Vol.9 Iss.1 pp.1-10 International Journal of Computer Science, Communication & Information Technology (CSCIT)

VLM, OCR et Vision par Ordinateur au Service de la Reconnaissance Client et de l'Extraction d'Informations en Langue Arabe

Hana BenAli[#1], Ines Abdeljaoued-Tej[#1,2], Mounir Melliti[*3]

[1]*Université de Carthage, ESSAI, Ariana, Tunisie*

[1]hana.ben.ali@ssai.ucar.tn

[2]*Laboratoire BIMS (LR16IPT09), Institut Pasteur de Tunis, Université Tunis-El-Manar, Tunis, Tunisie*

[2]ines.tej@essai.ucar.tn

[3]*Orange Tunisie, Avenue du Cheikh Mohamed Fadhel Ben Achour, Centre Urbain Nord, 1003 Tunis, Tunisie*

[3]mounir.melliti@orange.tn

ISSN: 1737-930X

***Abstract*— Le processus de vente de cartes SIM dans les boutiques des opérateurs de télécommunications nécessite la vérification d'un document d'identité. Nous avons développé une preuve de concept (POC) pour la vérification des cartes d'identité nationales tunisiennes (CIN) et la reconnaissance faciale dans les télécommunications. En numérisant entièrement le processus d'achat, nous visons à créer un processus de commande en ligne rationalisé et efficace. Le modèle YOLOv8n s'est avéré très performant dans l'identification et la localisation des caractéristiques du document pour la détection des éléments clés de la CIN. Nous avons entraîné ce modèle à sélectionner spécifiquement la photo d'identité. Nous l'avons ensuite comparé à une image capturée en temps réel, en utilisant FaceNet512 en conjonction avec le détecteur RetinaFace. Pour permettre le remplissage automatique des formulaires de contrat d'achat de carte SIM, nous avons utilisé le modèle de langage visuel (VLM) PaliGemma pour extraire le texte arabe**

de la CIN. Notre solution combine des techniques avancées de vision par ordinateur pour garantir une expérience fluide et conviviale. Le produit final est une application web Flask qui fournit une preuve de concept robuste (POC) pour un processus efficace de connaissance du client (KYC), intégrant la détection d'objets, la reconnaissance optique de caractères (OCR) combiné à un modèle de langage de vision et la reconnaissance faciale.

***Keywords*— Vision par Ordinateur ; Optical Character Recognition (OCR) ; Reconnaissance Faciale ; Know Your Customer (KYC) ; Image Processing, Digitalisation**

I. INTRODUCTION

Nous vivons à l'ère de l'intelligence artificielle (IA) générative (GenAI), où des modèles avancés comme les transformateurs révolutionnent divers domaines, de la génération de texte à la vision par ordinateur[1]. Ces outils permettent de créer des solutions de pointe qui surpassent les méthodes traditionnelles par leur précision, leur rapidité et leur adaptabilité. Dans ce contexte, il devient impératif d'utiliser ces technologies pour rester compétitif et répondre aux attentes croissantes des clients. La digitalisation des processus commerciaux représente une avancée majeure dans l'amélioration de l'efficacité, la réduction des coûts et l'enrichissement de l'expérience client[2,3]. Dans un monde de plus en plus connecté, les entreprises cherchent continuellement à adopter des technologies de pointe pour répondre aux attentes des consommateurs modernes. Parmi ces technologies, la reconnaissance optique de caractères (OCR) et la reconnaissance faciale se distinguent par leur capacité à automatiser et sécuriser les processus d'identification et de vérification[4,5]. Les applications de l'OCR sont nombreuses et variées. Par exemple, l'OCR est utilisé pour la lecture des plaques d'immatriculation et l'automatisation de la saisie de données dans les aéroports. De nombreuses organisations et entreprises peuvent économiser du temps et de l'argent en utilisant l'OCR avec l'IA pour automatiser des tâches telles que la saisie de données, le traitement des factures et le remplissage des formulaires. L'OCR avec l'IA jouent également un rôle crucial dans les applications de lecteur d'écran, convertissant le texte en audio ou en braille pour les personnes malvoyantes[6]. De plus, ces technologies permettent la traduction automatique de documents dans plusieurs langues, comme le démontre l'application Google Lens[7]. En logistique, l'OCR est utilisé pour lire les identifiants des conteneurs de fret et maintenir un inventaire précis des conteneurs dans les installations [8]. Il est également employé pour lire les caractères sur les colis, facilitant le routage et l'acheminement des envois. Le secteur des télécommunications n'échappe pas à cette tendance. L'achat de cartes SIM, étape cruciale pour l'accès aux services téléphoniques et Internet, reste souvent un processus manuel et sujet à des erreurs.Actuellement, en Tunisie, les clients doivent fournir des documents d'identité et les conseillers remplissent manuellement des formulaires électroniques (ce qui est long et peu pratique). Cette méthode traditionnelle comporte des risques de fraude et d'erreurs humaines. Face à ces défis, nous avons proposé une solution innovante visant à digitaliser le processus d'achat des cartes SIM en utilisant des technologies avancées telles que l'OCR et la reconnaissance faciale, tout en intégrant la démarche Know Your Customer (KYC)[9].

Nous visons à simplifier et sécuriser ce processus en automatisant l'extraction des informations contenues sur les cartes d'identité nationales, tout en vérifiant l'identité du client ainsi que l'authenticité du document fourni. La solution envisagée repose sur trois modèles puissants : un modèle multimodal pour l'OCR[10], un autre modèle de pointe pour la reconnaissance faciale, ainsi qu'une technique de détection d'objets[11]. L'utilisation de ces modèles peut remplacer les approches classiques et améliorer considérablement les résultats de l'OCR, en particulier pour des langues complexes comme l'Arabe[12,13,14]. Notre objectif est de créer un premier prototype ou proof of concept (POC) du processus complet de KYC avec les cartes d'identité tunisiennes (CIN), facilitant le

ISSN: 1737-930X

remplissage automatique des contrats d'achat des cartes SIM chez un opérateur téléphonique tunisien. Pour cela, nous utilisons un modèle de langage multimodal à grande échelle pour l'extraction de mots en arabe, comme ceux étudiés dans[15]. Les modèles multimodaux avec vision (VLM) représentent une avancée révolutionnaire dans le domaine de la reconnaissance optique de caractères (OCR)[16,17]. Nous parlons de modèle de langage multimodal à grande échelle pour la vision. Ces modèles combinent les capacités visuelles et linguistiques pour améliorer la performance de la reconnaissance de texte[18]. Contrairement aux systèmes OCR traditionnels, qui se concentrent principalement sur la reconnaissance de texte à partir d'images sans contexte, les VLMs intègrent à la fois des informations visuelles et textuelles pour améliorer la précision et la robustesse du processus de lecture[19,20]. Cette intégration permet, non seulement une meilleure compréhension des images, mais aussi une extraction plus précise des informations textuelles contenues dans celles-ci. Le processus commence par la vérification de la présence de la carte, suivi de contrôles pour améliorer la précision de l'OCR, la reconnaissance faciale et se termine par l'extraction des données en langue arabe.

II. CONTEXTE

Le KYC, ou Know Your Customer, est une composante essentielle utilisée par les banques, les assurances, et d'autres entreprises nécessitant la vérification de l'identité de leurs clients. Ce processus vise à prévenir la fraude, le blanchiment d'argent, et d'autres activités illicites en s'assurant que les clients soient bien ceux qu'ils prétendent être. Dans le cadre de ce travail qui vise à digitaliser l'achat de cartes SIM depuis le domicile des clients, l'intégration d'une solution KYC fiable est cruciale. L'objectif est d'offrir une expérience fluide et sécurisée, permettant au client de finaliser son achat sans se déplacer physiquement. Par exemple, des solutions comme Trust Swiftly et Trustmatic se distinguent par des tarifs compétitifs par vérification[21], ce qui peut être avantageux pour les entreprises qui traitent de grands volumes. Cependant, l'absence de tests spécifiques pour l'OCR en Arabe peut poser problème pour celles qui doivent traiter des documents dans cette langue. D'autre part, Shufti Pro, bien que coûteux, promet des vérifications rapides et précises, ce qui pourrait convenir aux grandes entreprises nécessitant des performances élevées[22]. Toutefois, ce coût élevé pourrait être un frein pour les petites entreprises de pays comme la Tunisie. D'autres solutions comme Regula[23] présentent des limites en termes de précision dans l'extraction OCR Arabe, ce qui pourrait être problématique pour les utilisations nécessitant une extraction rapide et fiable des informations. À l'inverse, l'API Image to Text OCR offre une excellente performance à un coût très abordable, mais reste limitée par le nombre d'images pouvant être traitées par mois[24]. Pour l'utilisation de la langue arabe, nous pouvons considérer Sumsub[25]. Il y a un processus de vérification de la carte d'identité tunisienne, mais on ne peut déterminer la précision qu'après un premier test (qui n'est pas gratuit). Notons que les solutions OCR actuelles montrent des faiblesses notables, notamment dans la reconnaissance des caractères en alphabet arabe. De plus, l'utilisation d'API OCR soulève des préoccupations de sécurité et de confidentialité, en raison du stockage des données dans le Cloud. Le coût élevé des API disponibles constitue un autre obstacle majeur à leur adoption généralisée, notamment pour des projets comme le nôtre, où l'objectif final est de permettre aux clients de conclure l'achat de cartes SIM depuis chez eux. Pour toutes ces raisons, et afin de garantir une vérification d'identité robuste, nous avons développé un algorithme avancé de reconnaissance faciale afin de prévenir les fraudes.

III. METHODOLOGIE

Nous avons créé une application web fonctionnelle, intégrant l'OCR pour l'extraction d'informations d'identité, la reconnaissance faciale pour la vérification de l'utilisateur, et un modèle de détection d'objets pour identifier les éléments clés d'une CIN. Dans ce qui suit, nous détaillons l'architecture robuste et les améliorations progressives, que nous avons réussi à implémenter. Notre solution doit répondre aux exigences du KYC tout en assurant une expérience fluide pour l'utilisateur.

A. Problématique de l'application web Flask

La solution que nous avons mise en place est une application web développée avec Flask[26], qui assure l'ensemble des étapes nécessaires à la vérification de l'identité du client via sa carte d'identité nationale (CIN) et son visage. Cette solution intègre des technologies avancées telles que la détection d'objets, la reconnaissance optique de caractères (OCR) et la reconnaissance faciale, constituant ainsi unprototype (POC) solide pour un processus complet de KYC. L'application web répond à plusieurs objectifs clés :

ISSN: 1737-930X

1) Automatisation de l'extraction de texte en Arabe Amélioration de l'expérience client:

Notre solution utilise des technologies OCR de pointe pour une extraction précise et fiable des informations contenues sur les CIN. Ces informations, pour la plupart textuelles, sont en langue arabe. Notre objectif est de surpasser les limitations des solutions existantes en matière de reconnaissance des caractères en alphabet arabe. Pour cela, il faudrait obtenir une précision d'OCR la plus élevée possible, ce qui revient à obtenir le taux d'erreur des caractères (CER) ou des mots (WER) le plus faible possible.

Il faut que la solution soit ergonomique afin de simplifier et d'accélérer le processus d'achat de cartes SIM. Ceci passe par la possibilité de compléter l'opération en ligne, ce qui réduit considérablement le besoin de se déplacer en magasin.

2) Conformité réglementaire et Vérification de l'identité par reconnaissance faciale:

Notre solution permet d'automatiser la vérification des informations, pour minimiser les risques de ventes non réglementaires, tout en assurant que toutes les transactions soient conformes aux normes en vigueur. Ainsi, nous devons vérifier que les documents soient fiables et conformes (détecter si le document est flou, falsifié, expiré, altéré, etc).

Nous avons intégré une solution de reconnaissance faciale pour vérifier de manière précise que la personne scannant la CIN est bien la propriétaire légitime, garantissant ainsi une validation rigoureuse de l'identité.

3) *Renforcement de la sécurité et de l'efficacité :*

Dans ce projet, la protection des données traitées est primordiale. Il faut s'assurer que notre prototype modernise le processus d'achat et améliore la sécurité et l'efficacité opérationnelle, tout en offrant une expérience client plus fluide et rapide.`

B. Utilisation de YOLOv8n pour la détection d'éléments clés dans la CIN et reconnaissance faciale

La détection d'objets est un élément fondamental de la vision par ordinateur, qui implique à la fois l'identification et la localisation précise des objets dans une image ou une vidéo. Dans le cadre de ce travail, la détection d'objets a été utilisée pour identifier et vérifier différents éléments présents sur une CIN. Pour cela, nous avons implémenté YOLOv8n (pour You Only Look Once). C'est un modèle de détection d'objets de pointe, basé sur des réseaux de neurones convolutifs (CNN), conçu pour détecter plusieurs objets dans une image[27,28]. Il se distingue par sa rapidité et sa précision, ce qui le rend idéal pour les applications en temps réel telles que la détection d'éléments spécifiques sur une carte d'identité. L'algorithme examine l'image entière en une seule fois, générant des prédictions rapides, ce qui le rend adapté aux environnements nécessitant des réponses instantanées, comme la vérification d'identité. Nous avons utilisé des données comprenant un total de 292 images, représentant diverses CIN. Chaque image a été annotée pour indiquer onze zones spécifiques. Ce sont quatre coordonnées pour chacune des classes suivantes : l'adresse, la date de naissance dob, le prénom first_name, le drapeau flag1, l'emblème de la République flag2, le nom complet full_name, le numéro de la carte d'identité id, la photo image, le nom de famille last_name, le lieu de naissance et la profession. Les annotations ont été effectuées à l'aide de l'outil Roboflow[29], et sont exprimées dans le format YOLO avec des coordonnées normalisées pour chaque élément détecté. L'ensemble des données a été divisé en trois sous-ensembles, avec 200 images allouées pour l'entraînement (soit environ 70% des données), 50 images pour la validation (environ 20%), et enfin 42 images pour les tests (10%).

ISSN: 1737-930X

Fig .1 Exemple de détection des éléments clés d'une CIN tunisienne

Parmi les zones extraites de la CIN par YOLOv8n, nous extrayons la photo de l'identité.Cette photo est comparée à une capture du visage prise par la solution Flask. Pour la reconnaissance faciale, nous avons choisi d'utiliser le modèle FaceNet512. C'est une solution avancée pour la reconnaissance faciale qui excelle dans l'extraction d'encodage de haute qualité[30]. Le détecteur RetinaFace, connu pour sa robustesse et sa précision dans la détection des visages[31]. Il complète ce choix en fournissant des détections fiables qui alimentent le modèle FaceNet512. Nous avons opté pour le modèle pré-entraîné FaceNet512 avec le détecteur RetinaFace. Les résultats obtenus confirment l'importance de sélectionner des modèles et des détecteurs adaptés pour atteindre des niveaux élevés de précision dans les applications de reconnaissance faciale.

1) VLM pour la mise en œuvre de l'OCR :

Le modèle PaliGemma est un modèle de langage de la vision (VLM) en libre accès (open-source) contient environ 3 milliards de paramètres, inspiré de l'architecture PaLI-3 Vision-Language Model[32]. Il est composé d'éléments open-source, notamment le modèle de vision SigLIP et le modèle de langage Gemma. Le VLM PaliGemma est capable de traiter à la fois des images et du texte en entrée, et de générer du texte en sortie, prenant en charge plusieurs langues. Il peut être affiné pour obtenir des performances élevées sur diverses tâches vision/langage telles que la légende d'images et de courtes vidéos, la réponse à des questions visuelles, la lecture de texte, la détection d'objets et la segmentation d'objets. Pour évaluer les performances de PaliGemma dans l'extraction des informations des CIN, nous avons procédé à une série de tests sur un ensemble varié d'images de CIN. Les informations clés à extraire comprenaient le nom, le surnom, le nom des grands parents, la date de naissance, et le numéro de la carte d'identité nationale. Le processus de test a inclus les étapes suivantes : Collecte d'un ensemble diversifié d'images de CIN avec différentes qualités et conditions de lumière (20 images) ; Application de PaliGemma sans entraînement afin d'extraire les informations textuelles des images ; Évaluation des résultats en comparant des informations extraites avec les données réelles, pour mesurer la précision et la fiabilité du modèle. La solution PaliGemma est pré entraînée pour la tâche OCR. Elle permet de lire les images de gauche à droite et de haut en bas. Cette méthode facilite l'extraction d'informations à partir de chaînes de caractères.

2) Métriques et performances de détection d'objets, reconnaissance faciale et OCR :

Lors de la détection d'objets, l'objectif est d'identifier les objets dans une image et de tracer des boîtes de délimitation autour d'eux. Le modèle doit non seulement reconnaître l'objet mais aussi le localiser correctement dans l'image. L'Intersection sur Union (IoU) est une mesure de la superposition entre la boîte de délimitation prédite et la boîte de délimitation observée. C'est un ratio entre les deux valeurs suivantes : L'intersection est l'aire de la zone où la boîte prédite et la boîte observée se chevauchent. L'union est l'aire de la zone totale couverte par les deux boîtes (prédite et observée).

L'IoU est calculé comme l'aire de l'intersection divisée par l'aire de l'union. Si l'IoU est supérieur à un certain seuil, la prédiction est considérée comme un vrai positif. Sur toutes les prédictions faites par le modèle, la Précision (P) indique le nombre de celles qui étaient correctes (vrais positifs). Parmi tous les objets réels dans l'image, le Rappel

ISSN: 1737-930X

(R) indique le nombre d'objets correctement détectés par le modèle. L'aire sous la courbe Précision vs. Rappel pour chaque classe donne la Précision Moyenne (AP). La Précision Moyenne (mAP) est la moyenne des APs pour toutes les classes de l'ensemble de données. Cela donne un seul chiffre pour comparer la performance de différents modèles à travers plusieurs classes. La Précision Moyenne (mAP) est une métrique clé utilisée dans la détection d'objets pour évaluer la qualité d'un modèle à prédire l'emplacement et la classification des objets dans une image. La mAP donne une mesure globale de la précision et de la cohérence avec lesquelles un modèle détecte et localise les objets à travers différentes classes dans un ensemble de données d'images. Nous notons mAP@50 le mAP calculé au seuil de l'IoU égal à 0.5 ; mAP@50-95 désigne la moyenne des mAP pour différents seuils de l'IoU, de 0.5 à 0.95.

Pour la reconnaissance faciale, la métrique utilisée est la distance cosinus. Elle mesure la distance angulaire entre deux vecteurs. Elle est souvent utilisée pour les recherches de similarité, comme la recherche d'images similaires dans une base de données. La valeur de la distance cosinus varie de zéro à un, 0 indiquant des vecteurs identiques et 1 indiquant des vecteurs complètement différents.

Pour l'OCR ou précisément l'évaluation du la précision du texte généré par le modèle de langage de vision VLM PaliGemma, les deux métriques utilisées sont le CER et le WER. Le Character Error Rate (CER) est calculé en fonction du nombre d'erreurs de caractères par rapport au nombre total de caractères dans la référence. Le Word Error Rate (WER) se concentre sur les mots.

VI. Deploiement et Resultats

Cette section présente une vue détaillée du pipeline développé pour digitaliser le processus d'achat de cartes SIM chez une entreprise de télécommunication. Fig.2 donne le diagramme de flux décrivant le processus d'authentification basé sur la reconnaissance faciale et l'extraction de données à partir d'une CIN. Elle illustre les composants de la solution et leur interaction. Elle est constituée de plusieurs étapes clés interconnectées et combinant les trois volets principaux du projet : l'OCR, la reconnaissance faciale, et la détection d'éléments clés de la CIN. Pour chacune des composantes de la Fig.2, différentes stratégies ont été testées afin d'aboutir à une approche optimale, prête à être déployée en tant que POC.

C. Architecture de la solution

Les principales étapes du pipeline de la solution sont les suivantes : détecter la CIN en temps réel ; extraire le visage depuis la CIN ; capturer le visage en temps réel ; reconnaître le visage ; extraire les données de la CIN.

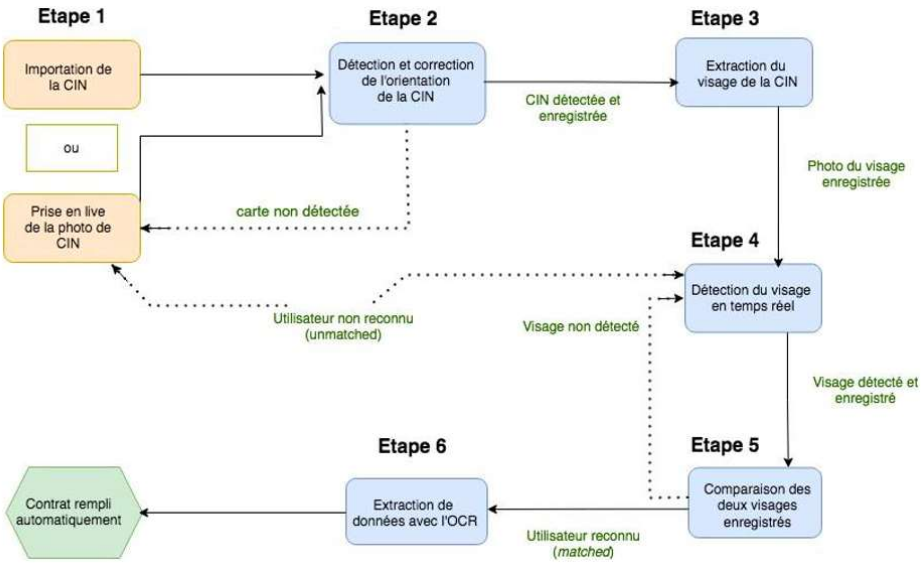

Fig. 2Diagramme du processus KYC de vérification de la CIN

ISSN: 1737-930X

1) Détection de la carte d'identité et extraction du visage depuis la carte d'identité :

Lors de cette première étape, le client est invité à importer ou à capturer en temps réel une image de sa carte d'identité, en s'assurant que la carte soit placée précisément à l'intérieur d'un rectangle spécifié à l'écran. C'est au niveau de l'interface utilisateur de Flask que cette étape est réalisée. Le système impose des critères stricts : l'image doit être claire, bien cadrée, et non déviée. Un modèle de détection pré-entraîné sur un jeu de données de 292 images de CIN est utilisé pour valider ces critères. Si l'une des conditions n'est pas respectée, le client est invité à capturer une nouvelle image. Pour la correction de l'orientation de la CIN, nous avons implémenté un script Python (voir Étape 2 de la Fig.2).Une fois que l'image de la carte a été validée, le système procède à l'extraction du visage apparaissant sur la carte d'identité. Ce visage est ensuite enregistré pour être comparé à l'image du visage du client, capturée dans les étapes suivantes.

2) Capture du visage du client en temps réel et reconnaissance faciale:

Le client doit ensuite capturer une image de son visage en temps réel à l'aide de la caméra. Cette image est stockée pour la phase de reconnaissance faciale qui suit.

Le modèle de reconnaissance faciale encode les deux images (le visage extrait de la carte d'identité et celui capturé en temps réel) et effectue une comparaison. Si une correspondance est trouvée avec un score de similarité suffisant, le pipeline continue avec l'extraction des données de la carte. Dans le cas contraire, le client est invité soit à soumettre une nouvelle carte d'identité, soit à refaire la capture de son visage. Nous avons utilisé FaceNet512 associé au détecteur RetinaFace qui montre une précision élevée[33] pour la reconnaissance faciale. Voir Étape 5 de la Fig.2.

3) Extraction des données *de la CIN :*

Lorsque la reconnaissance faciale est validée, le modèle OCR entre en action pour extraire les informations textuelles de la carte d'identité. Ces données sont ensuite utilisées pour pré-remplir automatiquement les champs du formulaire de contrat. Certaines informations, comme le numéro de la carte d'identité, sont obligatoires et non modifiables par le client, tandis que d'autres champs peuvent être révisés par ce dernier avant de finaliser le processus.

D. Détection des éléments clés de la CIN

Le processus d'entraînement du modèle YOLOv8n [34]a suivi plusieurs étapes. Nous avons utilisé une version pré-entraînée du modèle afin de bénéficier des connaissances acquises à partir de grands ensembles de données. Le modèle a ensuite été configuré pour détecter les onze classes d'éléments spécifiques à une CIN. La Fig.1 donne un aperçu de l'exécution de YOLOv8n avec par exemple la détection du numéro de la CIN id, du nom last_name, etc. L'entraînement s'est déroulé sur 200 epochs, avec un learning-rate=0.001 et une stratégie configurée pour s'arrêter automatiquement après 10 époques si aucune amélioration n'était observée. L'ensemble du processus a été exécuté sur un GPU, permettant ainsi d'accélérer le calcul et de suivre l'évolution des performances à chaque epoch.

À l'issue des 200 epochs, les performances du modèle ont été évaluées selon plusieurs métriques clés. Le modèle a atteint une Précision de 0.669, un Rappel de 0.861, un mAP@50 de 0.806 et un mAP@50-95 de 0.532. Ces résultats démontrent que le modèle est capable de détecter et de localiser les différents éléments d'une carte d'identité avec un niveau de précision et de rappel satisfaisant. Voir par exemple, la Table1 pour une sortie YOLOv8n de la CIN de la Fig.1.

VI. Discussion

La complexité de la langue arabe, caractérisée par son alphabet unique, présente à la fois des opportunités et

ISSN: 1737-930X

des obstacles pour une mise en œuvre efficace. En tirant parti des techniques avancées de vision par ordinateur associées au traitement du langage naturel, nous pouvons améliorer la précision et l'efficacité de l'extraction d'informations pertinentes à partir des documents d'identité tout en garantissant des capacités robustes de reconnaissance faciale. Lors de la mise en place de notre solution pratique, nous avons rencontré plusieurs défis, nécessitant de trouver des solutions pratiques.

Pour la détection de la CIN et de ses éléments, un modèle YOLOv8n a été utilisé, atteignant une mAP@50 de 80%. La reconnaissance faciale est effectuée avec le modèle RetinaFace qui a montré ses preuves dans[35]. Les résultats obtenus montrent que le modèle YOLOv8n est performant pour la détection des éléments d'une CIN. Le modèle YOLOv8n a également démontré sa capacité à détecter efficacement les éléments dans une séquence vidéo, offrant ainsi la possibilité d'étendre le projet vers une approche vidéo en temps réel.Cette extension sera explorée dans les futures étapes du projet, avecl'intégration du Multithreading pour traiter simultanément plusieurs flux vidéo et améliorer ainsi les performances globales du système de détection.Cependant, certaines améliorations peuvent être apportées pour optimiser davantage ses performances. Par exemple, en enrichissant la base de données avec des données plus variées, le modèle pourrait mieux capturer des éléments complexes ou changeants (ce qui permettrait d'améliorer le Rappel). Pour augmenter la Précision, une réduction des faux positifs pourrait être envisagée en affinant les hyperparamètres. L'optimisation de certains paramètres, comme le *learning-rate* ou la taille des lots, peut également conduire à une amélioration significative des résultats finaux.Nous avons opté pour l'utilisation d'un VLM avec un déploiement local pour l'extraction des informations. Cette approche, combinée avec une méthode de *zero-shot learning*, a permis d'obtenir une précision moyenne en Arabe pour servir un premier prototype (POC) pour le KYC. Elle fournit une méthode fonctionnelle pour l'extraction d'informations structurées à partir des cartes d'identité. L'utilisation de PaliGemma lors de l'extraction des informations de la carte d'identité. C'est une approche multimodale qui combine les capacités visuelles et linguistiques pour offrir une performance relativement élevée dans l'extraction du texte.En plus, c'est un modèle open-source et local assurant la confidentialité des données. L'extraction des données en Arabe a été réalisée par un modèle VLM nommé PaliGemmaatteignant un CER de 9% et un WER de 19% pour 20 cartes d'identité réelles testées.

Label	x	y	Largeur	Hauteur
image	73.18	201.82	137.27	167.27
id	231.82	140.91	145.45	27.27
last_name	323.64	169.09	149.09	32.73
first_name	308.33	195.33	172.67	21.33
full_name	295	219.67	275.33	27.33
dob	263	247	192.67	23.33
pob	363.67	277.33	84.67	33.33
flag1	56.15	62.82	73.38	50.71
flag2	399.52	67.62	42.77	60.31

Tableau 1 *EXEMPLE DE COORDONNEES DE ZONES SELECTIONNEES PAR YOLOV8N*

VI. CONCLUSION ET PERSPECTIVES

Ce travail a démontré la faisabilité d'une solution automatisée du processus de vérification d'identité dans le cadre des achats de cartes SIM. Malgré les défis techniques, l'utilisation d'approches innovantes et de technologies avancées a permis d'obtenir un prototype prometteur pour les futures applications. Le modèle de langage de vision PaliGemma ouvre la voie à des applications pratiques et évolutives dans divers domaines, permettant aux organisations d'améliorer leurs processus, de réduire les coûts et de fournir des solutions à leurs défis opérationnels. En mettant en évidence les innovations dans ce domaine, des exemples tels que PaliGemma montrent comment ces évolutions technologiques contribuent à l'amélioration continue des systèmes OCR, permettant des applications plus

ISSN: 1737-930X

précises et adaptées aux besoins spécifiques des utilisateurs. Le projet aboutit à la création d'une application web fonctionnelle, intégrant l'OCR pour l'extraction d'informations d'identité, la reconnaissance faciale pour la vérification de l'utilisateur, et un modèle de détection d'objets pour identifier les éléments clés d'une carte d'identité. Grâce à une architecture robuste et à des améliorations progressives, nous avons réussi à proposer une solution répondant aux exigences du KYC tout en assurant une expérience fluide pour l'utilisateur.

REFERENCES

1. Ashish Vaswani et al. « Attention is all you need ». In : Advances in neural information processing systems (2017). 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
2. Samia Oubraham et Rim Kaci Aissa. « Le lancement d'un nouveau produit à l'ère de la digitalisation Cas : Optimum Telecom Algérie ». Thèse de doctorat. Université Mouloud Mammeri, 2018.
3. Tonye Cottavoz - Paralegal. La digitalisation des entreprises, réussir la transformation numérique. [Online]. Accessed on 08/30/2024. Available:https://www.axiocap.com/blog/digitalisation-entreprises-reussir-transformation-ere-numerique. Jan. 2023.
4. Muhammad Maaz et al. « Video-chatgpt : Towards detailed video understanding via large vision and language models ». In : Preprint arXiv :2306.05424 (2023).
5. Jingyi Zhang et al. « Vision-Language Models for Vision Tasks : A Survey ». In : IEEE Transactions on Pattern Analysis and Machine Intelligence 46.8 (2024), p. 5625-5644. doi :10.1109/ TPAMI.2024.3369699.
6. IBM. [Online]. Available :https://www .ibm .com /fr- fr/topics/optical- character- recognition.
7. Irene Hou et al. « More robots are coming : large multimodal models (ChatGPT) can solve visually diverse images of Parsons problems ». In : Proceedings of the 26th Australasian Computing Education Conference. 2024, p. 29-38.
8. OCR and Logistics : An Ultimate Guide. [Online]. Available : https://packagex.io/blog/ocr- and-logistics. Accessed on 08/31/2024.
9. shuftiPro. All You Need To Know About KYC Compliance. [Online]. Available : https:// shuftipro.com /blog /all-you - need - to- know - about - kyc - compliance/. Accessed on 08/31/2024.
10. Aashi Jain et al. « MURAL : Multimodal, Multitask Representations Across Languages ». In : Findings of the Association for Computational Linguistics : EMNLP 2021. 2021, p. 3449-3463.
11. Deyao Zhu et al. « Minigpt-4 : Enhancing vision-language understanding with advanced large language models ». In : Preprint arXiv :2304.10592 (2023).
12. Simon JD Prince. Computer vision : models, learning, and inference. Cambridge University Press, 2012.
13. Geewook Kim et al. « OCR-free document understanding transformer ». In : European Conference on Computer Vision. Springer. 2022, p. 498-517.
14. Musa Dildar Ahmed Cheema et al. « Adapting multilingual vision language transformers for low-resource Urdu optical character recognition (OCR) ». In : PeerJ Computer Science 10 (2024), e1964.
15. Zijing Liang et al. « A Survey of Multimodel Large Language Models ». In : Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering. 2024, p. 405-409.
16. Daizong Liu et al. « A survey of attacks on large vision-language models : Resources, advances, and future trends ». In : Preprint arXiv :2407.07403 (2024).
17. Yifan Li et al. « Evaluating Object Hallucination in Large Vision-Language Models ». In : The 2023 Conference on Empirical Methods in Natural Language Processing.
18. Xi Chen et al. « Pali-x : On scaling up a multilingual vision and language model ». In : Preprint arXiv :2305.18565 (2023). https ://arxiv.org/pdf/2305.07895.
19. Jiayu Wang et al. « Is a picture worth a thousand words ? delving into spatial reasoning for vision language models ». In : Preprint arXiv :2406.14852 (2024).
20. Zheyuan Liu et al. « Image retrieval on real-life images with pre-trained vision-and-language models ». In : Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021, p. 2125-2134.
21. Identity Verification Pricing Comparison and Alternatives | Trust Swiftly. [Online]. Available :https://trustswiftly.com/blog/identity- verification- pricing- comparison- and- alternatives/. Accessed on 08/31/2024.
22. Shufti Pro Pricing, Alternatives & More 2024 | Capterra. [Online]. Available :https://www. capterra.com/p/181317/Shufti-Pro/. (Accessed on 08/31/2024).
23. ID Verification with SDK. [Online]. Accessed on 08/31/2024. Available : https://regulaforensics.com/products/ document-reader-sdk/.
24. Image To Text - Pricing. [Online]. Available : https://www.imagetotext.com/en/pricing. Accessed on 08/31/2024.
25. Sumsub Supported languages. [Online]. Available : https://docs.sumsub.com/docs/supported- languages. Accessed on 08/31/2024.
26. Miguel Grinberg. Flask web development. "O'Reilly Media, Inc.", 2018.
27. Muhammad Hussain. « YOLOv5, YOLOv8 and YOLOv10 : The Go-To Detectors for Real-time Vision ». In : Preprint arXiv :2407.02988 (2024).
28. Mupparaju Sohan et al. « A review on yolov8 and its advancements ». In : International Conference on Data Intelligence and Cognitive Informatics. Springer. 2024, p. 529-545.
29. Floriana Ciaglia et al. « Roboflow 100 : A rich, multi-domain object detection benchmark ». In : Preprint arXiv :2211.13523 (2022).
30. Florian Schroff, Dmitry Kalenichenko et James Philbin. « Facenet : A unified embedding for face recognition and clustering ». In : Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, p. 815-823.
31. Jiankang Deng et al. « Retinaface : Single-shot multi-level face localisation in the wild ». In : Proceedings of the IEEE/CVF

ISSN: 1737-930X

conference on computer vision and pattern recognition. 2020, p. 5203-5212.
32. PaliGemma – Google's Cutting-Edge Open Vision Language Model. [Online]. Available : https://huggingface.co/blog/paligemma. (Accessed on 09/01/2024).
33. Sefik Ilkin Serengil et Alper Ozpinar. « A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules ». In : Bilisim Teknolojileri Dergisi 17.2 (2024), p. 95-107. doi :10.17671/gazibtd.1399077.
34. Mariia Nazarkevych et al. « A YOLO-based Method for Object Contour Detection and Re- cognition in Video Sequences. » In : CPITS. 2024, p. 49-58.
35. Deepface/benchmarks at master . serengil/deepface. [Online]. Available : https://github.com/serengil/deepface/tree/master/benchmarks

ISSN: 1737-930X