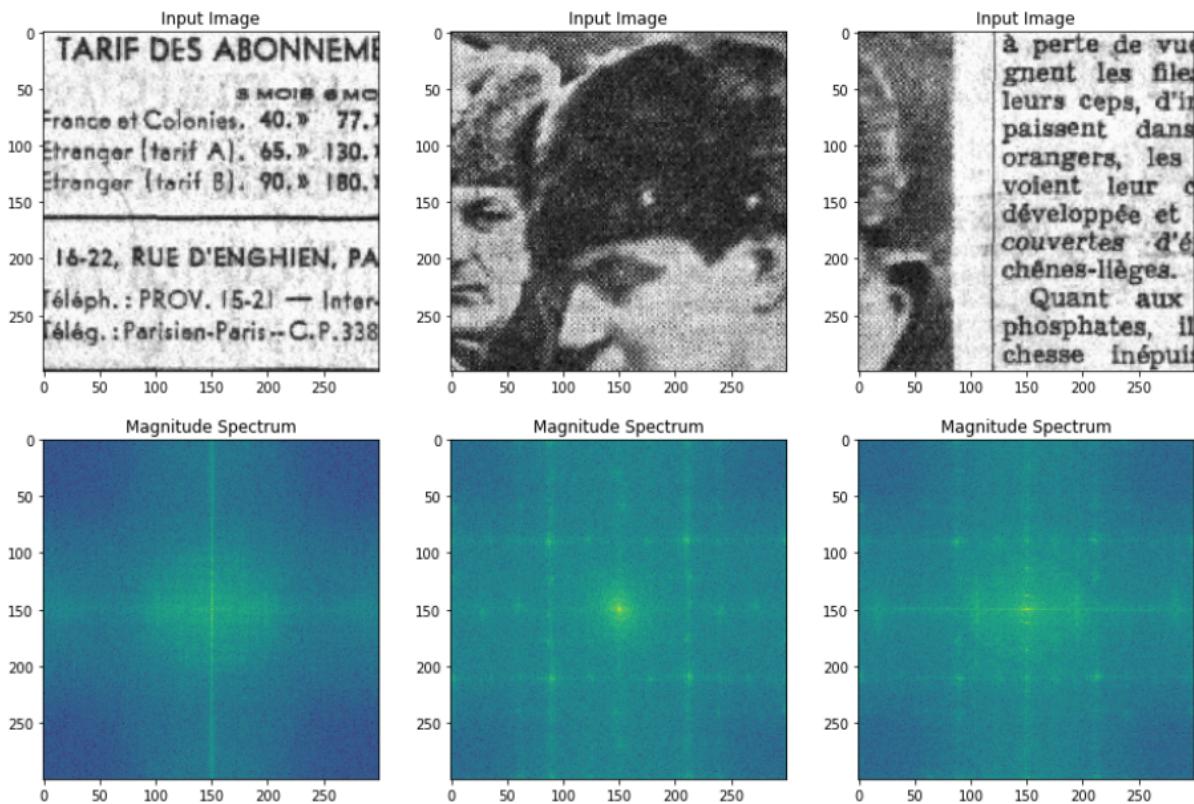
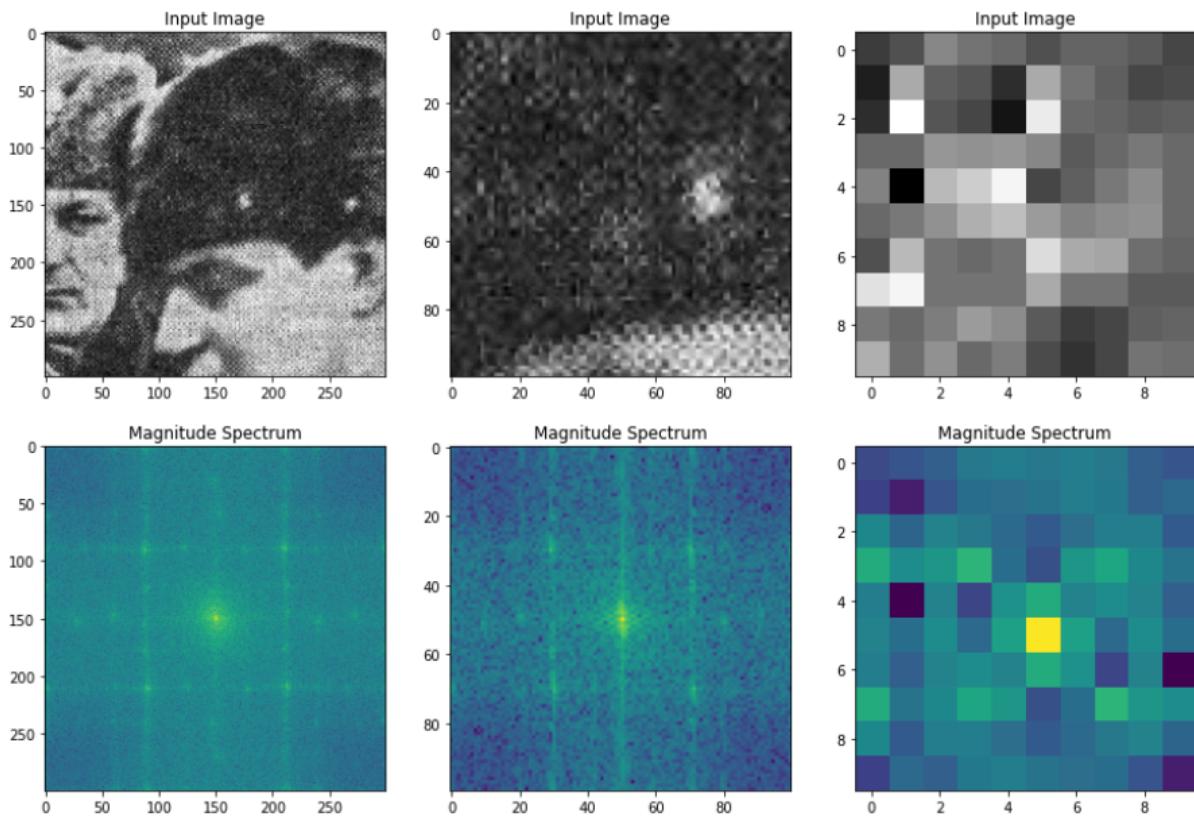
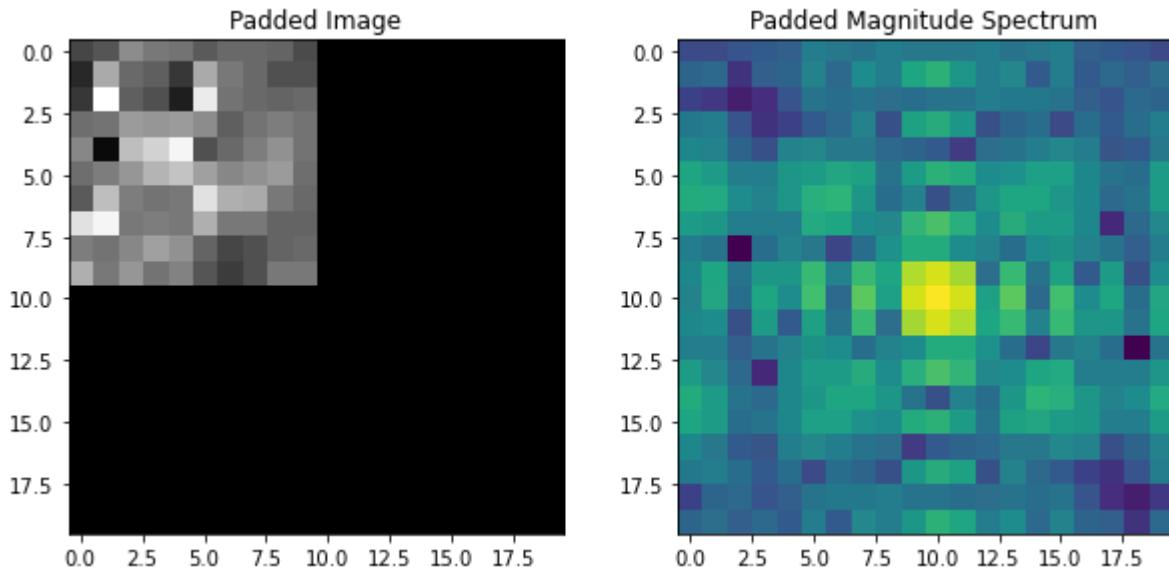


Figure 6. Réponses fréquentielles d'une image de texte (à gauche), une photo en demi-teinte (au milieu) et une de texte contenant une petite zone de la photo (à droite).

En pratique, en observant le domaine fréquentiel des photos on repère bien les pics fréquentiels caractéristiques de la demi-teinte, à la différence des zones de texte. On remarque aussi que dès lors qu'une zone de demi-teinte (même réduite) est inclue dans la région analysée, les pics apparaissent.

Maintenant, la question des limites de la taille de la fenêtre d'étude s'impose. La taille maximale de la fenêtre est la page entière, on observe alors bien les pics fréquentiels. En diminuant progressivement la taille de la fenêtre, on s'aperçoit qu'il devient difficile de distinguer les pics. Pour pallier cette limite en résolution, on peut augmenter artificiellement la fenêtre comme illustré ci-dessous. La résolution dans le domaine spatial reste réduite, mais celle dans le domaine fréquentiel est augmentée. Cependant l'augmentation artificielle du domaine fréquentiel ne permet pas de mieux distinguer les pics à un certain point puisque les patterns de demi-teinte ne sont même plus visibles dans le domaine spatial.

*Figure 7. Réponse fréquentielle à différentes échelles**Figure 8. Amélioration artificielle de la résolution fréquentielle.*

Ainsi il est établi que la détection des pics dans le domaine fréquentiel est affectée par la taille de la fenêtre d'analyse. Si la fenêtre d'analyse est trop petite, il n'est plus possible d'observer les pics avec certitude puisque les patterns répétitifs de la demi-teinte disparaissent.

3.3 Approche locale

3.3.1 Description générale

Cette première approche consiste à parcourir la page avec une fenêtre d'analyse caractérisant les zones de demi-teinte. À chaque position survolée, le contenu de la fenêtre est transformé dans le domaine fréquentiel. Si les pics de demi-teinte sont détectés, les pixels de la fenêtre sont marqués.

Pour détecter les pics de demi-teinte dans le domaine fréquentiel, j'applique un seuil pour garder uniquement les 5% des valeurs les plus hautes. Ensuite on s'intéresse aux pixels supérieurs au seuil dans les quatre quadrants susceptibles de contenir les pics de demi-teinte. Si chaque quadrant contient au moins une valeur non nulle, la zone de la fenêtre est marquée dans la carte.

Les valeurs de la carte de demi-teinte sont connectées en composantes connexes, ainsi, chaque région isolée (potentielle photo) est traitée indépendamment. Les composantes connexes trop petites sont supprimées, les autres sont considérées comme contenant une photo. Les bounding boxes (boîtes englobantes) des composantes connexes restantes représentent les zones des photos.

Optionnellement, on peut ajouter des relations logiques pour traiter les dernières erreurs, par exemple une boîte englobante contenue dans une autre est supprimée etc.

3.3.2 Choix des paramètres

Taille de la fenêtre : La fenêtre est définie comme un carré de 40×40 pixels. Ceci permet d'avoir une fenêtre relativement précise (car elle est directement liée à la précision du contour des régions détectées) mais assez grande pour détecter les pics de demi-teinte.

Pas : Le pas de déplacement de la fenêtre peut être modifié. Le pas ici est défini à 40 pixels, c-a-d qu'il n'y a jamais de chevauchement. Cette taille de pas est la plus intuitive pour réduire le temps de calcul et couvrir la totalité de la page. Il est intéressant de réduire le pas et forcer les chevauchements pour avoir des contours plus précis, mais ceci nécessite une étape additionnelle avant de générer les boîtes englobantes (voir la section Discussion).

Seuil des fréquences : Le seuil des fréquences est défini à 95%. Il est important d'avoir un seuil relatif aux fréquences et non-pas un seuil arbitraire.

Seuil des composantes connexes : Ce seuil a été sélectionné arbitrairement pour éviter les zones trop petites souvent liées au bruit, ici 10000 pixels. Si il est trop grand, des zones de photos risquent d'être supprimées. À l'avenir, il serait intéressant d'avoir

une valeur relative à la taille de la page. Les pages étant de tailles similaires, ceci n'a pas d'impact pour cette étude.

3.3.3 Exemple

L'algorithme permet de segmenter les trois photos visibles sur la page. On remarque que les zones sombres des photos ne sont pas marquées comme "demi-teinte". Malgré tout, les photos sont bien segmentées puisque la boîte englobante est calculée.

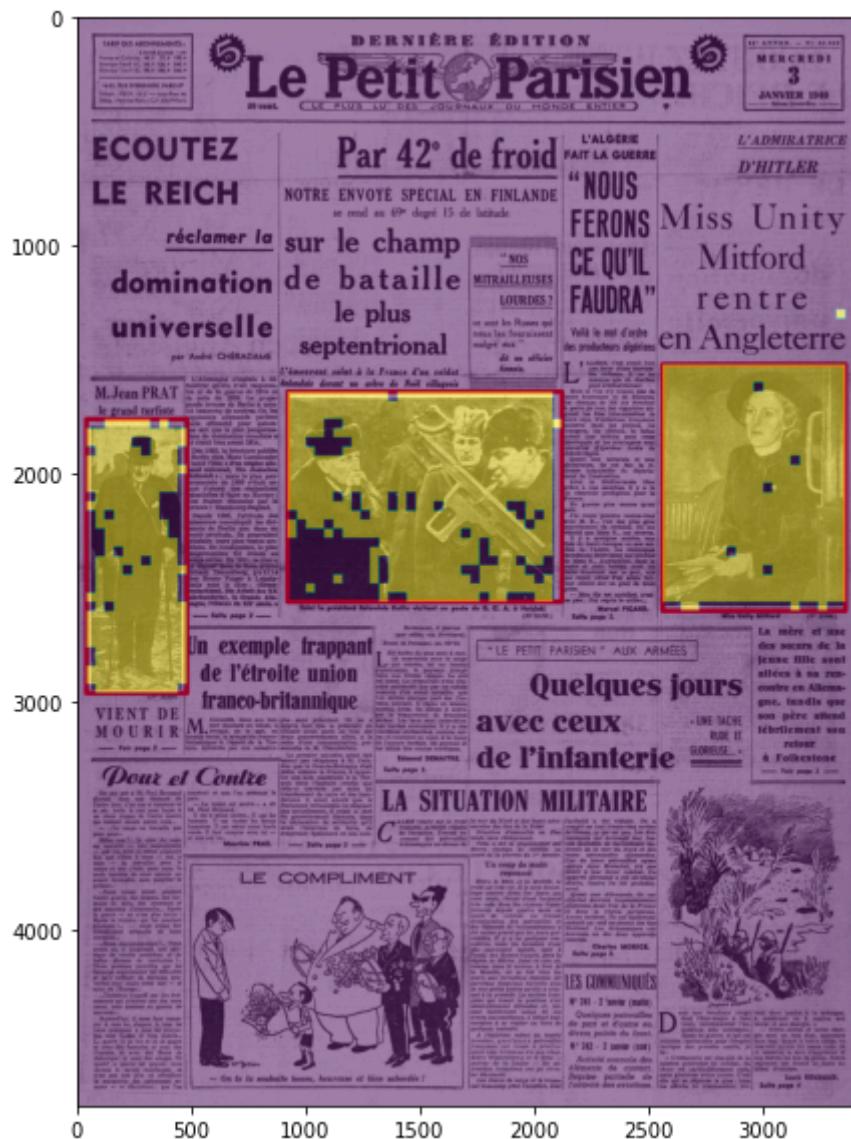


Figure 9. Résultats de la méthode locale. Zones de demi-teinte (jaune), zones sans demi-teinte (violet), photo détectées (cadres rouges).

Sur cet exemple, l'algorithme ne détecte pas les illustrations comme demi-teinte, il n'y a pas d'ambiguïté. Seule une fenêtre est détectée en dehors des régions segmentées (en haut à droite), celle-ci n'est pas prise en compte car elle fait partie d'une composante connexe trop petite.

3.4 Approche globale

3.4.1 Description générale

L'approche globale consiste à inverser la méthode de débruitage des images en demi-teinte pour détecter ces zones. On applique une FFT pour travailler dans le domaine fréquentiel. Un masque est appliqué au spectre des magnitudes pour ne garder que les pics de demi-teintes, puis l'image est reconstruite avec une transformée inverse. En appliquant un seuil et filtres morphologiques, les régions de demi-teinte de la page sont marquées.

Le masque du domaine fréquentiel consiste en quatre disques centrés sur les zones des pics de demi-teinte observés lors des expérimentations. Il est important de noter que les positions des pics peuvent varier en fonction du tramage. Dans notre cas je fais la supposition que les pics restent au même endroit puisque nous avons vraisemblablement à faire au même type d'impression et de numérisation pour ces archives.

Après la transformation inverse, un seuil binaire est appliqué à l'image reconstruite puis une fermeture et une ouverture sur toute la page. La fermeture permet de lier les pixels proches entre eux pour créer des régions d'intérêt plutôt que des pixels individuels séparés. L'ouverture permet de supprimer le bruit restant.

Le masque d'extraction se calcul de la manière suivante:

$$((IFFT(FFT(I) \cdot M)_b)^{B_F})_{B_O}$$

où I est l'image originale, M le masque dans le domaine fréquentiel, \cdot_b représente la binarisation de la matrice avec un seuil, B_F et B_O représentent les éléments structurants pour la fermeture et l'ouverture.

Enfin, comme pour la méthode locale, les composantes connexes sont marquées pour distinguer les régions, leurs boîtes englobantes représentent les zones à extraire.

3.4.2 Choix des paramètres

Seuil valeurs image reconstruite : Ce seuil permet de supprimer les valeurs basses de l'image reconstruite et conserver seulement les pixels dans les zones de demi-teinte. Il est défini à 5 ici.

Élément structurant fermeture : L'élément structurant a une taille de 20×20 pixels. Lorsque l'image est reconstruite on observe des zones dense de pixels avec des valeurs non-nulles mais ceux-ci ne sont pas connectés entre eux. La fermeture permet de connecter ces pixels.

Élément structurant ouverture : L'élément structurant a une taille 30×30 pixels. Après la fermeture, certaines zones de bruit sont dilatées. L'ouverture permet de supprimer ces zones. Il est important que la taille de l'élément structurant de l'ouverture soit supérieur à celui de la fermeture.

Seuil des composantes connexes : De la même manière que pour la méthode locale, un seuil de taille minimum est imposé aux composantes connexes. Celles inférieures à 20000 pixels sont supprimées.

3.4.3 Exemple

De la même manière que pour le cas local, on observe à nouveau que les zones sombres des photos ne sont pas marquées comme "demi-teinte". Certaines petites zones sur le texte et les illustrations sont marquées comme "demi-teinte" mais ne sont finalement pas segmentées car trop petites.



Figure 10. Résultats de la méthode globale. Zones de demi-teinte (jaune), zones sans demi-teinte (violet), photo détectées (cadres rouges).

L'évolution de la carte de demi-teinte est illustrée ci-dessous, d'abord la carte est construite comme l'image restituée avec la transformée inverse, puis un seuil est appliqué pour réduire le bruit, fermeture pour connecter les éléments entre eux, enfin ouverture pour supprimer certains éléments.

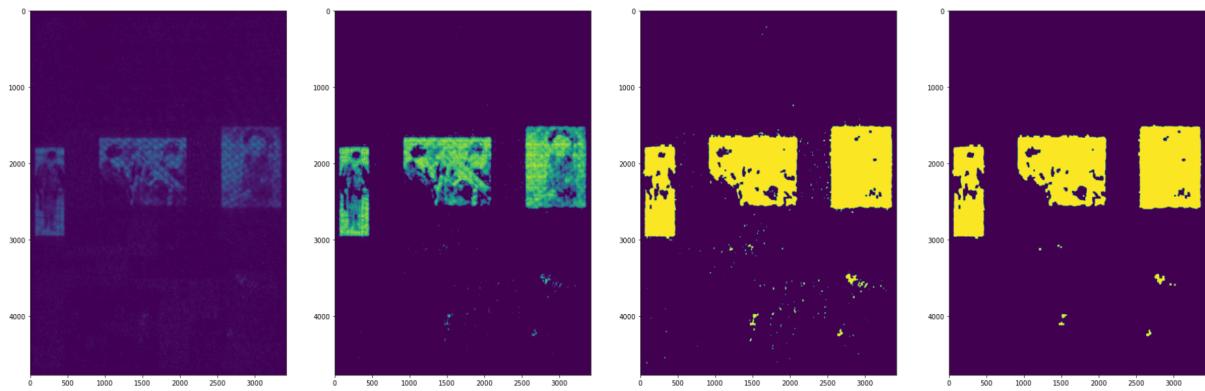


Figure 11. Évolution du masque avant extraction. De gauche à droite: réponse de l'image reconstruite, seuillage, fermeture morphologique, ouverture morphologique.

3.5 Évaluation des méthodes

3.5.1 Annotation des données

Pour évaluer les méthodes, j'ai construit un jeu de données de test. Celui-ci contient 50 pages choisies aléatoirement (voir fichier *utils.ipynb*). J'ai annoté manuellement chaque page avec le logiciel *VGG Image Annotator* (Dutta & Zisserman, 2019) en traçant les boîtes englobantes de chaque photo. Il est important de souligner que les illustrations et les publicités ne sont pas considérées comme des photos dans cette étude et donc pas annotées.



Figure 12. Annotation des données (boîtes englobantes) sur VGG Image Annotator.

En tout, 93 photos réparties sur 30 pages ont été annotées, 20 pages ne contiennent pas de photo. Le fichier d'annotation est exporté au format .csv pour être comparé aux résultats des deux méthodes.

3.5.2 Critères d'évaluation

Les boîtes englobantes générées par les algorithmes sont inscrites dans un fichier .csv au même format que les annotations. J'utilise la métrique *Intersection over Union* (IoU), elle mesure le degré de recouvrement entre deux régions: la boîte de prédiction générée et la boîte de l'annotation (ground truth). L'IoU est calculée comme le rapport entre l'intersection des deux régions et leur union. Ainsi, la valeur de l'IoU varie entre 0 (aucun recouvrement) et 1 (recouvrement parfait) (Rezatofighi et al., 2019). Pour adapter cette métrique à notre cas, je compare les pages entre elles c'est à dire que l'ensemble des boîtes générées d'une page est évalué avec l'ensemble des boîtes annotées de cette page. Dans le cas où les deux pages sont vides (annotation et génération), le score est de 1. Si seulement une des pages est vide, le score est de 0.

$$IoU = \frac{|B_p \cap B_a|}{|B_p \cup B_a|}$$

où B_p est la boîte de prédiction générée et B_a est la boîte d'annotation (ground truth).

Pour adapter cette métrique aux pages complètes, on définit le score de similarité entre deux pages P_1 et P_2 comme suit :

$$S(P_1, P_2) = \begin{cases} 1, & \text{si } P_1 \text{ et } P_2 \text{ sont vides} \\ 0, & \text{si seulement l'une des deux pages est vide} \\ \frac{\sum_{i,j} IoU(B_{p_i}, B_{a_j})}{N} & \text{sinon} \end{cases}$$

Où N correspond au nombre de couples (B_{p_i}, B_{a_j}) dont l'intersection est non nulle.

3.5.3 Performances

Les performances sont mesurées avec la métrique introduite précédemment (IoU) sur le jeu de données annoté (50 pages choisies aléatoirement). Les deux méthodes ont nécessité un temps d'exécution similaire, entre 3 et 4 minutes, pour parcourir les données de test. Les performances individuelles des algorithmes sont très proches. Les tableaux ci-dessous récapitulent les scores.

Méthode	Moyenne	Minimum	25%	50%	75%	Maximum
Locale	0.78	0.0	0.81	0.94	0.99	1.0
Globale	0.83	0.0	0.91	0.96	1.0	1.0

Table 1. Performances sur l'ensemble des pages annotées.

Méthode	Moyenne	Minimum	25%	50%	75%	Maximum
Locale	0.86	0.25	0.85	0.93	0.95	0.98
Globale	0.89	0.27	0.91	0.95	0.97	0.99

Table 2. Performances sur les pages annotées contenant au moins une photo.

On remarque que dans les deux cas la méthode globale est légèrement plus performante. De plus les scores moyens sont relativement proches dans le premier cas en comparaison avec le second tableau dans lequel les pages vides ne sont pas prises en compte. On peut donc supposer que les performances de la méthode locale sur une page vide sont mauvaises (ce qui a tendance à rapidement baisser le score moyen) et donc que la méthode locale détecte de nombreux faux positifs.

Pour aller plus loin, il est intéressant d'évaluer les résultats qualitativement. Les images générées sont accessibles dans le dossier "output". Quelques observations :

Toutes les photos annotées manuellement ont été extraites (au moins en partie) avec les deux méthodes.



Figure 13. Quelques exemples d'extractions réussies (méthode globale).

Les photos extraites sont généralement mieux centrées et les bordures plus précises avec la méthode globale. Ceci s'explique du fait que la précision de la méthode locale est fortement contrainte par la taille de la fenêtre d'exploration alors que la méthode globale a des contraintes plus souples liées aux éléments structurants.



Figure 14. Résultats de l'extraction, méthode globale à droite, méthode locale à gauche. La méthode locale tend à décentrer les photos à cause de la taille de la fenêtre et du pas utilisé.

Les photos non rectangulaires sont bien reconnues et extraites. Les architectures des deux méthodes permettent de bien détecter les contours circulaires.

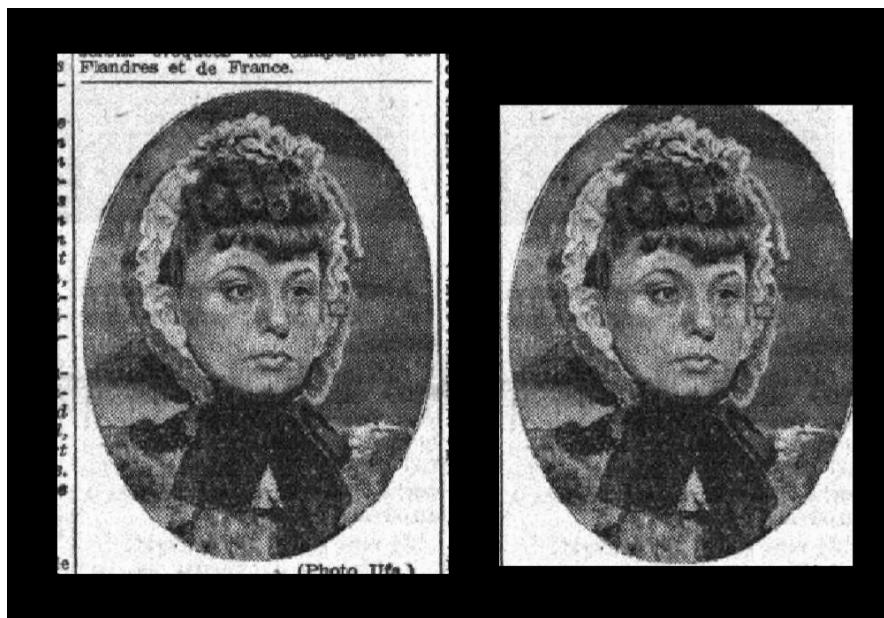


Figure 15. Résultats d'extraction d'une photo non rectangulaire, méthode globale à droite, méthode locale à gauche.

Des photos proches spatialement seront souvent extraites en un seul bloc, ceci est dû à la taille de la fenêtre pour la méthode locale et la taille des éléments structurants des filtres morphologiques pour la méthode globale.



Figure 16. Résultats de l'extraction. Les photos spatialement proches sont considérées comme une seule et même photo par les deux méthodes.

Les deux méthodes détectent mal les zones de demi-teinte sombres ce qui entraîne parfois l'extraction plusieurs zones d'extraction pour une même photo, ou simplement une zone plus petite.

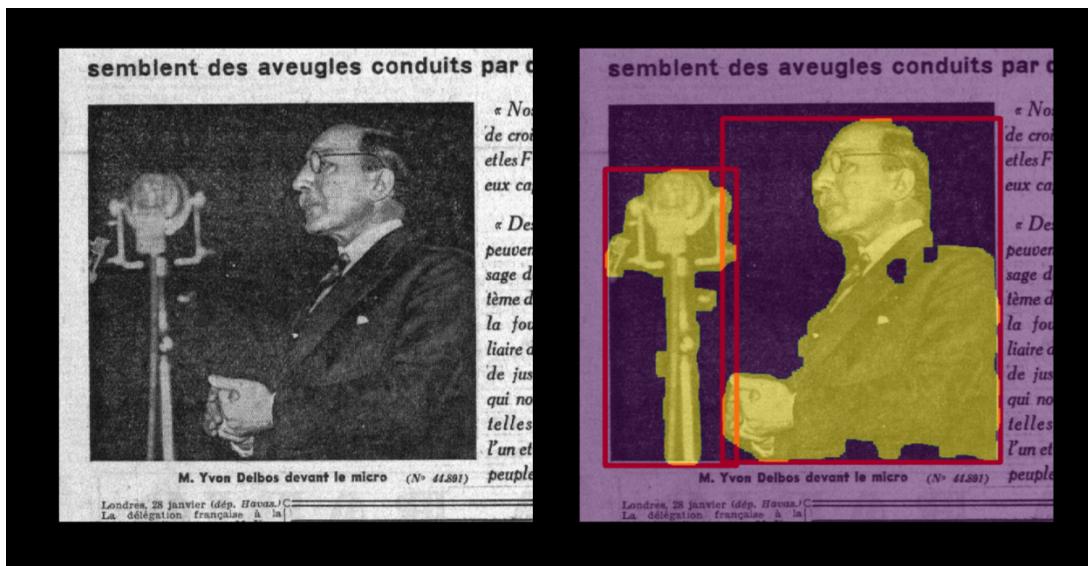


Figure 17. Processus de segmentation (méthode globale). Les zones sombres de la photo ne sont pas bien détectées, deux composantes connexes sont détectées pour une seule photo. Les cadres rouges représentent les régions extraites.

Les faux positifs sont souvent des illustrations. Certaines parties d'illustrations sont imprimées en demi-teinte et sont donc détectées comme des photos.

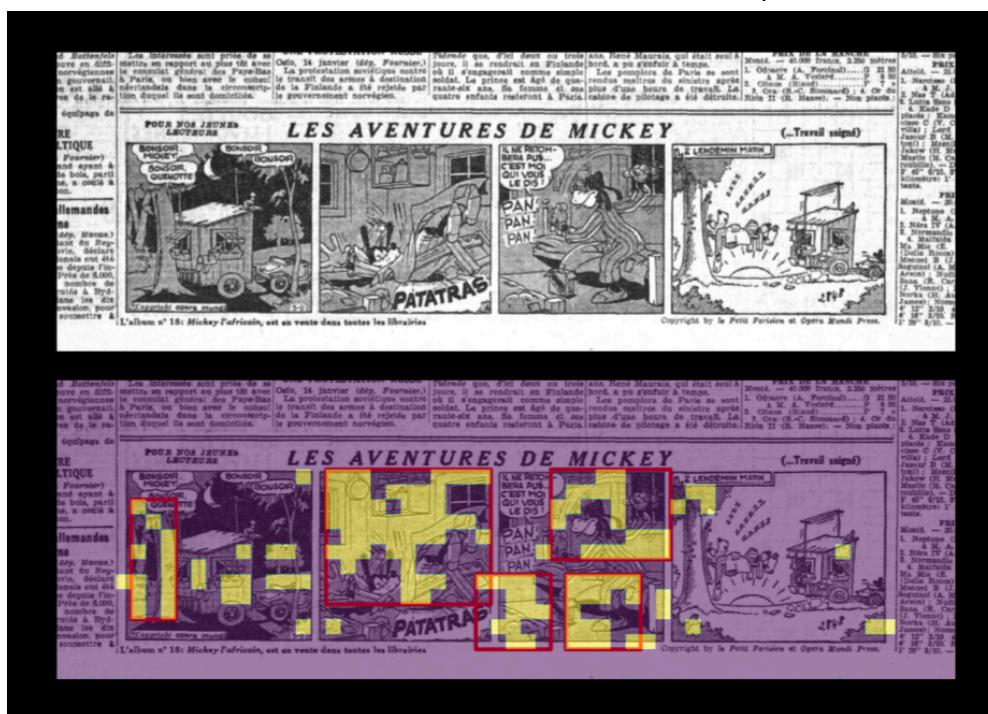


Figure 18. Processus de segmentation (méthode locale). Certaines zones d'illustrations imprimées en demi-teinte sont détectées comme des photos. Les cadres rouges représentent les régions extraites.

3.6 Résultats de l'extraction

Le processus d'extraction de l'ensemble de l'archive a été réalisé avec la méthode globale car elle présente de meilleures performances sur les données de test. Le processus complet (détection et extraction des photos) a duré 80 minutes. En tout 2043 photos ont été détectées. Le graphe ci-dessous représente le nombre de détection par page dans l'ordre chronologique. On observe un creux entre 170 et 250 environ durant lequel peu (voir pas) de photos ont été publiées...

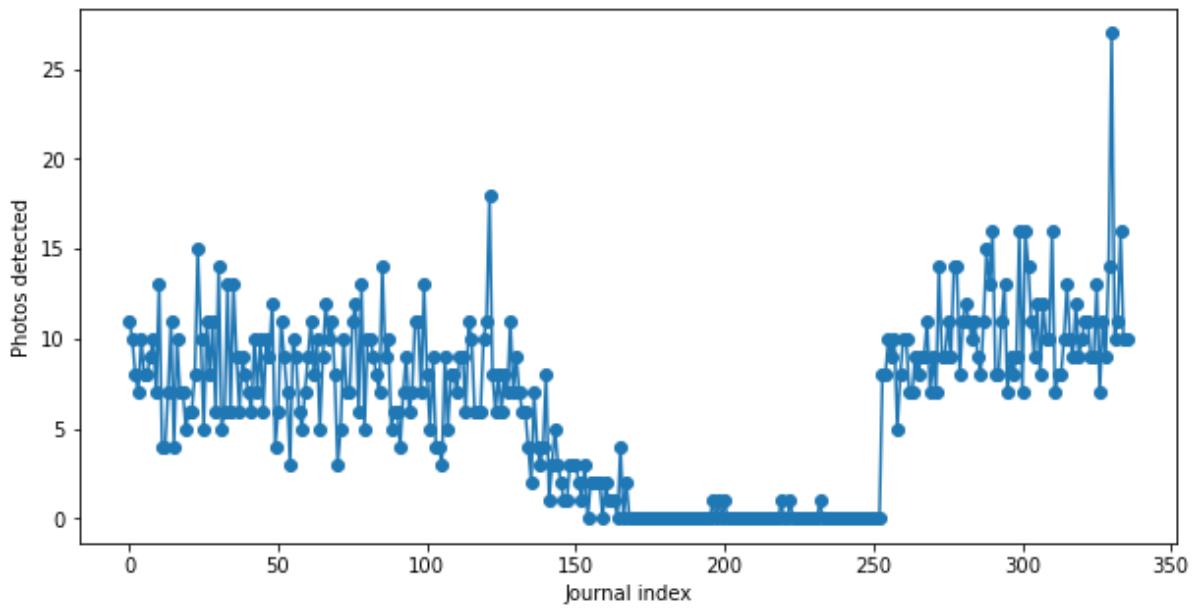


Figure 19. Nombre de photos détectées par journal. On remarque un creux de l'index 170 à 250 environ.

4 Discussion: limites et améliorations

Précision des contours

On a vu précédemment que des photos proches spatialement sont considérées comme un unique bloc pour l'extraction. Dans le cas de la méthode locale, cela est dû à la taille de la fenêtre de recherche et au pas. Pour la méthode globale, cela est dû aux dimensions des éléments structurants.

Ici je propose une piste de recherche pour améliorer les résultats dans le cas local. La résolution de l'extraction est naturellement limitée à la taille de la fenêtre lorsque le pas est aussi de la taille de cette fenêtre. Or on a vu qu'il est difficile de réduire la taille de la fenêtre car il devient impossible de détecter les pics fréquentiels avec une fenêtre trop petite. Néanmoins, en réduisant le pas de déplacement, la carte de demi-teinte peut saisir plus de subtilité dans les contours. Ci-dessous est illustrée la carte de demi-teinte en 3D avec un pas d'une moitié de fenêtre.

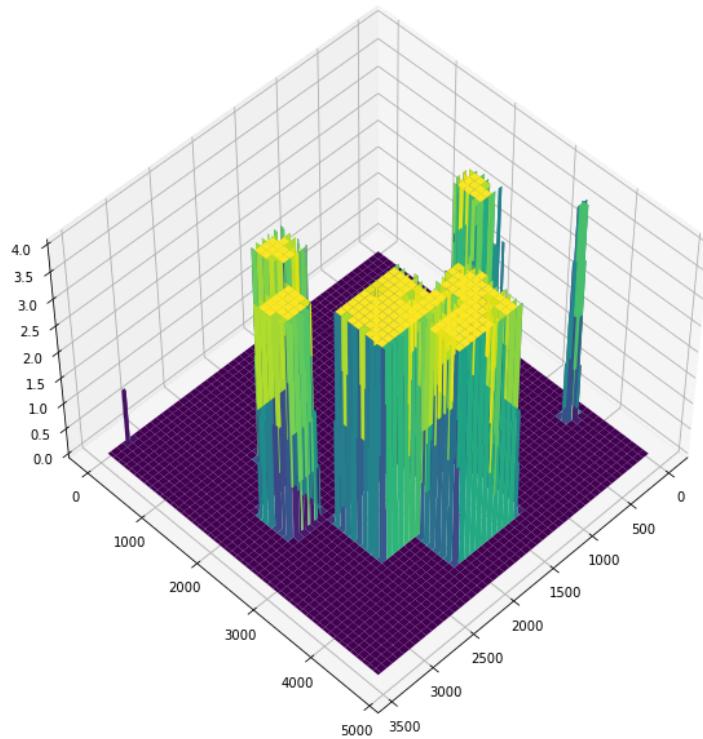


Figure 20. Carte de demi-teinte vue en 3D, construite avec superpositions des fenêtres (taille de pas inférieur à la taille de la fenêtre) pour la méthode locale.

Il est maintenant possible de séparer les zones connectées avec watershed par exemple et ainsi éviter de connecter deux images entre elles. Cette direction de recherche peut aussi permettre d'avoir des contours plus précis dans le cas général (une seule photo), la précision de l'algorithme n'est plus limitée par la taille de la fenêtre mais uniquement par le pas. Cela nécessitera naturellement plus de temps de calcul.

Choix des paramètres

Certains paramètres ont été choisis arbitrairement pour s'adapter au mieux aux données. Il serait judicieux de proposer un protocole détaillé pour choisir les meilleurs paramètres en fonction des cas, ou même automatiser la sélection des paramètres en amont.

Faux positifs

Certaines illustrations et publicités sont détectées comme des photos. La raison de cela est que la demi-teinte est aussi utilisée dans certains cas. Pour améliorer l'extraction et la limite aux photos, nous pourrions étudier plus finement les réponses fréquentielles des illustrations et publicités. Il est probable que les trames utilisées, ou d'autres éléments d'indices, soient différentes de celles pour les photos. Ainsi, nous pourrions discriminer ces différents blocs en adaptant simplement les méthodes proposées.

Détection des zones sombres

L'analyse qualitative des résultats a montré que les zones de demi-teintes sombres sont mal détectées. Ceci est vraisemblablement dû au manque de contraste dans ces zones. Je fais la supposition que les zones très claires donnent des réponses similaires. Dans notre cas, cela n'a affecté qu'en partie les performances de l'algorithme car la grande majorité des photos sont construites de telle sorte que les éléments de la photo (par exemple les personnes ou les objets) soient visibles: les photographes évitent la sous-exposition (trop sombre) ou la sur-exposition (trop clair).

5 Conclusion

Deux méthodes analytiques pour détecter et extraire les photos publiées dans le Petit Parisien au cours de l'année 1940 ont été proposées et évaluées. Les deux approches se basent sur la détection de zones de demi-teinte sur les pages numérisées. L'évaluation montre que l'approche globale donne de meilleurs résultats que l'approche locale. L'ensemble des pages publiées en 1940 ont été traitées. Des limites et pistes d'améliorations sont discutées en dernière partie du rapport.

6 Annexe

Le projet est disponible sur <https://github.com/LUKALAVAL/photo-extraction>

7 Bibliographie

- Acharyya, M., & Kundu, M. K. (2002). Document image segmentation using wavelet scale-space features. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(12), 1117–1127. <https://doi.org/10.1109/TCSVT.2002.806812>
- Alexander, D. (n.d.). *Pyllica*. GitHub. Retrieved January 25, 2025, from <https://github.com/Dorialexander/Pyllica>
- Bloomberg, D. S. (1991). Multiresolution Morphological Approach to Document Image Analysis. *Proc. of the International Conference on Document Analysis and Recognition*.
- Bukhari, S. S., Shafait, F., & Breuel, T. M. (2011). *Improved document image segmentation algorithm using multiresolution morphology* (G. Agam & C. Viard-Gaudin, Eds.; p. 78740D). <https://doi.org/10.1117/12.873461>
- Chingyu Yang, J., & Tsai, W.-H. (1998). Suppression of moiré patterns in scanned halftone images. *Signal Processing*, 70(1), 23–42. [https://doi.org/10.1016/S0165-1684\(98\)00111-X](https://doi.org/10.1016/S0165-1684(98)00111-X)
- Dutta, A., & Zisserman, A. (2019). The VIA Annotation Software for Images, Audio and Video. *Proceedings of the 27th ACM International Conference on Multimedia*, 2276–2279. <https://doi.org/10.1145/3343031.3350535>
- Halimi, A. (2024). *Identification de la Circulation d'Images Historiques* [Rapport de stage (M1)]. LIP6, Sorbonne Université.
- Hidayati, S. C., Che-Hao Hsu, Shih-Wei Sun, Wen-Huang Cheng, & Kai-Lung Hua. (2015). An efficient algorithm for periodic halftone identification. *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 1–5. <https://doi.org/10.1109/ICMEW.2015.7169843>
- Lee, B. C. G., Mears, J., Jakeway, E., Ferriter, M., Adams, C., Yarasavage, N., Thomas, D., Zwaard, K., & Weld, D. S. (2020). The Newspaper Navigator Dataset: Extracting Headlines and Visual Content from 16 Million Historic Newspaper Pages in Chronicling America. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 3055–3062. <https://doi.org/10.1145/3340531.3412767>
- Lombardi, F., & Marinai, S. (2020). Deep Learning for Historical Document Analysis and Recognition—A Survey. *Journal of Imaging*, 6(10), 110. <https://doi.org/10.3390/jimaging6100110>
- ModOAP. (n.d.-a). *ModOAP_Telechargement_documents_Gallica-IIIF-v3.ipynb*. Google Colab. Retrieved January 25, 2025, from <https://colab.research.google.com/drive/1aGm1eNGXcQpYo-GiZo3DMZ8bmaHVscJ2#scrollTo=hQ2QjFUXJpMz>

- ModOAP. (n.d.-b). *Telechargement de périodiques sur Gallica*. GitHub. Retrieved January 25, 2025, from <https://github.com/MODOAP/Telechargement-documents-Gallica/tree/main/tel echargevement-periodiques>
- Ok, J., Han, S. W., Jarno, M., & Lee, C. (2011). *Low complexity pixel-based halftone detection*. 81570N. <https://doi.org/10.1117/12.895089>
- Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 658–666. <https://doi.org/10.1109/CVPR.2019.00075>
- Schade, O., & Bloch, I. (2024). *PRAT Project Report*.
- Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., & Li, W. (2021). *LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis* (No. arXiv:2103.15348). arXiv. <http://arxiv.org/abs/2103.15348>
- Tholeti, T., Ganesh, P., & Ramanujam, P. (2015). Frequency domain filtering techniques of halftone images. *2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN)*, 427–430. <https://doi.org/10.1109/SPIN.2015.7095255>
- Wang, D., & Srihari, S. N. (1989). Classification of newspaper image blocks using texture analysis. *Computer Vision, Graphics, and Image Processing*, 47(3), 327–352. [https://doi.org/10.1016/0734-189X\(89\)90116-3](https://doi.org/10.1016/0734-189X(89)90116-3)
- Won, C. S. (2008). Image extraction in digital documents. *Journal of Electronic Imaging*, 17(3), 033016. <https://doi.org/10.1117/1.2970151>