MP – Système d'Extraction d'Informations de Documents d'Identité

Dans le contexte de la transformation numérique, de nombreuses organisations doivent traiter automatiquement des documents d'identité (passeports, cartes d'identité) pour des applications de vérification, d'enregistrement ou de contrôle d'accès. Le traitement manuel de ces documents est chronophage et sujet aux erreurs.

Objectifs

Base de données idéale pour stocker des documents de recettes avec des structures

Développer un système complet d'extraction d'informations utilisant :

- Détection et localisation intelligente des champs d'information via des réseaux de neurones convolutionnels
- La reconnaissance optique de caractères (OCR) pour extraire le texte
- L'apprentissage automatique pour classifier et valider les informations
- Une interface de visualisation des résultats

Fonctionnalités

Préparation des données :

- Téléchargez un dataset de documents d'identité depuis les ressources fournies
- Annotez manuellement +100 images minimum
- Divisez votre dataset en train/validation/test (70/15/15)
- Convertissez les annotations au format compatible avec votre modèle d'IA

Champs à détecter :

- Nom/Surname
- Prénom/Given Name
- Date de naissance
- Numéro de document
- Nationalité
- Date d'expiration

- Lieu de naissance
- Sexe

Détection

- Implémentez un système de détection d'objets avec TensorFlow Object Detection API
- Entraînez le modèle pour localiser automatiquement les champs d'information
- Utilisez le transfer learning sur modèles pré-entraînés
- Optimisez les performances avec validation croisée

Extraction OCR

- Implémentez un système OCR avec EasyOCR ou Keras-OCR pour lire le texte des documents
- Prétraitement des images pour améliorer la qualité
- Extraction du texte avec scores de confiance
- Gestion de textes multilingues (anglais, français)

Classification et Structuration

- Classification automatique des textes extraits avec réseaux de neurones
- Validation des formats (dates, codes pays, etc.) par apprentissage automatique
- Export des données au format JSON et CSV
- Calculez un score de confiance global pour évaluer la fiabilité de l'extraction

Interface de Visualisation

Affichage des images avec bounding boxes colorées

Visualisation des résultats d'extraction

Interface pour corriger manuellement les erreurs

Génération de rapports de traitement

Fonctionnalités Optionnelles

- Traitement par lot de multiples documents
- API REST pour l'intégration
- Interface web avec Flask/Django
- Support de formats de documents supplémentaires
- Détection de documents falsifiés
- Export vers base de données

Livrables Attendus

- Dataset annoté avec au minimum 100 images et Modèle entrainé
- Rapport d'analyse des performances
- Démonstration du système
- Présentation orale (15-20 minutes)

Ressources

Datasets

Kaggle - Generated USA Passports :
https://www.kaggle.com/datasets/tapakah68/generated-usa-passeports-dataset

Documentation Technique

https://tensorflow-object-detection-api-tutorial.readthedocs.io/en/latest/

https://tensorflow-object-detection-api-tutorial.readthedocs.io/

https://keras.io/

https://github.com/JaidedAI/EasyOCR

OpenCV Python Tutorials : https://opencv-python-tutroals.readthedocs.io/

Outils d'Annotation

https://cvat.org/

https://roboflow.com/

https://github.com/tzutalin/labelImg

Tutoriels et Guides

TensorFlow Object Detection Tutorial : https://tensorflow-object-detection-api-tutorial.readthedocs.io/

OCR with Python : https://towardsdatascience.com/optical-character-recognition-ocr-with-python-and-easyocr-4f6e5c42de4b