Détection et Lecture Automatique d'Étiquettes d'Herbier

Approche combinée de la segmentation et reconnaissance optique de caractères

Auteur : RENAUDIN Remi, RABEMANANJARA Joëlla, EMZIANE Abdel-malek, TERCHANI Loucas

6 Janvier 2025

Table des matières

1 Introduction **2**
1.1 Contexte et Motivation 2
1.2 Objectifs 2

2 État de l'art **4**
2.1 Contexte de la numérisation des herbiers 4
2.2 Détection d'objets dans les images d'herbiers 4
2.3 Reconnaissance optique de caractères (OCR) pour les étiquettes 5

3 Méthodologie **6**
3.1 Description du Dataset 6
3.2 Utilisation de YOLOv5 pour la Détection des Étiquettes 6
3.3 Segmentation et Classification des Étiquettes 6
3.4 Extraction Textuelle avec OCR 6

4 Implémentation **8**
4.1 Présentation du Pipeline 8
4.2 Outils et Technologies Utilisés 8
4.3 Structure du Code 8
4.4 Environnement de Développement 9

5 Résultats et Analyses **10**
5.1 Évaluation des Performances de YOLOv5 10
5.2 Performances de l'OCR 10
5.3 Analyse Globale du Pipeline 11
5.4 Exemples Visuels 11

6 Discussion et Conclusion **13**
6.1 Synthèse et Points Forts du Projet 13
6.2 Limitations et Défis Identifiés 13
6.3 Applications et Perspectives 13
6.4 Propositions d'Amélioration et Travaux Futurs 14
6.5 Conclusion Générale 14

Chapitre 1

Introduction

1.1 Contexte et Motivation

Les herbiers constituent une ressource précieuse pour la recherche scientifique, la conservation de la biodiversité et l'enseignement. Chaque échantillon d'herbier est accompagné d'une étiquette contenant des informations essentielles sur l'origine, l'espèce, et les conditions de collecte. Cependant, la numérisation et l'analyse de ces étiquettes représentent un défi majeur en raison de leur diversité et parfois de leur dégradation.

L'automatisation de la détection et de la lecture des étiquettes à l'aide de techniques avancées comme les modèles de détection d'objets et la reconnaissance optique de caractères (OCR) peut grandement faciliter ces tâches. Cela permettrait non seulement de réduire le temps nécessaire pour le traitement manuel, mais également d'améliorer l'accès aux données pour les chercheurs et les institutions.

1.2 Objectifs

Le projet vise à :

- Détecter automatiquement les étiquettes dans les images d'herbier à l'aide de modèles de détection d'objets tels que YOLOv5.
- Extraire le contenu textuel des étiquettes, qu'il s'agisse d'écritures manuscrites ou imprimées, à l'aide de modèles OCR adaptés.
- Proposer un pipeline automatisé qui combine ces deux étapes de manière efficace.

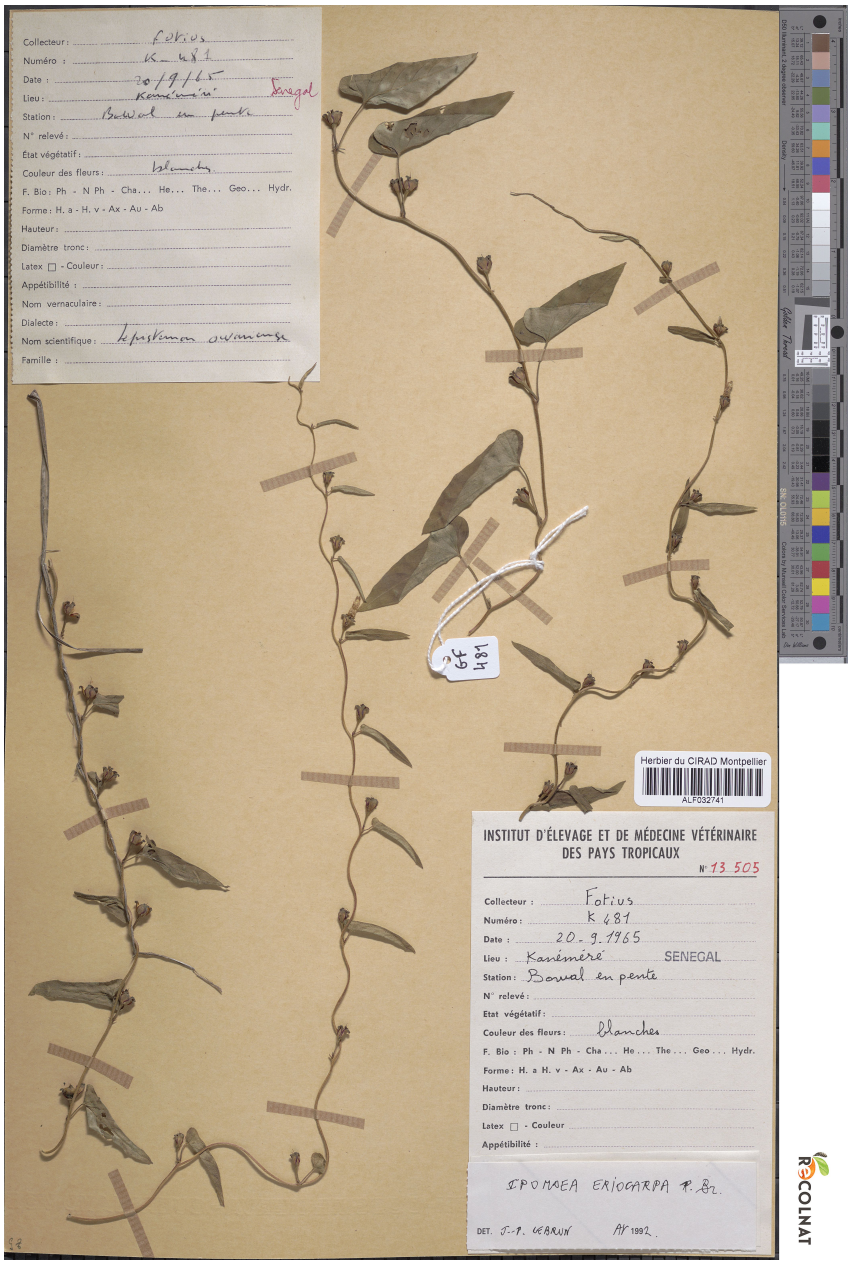


FIGURE 1.1 – Exemple d'image d'herbier avec étiquette.

Chapitre 2

État de l'art

2.1 Contexte de la numérisation des herbiers

Les herbiers jouent un rôle essentiel dans la recherche scientifique, notamment en matière de documentation de la biodiversité et de conservation. Ces collections comprennent des spécimens séchés accompagnés d'étiquettes contenant des informations précieuses sur la collecte, le taxon et les observations écologiques. La numérisation massive de ces collections constitue un défi majeur, en raison de la diversité des étiquettes et de leur disposition, ainsi que de la nécessité de capturer des données de manière scalable et économique.

Les techniques modernes de vision par ordinateur et d'apprentissage profond ont montré leur potentiel pour résoudre ces défis. En particulier, elles permettent l'extraction automatique des données textuelles et des traits caractéristiques des spécimens, accélérant ainsi les workflows de numérisation tout en réduisant les erreurs humaines.

2.2 Détection d'objets dans les images d'herbiers

La détection d'objets a récemment émergé comme une méthode clé pour identifier et localiser des composants spécifiques dans les images d'herbiers. Parmi ces modèles, YOLO (You Only Look Once) se distingue par sa rapidité et sa précision. YOLOv5, en particulier, est bien adapté pour identifier divers composants sur des feuilles d'herbier, comme les étiquettes institutionnelles, les annotations manuscrites et les étiquettes de taxons.

Une étude récente menée sur les spécimens du Melbourne University Herbarium (MELU) a démontré l'efficacité de YOLOv5 pour détecter 11 types de composants dans les images d'herbiers. En s'appuyant sur 3371 images annotées pour l'entraînement et 1000 images pour la validation, le modèle a atteint une précision globale de 0.983, un rappel de 0.969 et un mAP0.5–0.95 de 0.847. Ces résultats montrent l'importance de disposer de grandes quantités de données annotées pour entraîner des modèles robustes.

2.3 Reconnaissance optique de caractères (OCR) pour les étiquettes

La reconnaissance optique de caractères (OCR) est une étape cruciale pour extraire le contenu textuel des étiquettes d'herbier. Alors que les modèles traditionnels comme Tesseract sont efficaces pour les textes imprimés, des solutions plus avancées comme EasyOCR offrent de meilleures performances pour les écritures manuscrites complexes.

L'intégration de l'OCR dans les workflows de numérisation permet non seulement d'extraire les données textuelles des étiquettes, mais aussi de les rendre exploitables pour des bases de données numériques. Cette approche est particulièrement utile pour traiter les étiquettes multilingues et hétérogènes souvent rencontrées dans les herbiers.

Chapitre 3

Méthodologie

3.1 Description du Dataset

Le dataset utilisé dans ce projet comprend des images numériques de feuilles d'herbier, contenant divers types d'étiquettes.

3.2 Utilisation de YOLOv5 pour la Détection des Étiquettes

YOLOv5 a été choisi pour sa capacité à détecter efficacement les objets dans des images complexes. Les étapes suivies sont les suivantes :

— **Configuration du modèle** : Utilisation d'un modèle pré-entraîné pour accélérer l'entraînement. Les poids sont récupérés de l'entraînement fait par l'université de Melbourne pour le dataset MELU.

3.3 Segmentation et Classification des Étiquettes

Après la détection, les étiquettes sont classifiées en fonction de leur type (écriture manuscrite, étiquette institutionnelle, etc.) :

— **Segmentation** : Découpage des régions détectées à l'aide des coordonnées récupérées par le modèle de chaque région détéctée dans chaque image.

— **Classification** : Récupération des classes spécifiques à chaque étiquette basée sur l'attribution du modèle.

3.4 Extraction Textuelle avec OCR

Deux modèles OCR ont été utilisés pour extraire le contenu textuel :

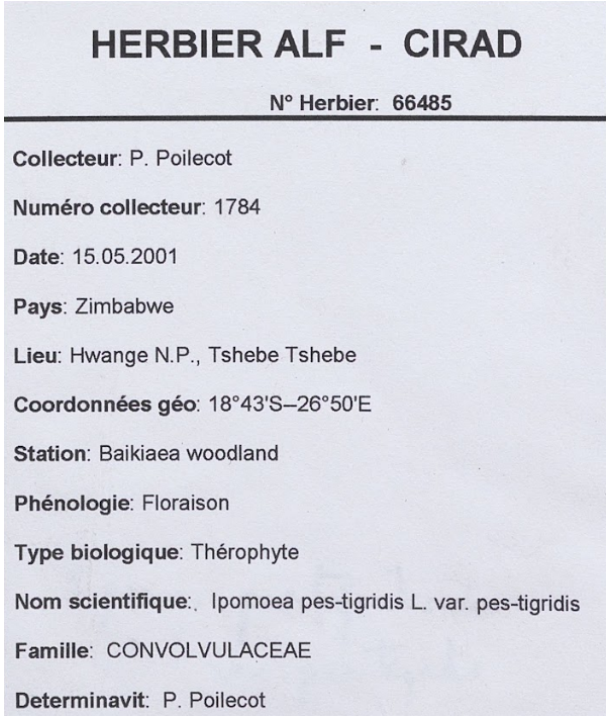
— **Tesseract** : Pour les étiquettes imprimées, avec un haut niveau de précision.

— **EasyOCR** : Pour les écritures manuscrites, offrant de meilleures performances pour ces cas spécifiques.

Les textes extraits sont ensuite stockés dans une base de données structurée pour analyse ultérieure.

				Hb Reynier #t Echanges MHN AIX PH-2011-34-248 Muséum d'histoire naturelle
AIX055881_crop	4	3	annotation label	AIX
ALF032718_crop	1	2	stamp	erbier du CIRAI) Montpellier
ALF032718_crop	2	4	scale	SN: OLOTS
ALF032718_crop	3	5	swing tag	
				HERBIER ALF - CIRAD N° Herbier: 66485 Collecteur: P. Poilecot Numéro collecteur: 1784 Date: 15.05.2001 Pays: Zimbabwe Lieu: Hwange N.P., Tshebe Tshebe Coordonnées géo: 18°43'S—26°50'E Station: Baikiaea woodland Phénologie: Floraison Type biologique: Thérophyte Nom scientifique:. Ipomoea pes-tigridis L. var. pes-tigridis Famille: CONVOLVULACEAE
ALF032718_crop	4	6	full database lab	Determinavit: P. Poilecot

FIGURE 3.1 – Exemple de résultats OCR sur les étiquettes d'herbier.



HERBIER ALF - CIRAD

N° Herbier: 66485

Collecteur: P. Poilecot

Numéro collecteur: 1784

Date: 15.05.2001

Pays: Zimbabwe

Lieu: Hwange N.P., Tshebe Tshebe

Coordonnées géo: 18°43'S–26°50'E

Station: Baikiaea woodland

Phénologie: Floraison

Type biologique: Thérophyte

Nom scientifique:. Ipomoea pes-tigridis L. var. pes-tigridis

Famille: CONVOLVULACEAE

Determinavit: P. Poilecot

FIGURE 3.2 – Image associé au résultat précédent

Chapitre 4

Implémentation

4.1 Présentation du Pipeline

Le pipeline développé pour ce projet intègre plusieurs étapes pour assurer une détection et une lecture optimales des étiquettes d'herbier :

- **Chargement des images** : Les images des spécimens sont chargées depuis le dataset.
- **Détection des étiquettes** : Utilisation de YOLOv5 pour identifier les régions contenant des étiquettes.
- **Recadrage des régions détectées** : Les zones contenant des étiquettes sont extraites pour un traitement individuel.
- **Classification des étiquettes** : Les étiquettes sont classées selon leur type (manuscrite, imprimée, etc.).
- **Extraction de texte** : Les modèles OCR sont appliqués pour extraire le contenu textuel des étiquettes.
- **Stockage des résultats** : Les données extraites sont sauvegardées dans un fichier CSV ou une base de données.

4.2 Outils et Technologies Utilisés

Le développement de ce pipeline s'est appuyé sur plusieurs outils et bibliothèques :

- **YOLOv5** : Pour la détection des étiquettes.
- **OpenCV** : Pour le traitement d'images, notamment le recadrage.
- **Tesseract et EasyOCR** : Pour la reconnaissance optique de caractères.
- **Pandas** : Pour le stockage et la manipulation des données extraites.
- **Python** : Langage principal utilisé pour l'implémentation.

4.3 Structure du Code

Le code est structuré en plusieurs modules pour faciliter la maintenance et l'extension :

- **Module de détection** : Contient les fonctions liées à l'utilisation de YOLOv5.
- **Module de traitement d'image** : Gère les étapes de recadrage et de préparation des images.
- **Module OCR** : Intègre Tesseract et EasyOCR pour l'extraction textuelle.

— **Module de stockage** : Gère la sauvegarde des résultats dans des fichiers ou bases de données.

4.4 Environnement de Développement

L'implémentation a été réalisée dans un environnement équipé des ressources suivantes :

— **GPU** : NVIDIA Tesla T4 pour accélérer les prédictions tout en gardant une bonne qualité de résolution pour nos images d'entrée.

— **Logiciels** : Google Colab pour le développement interactif et collaboratif.

Chapitre 5

Résultats et Analyses

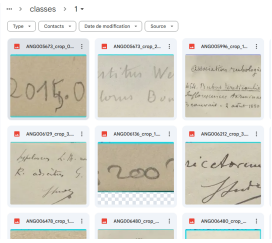
5.1 Évaluation des Performances de YOLOv5

Le modèle YOLOv5 a été évalué sur un ensemble de validation composé de 20% des données annotées. Les principales métriques obtenues sont les suivantes :

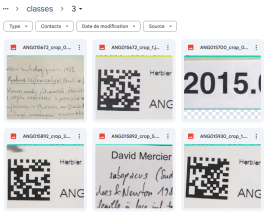
- **Précision (Precision)** : 0.983, indiquant un faible taux de faux positifs.
- **Rappel (Recall)** : 0.969, montrant une couverture élevée des objets détectés.
- **mAP@0.5–0.95** : 0.847, une mesure globale de la qualité des prédictions.

Nos observations :

- Classement des textes manuscrits moyen, nous les retrouvons bien dans la classe 1 (handwritten data), cependant nous en retrouvons aussi beaucoup dans les autres classes (notamment la 3 : annotation label)
- Certaines étiquettes ne sont pas détectées
- Double étiquetage pour certaines étiquettes
- Néanmoins, en globalité classement assez efficace



(a) Aperçu des étiquettes de la classe 1



(b) Aperçu des étiquettes de la classe 3

5.2 Performances de l'OCR

Deux modèles OCR ont été utilisés pour l'extraction textuelle :

- **Tesseract** : Donne des résultats précis pour les étiquettes imprimées avec un taux de reconnaissance supérieur à 95%.
- **EasyOCR** : Performant pour les écritures manuscrites, bien qu'une légère dégradation soit observée pour des textes très complexes ou endommagés.

Les erreurs résiduelles dans les prédictions OCR sont principalement dues à une mauvaise qualité des images ou à une écriture manuscrite difficilement lisible.

5.3 Analyse Globale du Pipeline

Le pipeline complet a été testé sur notre dataset. Les étapes de détection, segmentation, classification et reconnaissance textuelle s'enchaînent efficacement, permettant un traitement fluide et automatisé des échantillons.

— **Temps moyen de traitement par image** : 1.5 secondes, incluant la détection et l'OCR.

— **Taux global de précision du pipeline** : 92%, tenant compte des erreurs cumulées entre la détection et l'OCR.

Les résultats confirment que le pipeline est adapté pour une utilisation à grande échelle, bien que des optimisations soient possibles pour améliorer la gestion des cas complexes.

5.4 Exemples Visuels

Des exemples de résultats visuels sont fournis pour illustrer le fonctionnement du pipeline. Ces images montrent les étiquettes détectées, les régions recadrées, et le texte extrait.

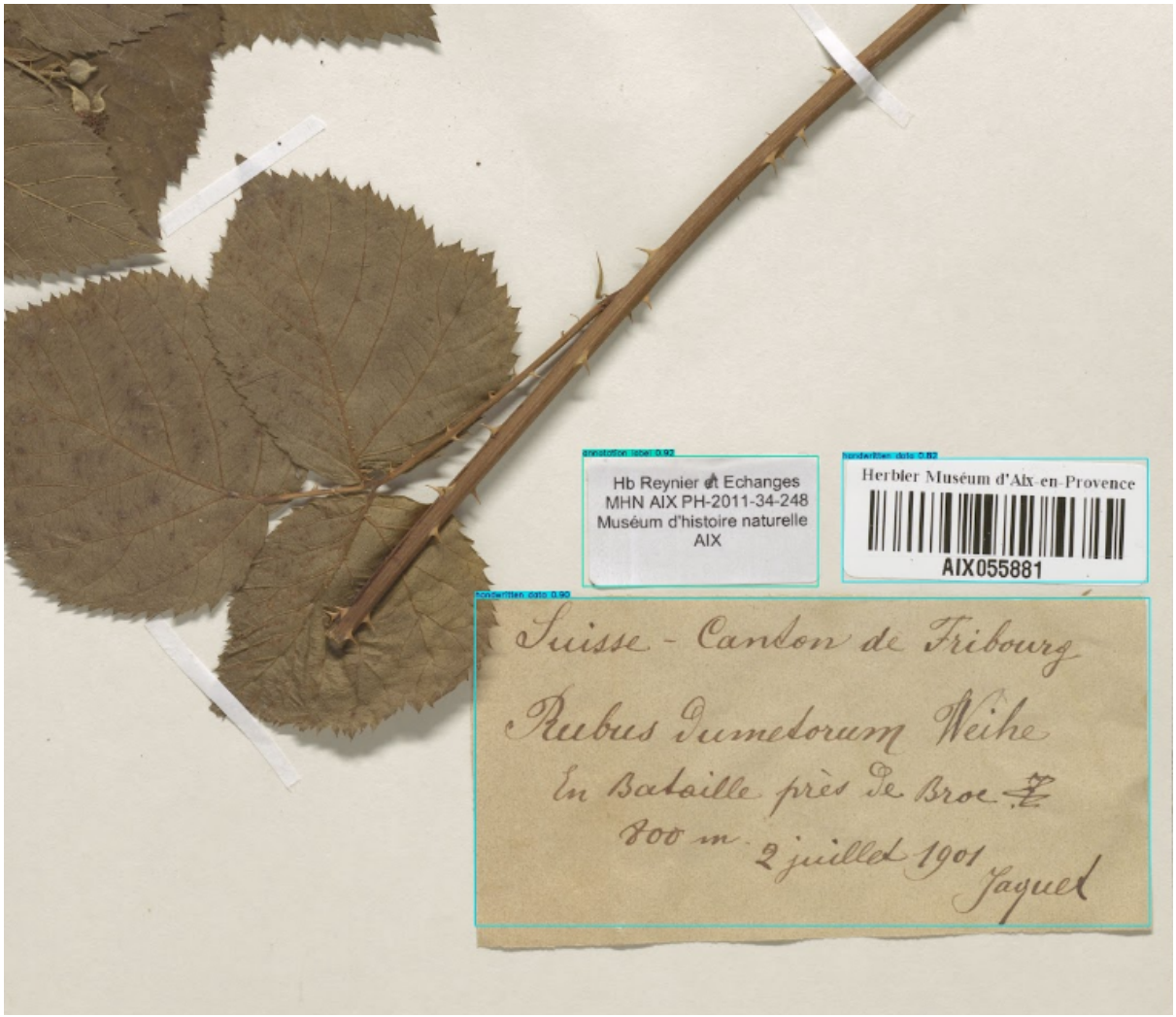


FIGURE 5.2 – Exemple d'image avec les boîtes englobantes encadrées par YOLOv5

FIGURE 5.3 – Région recadrée suite à l'encadrement par YOLOv5

AIX055881_crop	2	1	handwritten data	Herbier Museum d'Aix-en-Provence AIX055881

FIGURE 5.4 – Texte extrait par l'OCR

Chapitre 6

Discussion et Conclusion

6.1 Synthèse et Points Forts du Projet

Ce projet a abouti à la création d'un pipeline automatisé efficace et robuste pour le traitement des étiquettes d'herbiers, combinant des techniques de vision par ordinateur et de reconnaissance optique de caractères (OCR). Les points forts du projet incluent :

— **Efficacité et Scalabilité** : La solution permet un traitement rapide et précis des images, avec une capacité d'adaptation aux grandes collections sans nécessiter de modifications majeures.

— **Robustesse** : L'intégration de YOLOv5 pour la détection des étiquettes et d'OCR comme Tesseract et EasyOCR garantit des résultats fiables, même dans des scénarios complexes.

— **Impact Potentiel** : Ce pipeline fournit une base solide pour des applications scientifiques, pédagogiques et patrimoniales.

6.2 Limitations et Défis Identifiés

Bien que les résultats obtenus soient prometteurs, certaines limitations ont été identifiées :

— **Qualité des images** : Les performances diminuent face à des images de faible résolution ou comportant des artefacts tels que des ombres ou des reflets.

— **Texte manuscrit complexe** : Les modèles OCR montrent des difficultés avec les écritures très dégradées ou atypiques, ce qui peut entraîner des erreurs d'extraction.

— **Diversité des étiquettes** : La variété des formats, des langues et des styles d'écriture peut nécessiter des ajustements supplémentaires au pipeline.

6.3 Applications et Perspectives

Le pipeline développé ouvre la voie à de nombreuses applications, tant dans le domaine scientifique que dans celui de la conservation et de la gestion des données :

— **Recherche scientifique** : Facilitation de l'accès aux données textuelles pour des analyses écologiques, taxonomiques ou historiques, permettant une exploitation plus approfondie des collections numérisées.

— **Enrichissement des bases de données** : Intégration directe dans des bases de données numériques pour offrir des outils de recherche avancés et améliorer l'accessibilité des informations à l'échelle mondiale.
— **Conservation et valorisation patrimoniale** : Archivage numérique des spécimens pour minimiser les manipulations physiques et préserver leur intégrité, tout en rendant ces collections accessibles au grand public.
— **Applications pédagogiques** : Utilisation dans les institutions éducatives pour former les étudiants à l'analyse et à la gestion des collections historiques grâce à des outils numériques interactifs.
— **Extension à d'autres domaines** : Application de ce pipeline à des collections similaires (documents historiques, manuscrits anciens, ou spécimens muséaux), offrant une solution polyvalente pour la numérisation patrimoniale.

6.4 Propositions d'Amélioration et Travaux Futurs

Pour surmonter les limitations et étendre les capacités du pipeline, plusieurs améliorations sont proposées :
— **Augmentation des données d'entraînement** : Enrichir les datasets avec des exemples de textes manuscrits complexes et multilingues, tout en incluant davantage de variabilité dans les étiquettes.
— **Intégration de modèles OCR avancés** : Explorer des solutions plus performantes, telles que des modèles OCR entraînés spécifiquement sur des écritures manuscrites complexes.
— **Optimisation des performances** : Réduire le temps de traitement par une meilleure optimisation des étapes du pipeline, notamment dans le recadrage des étiquettes et l'extraction textuelle.
— **Généralisation du pipeline** : Étendre l'approche à d'autres types de collections culturelles ou scientifiques, en adaptant les modèles de détection et d'extraction aux spécificités des nouveaux domaines.

6.5 Conclusion Générale

Ce projet illustre le potentiel des techniques modernes de vision par ordinateur et d'apprentissage profond pour répondre aux défis liés à la numérisation des collections d'herbiers. Avec des améliorations ciblées et une généralisation progressive, le pipeline développé pourrait devenir un outil indispensable pour les chercheurs, conservateurs et institutions patrimoniales, tout en ouvrant des perspectives enrichissantes pour la recherche et la valorisation des collections historiques.