TP 7 – Rendu de projet

IA et deep learning

ESIEE – 2025

Projet réalisé en binôme
Classes de la figure 7 anonymisées.

Kévin FELTRIN
kevin.feltrin@edu.esiee.fr

***Résumé* — Ce rapport examine l'intégralité du processus d'un système de reconnaissance faciale, depuis la détection des visages jusqu'à la prédiction de l'identité, en utilisant à la fois des réseaux de neurones convolutifs et des classifieurs d'apprentissage automatique classiques. Les défis rencontrés incluent la variabilité de l'orientation des visages, les conditions d'éclairage et la qualité des images. Pour relever ces défis, plusieurs étapes de prétraitement ont été mises en œuvre, telles que la détection des visages, l'estimation de la pose et le codage des visages. Chaque étape a été évaluée séparément pour observer son impact sur les performances de reconnaissance. De plus, des données personnelles ont été utilisées pour évaluer la généralisabilité du modèle et analyser les biais potentiels. En combinant des techniques d'apprentissage profond avec des stratégies de normalisation des données, nous avons atteint une précision de test allant jusqu'à 97 %, tout en discutant du compromis entre la complexité du modèle et l'efficacité.**

***Mots-clés* — Deep Learning, Détection de visage, Estimation de la pose, Reconnaissance faciale**

I. INTRODUCTION

La reconnaissance faciale occupe une place centrale dans de nombreuses applications de vision par ordinateur, qu'il s'agisse de sécurité, d'accès personnalisé ou d'analyse d'images. Contrairement à la simple détection de visages, qui localise leur présence dans une image, la reconnaissance cherche à identifier avec précision l'individu représenté. Ce processus reste complexe car il doit composer avec de nombreux facteurs : variations d'angle, expressions faciales, conditions d'éclairage ou qualité de l'image.

Ce rapport présente la mise en œuvre progressive d'un système de reconnaissance faciale complet, intégrant des étapes essentielles telles que la détection, l'alignement (via l'estimation de la pose), l'encodage des visages, puis la classification. Chaque étape est analysée pour évaluer son impact sur la précision du système. Nous testons également la robustesse du modèle à l'aide de données personnelles variées, afin d'évaluer ses performances en conditions réelles et d'identifier d'éventuels biais.

II. TRAVAUX CONNEXES

Une méthode classique sans apprentissage profond est celle des « Eigenfaces » de Turk et Pentland (1991) [1], utilisant l'ACP pour réduire la dimensionnalité et extraire des caractéristiques faciales significatives. Testée sur la base ORL (40 individus, 10 images chacun), cette approche atteignait environ 90 % de reconnaissance dans des conditions contrôlées. Toutefois, elle reste sensible aux variations de lumière, de pose ou d'expression. Des techniques de prétraitement comme l'égalisation d'histogramme ou les LBP peuvent améliorer sa robustesse sans recourir à l'apprentissage profond.

Du côté des approches profondes, l'article « DeepFace » de Taigman et al. (2014) [2] a introduit un réseau à neuf couches, entraîné sur plus de 4 millions d'images Facebook, avec une normalisation 3D des visages pour réduire les variations de pose. Il a atteint 97,35 % de précision sur LFW, proche des performances humaines. Cependant, la dépendance à des données privées limite la reproductibilité. Utiliser des bases publiques et l'augmentation de données pourrait améliorer la généralisation.

Enfin, « FaceNet » de Schroff et al. (2015) [3] propose un réseau convolutif avec perte triplet, apprenant une représentation compacte de visages. Entraîné sur 200 millions d'images privées, il a atteint 99,63 % sur LFW. La sélection des triplets, essentielle à la performance, reste coûteuse. L'adoption de méthodes adaptatives pourrait rendre l'entraînement plus efficace.

III. PROPOSITION

A) Détection des visages

La détection de visages est une technologie permettant d'identifier la présence de visage sur une image ou un flux d'images. Contrairement à la reconnaissance faciale, elle ne cherche pas à reconnaître l'identité de la personne, mais plutôt à localiser les visages sur l'image. La localisation des visages est cependant nécessaire à tout processus de reconnaissance faciale. Elle repose sur divers algorithmes, allant de méthodes basées sur les traits du visage à des techniques plus avancées d'apprentissage profond. Comme l'explique Raymond Bruyer dans *Les Mécanismes de reconnaissance des visages* [4], cette fonction s'inspire des processus cognitifs humains qui permettent de distinguer un visage d'autres formes visuelles, et constitue un domaine à mis chemin entre la psychologie cognitive et l'intelligence artificielle.

Les données reçues sont des images de dimensions variées, avec différentes extensions possibles (jpg, JPG, jpeg, png). Chaque image représente un personnage issu de l'univers de la saga Jurassic Park. Les images sont regroupées dans six dossiers, chacun portant le prénom et le nom du personnage correspondant. Au total, 218 images sont réparties de manière inégale entre ces dossiers, avec un minimum de 22 images et un maximum de 53 par dossier.

La détection de visages est une étape essentielle dans ce projet, car elle permet d'isoler la zone du visage sur chaque image avant d'appliquer les algorithmes de reconnaissance faciale. Cela permet d'éviter de perturber le modèle de reconnaissance faciale par des éléments d'arrière-plan, et ainsi de garantir à la qualité des données d'entrées pour une meilleure performance du modèle.

Les images ont été soumises à un prétraitement visant à extraire automatiquement les visages grâce à un modèle de détection de visages. Deux modèles ont pût être utilisés, l'un basé sur la méthode HOG (Histogram of Oriented Gradients), plus rapide et adapté aux environnements peu complexe, et l'autre reposant sur un réseau de neurones convolutif (CNN), offrant une détection plus précise au prix d'un coût computationnel plus élevé. Le modèle HOG s'est révélé insuffisant pour détecter tous les visages présents dans les images, c'est pourquoi nous avons recouru au modèle CNN. Les visages détectés sont extraits en rognant l'image selon les

coordonnées fournies par le détecteur. Ils sont ensuite redimensionnés à 128×128 pixels afin de garantir une cohérence des dimensions et d'uniformiser les données. Enfin, les valeurs des pixels ont été normalisées par les générateurs d'images, en les ramenant dans l'intervalle [0;1] au lieu de [0;255].

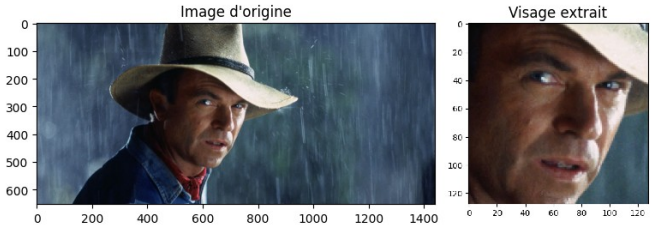


Figure 1 – Extraction d'un visage

Les images ont été réparties en trois ensembles distincts : entraînement, validation et test. Environ 65 à 70 % des images ont été allouées à l'ensemble d'entraînement, tandis que le reste a été réparti de manière approximativement égale entre les ensembles de validation et de test (soit environ 15 à 22,5 % pour chacun). Cette division permet d'entraîner le modèle sur une grande partie des données tout en conservant des échantillons indépendants pour évaluer ses performances et ajuster les hyperparamètres, réduisant ainsi le risque de surapprentissage.

Pour entraîner directement un modèle sur ces données, nous avons conçu un réseau de neurones convolutif adapté à la taille modeste du jeu de données. La base du modèle se compose de plusieurs couches convolutionnelles avec un nombre de kernels croissants afin de capturer des patterns de plus en plus complexes. Chacune de ces couches est suivie d'une couche de max-pooling 2×2 afin de réduire progressivement la dimension spatiale. L'utilisation de l'activation ReLU après chaque convolution permet d'introduire de la non-linéarité et d'accélérer la convergence. Les données sont ensuite aplaties et passer dans une couche dense de 512 neurones avant d'aboutir à la couche de sortie avec une activation softmax afin d'obtenir les probabilités d'appartenance aux six classes.

Le réseau de neurones convolutif entraîné a atteint une précision d'environ 74 % sur l'ensemble de test. Les images sont mélangées aléatoirement avant d'être répartie dans les différents dossiers, ce qui donnent un résultat différent à chaque exécution du programme. Ce résultat est plutôt satisfaisant compte tenu de la simplicité de l'architecture du modèle et de la taille relativement réduite du jeu de données, mais suggère qu'il reste une grande marge d'amélioration.

Les hyperparamètres ont un impact sur les performances du réseau convolutionnel. Réduire le nombre de couches dans la base a permis d'augmenter la précision du modèle, passant de 74 % à 83 % sur l'ensemble de test. Inversement, ajouter davantage de couches provoque un résultat similaire mais nécessite plus d'époques pour converger.

Une augmentation du nombre d'unités dans la couche dense a conduit à un surajustement entraînant une baisse de la précision, mais réduire excessivement ce nombre nuit à la capacité du modèle à apprendre des représentations suffisamment discriminantes.

Une valeur trop faible du taux d'apprentissage entraîne une baisse de performances car le modèle apprend plus lentement, empêchant une convergence satisfaisante en 30 époques. À l'opposé, l'augmenter conduit à des performances plus erratiques et peut également dégrader les résultats puisque le modèle oscille autour du minimum sans s'y stabiliser, ou même diverge complètement.

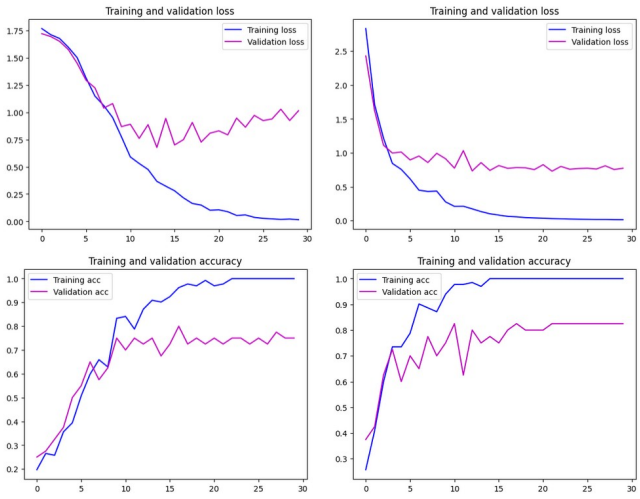


Figure 2 – 4 couches (gauche) vs 1 couche de convolution

B) Estimation de la pose

L'estimation de la pose faciale consiste à déterminer l'orientation d'un visage dans l'espace à partir de points de repère anatomiques. Comme le souligne Richard Hartley et Andrew Zisserman dans leur ouvrage de référence *Multiple View Geometry in Computer Vision (2004)* [5], l'estimation de la pose repose sur des méthodes de vision par ordinateur qui exploitent la géométrie projective et les repères tridimensionnels, souvent à partir de points clés détectés sur le visage (yeux, nez, bouche, etc.). Si le visage est pivoté ou incliné, les points de repères permettent de corriger son alignement. Cette correction est particulièrement utile dans les systèmes de reconnaissance faciale qui peuvent voire leur précision diminuer en raison de visages particulièrement tournés ou inclinés.

Les données utilisées pour l'estimation de la pose correspondent aux visages extraits lors de l'étape de détection précédente. Chaque visage est redimensionné à une taille standard de 128 × 128 pixels et sauvegardé au format JPEG, de manière à faciliter l'alignement et le traitement automatique. Ces images représentent des personnages issus de la saga Jurassic Park et sont réparties dans trois ensembles : entraînement, validation et test. Au sein de chaque ensemble, les images sont organisées en sous-dossiers nommés selon le prénom et le nom du personnage représenté. L'ensemble total comprend 218 images, réparties de manière inégale entre les différentes classes.

L'estimation de la pose est une étape importante dans le cadre de notre travail, car elle permet de corriger l'orientation des visages afin de les aligner de manière cohérente avant la reconnaissance faciale. En effet, les performances des algorithmes de reconnaissance sont fortement dépendantes de la régularité des données en entrée. Des variations d'angle, d'inclinaison ou de position peuvent entraîner des erreurs de classification, même pour des visages bien connus du modèle. Grâce à l'estimation de la pose, nous identifions des points repères du visage qui servent ensuite à appliquer une transformation géométrique pour repositionner le visage dans une posture standard. Cette normalisation garantit que toutes les images présentent des visages orientés de manière similaire, ce qui améliore la stabilité et la précision du modèle de reconnaissance.

Pour l'estimation de la pose, nous utilisons une approche basée sur la détection de points de repère (landmarks) du visage, grâce aux modèles pré-entraînés de la bibliothèque dlib. Deux prédicteurs sont mis en œuvre : un modèle à 68 points, qui offre une description fine de la structure faciale (contours du visage, yeux, sourcils, nez, bouche), et un modèle allégé à 5 points, plus rapide mais moins détaillé. Une fois les points détectés, nous appliquons une transformation affine à partir de trois d'entre eux (généralement les

deux coins internes des yeux et le centre de la lèvre inférieure), ce qui permet de recadrer et redresser le visage selon une position de référence.

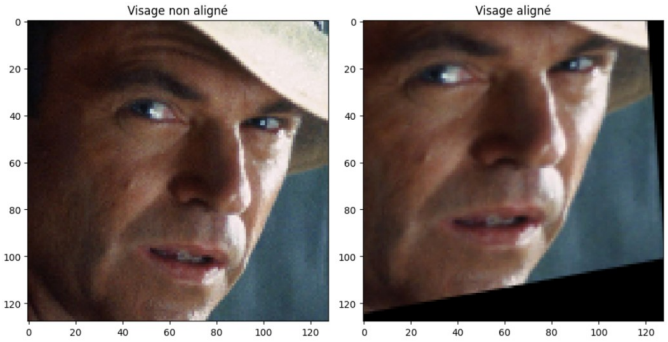


Figure 3 – Alignement d'un visage

Comme dans la section précédente, les images sont réparties en trois ensembles distincts : entraînement, validation et test. 65 % des images ont été allouées à l'ensemble d'entraînement, tandis que le reste a été réparti de manière approximativement égale entre les ensembles de validation et de test (soit environ 22,5 % pour chacun).

L'évaluation des performances de nos modèles convolutionnels (convnets) sur l'ensemble de test a permis de mesurer l'impact de la correction de la pose sur la qualité de la reconnaissance faciale. En réentraînant notre modèle convolutionnel de base sur les nouvelles données alignées (issues de l'estimation de la pose), la précision sur l'ensemble de test est passée de 74 % à 80 %.

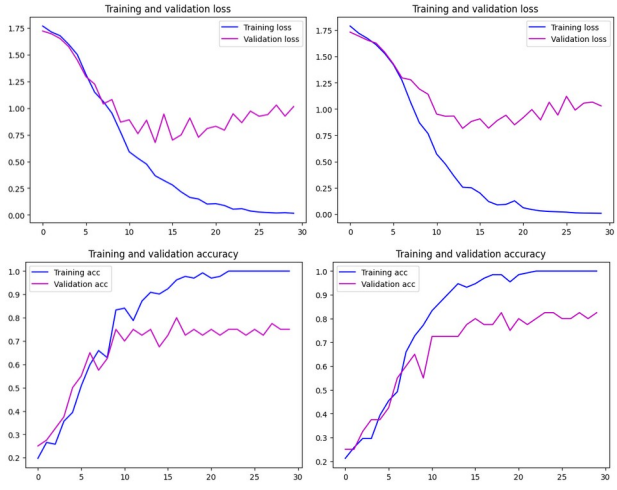


Figure 4 – Avec (droite) et sans alignement

Le modèle plus simple, constitué d'une seule couche de convolution, qui atteignait déjà une précision de 83 % sans alignement, ne bénéficie pas d'un gain supplémentaire après l'alignement. Ce résultat suggère que ce modèle était déjà adapté aux données non alignées ou qu'il avait atteint un plateau de performance.

L'amélioration des résultats peut être attribuée à la réduction de la variabilité dans les données d'entrée, liée à l'orientation des visages, grâce à la correction de la pose. En alignant les visages selon une orientation standard, le réseau convolutif reçoit des images plus homogènes. Cela améliore la qualité des représentations apprises et facilite l'apprentissage des caractéristiques discriminantes propres à chaque individu.

C) Encodage des visages

L'encodage des visages est une opération qui consiste à transformer une image de visage en un vecteur numérique, généralement de dimension fixe, qui résume les caractéristiques essentielles du visage. Ce vecteur, parfois appelé embedding, permet de représenter un visage de manière compacte et exploitable par un algorithme de reconnaissance. Comme l'expliquent Parkhi, Vedaldi et Zisserman dans *Deep Face Recognition (2015)* [6], cette représentation repose sur l'idée qu'un visage peut être projeté dans un espace vectoriel où les visages similaires sont proches les uns des autres, et les visages différents éloignés.

Ce procédé d'encodage des visages réduit considérablement la complexité des données visuelles tout en conservant l'information discriminante nécessaire à la reconnaissance. La représentation sous la forme d'un vecteur de dimension fixe permet de comparer efficacement des visages entre eux en calculant simplement des distances mathématiques entre vecteurs. Il permet ainsi d'améliorer la performance des systèmes en termes de rapidité, de mémoire et de précision, notamment lorsqu'il est appliqué à des bases de données de grande taille.

Les données utilisées pour l'encodage des visages proviennent des étapes précédentes, où chaque image a été préalablement traitée par la détection de visage et l'alignement basé sur l'estimation de pose. Il s'agit d'images de taille standard de 128 × 128 pixels, au format JPEG, représentant des visages extraits de personnages issus de la saga Jurassic Park. Ces images sont réparties en trois ensembles distincts : entraînement, validation et test. À l'intérieur de chaque ensemble, elles sont organisées dans des sous-dossiers, chacun portant le prénom et le nom du personnage représenté. L'ensemble total comprend 218 images, distribuées de manière inégale entre les différentes classes.

Après l'encodage des visages, les données sont transformées en vecteurs numériques, appelés des descripteurs, qui ont une dimension fixe de 128 éléments pour chaque visage. Ces vecteurs représentent les caractéristiques uniques et discriminantes des visages dans un espace vectoriel. L'algorithme utilisé, publié par OpenFace, mesure des valeurs qui peuvent refléter différentes propriétés du visage, telles que la distance entre les yeux, la forme du nez, ou encore la disposition générale des traits du visage, mais sans faire appel à des caractéristiques explicites comme la couleur des yeux ou la forme des oreilles. Cette transformation élimine la complexité des données visuelles brutes tout en conservant l'information pertinente pour la reconnaissance faciale.

Comme les données sont seulement des vecteurs de 128 valeurs (contre 3 × 128 × 128 avant encodage), nous n'avons pas eu besoin de recourir à une base convolutionnelle. Le réseau de neurones utilisé est simplement constitué de trois couches denses dont une couche caché de 512 neurones permettant au modèle d'extraire des relations complexes entre les caractéristiques des visages encodés. Bien entendu, la taille des couches d'entrée et de sortie correspondent respectivement à la taille des vecteurs encodé et au nombre de classes.

L'évaluation des performances du modèle simple sur l'ensemble de test des données encodées révèle une amélioration drastique par rapport à la phase préalable de l'encodage. En comparaison avec un modèle convolutionnel antérieur, qui utilisait le même nombre d'unités cachées dans la couche dense, la précision obtenue sur l'ensemble de test est passée de 80 % à 97 %. Le modèle est aussi beaucoup plus confiant dans ses prédictions, comme en témoigne la courbe de perte sur les données de validation.

L'amélioration des résultats peut être attribuée à la pertinence des données encodées. En effet, une représentation compacte en 128 valeurs réduit la variabilité inutile, comme l'arrière plan ou les variations d'éclairages, et se concentre sur les éléments discriminants du visage. L'encodage permet de se passer du traitement d'images brutes par un modèle, ce qui a éliminé le surapprentissage des

données. Le modèle arrive donc à mieux distinguer les différentes classes de visages.

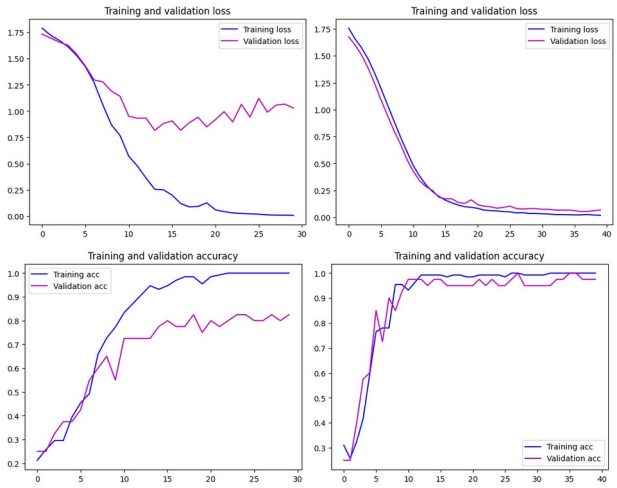


Figure 5 – Avec (droite) et sans encodage

D) Reconnaissance faciale

La reconnaissance faciale est une technologie qui permet d'identifier ou de vérifier l'identité d'une personne à partir de son visage. Contrairement à la détection des visages, qui se contente de localiser la position des visages dans une image, la reconnaissance faciale vise à associer chaque visage détecté à une personne spécifique à l'aide d'une base de données de visages connus. Comme l'expliquent Zhao, Chellappa et Rosenfeld dans *Face Recognition: A Literature Survey (2003)* [7], la reconnaissance faciale repose sur l'extraction et la comparaison de caractéristiques uniques présentes dans les traits du visage, telles que la distance entre les yeux, la forme du nez ou des joues, et d'autres traits biométriques distinctifs.

L'image à évaluer est d'abord traitée par un modèle de détection de visages, qui extrait la zone correspondant aux visages présents dans l'image. Ensuite, ces visages sont convertis en vecteurs numériques (les "embeddings" des visages), qui sont ensuite utilisés comme entrée pour le modèle de reconnaissance faciale, afin de déterminer l'identité associée à chaque visage.

La principale différence entre la détection des visages et la reconnaissance faciale réside dans leurs objectifs respectifs : la détection des visages consiste à localiser la présence d'un visage dans une image, tandis que la reconnaissance faciale cherche à identifier ou valider l'identité d'un visage. En d'autres termes, la détection est une étape préliminaire qui permet d'isoler les zones où se trouvent les visages, tandis que la reconnaissance utilise ces zones pour faire correspondre le visage détecté à une identité spécifique.

Pour effectuer de la reconnaissance faciale, plusieurs classificateurs ont été testés afin de comparer leurs performances. Le premier modèle testé est le réseau de neurones utilisé dans la section précédente. Ensuite, plusieurs classificateurs classiques ont été utilisés. Le premier, un classificateur de régression logistique, est une méthode linéaire qui cherche à trouver la meilleure frontière de décision entre les différentes classes en minimisant la fonction de coût. Le deuxième classificateur est un modèle de machine à vecteurs de support (SVM) avec un noyau linéaire, qui est connu pour sa capacité à maximiser la marge entre les classes et à gérer efficacement les problèmes de classification à haute dimensionnalité. Enfin, un classificateur basé sur les k-plus proches voisins (kNN) a été testé avec un paramètre k de 5, ce qui permet de classifier un visage en fonction de la majorité des voisins les plus proches dans l'espace des caractéristiques.

Ces classificateurs ont été choisis en raison de leur popularité et de leur efficacité dans des tâches de classification de données complexes telles que la reconnaissance faciale. Le choix d'un réseau de neurones permet d'exploiter la puissance de l'apprentissage profond, tandis que les autres modèles permettent de comparer des méthodes plus classiques et d'identifier celle qui performe le mieux sur les données spécifiques de notre projet.

Pour évaluer les modèles, deux métriques sont utilisés, à savoir la précision et le temps de prédiction. Pour avoir une comparaison équitable, chaque modèle a été testé sur les mêmes données d'entraînement et de test. L'ensemble des modèles ont obtenu une précision de 97 %. Bien que le modèle de régression logistique utilise une méthode simple, il reste compétitif et atteint un temps de prédiction très court de 0,4 ms. Le modèle SVM présente un temps de prédiction légèrement plus long, autour de 1 ms. Le classificateur kNN a obtenu un temps de prédiction encore plus élevé, d'environ 4 ms. Cette méthode est peut-être plus lente à mesure que la taille de l'ensemble de données augmente car elle nécessite le calcul des distance entre les points à chaque prédiction. Enfin, le réseau de neurones obtient un temps de prédiction notablement plus long, autour de 60 ms. Bien qu'il soit capable de capturer des relations non linéaires complexes dans les données, son temps de prédiction élevé peut être un inconvénient par rapport aux autres modèles.

En terme de précision, tous les classifieurs ont montré une performance équivalente de 97 %. Toutefois, la régression logistique et le SVM se distinguent par leur rapidité, ce sont donc les choix les plus appropriés.

```
Logistic Regression - Accuracy: 0.97, Prediction Time: 0.0004s
SVM - Accuracy: 0.97, Prediction Time: 0.0012s
kNN - Accuracy: 0.97, Prediction Time: 0.0048s
Neural Network - Accuracy: 0.97, Prediction Time: 0.0615s
```

Figure 6 – Comparaison des modèles

E) Ensemble de données personnelles

Pour cette partie du projet, nous avons constitué notre propre ensemble de données regroupant des photos de dix individus. Ce groupe inclut nous-mêmes, un collègue, plusieurs membres de nos familles (célébrités ou non), ainsi que deux personnalités publiques additionnelles. La collecte des images a été réalisée selon plusieurs méthodes.

Pour les personnes connues, les images ont été soigneusement sélectionnées une à une à partir de moteurs de recherche, en privilégiant la variation des angles, expressions et conditions d'éclairage. Concernant notre collègue, les images ont été extraites d'une vidéo dans laquelle il présente son visage sous différents angles et expressions, avec une netteté d'image variable. Enfin, les autres images ont été obtenues soit en les prenant spécifiquement pour ce projet, soit en les sélectionnant dans nos galeries personnelles ou parmi des clichés de vacances.

Dans les cas où une image comportait plusieurs visages, les visages non concernés ont été masqués afin de ne pas polluer le jeu de données lors de l'extraction. Nous avons également pris soin de convertir les fichiers dans des formats compatibles (jpg, jpeg, png), et d'orienter correctement les visages pivotés à plus de 90°, spécifiquement ceux que le détecteur de visages ne reconnaissait pas.

Au final, notre base de données personnelles contient entre 50 et 110 images par individu, avec une diversité appréciable en termes de poses, d'expressions faciales et de qualité.

Le modèle de régression logistique, qui avait montré de très bonnes performances sur le jeu de données initial, atteint une précision comprise entre 94 % et 98 % sur notre propre ensemble de test. Cette variation s'explique principalement par le caractère aléatoire du mélange des images avant leur répartition dans les ensembles d'entraînement et de test. En moyenne, bien que le

modèle conserve une précision élevée, ses résultats restent légèrement inférieurs à ceux obtenus précédemment (97 %), ce qui peut s'expliquer par plusieurs facteurs.

Notre jeu de données personnel présente en effet davantage de variabilité que le jeu de données utilisé initialement. Certaines images souffrent de fortes variations d'éclairage, d'autres sont de mauvaise qualité ou floues, ce qui peut compliquer la tâche de l'encodeur lors de l'extraction des descripteurs de visage. Dans certains cas, les visages sont partiellement masqués ou présentés de profil, ce qui rend leur représentation moins fiable et peut nuire à la performance du classificateur.

On observe par ailleurs que lorsque la précision atteint des valeurs élevées (jusqu'à 98 %) sur l'ensemble de test, elle est accompagnée d'une performance inférieur sur l'ensemble de validation. Cela suggère une possible sous-représentation des cas difficiles dans le test, due au tirage aléatoire lors du partitionnement des données.

En comparaison, le jeu de données précédent était plus propre et mieux contrôlé (photos issues de films, les deux yeux sont toujours visibles, résolution correcte), ce qui a probablement facilité la détection et l'encodage des visages. Ainsi, bien que notre modèle fonctionne globalement bien sur notre jeu de données, la baisse de performance souligne l'importance de la qualité et de la diversité des données d'entrée dans un système de reconnaissance faciale.

F) Analyse des biais

En apprentissage automatique, un biais désigne une distorsion ou un déséquilibre dans les données ou dans les hypothèses du modèle, qui entraîne une représentation inexacte de la réalité. Il peut s'agir de biais dans les données d'entraînement (ex : surreprésentation de certaines classes), dans la manière dont les données sont collectées, ou encore dans la façon dont le modèle apprend et généralise.

Les biais peuvent compromettre la fiabilité et la justice des modèles. Un modèle entraîné sur des données biaisées risque de mal généraliser sur de nouvelles données, notamment si ces dernières sont issues de groupes ou de situations sous-représentés. Cela peut conduire à des discriminations involontaires, des erreurs de prédiction, voire des conséquences graves dans des domaines sensibles comme la santé, la sécurité ou la justice.

Dans le cadre de la reconnaissance faciale, les biais sont particulièrement critiques. Par exemple, si un modèle est principalement entraîné sur des visages clairs, il risque d'avoir des performances nettement inférieures sur les visages plus foncés. Cela a été démontré dans plusieurs études (notamment celle du MIT Media Lab en 2018), où certains systèmes affichaient des taux d'erreur beaucoup plus élevés pour les femmes noires que pour les hommes blancs.

Si le jeu de données ne contient que des personnes d'un certain groupe ethnique, d'un âge similaire, ou prises dans des contextes similaires (éclairage, qualité, angles), notre modèle risque de mal reconnaître des personnes en dehors de ces caractéristiques. Il serait alors biaisé, car il n'aurait pas été exposé à une diversité suffisante de cas lors de l'apprentissage.

Nous avons pris soin de diversifier notre ensemble de données personnelles en y incluant des personnes de tranches d'âge variées, de genres différents et d'origines ethniques diverses. En examinant les scores F1 par classe, il ne semble pas exister de biais manifeste en faveur d'un groupe particulier : les scores sont globalement élevés et équilibrés, ce qui suggère une bonne capacité de généralisation du modèle sur notre ensemble hétérogène.

Plusieurs statistiques peuvent être calculées à partir du jeu de données original de détection de visages, comme par exemple : la répartition des angles de vue (profil, 3/4, face), la distribution des conditions d'éclairage, ou encore le nombre de visages masqués ou partiellement visibles. Ce type de statistiques permettrait de mieux comprendre pourquoi certaines classes sont plus difficiles à reconnaître, et d'identifier d'éventuelles sources de biais ou de bruit dans les données.

	precision	recall	f1-score	support
Classe 01	1.00	1.00	1.00	12
Classe 02	1.00	0.91	0.95	11
Classe 03	1.00	0.83	0.91	12
Classe 04	1.00	0.92	0.96	12
Classe 05	1.00	1.00	1.00	15
Classe 06	0.94	1.00	0.97	16
Classe 07	0.86	1.00	0.92	12
Classe 08	1.00	1.00	1.00	16
Classe 09	0.91	0.91	0.91	11
Classe 10	0.94	1.00	0.97	15
accuracy			0.96	132
macro avg	0.96	0.96	0.96	132
weighted avg	0.97	0.96	0.96	132

Figure 7 – Score F1 par classe

IV. CONCLUSION

Ce travail nous a permis de concevoir et d'évaluer une chaîne complète de reconnaissance faciale, allant de la détection de visages à l'identification de l'individu. À travers l'étude et l'expérimentation de différentes techniques comme l'estimation de la pose, l'encodage de visages ou encore l'utilisation de divers classificateurs, nous avons pu observer l'impact concret de chaque étape sur les performances globales du système. L'encodage s'est révélé être un levier particulièrement efficace, permettant d'atteindre une précision allant jusqu'à 97 %, tout en réduisant considérablement la complexité du traitement.

Au-delà des performances techniques, ce projet a aussi mis en lumière l'importance de la qualité et de la diversité des données, ainsi que les risques liés aux biais dans les jeux de données. L'expérimentation avec un jeu de données personnel a renforcé notre compréhension des limites et des exigences pratiques d'un tel système dans un contexte réel. Ce travail a donc été l'occasion d'acquérir une vision globale, critique et appliquée des technologies de reconnaissance faciale.

V. RÉFÉRENCES

[1] M. Turk and A. Pentland, "*Eigenfaces for recognition*", Journal of Cognitive Neuroscience, vol. 3, no. 1, pp. 71–86, Jan. 1991, doi: https://doi.org/10.1162/jocn.1991.3.1.71.

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "*DeepFace: Closing the Gap to Human-Level Performance in Face Verification*", in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[3] F. Schroff, D. Kalenichenko, and J. Philbin, "*FaceNet: A Unified Embedding for Face Recognition and Clustering*", in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 815-823.

[4] R. Bruyer, *Les Mécanismes de reconnaissance des visages*. PUG, 1987. Available: https://www.decitre.fr/livres/les-mecanismes-de-reconnaissance-des-visages-9782706102813.html

[5] R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, 2nd ed. Cambridge: Cambridge University Press, 2004. Available: https://www.cambridge.org/core/books/multiple-view-geometry-in-computer-vision/0B6F289C78B2B23F596CAA76D3D43F7A

[6] O. Parkhi, A. Vedaldi, and A. Zisserman, "*Deep Face Recognition*", 2015. Available: https://www.robots.ox.ac.uk/~vgg/publications/2015/Parkhi15/parkhi15.pdf

[7] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "*Face recognition*", ACM Computing Surveys, vol. 35, no. 4, pp. 399–458, Dec. 2003, doi: https://doi.org/10.1145/954339.954342.