N° d'ordre : 016 / STI / TCO Année Universitaire : 2019 / 2020

UNIVERSITE D'ANTANANARIVO

ECOLE SUPERIEURE POLYTECHNIQUE

MENTION TELECOMMUNICATION

MEMOIRE

en vue de l'obtention

du DIPLOME de Master

Domaine : **Sciences de l'Ingénieur**

Mention : **Télécommunication**

Parcours : **Système et Traitement de l'Information**

Présenté par : **RAKOTONARIVO Laza Manampisoa**

CONCEPTION D'UN SYSTEME INTELLIGENT POUR LA VIDEO SURVEILLANCE

Soutenu le **21 mars 2022** devant la Commission d'Examen composée de :

Président :

M. RATSIHOARANA Constant

Examinateurs :

Mme. RAMAFIARISONA Hajasoa Malalatiana

M. RATSIMBAZAFY Andriamanga

M. RATSIMBAZAFY Tsiory Harifidy

Directeur de mémoire :

M. RAVONIMANANTSOA Ndaohialy Manda-Vy

REMERCIEMENTS

Avant tout, nous tenons à glorifier Dieu tout puissant de m'avoir soutenue, de m'avoir donné force et santé durant l'élaboration de ce mémoire.

Ce présent travail a été achevé grâce à l'aide et au soutien des personnes que nous tenons tout particulièrement à remercier :

- Monsieur RAVELOMANANA Mamy Raoul, Professeur Titulaire, Président de l'Université d'Antananarivo.
- Monsieur RAKOTOSAONA Rijalalaina, Professeur et Responsable du domaine Science de l'ingénieur à l'École Supérieur Polytechnique d'Antananarivo.
- Monsieur RAKOTONDRAINA Tahina Ezéchiel, Maître de Conférences, Responsable de la mention Télécommunication.
- Monsieur RATSIHOARANA Constant, Maître de conférences, qui nous fait également l'honneur de présider le jury de la soutenance de mémoire.
- Nous tenons à témoigner notre reconnaissance et notre gratitude à Monsieur RAVONIMANANTSOA Ndaohialy Manda-Vy, Professeur, mon encadreur, qui m'a étroitement assisté depuis le début de la réalisation de ce présent mémoire, et qui s'est toujours montré à l'écoute tout au long de son élaboration.

Tous les membres du jury, à savoir :

- Madame RAMAFIARISONA Hajasoa Malalatiana, Professeur de l'Université d'Antananarivo.
- Monsieur RATSIMBAZAFY Andriamanga, Maître de Conférences de l'Université d'Antananarivo.
- Monsieur RATSIMBAZAFY Tsiory Harifidy, Docteur de l'Université d'Antananarivo.

TABLE DES MATIERES

NOTATIONS ET ABREVIATIONS

1. Minuscules latines

b	Biais pour la sortie du neurone
f	Fonction quelconque
h	Fonction quelconque
x_i	Entré du réseau de neurone

2. Majuscules latines

2D	2 Dimension
3D	3 Dimension
C	Colonne
C	Contraste
D	Dimension
D_{1H}	Matrice filtre dérivé horizontale
D_{2V}	Matrice filtre dérivé verticale
I	Image a traité
L_1	Degré de luminosité d'une zone A_1 de l'image
L_2	Degré de luminosité d'une zone A_2 de l'image
L_P	Matrice filtre de Laplace
L_F	Matrice filtre lissage fort
L_M	Matrice filtre lissage moyen
M	Nombre de ligne d'une image
N	Nombre de colonne d'une image
S_H	Matrice filtre de Sobel horizontale
S_V	Matrice filtre de Sobel verticale

3. Minuscules grecques

Δ	Delta
η	Pas de l'apprentissage
ω_{ji}	Poids synaptique entre les neurones ket j

ω_i	Poids synaptique de l'entrée x_i

4. Majuscules grecques

Δ_p	Correction à effectuer pour la valeur du poids

5. Abréviations spéciales

ADALINE	ADAptive LInear NEuron
API	Application Interface Programming
ART	Adaptive Resonance Theory
AVI	Audio Video Interleave
BN	Batch Normalisation
CPL	Courant Porteur en Ligne
CPU	Central Processing Unit
CSS	Cascading Style Sheets
CNN	Convolutional Neural Network
DCNN	Deep Convolutional Neuronal Network
DSCNN	Deep Separable Convolutional Neuronal Network
FLV	Flash Video
HTML	HyperText Markup Language
IA	Intelligence artificiel
Ip	Internet Protocol
KNN	K- Nearest Neighbor

LAN	Local Area Network
GPU	Graphic Processing Unit
JS	JavaScript
MKV	Matroska Video
MPEG-4	Moving Picture Expert Group 4
MPL	Multi-Layer-Perceptron
MSE	Minimum Square Error
MVC	Modèle Vue Conceptuel
NLP	Natural Language Processing
NTSC	National Television System Committee
PAL	Phase Alternating Line
Pc	Personal Computer
ReLU	Rectified Linear Unit
RBF	Radial Basis Function
ROI	Region Of Interest
RVB	Rouge Vert Bleu
SECAM	Séquentielle Couleur à Mémoire
SSD	Single Shot MultiBox Detector
SVM	Support Vector Machine
VJ	Viola-Jones
VPN	Virtual Private Network
VSI	Vidéo Surveillance Intelligente
Wi-Fi	Wireless -Fidelity
WMV	Windows Media Video

INTRODUCTION GENERALE

Au cours de la dernière décennie, la communauté de la recherche en vision par ordinateur a montré beaucoup d'intérêt pour l'analyse et la reconnaissance automatique d'image faciale. Initialement inspirée par les découvertes des chercheurs en sciences cognitives, la communauté de la vision par ordinateur et de la recherche scientifique envisageait de développer des systèmes capables de reconnaître l'identité, l'âge, le sexe et l'expression faciale d'une personne. Dans ce projet de mémoire, nous allons nous intéresser à ce système d'analyse appliquer dans la vidéo surveillance.

Dans la vie quotidienne, chacun de nous identifie tout au long de la journée différents visages. Ainsi lorsque nous rencontrons une personne, notre cerveau va chercher dans notre mémoire, analyser et vérifie si cette personne est répertoriée ou non, est-ce un homme ou une femme ? quel âge a-t-elle approximativement ? De quelle humeur est-elle ? C'est une tâche difficile pour les humains face à une masse personne. En est-il de même pour une machine ?

Dans ce projet, nous allons justement présenter les grandes lignes de notre travail, qui s'intitule « **Conception d'un système intelligent pour la vidéo surveillance** » qui permettra : la détection du repère facial et des 68 points de repère facial, la reconnaissance faciale, la reconnaissance des expressions faciales, l'estimation de l'âge et la reconnaissance du genre. Pour ce faire, notre travail est composé de quatre chapitres :

Dans le premier chapitre nous allons voir la surveillance et la notions sur l'apprentissage machine. Ensuite, en deuxième chapitre nous allons voir de manière générale la vidéo et le traitement d'image : nous expliquerons les notions importantes du traitement d'image.

La troisième partie se concentre sur l'apprentissage profond notamment les modèles de réseaux neuronaux et les classificateurs. Notre travail repose entièrement sur la mise en pratique de ces techniques.

Le quatrième et dernier chapitre va entamer la réalisation du projet. À la fin de ce chapitre, nous allons faire des tests d'évaluations afin de prouver l'efficacité et performance de notre système.

CHAPITRE 1

LA SURVEILLANCE ET LA NOTION SUR L'APPRENTISSAGE MACHINE

1.1 Introduction

Depuis quelque temps, l'apprentissage automatique tient un rôle plus que déterminant dans le domaine de l'ingénierie. Tous les outils d'ingénierie de nos jours tendent à utiliser cette technologie pour plus de performance et de précision dans les résultats. Dans ce mémoire nous allons le proposer dans la vidéo surveillance. La vidéo surveillance est, le plus souvent, implantée dans le cadre d'un programme de prévention de la criminalité ou de renforcement de la sécurité publique.

1.2 Vidéo surveillance

1.2.1 Définition

La vidéo surveillance est un système de caméras permettant de surveiller un espace privé ou public. Des images sont enregistrées avec ce système et sont par la suite visionnées et sauvegardées. Les systèmes de la vidéo surveillance sont composés de différents types de matériel en fonction des besoins de son utilisateur (les caméras de surveillance, l'écran de la vidéo surveillance, l'alimentation des caméras de la vidéo surveillance, les enregistreurs de la vidéo surveillance, les câbles de la vidéo surveillance ou les liaisons sans fil …etc.). [1]

1.2.2 Objectifs

L'objectif général d'un système de vidéo surveillance est de contribuer à la sécurité de biens et/ou de personnes. [1]

Cette contribution peut se focaliser sur diverses composantes :

- Prévention de la criminalité.
- Sécurité routière.
- Sécurité industrielle.
- Sûreté.
- Collecte de donnée.

1.2.3 Les problèmes liés à la vidéo surveillance

Le plus important des problèmes est lié à l'atteinte de la vie privée puisque nous sommes constamment surveillés. Les familles ne se sentent pas à l'aise en sachant qu'une caméra les surveille en permanence. Cela pose une restriction des libertés.

L'autre problème est celui de la mise en place de caméra qui est très coûteuse puisque si l'on souhaite un système sophistiqué de vidéo surveillance cela représentera un investissement considérable et à cela s'ajoutera la maintenance du système qui devra être régulière.

Enfin, cela enlève de l'emploi des vigiles dans la zone concernée. [2]

1.2.4 Les types de vidéo surveillance

Il existe deux types principaux de mode de vidéo surveillance :

- la vidéo surveillance analogique, plus traditionnelle.
- La vidéo surveillance IP.

1.2.4.1 Vidéo surveillance analogique

Le système de la vidéo surveillance analogique est équipé de caméras analogiques dont le seul rôle est de capturer les images et les envoyer à un enregistreur à durée limitée (type magnétoscope) via un signal analogique. Ce système est composé d'une ou plusieurs caméras, d'un moniteur (ou téléviseur), d'un enregistreur et d'un câblage (le transfert d'images se fait via un câble dit coaxial).

a) Avantages de la vidéo surveillance analogique

Les avantages de la vidéo surveillance analogique sont :

- La qualité et la fluidité des images ainsi que sa facilité d'utilisation.
- La grande diversité de caméras analogiques : dimensions, formes, applications.
- Le prix des caméras analogiques : plus économique que celui des caméras IP.

b) Inconvénients de la vidéo surveillance analogique

Par rapport à la vidéo surveillance en réseau, le système analogique est assez limité en termes de fonctionnalités :

- La capacité de stockage limitée dans la durée (cassettes) ;

- Le format peu flexible des images ;
- Pas d'accès en temps réel par l'Internet (sauf si les caméras sont connectées à un réseau) et l'Internet mobile ;
- L’évolutivité limitée : difficile d'ajouter des caméras supplémentaires dans le temps (longueur de câble, travaux...) ;
- Pas de gestion à distance : installation, maintenance.
- Dépasser par le temps. [2]

1.2.4.2 Vidéo surveillance IP

La vidéo surveillance IP (Internet Protocole) est venu compléter la vidéo surveillance analogique. Elle fonctionne avec les mêmes composants (caméras, moniteur, enregistreur, câbles) mais passe par un réseau informatique :

- Les caméras IP sont installées sur un réseau IP (Intranet, Internet, LAN -réseau local-, CPL - courant porteur en ligne, c'est-à-dire les prises électriques - ou VPN...) et reliées à un serveur de vidéo surveillance centrale : elles capturent les images et les acheminent vers le réseau ;
- Le serveur de vidéo surveillance est équipé d'un logiciel de vidéo surveillance : c'est lui le cœur du système, il récupère les images et les stocke sur disque dur.

Avantages de la vidéo surveillance IP :

- Bénéficie de toutes les fonctionnalités d'Internet.
- Multiple grâce à la technologie réseau : la vidéo surveillance IP peut s'intégrer à d'autres technologies qui relèvent de l'IP (système de sécurité, visioconférence...).
- Peut s'installer sur un réseau informatique qui existe déjà (économie).
- Compatible avec tous types de câblage : IP, coaxial... ou sans fil (Wi-Fi).
- Tout passe par un ordinateur : pas besoin de moniteur, d'enregistreur.
- Grande flexibilité d'installation des caméras : possibilité d'en ajouter facilement ou de les changer de place.
- Visualisation des images en temps réel.

- Réglage possible des images (dimension, zoom...).
- Les caméras peuvent être commandées à distance (selon les modèles).
- Verrouillage de l'accès aux images par mot de passe peut se coupler avec un système de sécurité (alarme...)**.**
- Possibilité d'intégration de système intelligent.

Inconvénients de la vidéosurveillance IP :

- Le prix des matériels et de son installation.
- Complexe : nombreux réglages, présence d'un informaticien indispensable. [2]

1.2.5 Domaines d'application de la vidéo surveillance

De nos jours, la vidéosurveillance est utilisée par un grand nombre de commerces, d'entreprises, de résidences et d'institutions. La vidéo surveillance a un champ d'application illimité : [2]

- L'industrie.
- Le transport privé ou collectif.
- Le commerce/marketing et la distribution.
- Les administrations et les services publics.
- La santé.
- Les lieux publics.
- L'enseignement.
- Les banques.
- Les loisirs.

1.3 Avantage d'une vidéo surveillance intelligente

Une caméra d'une vidéo surveillance est classée comme intelligente dès lors que son système lui permet d'analyser une image avec plus d'approfondissement que les fonctions classiques de détection de mouvement par exemple.

La vidéo surveillance intelligente présente plusieurs avantages par rapport à son domaine d'utilisation et les fonctionnalités qui lui sont inclus. Nous allons citer quelques avantages globaux et ensuite citer les avantages par rapport à notre système intelligent (la détection de repère faciale et des points de repère faciaux, la reconnaissance faciale, la reconnaissance d'émotion, la reconnaissance du genre et l'estimation de l'âge). [2] [3]

1.3.1 Avantages globaux

Les systèmes de la VSI (vidéo surveillance intelligente) sont capables d'analyser, de diagnostiquer (trier, prévenir, estimer, comptabiliser, détecter), d'assurer et alerter en live pour l'utilisation public et privé. Voici ces avantages illustrer avec quelques exemples :

- La **détection rapide d'anomalies** qui permet d'alerter et d'intervenir efficacement : alerter les comportements suspects d'un individu mal intentionné et contacter automatiquement la police, alerter un accident routier et les incendies, contrôler la sureté dans un hôpital (alerter les médecins pour le malaise ou la crise d'un patient), suivit automatique du déplacement d'une silhouette dans une scène tout en éliminant tous les éléments annexes perturbateurs et sans intérêt (vibrations, variations de la luminosité, mouvements occasionnés par le vent dans les feuillages, passage d'oiseaux…), la détection du non-port des équipements de sécurité d'un employé, le respect des gestes barrières (distance 1 mètre et le port de cache bouche) ;

- **Se doter de data** : pour les prises de décisions commerciales par rapport aux personnalités des clients intéresser par un produit ou une boutique pour accroître les ventes par exemple, le pointage biométrique faciale de chaque employé, le nombre d'entrées et de sorties dans un local fait par un individu, comptabiliser et automatiser les trafics routiers, collecter et identifier les plaques d'immatriculation sur les trafics et les parkings, la sûreté ou les problématiques métiers propres à chaque entreprise, les enquêtes et les recherches d'individu ou autres qui peut faire office de preuves tangibles, la prédiction d'un danger potentiel comme une collision des trains ou des voitures sur les voies publiques et prendre des précautions.

- **Gagner du temps** grâce aux analyses intelligentes des images vidéo (divers types de reconnaissance intelligent) destinées à assister les agents de sécurité ou les personnels : l'humain peut se concentrer sur des tâches à valeurs ajoutées pendant que son système de vidéosurveillance intelligent trie et analyse.

- La **vision globale et la gestion simultanée des espaces à distant :** des points considérés comme stratégiques pour une entreprise ou le gouvernement ou un particulier, même ceux

difficiles d'accès. Il peut s'avérer fastidieux d'organiser une équipe de sécurité de manière efficace et fiable, le système de caméra surveillance permet, par un simple positionnement stratégique, de quadriller une zone étendue avec d'excellents angles de vue depuis des points d'observation surélevés.

- **Influencer les concernés** : la simple mise en place d'une telle installation permet de dissuader tout individu mal intentionné, les employer ou simple individu dans un local privé ou public (écarter et réduire les risques et les dangers). [3] [4]

1.3.2 Avantage par rapport à notre projet de mémoire

Dans ce projet de mémoire nous avons fait un système de VSI qui est la combinaison de sept (07) fonctionnalités. Nous avons pu inclure les basiques requis pour qu'un VSI soit considéré comme intelligent pour notre simulation : ce sont la détection de repère faciale et la détection des points de repère faciale, la reconnaissance faciale, la reconnaissance d'émotion, la reconnaissance du genre, l'estimation de l'âge.

Les avantages que présentent ses fonctionnalités basiques dépendent du domaine de son application, donc se réfèrent aux avantages globaux cités au-dessus.

Cependant nos fonctionnalités actuelles nous permettent les avantages spécifiques suivants :

- **Identifier une ou plusieurs personnes en même temps ou simultanément** et de savoir leurs personnalités (nom, sexe, émotion, âge estimer) ;

- **Gagner du temps** grâce aux analyses intelligentes des images vidéo (divers types de reconnaissance intelligent) destinées à assister les agents de sécurité ou les personnels : l'humain peut se concentrer sur des tâches à valeurs ajoutées pendant que son système de vidéosurveillance intelligent trie et analyse. Elle vient assister l'opérateur humain et le rend plus performant.

- **Influencer les concernés** : la simple mise en place d'une telle installation permet de dissuader tout individu mal intentionné, les employer ou simple individu dans un local privé ou public (écarter et réduire les risques et les dangers).

1.3.3 Limites et contrainte globale d'un système VSI

- L'intelligence artificielle, tel qu'elle existe aujourd'hui, permet uniquement d'automatiser des process informatiques et électroniques : détections dans les images vidéo, déclenchements

d'alarme ou d'actions telles que des fermetures automatiques, analyse et prédiction, etc. Elle n'ira pas plus loin que ce qu'on lui a appris et elle ne sera pas capable de faire preuve de bon sens. C'est pourquoi un dispositif de vidéosurveillance intelligente nécessite toujours l'analyse minimum d'un opérateur humain.

- La surveillance vidéo est évidemment un système cadré par la loi et soumis à plusieurs règles et normes. Ainsi, l'entreprise doit répondre à l'obligation d'information et au respect du droit à l'image et à l'atteinte de la vie privée. Donc, le responsable des zones sous VSI doit informer son public qu'il se trouve dans un lieu sous vidéosurveillance.

Vu que la vidéosurveillance intelligente est rendue intelligence grâce à l'intelligence artificielle, nous allons maintenant parler des notions de celle-ci. [3] [4]

1.4 Historique et contexte de l'apprentissage machine

L'apprentissage automatique tient son histoire dans l'évolution de l'intelligence artificielle. La publication en 1950 de l'article scientifique « Computing Machinery and Intelligence » par Alan Turing est considérée par certains comme l'acte de naissance de l'Intelligence artificielle, Turing étant reconnu comme le père de cette discipline à convergence de plusieurs champs de connaissances technique et scientifique.

Les tout débuts de l'intelligence artificielle datent donc de 70 ans, même si la table avait été mise avant Turing par d'autres visionnaires. C'est notamment dans l'article cité en haut que le brillant mathématicien britannique inventera le célèbre « test de Turring » qui, pour résumer, consiste à établir une conversation par messages écrits et à l'aveugle entre un humain, un autre humain et une machine afin que le premier détecte lequel de ses 2 interlocuteurs est la machine.

Depuis, l'intelligence artificielle s'est beaucoup évoluée et englobe plusieurs disciplines présentées par la figure dans la page suivante. [5]

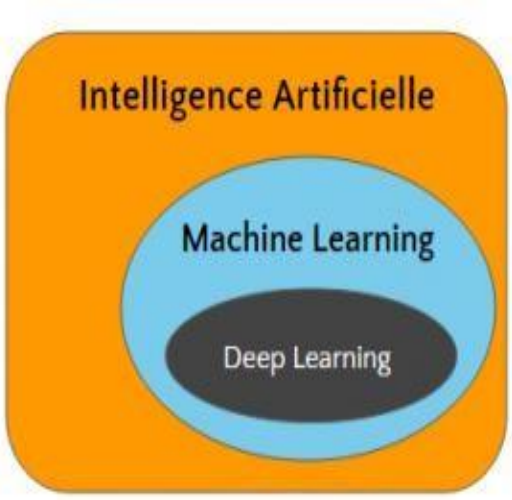


Figure 1.1 : *Schéma des grandes disciplines de l'intelligence artificielle*

1.5 Principe de l'apprentissage automatique

Discipline de l'intelligence artificielle, l'apprentissage automatique (Machine Learning) consiste à générer des connaissances de façons automatiques en exploitant des données brutes. À partir des connaissances acquises, il est alors possible pour la machine de créer un modèle permettant de prendre des décisions [5].

1.5.1 Cycle de l'apprentissage automatique

L'ensemble du cycle d'apprentissage automatique peut être résumé comme suit :

- **Acquisition des données**

La première étape consiste à obtenir des données pertinentes pour l'application à développer. Les données doivent être de grande qualité et détaillées.

- **Préparation des données**

Cette étape est également appelée nettoyage des données. Les données doivent être précises, propres et sécurisées.

- **Sélection de l'algorithme**

L'algorithme le plus approprié pour l'application à développer doit être choisi.

- **Entraînement du modèle**

Un modèle d'apprentissage automatique est une représentation mathématique d'un processus réel.

L'algorithme retenu doit être entraîné sur les données pour créer le modèle. Le processus d'entraînement peut être supervisé, non supervisé ou renforcé.

- **Évaluation**

Le modèle doit être évalué pour s'assurer que l'algorithme retenu est le mieux adapté.

- **Déploiement**

Il faut décider si le modèle doit être déployé dans le nuage informatique ou sur place.

- **Test**

Le modèle doit être testé avec des données nouvelles et pour faire des prédictions.

- **Évaluation**

La validité des prédictions établies par le modèle doit être évaluée, et le raffinement des données, du modèle et de l'algorithme doit être mis en œuvre selon qu'il convient. [5] [6]

1.6 Type d'apprentissage automatique

L'apprentissage automatique est un sous-ensemble de l'intelligence artificielle abrégé IA, qui est elle-même un sous-ensemble de la science des données. Il concerne les analyses descriptives, diagnostiques, prédictives et prescriptives. L'analyse descriptive se rapporte à ce qui s'est passé ; l'analyse diagnostique explique pourquoi c'est arrivé ; l'analyse prédictive permet de prévoir ce qui est le plus susceptible de se produire à l'avenir ; et l'analyse prescriptive recommande le plan d'action le plus logique pour atteindre le résultat souhaité. [7]

1.6.1 Apprentissage supervisé

1.6.1.1 Principe

Les algorithmes d'apprentissage supervisé font des prévisions en fonction d'exemples. Exemple un historique de vente pour déterminer des prix futurs. Dans un tel cas, il y a une variable d'entrée composée de données d'entraînement étiquetées et d'une variable de sortie souhaitée. Un algorithme est utilisé pour analyser les données d'entrainement afin d'apprendre la fonction qui associe l'entrée à la sortie. Cette fonction permet de procéder à une mise en correspondance de nouveaux exemples en généralisant à partir des données d'entrainement pour anticiper les résultats de situation non connue. [8]

1.6.1.2 Problèmes utiliser pour l'apprentissage supervisé

Les problèmes utilisés pour l'apprentissage supervisé sont :

- **Classification**

Lorsque les données servent à prédire une variable catégorielle, l'apprentissage supervisé est également appelé classification. C'est le cas par exemple lorsqu'une étiquette ou un indicateur par exemple « chien » ou « chat ») est attribué à une image. Lorsqu'il n'y a que deux étiquettes, on parle de classification binaire. Lorsqu'il y a plus de deux catégories, on parle de classification en classes multiples.

- **Régression**

Lorsqu'on procède à la prédiction de valeurs continues, on parle de régression.

- **Prévision**

Il s'agit de faire des prédictions à partir de données passées et présentes. Ce type de processus sert le plus souvent à analyser des tendances, p. ex. estimer les ventes de l'année prochaine à partir des ventes de l'année en cours et des années précédentes. [8] [9]

1.6.1.3 Algorithme d'apprentissage supervisé

Les exemples d'algorithme d'apprentissage supervisé sont : le boosting, machine à vecteur de support (SVM), mélanges de lois, réseau de neurones artificiels, méthode des k plus proches voisins, arbre de décision, classification naïve bayésienne, inférence grammaticale, espace de versions. [8] [9]

1.6.1.4 Domaines d'application

À ces jours, les algorithmes d'apprentissage supervisé couvrent des centaines d'applications réparties sur plusieurs domaines. Les paragraphes ci-dessous présentent quelques applications basiques. [9] [10]

a. Reconnaissance faciale

Le système de la reconnaissance faciale est une application logicielle visant à reconnaître une personne grâce à son visage de manière automatique. À l'aide d'algorithmes, cette application analyse toutes les caractéristiques faciales telles que l'écartement des yeux, des arêtes du nez, des commissures des lèvres, des oreilles, du menton, à partir d'une image de son visage qui peut provenir à la fois d'une photo ou d'une vidéo. [9] [10]

b. Reconnaissance vocale

Tous les logiciels de reconnaissances vocales utilisent tous la machine Learning. Les reconnaissances vocales évoquent 2 phases d'apprentissages : le premier, avant que le logiciel soit distribué (phase d'entrainement générale du système indépendamment de l'utilisateur), et le second, à l'acquisition du logiciel (pour obtenir un meilleur résultat en fonction de l'utilisateur) [9] [10].

1.6.2 Apprentissage semi-supervisé

Dans l'apprentissage semi-supervisé, l'étiquetage des données peut être long et coûteux. Si les étiquettes sont limitées, il est possible d'utiliser des exemples non étiquetés pour améliorer l'apprentissage supervisé. Étant donné que la machine n'est pas entièrement supervisée, on emploie le terme « semi-supervisé ». En ce qui concerne l'apprentissage semi-supervisé, on l'utilise des exemples non étiquetés et une petite quantité de données étiquetées pour améliorer la précision de l'apprentissage.[11].

1.6.3 Apprentissage non supervisé

1.6.3.1 Principe

L'apprentissage non supervisé constitue un processus itératif d'analyse de données sans intervention humaine. Contrairement à l'apprentissage supervisé, l'apprentissage non supervisé consiste à inférer des connaissances sur les données. Elle recherche la structure naturelle dans les données. La machine reçoit des données non étiquetées et on lui demande de découvrir les schémas qui sous-tendent les données, p. ex. une structure en grappes, une variété en basses dimensions, ou un arbre et un graphique de faible densité. [10]

1.6.3.2 Problèmes utiliser pour l'apprentissage non supervisé

L'apprentissage non supervisé convient lorsqu'un problème nécessite une quantité considérable de données non étiquetées.

Différentes tâches sont associées à l'apprentissage non supervisé :

- **Le Clustering** (segmentation, regroupement ou mise en grappe) :

  Il s'agit de regrouper des exemples de données afin que les exemples d'un groupe (ou d'une grappe) ressemblent plus (selon certains critères) aux exemples d'un autre groupe. Ce processus est souvent utilisé pour segmenter un ensemble complet de données en plusieurs groupes. Une analyse peut être effectuée dans chaque groupe afin de trouver les modèles intrinsèques. Il construit des classes automatiquement en fonction des exemples disponibles

- **Règles d'association** :

  Analyser les relations entre les variables ou détecter des associations

- **Réduction de dimensions** :

  Il s'agit de réduire le nombre de variables examinées. Dans de nombreuses applications, les données brutes possèdent de nombreuses caractéristiques dimensionnelles, dont certaines sont superflues ou non pertinentes. Réduire les dimensions permet donc de trouver la véritable relation latente.

On obtient des modèles descriptifs qui permettent de mieux connaitre ses données, de découvrir des informations cachées dans la masse de données. [10] [12]

1.6.3.3 Domaines d'applications

L'apprentissage non supervisé est surtout utilisé dans les systèmes qui gèrent les données massives non structurées. L'apprentissage non supervisé peut constituer la première étape d'un traitement avant de soumettre les données à l'apprentissage supervisé : étant donné que le développeur ne connaît pas le contexte des données en cours d'analyse, l'étiquetage n'est pas possible à ce stade.

Bien qu'elle soit de nos jours en plein essore et s'étend dans plusieurs domaines, voici quelques applications les plus fréquentes de cet algorithme. [10]

a. Analyse de réseaux sociaux

L'apprentissage non supervisé peut automatiquement identifier des amis dans un cercle d'amis Facebook ou Google, et peut aussi identifier le nombre maximum d'e-mails envoyé par une personne particulière dans une catégorie de groupe d'amis. Il peut aussi identifier quel groupe de personnes se connaît [10].

b. Segmentation de marché

Beaucoup de compagnies possèdent une très grande base de données d'information sur sa clientèle. L'algorithme d'apprentissage non supervisé analyse ces données client puis permet automatiquement d'assigner à chaque utilisateur les espaces de marché qui peuvent leur convenir [10].

1.6.4 Apprentissage par renforcement

1.6.4.1 Principe

L'apprentissage par renforcement se caractérise par les objectifs suivants : l'acquisition automatisée de compétences pour la prise de décisions en milieu complexe et incertain, on dit qu'elle apprend par l'expérience qui est une stratégie comportementale (appelée politique) en fonction des échecs ou succès constatés.

Ce qui distingue l'apprentissage par renforcement des autres techniques, c'est son l'apprentissage par essais et erreurs et la récompense différée. Par conséquent, une série de décisions a pour effet de « renforcer » le processus, car celui-ci convient le mieux pour résoudre le problème. [13]

1.6.4.2 Problèmes utiliser pour l'apprentissage par renforcement

Contrairement à l'apprentissage supervisé, l'apprentissage par renforcement opère dans un environnement incertain (complexe et partiellement observable) pour ensuite renforcer les capacités de l'Agent décisionnel. Le processus d'apprentissage est illustré par la figure 1.02. [13]

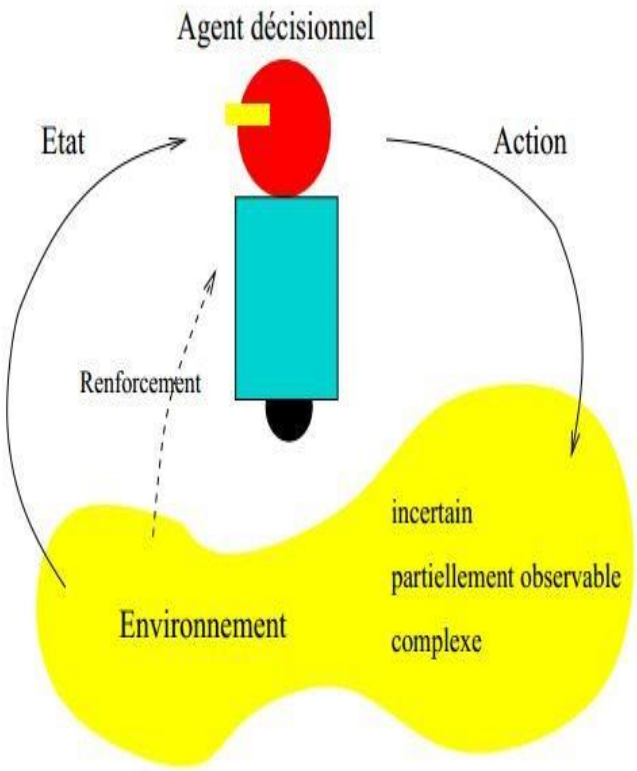


Figure 1.2 : *Processus décisionnel de l'apprentissage par renforcement*

1.6.4.3 Domaines d'applications

- Jeu vidéo d'ordinateur :

  L'industrie du jeu vidéo est très développée depuis ces quelques années. Les intelligences artificielles contrôlent les agents des jeux pour que les joueurs puissent avoir une expérience interactive, ces agents peuvent prendre des variétés de rôles.

- Application pour les machines (robotique) :

  Il y a certaines applications qui ne peuvent pas être programmées à la main. Par exemple, les hélicoptères qui peuvent apprendre par eux même comment voler stablement ou les voitures autonomes. [13]

1.6.5 Apprentissage profond

1.6.5.1 Principe

L'apprentissage profond est une méthode d'apprentissage automatique qui intègre des réseaux de neurones à des couches successives afin d'apprendre de manière itérative à partir de données. Les méthodes d'apprentissage profond sont conçues pour imiter le fonctionnement du cerveau humain (réseaux de neurones complexes) afin que les ordinateurs puissent être entraînés à traiter des abstractions et des problèmes mal définis.

Le terme « apprentissage profond » est utilisé lorsqu'il y a de multiples couches cachées dans un réseau de neurones. Un réseau de neurones s'ajuste continuellement de manière itérative et fait des déductions jusqu'à ce qu'un stade spécifique soit atteint. La machine apprend à partir de données non étiquetées et non structurées. [14]

1.6.5.2 Problèmes utiliser pour l'apprentissage profond.

L'apprentissage profond est particulièrement utile pour détecter des schémas dans des données non structurées. Bien que l'apprentissage profond soit très similaire à un réseau de neurones traditionnel, les couches cachées sont beaucoup plus nombreuses. Plus le problème est complexe, plus le modèle contient de couches cachées. [14]

1.6.5.3 Domaine d'application

Les réseaux de neurones et l'apprentissage profond servent souvent aux applications de reconnaissance d'images et de la parole ainsi que de vision par ordinateur et à prévoir le dysfonctionnement d'une machine dans l'Internet des objets et les applications de fabrication. [4]

Dans cet ouvrage l'apprentissage profond nous sera utile dans la détection de repère facial, la reconnaissance faciale, la reconnaissance d'expression faciale, la reconnaissance du genre, l'estimation d'âge. [14]

1.6.6 Apprentissage par transfert (transfert Learning)

1.6.6.1 Principe

Le Transfer Learning désigne l'ensemble des méthodes qui permettent de transférer les connaissances acquises à partir de la résolution de problèmes donnés pour traiter un autre problème dans l'apprentissage profond.

1.6.6.2 Problème utiliser pour l'apprentissage par transfert

Le Transfer Learning désigne l'ensemble des méthodes qui permettent de transférer les connaissances acquises à partir de la résolution de problèmes donnés pour traiter un autre problème. C'est donc une méthode hybride permettant de résoudre plusieurs types de problèmes pour avoir un système performant et efficace. En utilisant des modèles pré-entraînés comme point de départ, le Transfer Learning permet de développer rapidement des modèles performants et résoudre efficacement des problèmes complexes. [14] [15] [16]

1.6.6.3 Domaine d'application

Le transfert learning est utiliser dans de nombreux domaines complexes tels qu'en vision par ordinateur ou Natural Language Processing (NLP)

Son utilisation consiste principalement à exploiter des réseaux de neurones pré-entraînés, comme dans la page suivante.

- **Extracteurs de features ou de caractéristiques :**

L'idée est de réutiliser un réseau préentraîné sans sa couche finale (MLP). Ce nouveau réseau fonctionne alors comme un extracteur de features fixes pour la réalisation d'autres tâches.

- **Ajustement de modèles pré-entrainés :**

Il s'agit d'une technique plus complexe, dans laquelle non seulement la dernière couche est remplacée pour réaliser la classification ou la régression, mais d'autres couches sont également réentraînées de manière sélective. L'idée est donc de fixer les poids de certaines couches pendant l'entraînement et affiner le reste pour répondre à la problématique et d'obtenir de meilleures performances avec un temps d'entraînement plus court. [16]

1.7 Conclusion

On a présenté brièvement dans ce chapitre la définition d'une simple vidéosurveillance et d'une vidéosurveillance intelligente ainsi que leurs objectifs, les problèmes liés à la vidéo surveillance, les types de vidéosurveillance, les domaines de son application, ses avantages globales et ses avantages par rapport aux fonctionnalités de notre projet, les limites et les contraintes liés à son utilisation. On a aussi vu la généralité de l'apprentissage automatique concernant son concept et les différents types d'apprentissages, notamment : l'apprentissage supervisé, l'apprentissage semi-supervisé l'apprentissage non supervisé, l'apprentissage par renforcement, l'apprentissage profond et l'apprentissage par transfert. Nous avons développé ces principes et les applications de ces différents types d'apprentissages dans quelques domaines d'activité et les problèmes spécifiques à leurs utilisations. On a aussi passé en revue le poids de l'intelligence artificielle utilisée dans le monde d'internet et les avantages qu'il présente dans la vidéo surveillance.

Nous allons maintenant dans le deuxième chapitre, parler de la notion sur la vidéo et le traitement d'image qui permettra la réalisation finale de notre projet.

CHAPITRE 2

GÉNÉRALITÉS SUR LA VIDÉO ET LE TRAITEMENT D'IMAGE

2.1 Introduction

Le support d'information qui régit le mieux la description de notre environnement est sans doute la capture vidéo et la capture d'image. Une vidéo est une succession image. En réalité, l'image n'est qu'une illusion ou tout au plus notre abstraction de ce qui est réel. À l'état brut, l'image n'a aucun intérêt, ce n'est qu'après certains traitements qu'on peut en extraire les informations utiles pour son exploitation. Ce chapitre nous parlera des notions sur la vidéo et le traitement d'image.

2.2 Vidéo

2.2.1 Définition

Un flux de vidéo est une suite d'images 2D dont la résolution est exprimée en nombre de pixels qui définit la dimension de ces images. La durée du temps entre deux images (Δt) est très petite, parce en général un flux de vidéo à la vitesse de 24 à 60 images par seconde (frame par seconde ou FPS).

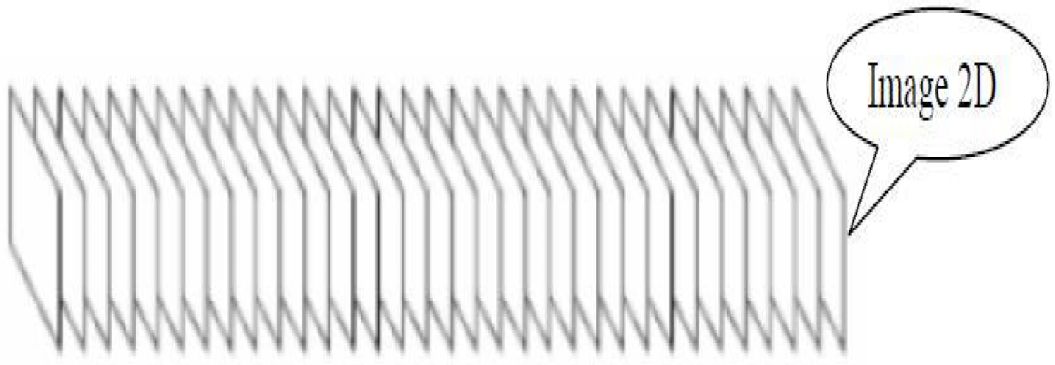


Figure 2.01 : *Représentation d'un flux de vidéo*

2.2.2 Frame

Les frames sont l'ensemble des images composant une séquence ou un flux de vidéo ou photogramme.

2.2.3 Objet vidéo

Les objets physiques sont les objets du monde réel qui apparaissent dans les scènes observées par les caméras. Les objets physiques sont divisés en deux types : les objets de contexte et les objets mobiles.

- **Les objets de contexte :**

Ce sont des objets physiques qui sont habituellement statiques (p. ex. les murs). Dans le cas où ils ne sont pas statiques, leurs mouvements peuvent être prédits par les informations contextuelles p.ex. les chaises, les portes sont des objets de contexte.

- **Les objets mobiles :**

Ce sont des objets physiques qui peuvent être perçus dans les scènes par leurs mouvements.

2.2.4 Types de vidéo

Le signal vidéo est le signal qui permet de transporter une séquence d'images de la source à un dispositif d'affichage sous forme électrique. Selon la façon dont les signaux sont traités, on peut distinguer les deux modes :

2.2.4.1 La vidéo analogique

Décris le signal analogique comme un signal électrique dont l'intensité varie dans le temps de façon continue. La qualité du signal final dans ce mode est plus faible, car le bruit rajouté au signal lors son traitement altère sa qualité.

2.2.4.2 La vidéo numérique

Décris un signal qui porte une information représentée par une suite de valeurs minimales ou maximales correspondant respectivement à 0 et à 1. L'un des facteurs qui avantagent le signal numérique par rapport au signal analogique est la facilité de distinguer l'information émise du bruit.

2.2.5 Les paramètres clés d'une vidéo

Le stockage et la diffusion d'une vidéo exigent un espace volumineux et un taux de transfert plus élevé. Le contrôle de qualité, et la taille d'une séquence vidéo sont déterminés par deux paramètres clés :

- **Le nombre d'images par seconde** :

  Le nombre d'images du système visuel humain exigé en général 25 ou 30 images par seconde.

- **La résolution** :

  Ce terme désigne que la quantité de l'information est limitée dans l'image. Autrement, c'est le nombre de pixels qui peuvent être affichés par un dispositif d'affichage.

Trouver le compromis entre ces paramètres et les limitations imposées par la technologie permet d'obtenir une qualité de vidéo optimale.

2.2.6 Exemple de format vidéo

Il existe plusieurs codages de format vidéo, par exemple :

- **Pour le format analogique, on peut citer** :

  Le NTSC (National Television System Committee), le PAL (*Phase Alternating Line*) et le SECAM (Séquentielle Couleur à Mémoire) désigne un standard international de codage couleur du signal vidéo analogique.

- **Pour le format numérique, on peut citer** :

  L'AVI (Audio Video Interleave), le WMV (Windows Media Video), le MKV(Matroska Video), le FLV (Flash Video), le MP4 ou MPEG-4 Part 14 (Moving Picture Expert Group 4 part 14).

2.3 Notion d'image et types d'images

2.3.1 Définition

Une image est une représentation d'une personne ou d'un objet par la peinture, la sculpture, le dessin, la photographie, le film, etc. C'est aussi un ensemble structuré d'informations qui, après affichage sur l'écran, ont une signification pour l'œil humain. [17]

2.3.2 Image numérique et numérisation

L'image numérique est l'image dont la surface est divisée en éléments de taille fixes appelés cellules ou pixels, ayant chacun comme caractéristique un niveau de gris ou de couleurs prélevées à l'emplacement correspondant dans l'image réelle, ou calculée à partir d'une description interne de la scène à représenter.

La numérisation d'une image est la conversion de celle-ci de son état analogique (distribution continue d'intensités lumineuses dans un plan x-o-y, en une image numérique représentée par une matrice bidimensionnelle de valeurs numériques f(x, y).

L'ensemble x, y sont les coordonnées cartésiennes d'un point de l'image et f(x, y) le niveau de gris (couleur) en ce point.

Pour des raisons de commodité de représentation pour l'affichage et l'adressage, les données images sont généralement rangées sous forme de tableau I de n lignes et p colonnes. Chaque élément I(x,y),représente un pixel de l'image et à sa valeur est associé un niveau de gris codé sur m bits (2^m niveaux de gris ; 0 = noir ; 2^m-1 = blanc). La valeur en chaque point exprime la mesure d'intensité lumineuse perçue par le capteur. [17]

2.3.3Pixel

C'est la contraction de l'expression anglaise « picture elements » ou élément d'image, le pixel est le plus petit point de l'image, c'est une entité calculable qui peut recevoir une structure et une quantification. Si le bit est la plus petite unité d'information que peut traiter un ordinateur, le pixel est le plus petit élément que peuvent manipuler les matériels et logiciels d'affichage ou d'impression. [17]

2.3.4 Image binaire

L'image binaire est la représentation de l'image numérique où chaque pixel ne peut prendre que 2 valeurs possibles (binaire). Quand la valeur d'un pixel franchi un seuil établi elle prend la valeur 1 sinon elle a pour valeur 0. [17]

Figure 2.02 : *Image binaire avec un seuil de 70*

2.3.5 Image à niveau de gris

Le niveau de gris est la valeur de l'intensité lumineuse en un point. La couleur du pixel peut prendre des valeurs allant du noir au blanc en passant par un nombre fini de niveaux intermédiaires. Donc pour représenter les images à niveaux de gris, on peut attribuer à chaque pixel de l'image une valeur correspondant à la quantité de lumière renvoyée. Cette valeur peut être comprise par exemple entre 0 et 255. Chaque pixel n'est donc plus représenté par un bit, mais par un octet.

Le nombre de niveaux de gris dépend du nombre de bits utilisés pour décrire la « couleur » de chaque pixel de l'image. Plus ce nombre est important, plus les niveaux accessibles sont nombreux. [17]

Figure 2.03 : *Image à niveau de gris*

2.5.6 Image True-color (RVB)

Les applications multimédias utilisent le plus souvent des images en couleurs. La représentation des couleurs s'effectue de la même manière que les images monochromes avec cependant quelques particularités. En effet, il faut tout d'abord choisir un modèle de représentation.

Elle consiste à utiliser 24 bits pour chaque point de l'image. 8 bits sont employés pour décrire la composante rouge (R), 8 pour le vert (V), et les 8 autres pour le bleu (B). Il est ainsi possible de représenter 256^3 = 16,7 millions de couleurs différentes simultanément. Cela est cependant théorique, car aucun écran n'est capable d'afficher 16 millions de points. Dans la plus haute résolution actuelle, l'écran affiche 1 920 000 de points. [17]

2.4 Propriété d'une image

2.4.1 Luminance

C'est le degré de luminosité des points de l'image. Le mot luminance est substitué au mot brillance, qui correspond à l'état d'un objet. Une bonne luminance se caractérise par : des images lumineuses(brillantes), l'absence de parasites.

Elle est définie aussi comme étant le quotient de l'intensité lumineuse d'une surface par l'aire apparente de cette surface [17].

$$\text{Lum(I)} = \frac{1}{I_{L} \times I_{C}} \sum_{x=0}^{I_{L}-1} \sum_{y=0}^{I_{C}-1} I(x, y) \qquad (2.01)$$

2.4.2 Dimension

C'est la taille de l'image, ou encore le nombre total de pixels dans l'image. Cette dernière se présente sous forme de matrice dont les éléments sont des valeurs numériques représentatives des intensités lumineuses (pixels). Donc la dimension D d'une image est le nombre de lignes M multiplié par le nombre de colonnes N. de cette matrice. [17]

$$D = M^{*} N \qquad (2.02)$$

2.4.3 Résolution

C'est la clarté ou la finesse de détails atteinte par un moniteur ou une imprimante dans la production d'images. Sur les moniteurs d'ordinateurs, la résolution est exprimée en nombre de pixels par unité de mesure (pouce ou centimètre). On utilise aussi le mot résolution pour désigner le nombre total de pixels affichable horizontalement ou verticalement sur un moniteur ; plus grand est ce nombre, meilleure est la résolution. [17]

2.4.4 Chrominance

La chrominance est l'information qui porte la couleur dans le signal image. Généralement ce signal couleur est obtenu à partir de la synthèse des trois couleurs. [17]

2.4.5 Contraste

C'est l'opposition marquée entre deux régions d'une image, plus précisément entre les régions sombres et les régions claires de cette image. Le contraste est défini en fonction des luminances de

deux zones d'images. Si L1 et L2 sont les degrés de luminosité respectivement de deux zones voisines A1 et A2 d'une image, le contraste C est défini par le rapport : [17]

$$C=\frac{L_1-L_2}{L_1+L_2} \tag{2.03}$$

2.4.6 Contour

Les contours représentent la frontière entre les objets de l'image, ou la limite entre deux pixels dont les niveaux de gris représentent une différence significative. [17]

2.4.7 Texture

Les textures décrivent la structure de ceux-ci. L'extraction de contours consiste à identifier dans l'image les points qui séparent deux textures différentes. [17]

2.4.8 Bruit

Le bruit (parasite) dans une image est considéré comme un phénomène de brusque variation de l'intensité d'un pixel par rapport à ses voisins, il provient de l'éclairage des dispositifs optiques et électroniques du capteur. [17]

2.3.9 Histogramme d'une image

L'histogramme des niveaux de gris ou des couleurs d'une image est une fonction qui donne la fréquence d'apparition de chaque niveau de gris (couleur) dans l'image. Pour diminuer l'erreur de quantification, pour comparer deux images obtenues sous des éclairages différents, ou encore pour mesurer certaines propriétés sur une image, on modifie souvent l'histogramme correspondant.

Il permet de donner un grand nombre d'informations sur la distribution des niveaux de gris (couleur) et de voir entre quelles bornes est repartie la majorité des niveaux de gris (couleur) dans les cas d'une image trop claire ou d'une image trop foncée. Il peut être utilisé pour améliorer la qualité d'une image (Rehaussement d'image) en introduisant quelques modifications, pour pouvoir extraire les informations utiles de celle-ci. [17]

Voici quelques exemples d'histogramme :

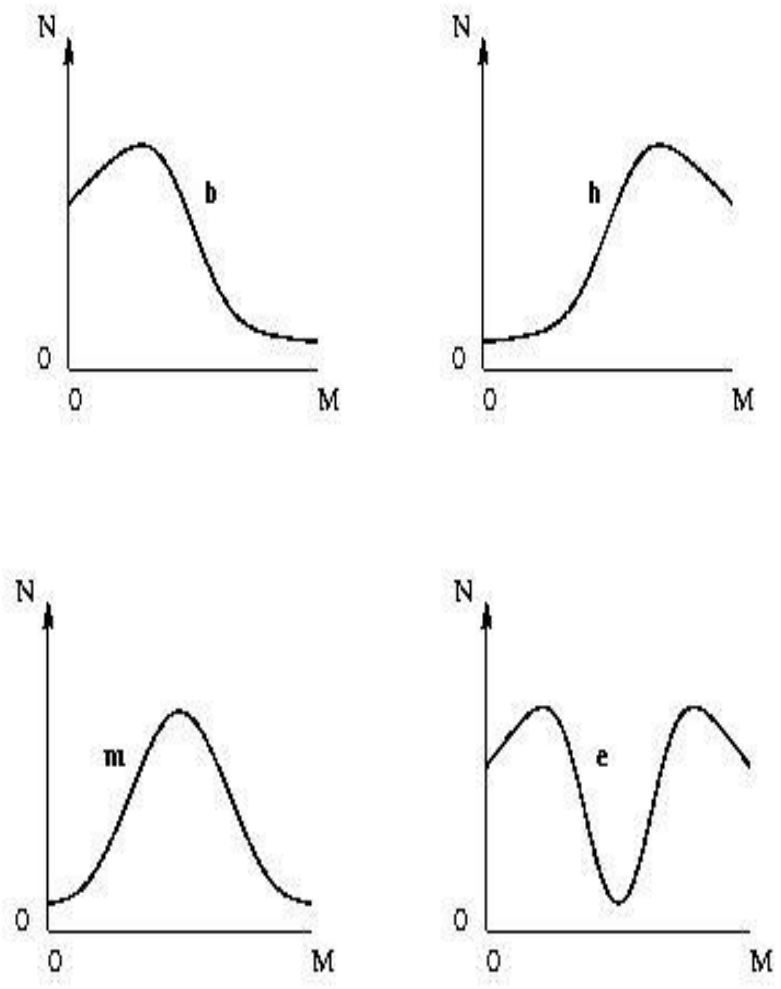


Figure 2.04 : *Quelques exemples d'histogramme*

Avec N. est le nombre de pixel et M est le niveau de gris

- Si l'histogramme b est concentré sur les bas niveaux de gris, c'est à dire l'image est trop sombre.
- Si l'histogramme h est concentré sur les hauts niveaux de gris, c'est à dire l'image est trop claire.
- Si l'histogramme m est concentré sur les niveaux de gris moyens, c'est à dire l'image est floue et manque de contraste.
- Si l'histogramme e est concentré sur les niveaux de gris extrêmes c'est à dire l'image manque de nuances intermédiaires entre le sombre et le clair. [17]

2.5 Opération dans le traitement d'image

2.5.1 Transformation spatiale

2.5.1.1 Redimensionnement

Le redimensionnement de l'image consiste à lui attribuer une nouvelle dimension à partir d'un coefficient d'amplification. Si on veut que l'image grossie, on utilise un coefficient supérieur à 1. Si on veut qu'il diminue, on choisit un coefficient entre 0 et 1. [17]

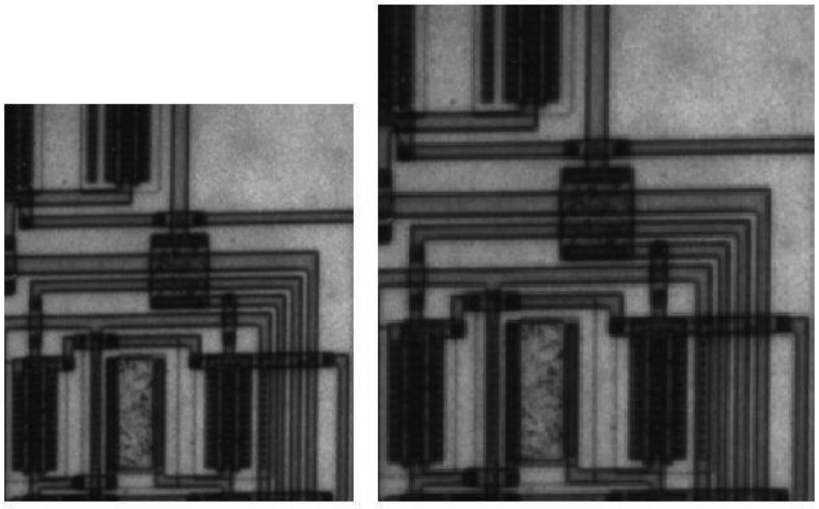


Figure 2.05 : *Agrandissement de l'image avec un coefficient de 1.25*

2.5.1.2 Rotation

La rotation d'une image se fait par la définition de l'angle en degré de sa rotation. [17]

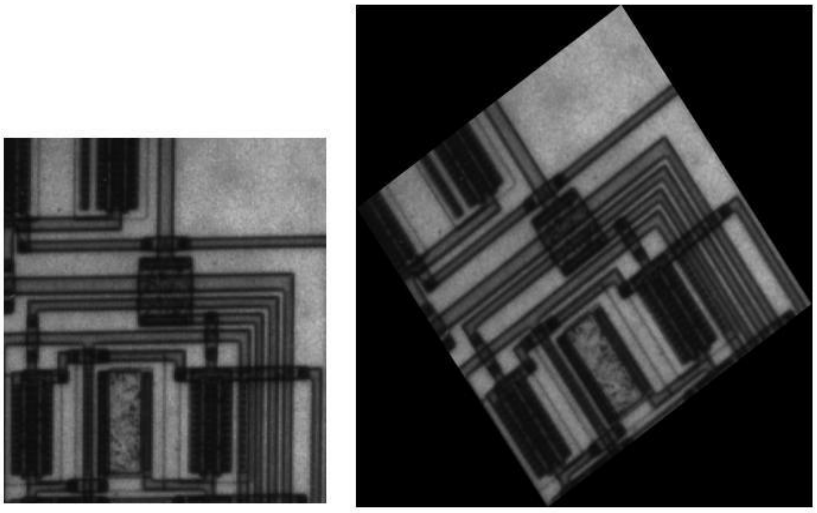


Figure 2.06 : *rotation d'image à 35°*

2.5.1.3 Rognage

Le rognage d'une image est l'extraction d'une portion rectangulaire. En définissant, les coordonnées du point de départ et la taille du rognage (rectangle). [17]

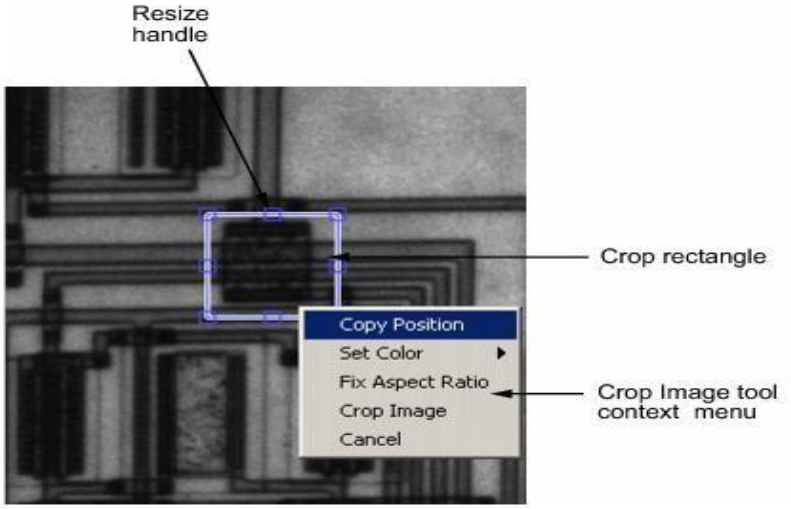


Figure 2.07 : *Rognage d'une image dans Matlab*

2.5.2 Filtrages et filtrage linéaire

2.5.2.1 Définition

L'amélioration d'une image peut être obtenue par ce que l'on appelle une opération de filtrage. Le filtrage est utilisé pour modifier ou pour mettre en valeur une image, on pourra par exemple filtrer une image pour accentuer certains attributs ou pour en supprimer d'autres.

L'objectif avoué du filtrage est d'éliminer les perturbations induites par les procédés d'acquisitions d'image et aux problèmes de transmission ainsi de réduire les variations d'intensité au sein de chaque région de l'image tout en respectant l'intégrité de la chaine originale comme les transitions entre les régions homogènes. Les éléments significatifs de l'image doivent être préservés au mieux.

Le filtrage linéaire d'une image peut s'envisager de 2 manières : le filtrage peut tout d'abord se réaliser dans le domaine spatial en effectuant un produit de convolution ou dans le domaine fréquentiel en multipliant la transformée de Fourier de l'image par la fonction de Transfert du filtre. L'image originale est alors obtenue par la transformée de Fourier inverse. [17]

2.5.2.2 Convolution

Le filtrage est un produit de convolution qui met en jeu l'environnement où le voisinage de chaque pixel. [17]

La convolution est définie par :

- A une dimension :

$$g(t) = f(t)*h(t) = \int_{-\infty}^{+\infty} f(t-\tau)\, h(\tau)\, d(\tau) = \int_{-\infty}^{+\infty} h(t-\tau)\, f(\tau)\, d\tau \qquad (2.04)$$

- A deux dimensions :

$$g(x,y) = h(x,y) * f(x,y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(\propto,\beta) f(x-\alpha, y-\beta) d_\alpha d_\beta \qquad (2.05)$$

- Pour le cas d'une image numérique, la convolution numérique est donnée par :

$$g(x,y) = \sum_u \sum_v h(u,v)\, f(x-u, y-v) = \sum_u \sum_v f(u,v)\, h(x-u, y-v) \qquad (2.06)$$

L'image ci-dessous est une illustration plus claire d'une image après filtrage par convolution :

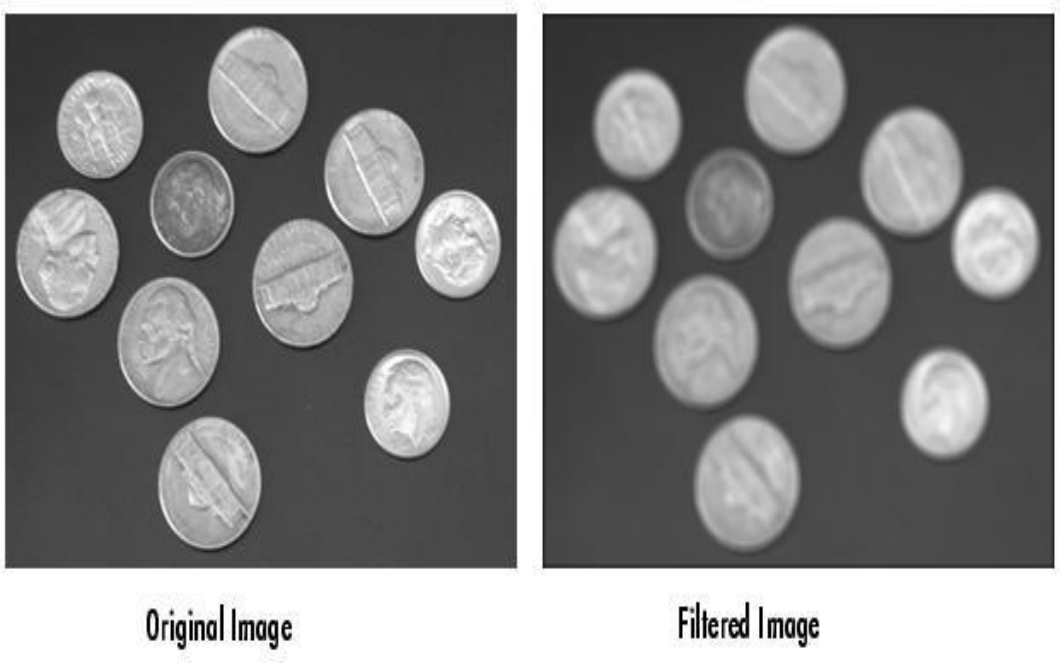


Figure 2.08 : *Image après filtrage par convolution*

2.5.2.3 Filtre Dérivée première

En développant au premier ordre, avec une approximation, la dérivée en un point x,f(x) peut s'exprimer par : [17]

$$f'(x) = \frac{1}{2}[f(x+1) - f(x-1)] \quad (2.07)$$

Appliquée la dérivée partielle à une image numérique, on peut définir :

- Dérivée horizontale, suivant les colonnes de l'image qui est alors définie par la matrice :

$$D_{1H} = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \quad (2.08)$$

- La dérivée verticale, dérivée partielle suivant les colonnes de l'image est alors définie par la matrice :

$$D_{2v} = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad (2.09)$$

2.5.2.4 Filtre de Sobel

Les opérateurs de Sobel sont aussi des opérateurs de dérivation qui accentuent les bruits dans une image c'est-à-dire, les pixels de valeurs parasites et de répartition aléatoire.

Les opérateurs de dérivation effectuent de contour dans une image. [17]

- Dérivation horizontale :

$$S_H = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad (2.10)$$

- Dérivation verticale :

$$S_v = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \quad (2.11)$$

- Dérivation oblique :

$$S_o = \begin{bmatrix} 0 & 1 & 2 \\ -1 & 0 & 1 \\ -2 & -1 & 0 \end{bmatrix} \quad (2.12)$$

2.5.2.5 Filtre Laplacien

L'opérateur Laplace donne une approximation directe de la somme des dérivées secondes, ce qui peut être obtenu avec une matrice. [17]

$$L_P = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad (2.13)$$

2.5.2.6 Filtre passe-bas ou lissage

Le lissage est une opération destinée à éliminer les bruits dans une image, les lissages sont des filtres passe-bas ce qui signifie qu'ils éliminent les signaux de haute fréquence caractérisés par des grandes variations des niveaux de gris entre pixels voisins. [17]

- Lissage fort : le lissage fort est caractérisé par la matrice L_F donnée par :

$$L_F = 1/9 \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad (2.14)$$

- Lissage moyen : ce lissage est caractérisé par la matrice L_M donné par :

$$L_M = 1/16 \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \quad (2.15)$$

2.5.3 Les transformées

2.5.3.1 Transformée de Fourier

La transformée de Fourier permet la décomposition d'un signal (image) f en combinaison linéaire de sinusoïde complexe, dont les coefficients F[p,q] dit coefficients de Fourier, fournissent des informations sur les fréquences (p,q)et permettent des manipulations dans le domaine fréquentiel. [17]

Elle est définie par la formule suivante :

$$F(p,q) = \sum_{m=0}^{M-1} \sum_{n=0}^{M-1} f(m,n) e^{-j2\pi pm/\, M_e - j2\pi qn\, /\, N} \quad (2.16)$$

Et son inverse est :

$$f(m,n) = \frac{1}{MN} \sum_{P=0}^{M-1} \sum_{Q=0}^{N-1} F(p,q) e^{j2\pi pm/Me^{j2\pi} qn} \quad (2.17)$$

2.5.3.2 Transformer en Ondelette

Une odelette est une possibilité de représentation d'un signal. On le représente comme une somme pondérée de ces petites ondes translatées ou dilatées. L'odelette peut servir pour le dé bruitage des images (voir la figure dans la page suivante). [17]

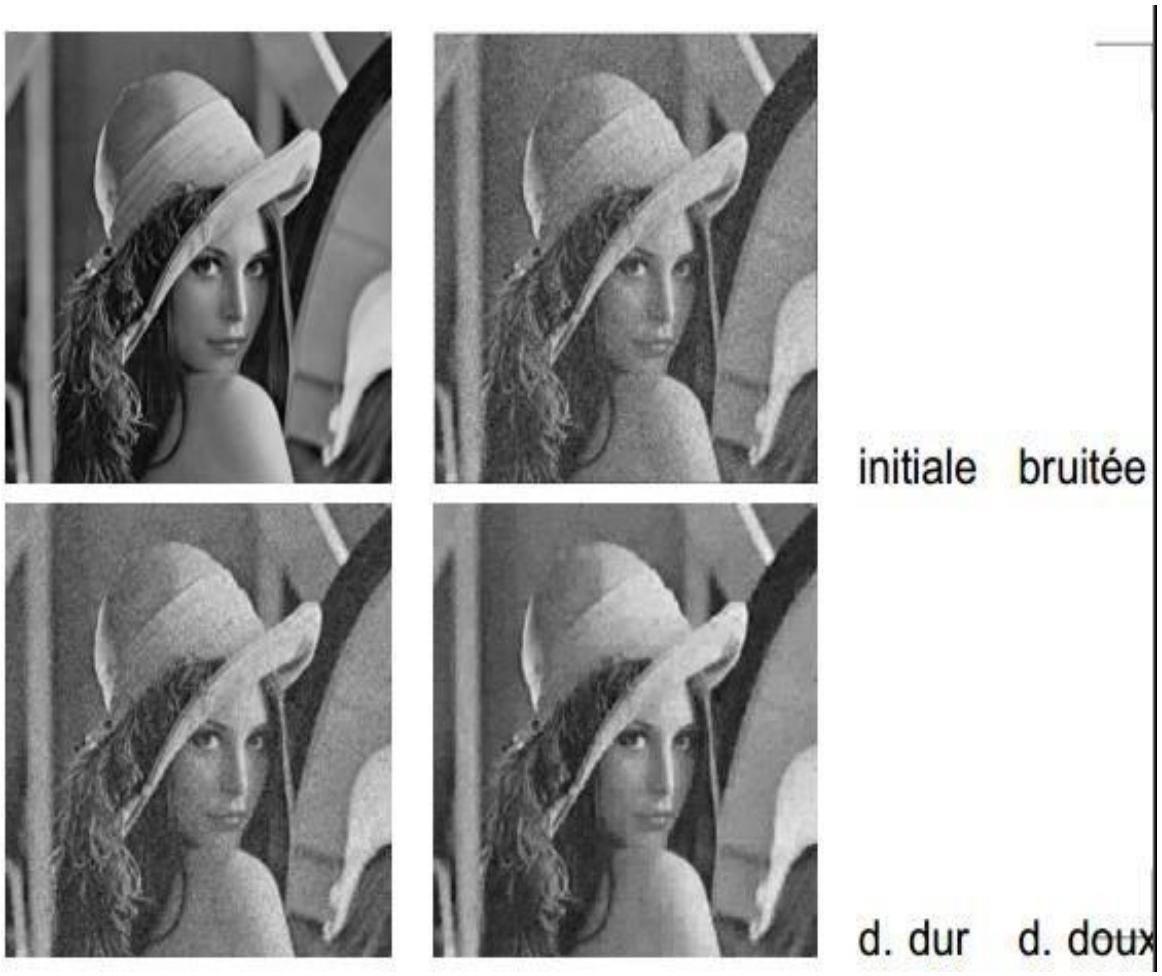


Figure 2.09 : *Débruitage par ondelette*

2.5.3.3 Transformée en cosinus discret

La transformée en cosinus discrète est surtout utilisée dans la compression d'image puis sa reconstitution. Elle se fait en 4 étapes : [17]

- Extraction du DCT en 2 dimensions de l'image
- Quantification de la DCT
- Compression de l'image
- Reconstitution de l'image

Figure 2.10: *Reconstitution à 75% d'une image compresser par la transformer en cosinus discret*

2.6 Conclusion

Dans ce chapitre nous avons parlé de la généralité sur la vidéo. Ainsi, nous avons vu que la vidéo est un ensemble de frames qui est constitué de plusieurs images composantes. On s'est concentré sur diverses propriétés de l'image et sur quelques opérations pour son traitement. On a défini l'image numérique, puis on a donné les différents types d'images numériques. Ensuite on a présenté les différentes techniques de filtrages et de transformations appliquées à l'image. La connaissance de ces différentes opérations de base dans le domaine de l'image sera nécessaire pour la compréhension de notre travail final.

CHAPITRE 3

L'APPRENTISSAGE PROFOND

3.1 Introduction

Ces dernières années sont témoins de nombreuses avancées en matière d'exploitation des données. Parmi ces données, il y a l'image qui nous intéresse plus particulièrement. En effet, les domaines d'application de l'apprentissage machine et du traitement de l'image dans le domaine de l'imagerie sont déjà très vaste : dont les reconnaissances faciales, les classifications d'image par contenues, et autres encore. Dans ce chapitre, nous allons nous concentrer sur les modèles de l'apprentissage profond (deep-learning) qui est la branche de l'apprentissage automatique utilisée pour la classification d'image : c'est-à-dire le CNN et ses variants. Puis nous allons voir la régression linéaire, la classification de donnée par la méthode des K-plus proches voisins et le SVM ou séparateur à vaste marge.

3.2 Réseau de neurones

3.2.1 Classification topologique de réseaux de neurones

Il existe deux classes :

- **Les réseaux de feed-forward (non bouclés)** qui ce sont des réseaux dans lesquels l'information se propage de couche en couche sans retour en arrière possible, ce sont : le perceptron monocouche, le perceptron multicouche et les réseaux à fonction radiale
- **Les réseaux de feed-back (récurrents)** qui peuvent contenir des chemins bouclés, passant plusieurs fois par un même neurone, ce sont : les cartes auto-organisatrices de Kohonen, les réseaux de Hopfield, les ART (Adaptive Resonance Theory), les réseaux à compétition.

Dans ce chapitre nous allons parler surtout des réseaux de feed-forward qui nous intéresse dans la compréhension de notre travail. [17]

3.2.1.1 Perceptron-Neurone formel

Les neurones artificiels sont inspirés du fonctionnement des neurones biologiques. Elle est composée des entrées, du neurone et de la sortie. On associe un poids (poids synaptique) à chacune des entrées x. L'ensemble des entrées-poids est ensuite appliqué à un sommateur suivi d'un comparateur [17].

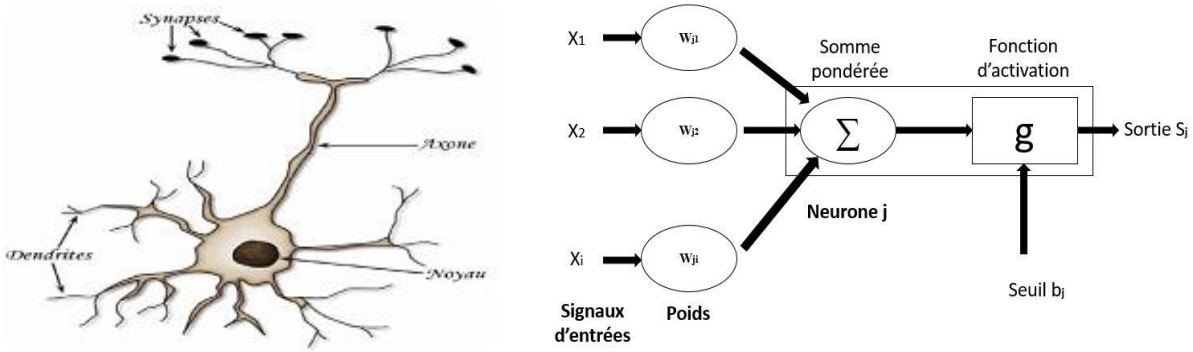


Figure 3.01 : *Neurone biologique* et *neurone formel*

L'apprentissage du perceptron repose sur la mise à jour automatique des poids du réseau entrainant le développement de l'algorithme selon une sortie désirée d(i) durant un processus d'apprentissage en 6 étapes : [17]

- *1ère Étape* : Choisir des valeurs aléatoires pour les poids W_i et le biais b ou seuil à des valeurs choisis au hasard.
- *2ème Étape* : Appliquer le vecteur d'entrée $X_1 = (x_1, x_2, \dots, x_i)$ de la base d'apprentissage
- *3ème Étape* : Calculer la valeur de la sortie x pour cette entrée E (la valeur de de seuil est introduite ici dans le calcul de la somme pondérée) :

$$\boldsymbol{a = \sum(w_i . x_1) - b_i} \quad (3.01)$$

$S = signe(a)$ (si $a > 0$ alors x = +1 sinon $a \leq 0$ alors x = −1)

- *4ème Étape* : Si S = d(i) on retourne à la deuxième étape sinon on passe à la prochaine étape
- *5ème Étape* : Si la sortie du Perceptron est différente de la sortie désirée d_1, c'est l'étape où le perceptron met à jour les anciens poids, elle est appelée règle du perceptron ou encore « règle delta » (avec μ le pas de modification) :

$$\boldsymbol{dw_0 = d(i)\ et\ w_i = x(i)d(i)\ \ et\ \ w_0 = w_0 + dw_0\ ,\ \ w_i = w_i + dw_i\ pour\ i = 1 \dots n} \quad (3.02)$$

Ou :

$$\boldsymbol{w_{ij}(t+1) = w_{ij}(t) + \mu.((d_1 - x).e_i)}$$

Avec $d_1 = +1$ si E est de la classe 1, $d_1 = -1$ si E est de la classe 2 et $(d_1 - x)$ est une estimation de l'erreur.

- Retourner à l'étape 2
- Enfin la sortie du neurone est présentée à une fonction d'activation (sigmoïde, linéaire…).

3.2.1.2 Multi-Perceptron Layer

Le MPL (Multi-Perceptron Layer) est un ensemble de perceptron organisée en plusieurs couches cachées pour résoudre les problèmes non linéairement séparables telle que la porte logique XOR.

Les entrées des couches cachées sont la sortie des couches précédentes. [17]

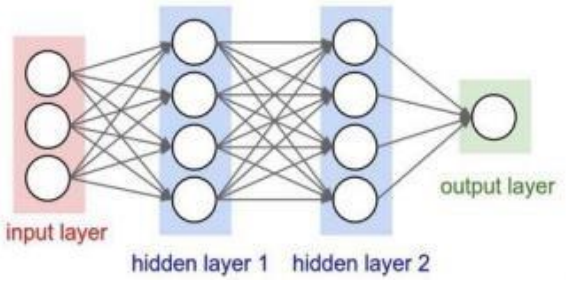
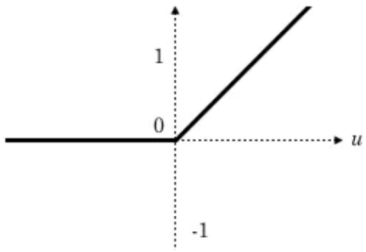


Figure 3.02 : *Perceptron Multicouche*

3.2.2 Fonction d'activation

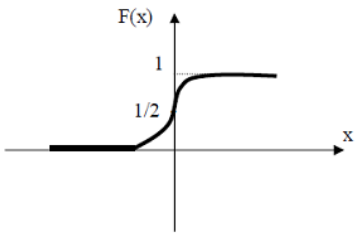
La fonction d'activation est une fonction de transfert analysant les entrées d'un neurone pour donner un résultat ou une réponse. Dans ce cas, il existe plusieurs types de fonctions comme étant une fonction de transfert, mais les plus fréquentées sont la fonction a seuil, la fonction sigmoïde, la fonction linéaire ou linéaire par morceaux, la fonction gaussienne. Pour cela, chaque fonction de transfert à sa spécificité qui correspond aux types de neurones à modéliser et à exploiter.

Nous allons seulement voir ici la fonction d'activation qui nous intéresse, la fonction d'activation ReLu (Unité de Réctificaton Linéaire) et la fonction d'activation sigmoïde. [17] [18]



$$f(x) = \begin{cases} 0 & \text{si} \quad x < 0 \\ x & \text{si} \quad x \geq 0 \end{cases} \tag{3.03}$$

Figure 3.03 : *Fonction de transfert ReLu*



$$F(x) = \frac{1}{1+e^{-x}} \tag{3.04}$$

Figure 3.04 : *Fonction d'activation sigmoïde*

(Le résultat obtenu par la fonction sigmoïde peut être interprété comme la **probabilité que l'observation X soit un label ou étiquette)**

3.2.3 Rétropropagation

En anglais, back propagation est une généralisation de la règle delta : c'est une méthode pour entraîner un réseau de neurones, consistant à mettre à jour les poids de chaque neurone de la dernière couche vers la première. Elle vise à corriger les erreurs selon l'importance de la contribution de chaque élément à celles-ci. Dans le cas des réseaux de neurones, les poids synaptiques qui contribuent plus à une erreur seront modifiés de manière plus importante que les poids qui provoquent une erreur marginale.

La méthode de calcul d'erreur pour toutes les couches du réseau de neurones. Posons j l'indice des neurones de la couche cachée, i et k, pour les neurones de la couche d'entrée et de la couche de sortie respectivement. On pose aussi y_j la couche de sortie du j ème neurone de la couche cachée et y_k la couche de sortie k ème neurone de la couche de sortie p étant l'indice du prototype.

Entre la couche cachée et la couche d'entrée, pour mettre à jour les poids, on utilise : [17] [18]

$$\Delta_p \boldsymbol{w_{ji}} = \boldsymbol{n}\ \boldsymbol{\delta_j^p}\ \boldsymbol{\hat{y}_i^p} \tag{3.05}$$

3.2.3.1 Algorithme de la descente de gradient

La descente de gradient est un algorithme d'optimisation qui est utilisé de manière itérative pour trouver le minimum local d'une fonction. Pour entraîner notre modèle, c'est-à-dire trouver les poids adaptés, il est nécessaire de **définir une fonction qui va quantifier l'erreur de notre modèle ou optimisation de la fonction de perte ou fonction de coût**. En général, pour les problèmes de classification, nous utilisons la fonction de perte « cross-entropy » pour mesurer l'erreur (MSE Minimum Square Error ou erreur quadratique moyenne) :

$$\boldsymbol{E} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{1}}^{\boldsymbol{n}} (\boldsymbol{y_i} - \bar{\boldsymbol{y}}_{\boldsymbol{i}})^{\boldsymbol{2}} \tag{3.06}$$

n est le nombre de points de données. **y** est la valeur désirée et $\bar{\mathbf{y}}$ tiret est celui calculé à la sortie.

Notre problème d'optimisation consiste à trouver les meilleures valeurs des poids de notre modèle qui minimise notre fonction d'erreur E(w) : **Plus la valeur de l'erreur calculée à partir de la fonction de coût est faible, meilleure est la valeur des poids donc, meilleurs sont les résultats de l'algorithme**. La descente de gradient fonctionne en calculant la dérivée (concept de calcul) de la fonction d'erreur ci-dessus. [18]

3.2.3.2 Principe de la descente de gradient

- La formule générale est la suivante :

$$\boldsymbol{X_{t+1} = X_t - \eta \Delta X_t} \quad (3.07)$$

Où η le taux d'apprentissage et Δx_t la direction de descente.

- Le but à chaque itération est d'avoir : $\boldsymbol{f(x_{t+1}) \leq f(x_t)}$, avec *f* une **fonction convexe** que l'on souhaite minimiser.
- L'algorithme de descente du gradient décide de suivre comme direction de descente l'opposé du gradient d'une **fonction convexe** f, ie – Δ*f* : parce que Le gradient d'une fonction indique sa croissance maximale à partir d'un point. Alors choisir l'opposé revient à prendre la pente la plus abrupte, dans l'objectif de minimiser la valeur de cette fonction. Voici son fonctionnement :

  - *1ère Étape* : Soit un point d'initialisation x_0 appartenant au domaine de f
  - *2ème Étape* : Calculer $f(x_t)$
  - *3ème Étape* : Mettre à jour les coordonnées :

$$\boldsymbol{X_{t+1} = X_t - \eta \Delta f(X_t)} \quad (3.08)$$

  - *4ème Étape* : Répéter 2 et 3 jusqu'au critère d'arrêt :

$$\boldsymbol{|fX_{t+1} - fX_t| \leq \varepsilon} \text{ Où } \varepsilon \text{ est très petit signe qu'on a convergé.} \quad (3.09)$$

Un mauvais point d'initialisation x_0 **ou un taux d'apprentissage** η **peu adapté peut empêcher l'algorithme de converger vers le minimum** :

- Si η est trop élevé, on peut osciller très longtemps autour du minimum, voir le manquer. Il peut faire diverger et aller à l'encontre du but voulu, i.e. atteindre le minimum.
- Tandis qu'un taux d'apprentissage trop petit peut mettre beaucoup de temps avant de converger.
- Le problème est le même pour le point d'initialisation. Deux points d'initialisation peuvent mener à deux résultats différents en fonction de la complexité de la fonction de perte.

Alors l'ajustement du taux d'apprentissage et le choix du point d'initialisation sont beaucoup plus compliqués dans la réalité. Dans la pratique, il est nécessaire d'appliquer à notre modèle divers taux d'apprentissage pour observer celui qui apporte les meilleures performances, de même pour le point d'initialisation. [19]

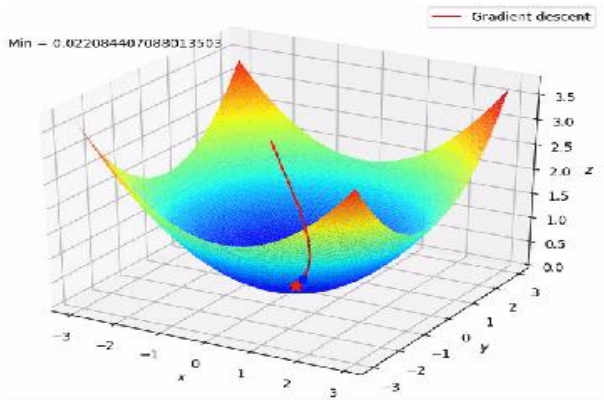


Figure 3.05 : *Représentation de l'algorithme de la descente de gradient*

3.3 Réseau de neurones à convolution CNN

Les CNN désignent une sous-catégorie de réseaux de neurones et sont à ce jour un des modèles de classification d'images réputés être les plus performant. Leur mode de fonctionnement est à première vue simple, l'utilisateur fournit en entrée une image sous la forme d'une matrice de pixels qui dispose de 3 dimensions :

- Deux dimensions pour une image en niveaux de gris.
- Une troisième dimension, de profondeur 3 pour représenter les couleurs fondamentales (Rouge, Vert, Bleu).

Contrairement à un modèle MLP (Multi Layers Perceptron) classique qui ne contient qu'une partie classification, l'architecture du Convolutional Neural Network dispose en amont d'une partie convolutive et comporte par conséquent deux parties bien distinctes :

- **Une partie convolutive :** Son objectif final est d'extraire des caractéristiques propres à chaque image en les compressant de façon à réduire leur taille initiale. En résumé, l'image fournie en entrée passe à travers une succession de filtres, créant par la même occasion de nouvelles images appelées cartes de convolutions. Enfin, les cartes de convolutions obtenues sont concaténées dans un vecteur de caractéristiques appelé code CNN.
- **Une partie classification :** Le code CNN obtenu en sortie de la partie convolutive est fourni en entrée dans une deuxième partie, constituée de couches entièrement connectées appelées perceptron multicouche (MLP pour Multi Layers Perceptron). Le rôle de cette partie est de combiner les caractéristiques du code CNN afin de classer l'image. [20]

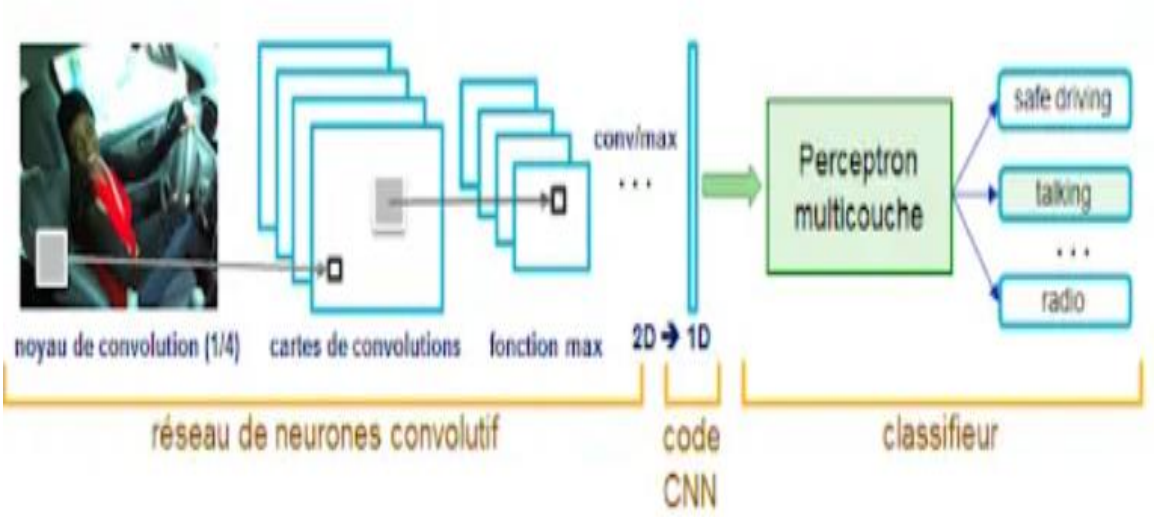


Figure 3.06 : *Représentation globale de l'architecture d'un CNN*

3.3.1 Fonctionnement général et but de la convolution

La convolution est une opération mathématique simple généralement utilisée pour le traitement et la reconnaissance d'images. Sur une image, son effet s'assimile à un filtrage dont voici le fonctionnement :

- Dans un premier temps, on définit la taille de la fenêtre de filtre située en haut à gauche.
- La fenêtre de filtre, représentant la feature, se déplace progressivement de la gauche vers la droite d'un certain nombre de cases défini au préalable (le pas) jusqu'à arriver au bout de l'image.
- À chaque portion d'image rencontrée, un calcul de convolution s'effectue permettant d'obtenir en sortie une carte d'activation ou feature map qui indique où est localisées les features dans l'image : plus la feature map est élevée, plus la portion de l'image balayée ressemble à la feature.

La convolution standard permet à chaque nombre N de filtre d'avoir une seule image en sortie avec une taille plus petite que l'original ayant une profondeur de N. (voir dans la page suivante pour plus de compréhension). [20]

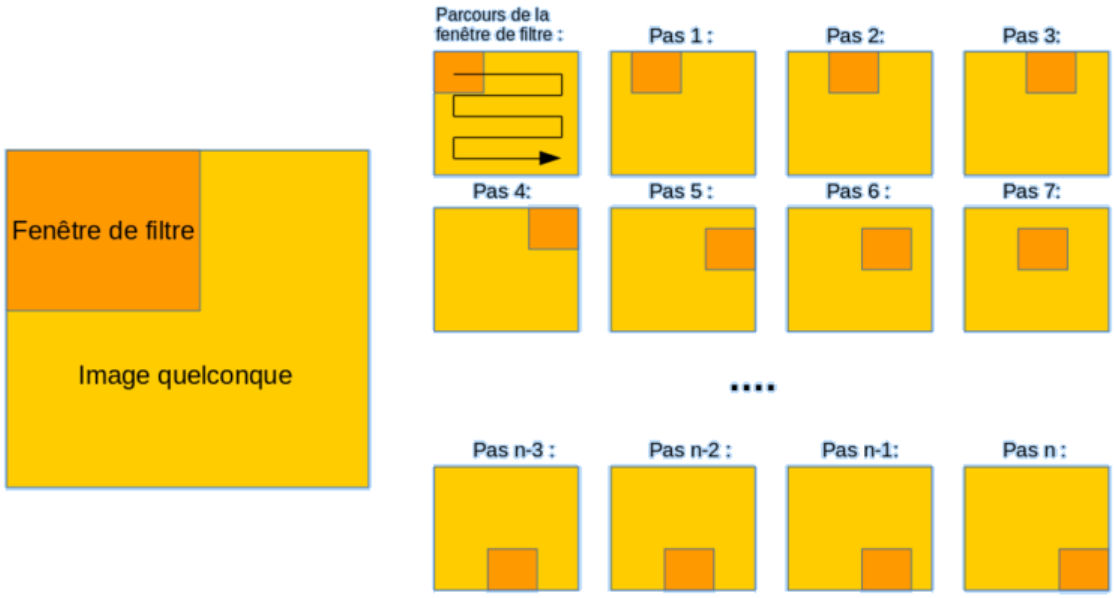


Figure 3.07 : *Schéma des parcours de la fenêtre de convolution*

Le but est de se servir des valeurs présentes dans le filtre à chaque pas. Par exemple si l'on définit une fenêtre 3 par 3, cela représentera 9 cases du tableau (c'est-à-dire 9 pixels). La convolution va effectuer une opération avec ces 9 pixels. Il peut s'agir de n'importe quelle opération (filtre d'image), par exemple on extrait la valeur la plus grande ou ArgMax ou Argument Maximal (soit le pixel avec la plus grande valeur) que nous pouvons voir dans la figure suivante : [20]

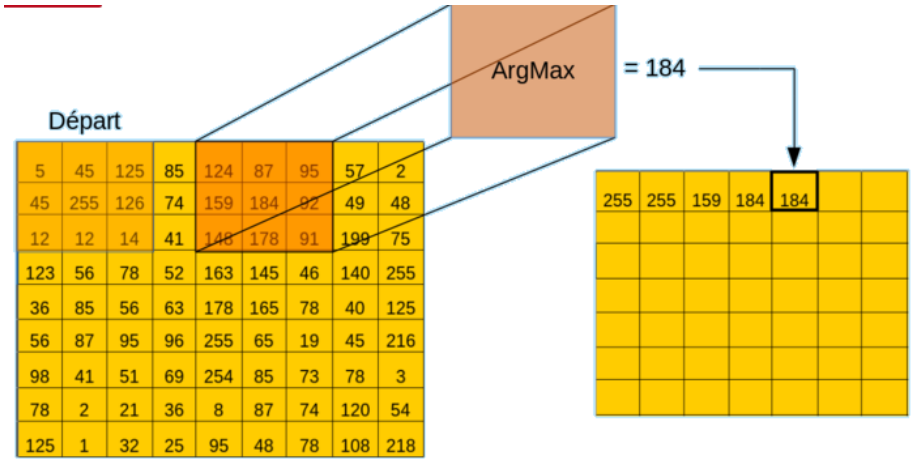


Figure 3.08 : *Représentation de la convolution ArgMAx avec pas horizontale = 1 pixel, pas vertical = 1 pixel.*

On fait glisser la fenêtre en orange et à chaque pas on récupère la valeur la plus grande parmi les 9 valeurs de pixels**.** On remarque que la sortie de la convolution, que l'on peut appeler « carte de caractéristiques », à des dimensions plus petites que celle de l'image en entrée. [20]

3.3.2 Avantage de la convolution

La convolution a pour avantage de réduit considérablement le nombre de calculs et d'extraire des caractéristiques propres à chaque image.

Prenons un exemple simple. Imaginons qu'en entrée de notre réseau de neurones, nous avons une image quelconque de 512 pixel**s** de côté. Cette image comporte donc 262.144 pixels (512*512). Imaginons maintenant que nous n'avions jamais entendu parler des **CNN** et que nous utilisons donc un réseau de neurones profond classique (MLP). Ce réseau, par exemple, présente sur sa première couche cachée 512 neurones. Cela a pour conséquences que sur la première couche cachée uniquement nous avons plus de 13 millions de poids à calculer (262.144 * 512). Ce n'est pas envisageable.

Le CNN et en particulier sa partie convolutive permet de pallier ce problème. Cela va grandement diminuer le nombre de poids à calculer dans le modèle. En effet, comme nous l'avons vu ci-dessus, la convolution va avoir pour effet de réduire la dimension de **«** la carte de caractéristiques » que l'on obtient après convolution (en comparaison avec la taille de l'image en entrée). Si l'on répète ce processus plusieurs fois, en prenant comme nouvelle entrée (sur laquelle nous allons effectuer la convolution) la sortie de la convolution précédente, nous allons diminuer de plus en plus la taille de la carte de caractéristiques, et donc nous diminuons également le nombre de calculs. [20]

3.3.3 Architecture du CNN

Le bloc de base du CNN est constitué de :

- La couche de convolutions
- L'opération de pooling ou la couche de max pooling
- La couche de la fonction d'activation ReLu
- Les couches entièrement connectées ou couche de softmax qui sont similaires aux couches cachées du Perceptron Multicouche (MPL).

L'ensemble des couches successives de convolution-fonction d'activation et de pooling forme ce qu'on appelle « l'extracteur de caractéristique » ou feature extraction. En effet, dans le CNN, il faut

d'abord extraire les caractéristiques de l'image avant d'en prédire la classe a la sortie des couches entièrement connectées pour que le système soit le plus efficace possible. [20]

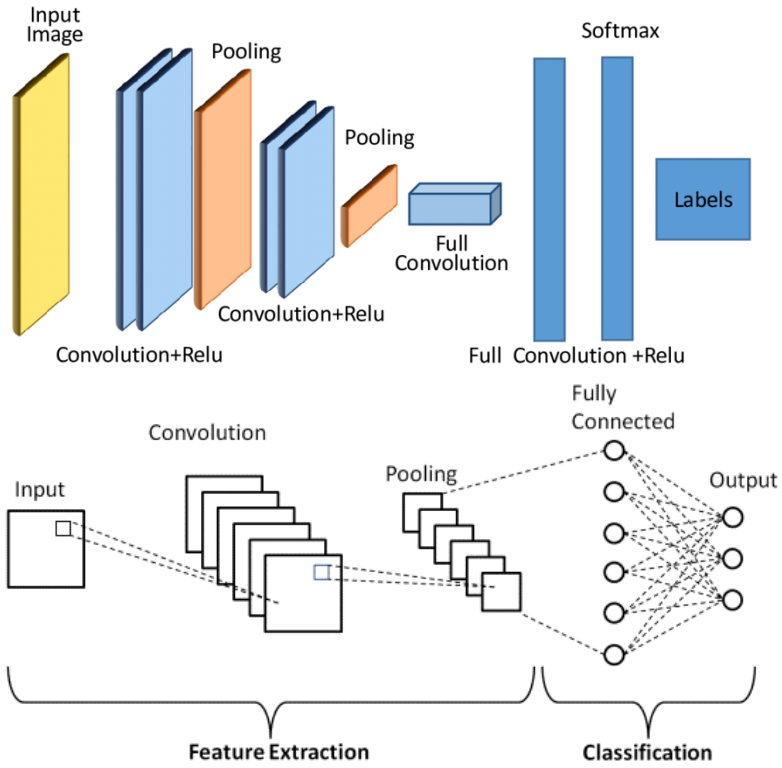
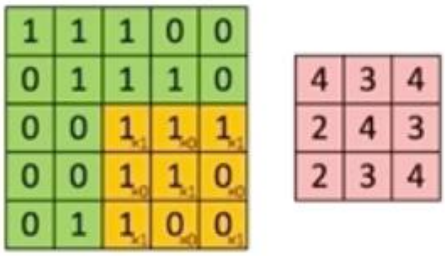


Figure 3.09 : *Architecture basique du CNN*

3.3.4 Principe de chaque couche du CNN

3.3.4.1 Couche de convolution

La couche est composée de plusieurs filtres, chacun est appliqué à la matrice (image) d'entrée. Chaque filtre n'est rien d'autre qu'une matrice k*k dont les poids sont W_i. Chaque poids est un paramètre dont le modèle apprend au fur de son entrainement. Le rôle de cette première couche est d'analyser les images fournies en entrée et de détecter la présence d'un ensemble de features. On obtient en sortie de cette couche un ensemble de features maps. (Voir le paragraphe du dessus concernant le fonctionnement de la convolution). **Concrètement la convolution est une somme de produits entre les valeurs RVB des pixels et les coefficients du kernel ou de la fenêtre de convolution.** [21]



À chaque passage du filtre sur l'image, on calcule une somme de produits entre les valeurs de chaque canal et le kernel.

$$Filtre: \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Figure 3.10 : *Architecture basique du CNN*

3.3.4.2 Couche d'activation

Après sortie de la couche de convolution, l'image filtrer est appliquer à une fonction dite « fonction d'activation » pour **accélérer l'apprentissage** selon les études. Généralement, on utilise la « Rectified linear function (ReLu) ». Cette couche remplace toutes les valeurs négatives reçues en entrées par des zéros, ainsi la sortie ne présente que des valeurs positives. L'intérêt de ces couches d'activation est de rendre le modèle non linéaire et de ce fait plus complexe. [17]

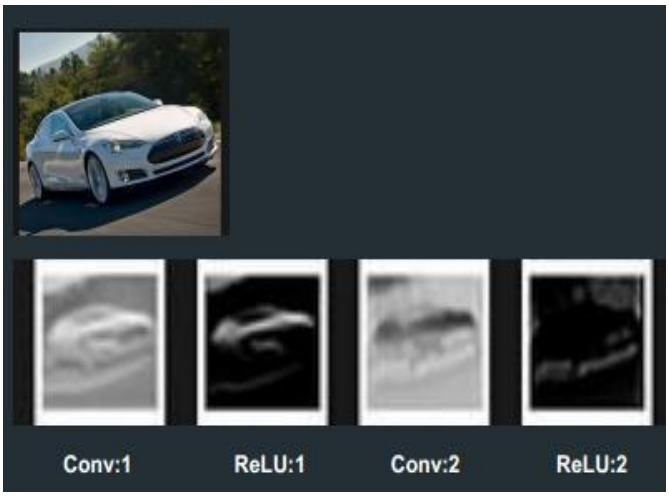


Figure 3.11 : *Exemple d'image à la sortie des couches de convolutions et de fonction d'activations Relu*

3.3.4.3 Couche de Pooling

La couche de Pooling est une opération généralement appliquée entre deux couches de convolution. Celle-ci reçoit en entrée les features maps formées en sortie de la couche de convolution et **son rôle est de réduire la taille des images, tout en préservant leurs caractéristiques les plus essentielles**. De plus, son intérêt est qu'il réduit le coût de calcul en **réduisant le nombre de paramètres à apprendre** et fournit une **invariance par petites translations**. Parmi les plus utilisés, on retrouve :

- Le max-pooling prend le maximum dans la fenêtre créant ainsi par la même occasion

une nouvelle matrice de sortie où chaque élément correspondra aux maximums de chaque région rencontrée

- L'average pooling dont l'opération consiste à conserver à chaque pas, la valeur moyenne de la fenêtre de filtre". [17] [20] [22]

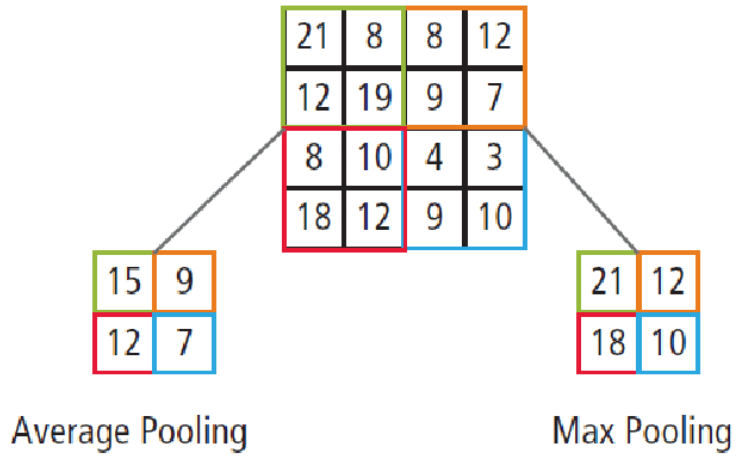


Figure 3.12 : *Illustration de l'opération de Average pooling et Max-Pooling avec une fenêtre de taille 2×2 avec pas de 2.*

3.3.4.4 Couche entièrement connectée

Ces couches sont placées en fin d'architecture du CNN et sont **entièrement connectées à tous les neurones de sorties** (d'où le terme fully-connected) qui fonctionnent similairement aux couches cachées des Perceptrons multicouches ou MPL. La matrice à la sortie de l'extracteur de caractéristique est vectorisée. Puis mis à l'entrée des couches entièrement connectée. Si on a par exemple à la sortie de l'extracteur de caractéristique une matrice 4*4*40, elle est transformée en vecteur de taille 1*640. Après avoir reçu un vecteur en entrée, la couche FC applique successivement une **combinaison linéaire** puis une **fonction d'activation** dans le but final de **classifier l'input image** (voir schéma suivant). Elle renvoie enfin en sortie un **vecteur de taille d correspondant au nombre de classes** dans lequel chaque composante représente la probabilité pour l'input image d'appartenir à une classe. [20] [22]

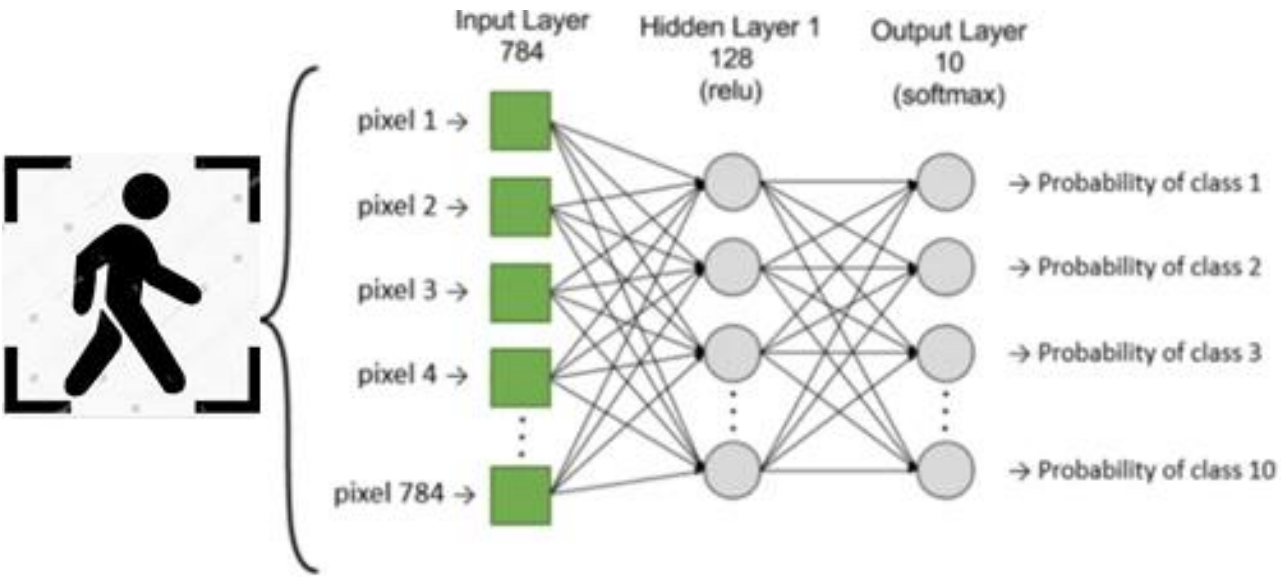


Figure 3.13 : *Représentation du modèle MPL pour la classification d'images facial.*

$$P(y = c|x; w; b) = softmax(X^T w + b) = \frac{e^{x^T w_c + b_c}}{\sum_j e^{x^T w_j + b_j}} \quad (3.10)$$

Où **y** est la class prédite, **x** est le vecteur sorti de la couche précédente, **w** et **b** sont respectivement les poids et les biais associés à chaque neurone de la couche de sortie et **P** la probabilité de la classe.

3.3.5 Principe de la CNN en profondeur

La principale différence entre les convolutions 2D et la convolution en profondeur est que les convolutions 2D sont effectuées sur tous/plusieurs canaux d'entrés, alors que dans la convolution en profondeur, chaque canal est séparé. Voici l'approche en trois (03) étapes :

- Le tenseur d'entrée de 3 dimensions est divisé en canaux séparés.
- Pour chaque canal, l'entrée est convoluée avec un filtre (2D).
- La sortie de chaque canal est ensuite empilée pour obtenir la sortie sur l'ensemble du tenseur 3D. [23] [24] [25]

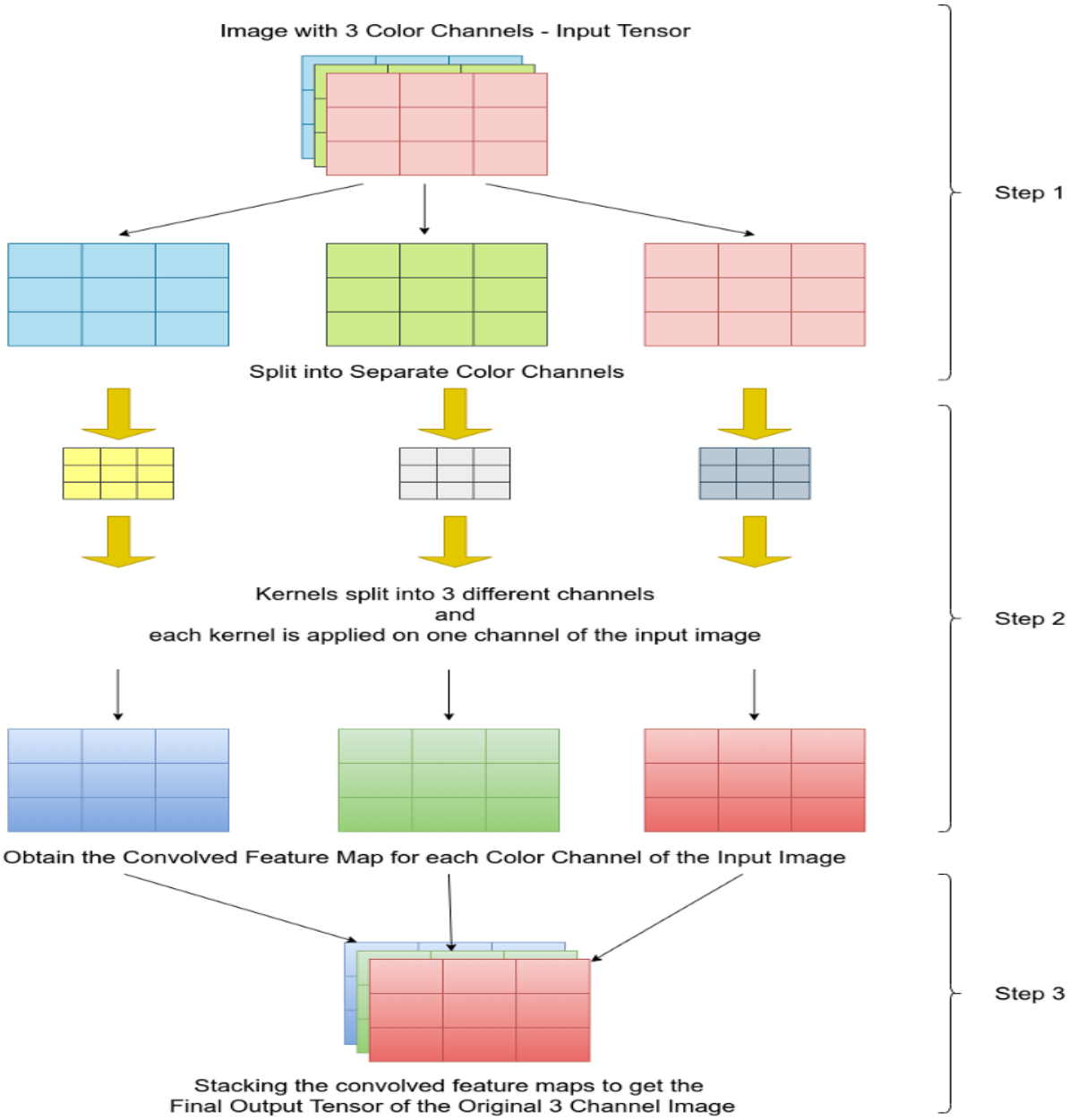


Figure 3.14 : *Représentation des trois étapes de la convolution en profondeur*

L'image initiale composée de trois canaux (RGB) subira simultanément pour chacun de ses canaux une convolution réalisée par trois kernels différents de taille fixe et de profondeur 1. Chaque kernel (ici trois) va générer une image représentant une caractéristique particulière de l'image initiale. On obtient alors en sortie 3 images de plus petite taille. **La Depthwise convolution ou convolution en profondeur n'augmente pas la profondeur de l'image.** [23] [24] [25]

3.3.6 Principe de la CNN séparable en profondeur

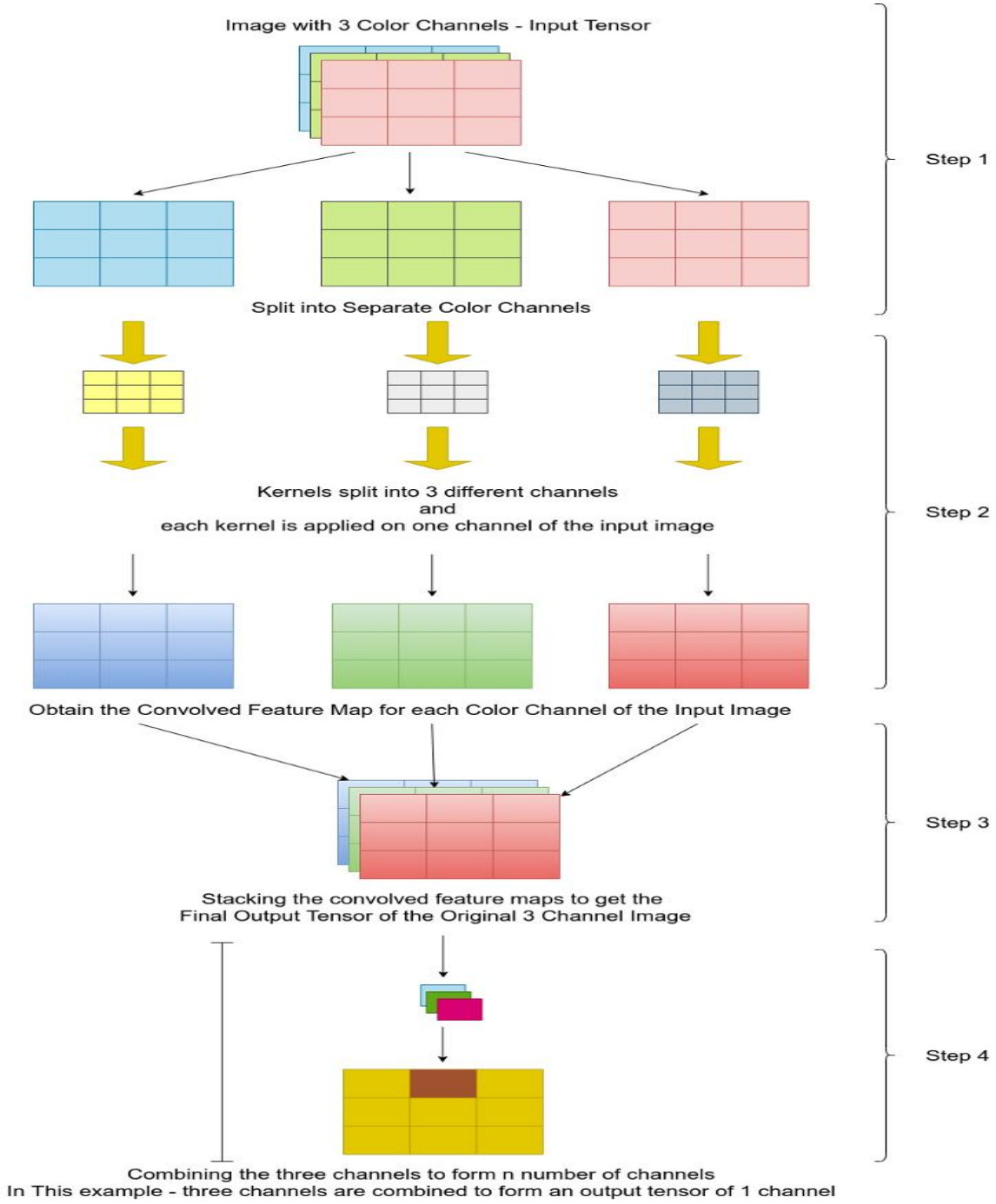


Figure 3.15 : *Représentation des étapes de la convolution séparable en profondeur*

Les convolutions en profondeur (voir la figure dans la page précédente) sont généralement appliquées en combinaison avec une autre étape, la convolution séparable en profondeur. Elle s'effectue en deux étapes :

- **La convolution en profondeur :** On applique un filtre sur chaque canal simultanément, contrairement à la convolution classique qui applique un filtre sur l'ensemble des canaux.

- **Pointwise convolution :** consiste à combiner les sorties de la convolution en profondeur pour former 'n' nombre de canaux que l'on souhaite, elle est aussi appelée convolution 1×1 ou point par point. [23] [24] [25]

3.3.7 Avantage du CNN séparable en profondeur

La convolution standard effectuer beaucoup plus d'opérations de multiplication que la convolution séparable en profondeur. La convolution séparable en profondeur permet d'économiser les ressources en termes de calcul et de mémoire (moins de paramètres) et donc, plus rapides avec des résultats nettement meilleurs. [23] [24] [25]

$$\text{Convolution standard} = N * {D_p}^2 * {D_k}^2 * M \tag{3.11}$$

$$\text{Convolution séparable en profondeur} = M * {D_p}^2 * ({D_k}^2 + N) \tag{3.12}$$

${D_k}^2$: taille de la fenêtre de convolution ; ${D_p}^2$: taille de l'image à la sortie du filtre ; M : nombre de canaux ; N : nombre du type de filtre appliquer.

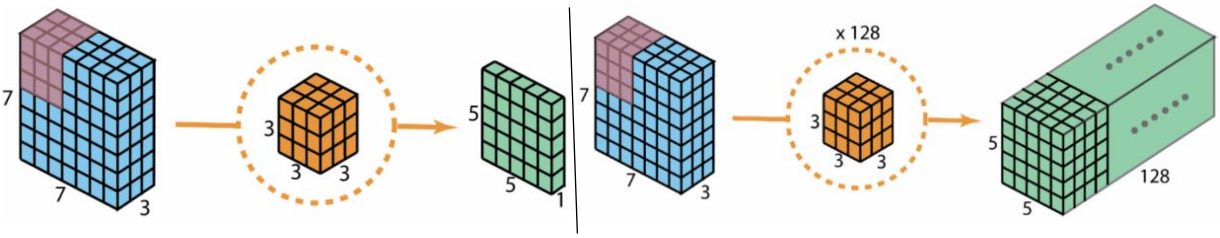


Figure 3.16 : *Représentation de la Convolution 2D standard pour créer une sortie avec 1 couche, en utilisant 1 filtre et Convolution 2D standard pour créer une sortie avec 128 couches, en utilisant 128 filtres*

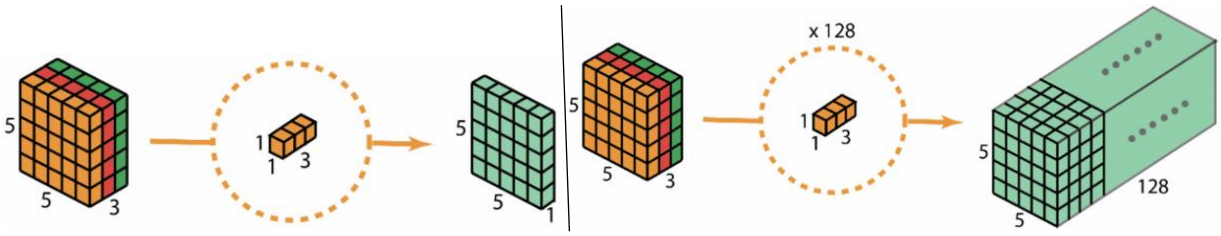


Figure 3.17 : *Représentation de la convolution séparable en profondeur pour créer une sortie avec 128 couches, en utilisant 128 filtres*

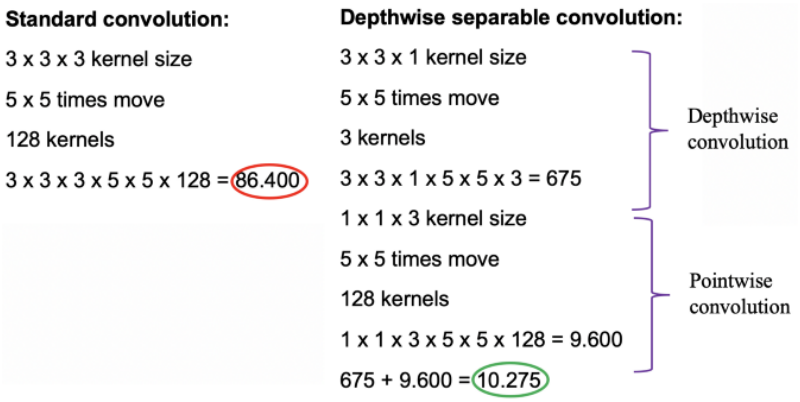


Figure 3.18 : *Comparaison des opérations de multiplication réaliser par la convolution standard et la convolution séparable en profondeur*

3.4 Régression linéaire

La régression linéaire est une approche ou un modèle permettant de prédire la relation entre différentes variables (principalement une variable dépendante et une ou plusieurs variables indépendantes). La régression linéaire est paramétrique, de paramètre m et b et ne va donc pas avoir besoin de conserver toutes les données pour effectuer des prédictions, mais seulement m et b.

. Soit x la variable indépendante, y la variable dépendante, b l'ordonnée à l'origine et m la pente de la droite, alors : **y = mx + b.** (3.13)

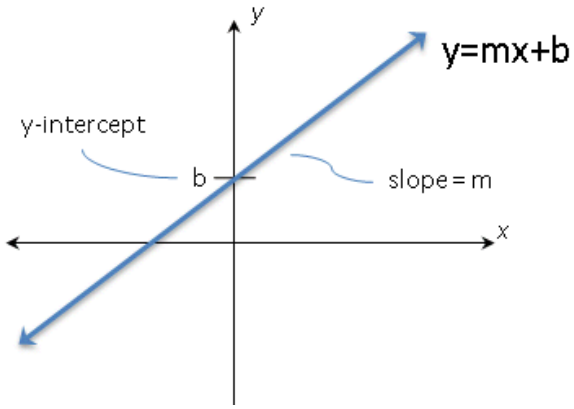


Figure 3.19 : *Représentation d'une droite de régression linéaire*

Si nous pouvons trouver la valeur de m et b, nous pouvons prédire la valeur de y pour x donné (en utilisant la formule y = mx + b). Mais, il peut y avoir de nombreuses valeurs possibles de m et b avec lesquelles nous pouvons former de nombreuses lignes linéaires. Pour trouver la meilleure ligne pour l'ensemble de données donné, nous utilisons la descente de gradient. [17] [26]

3.5 Classificateur K-NN ou K-plus proche voisin

K-Nearest Neighbours est un algorithme standard de classification qui repose exclusivement sur le choix de la métrique de classification. Il est « non paramétrique » (seul k doit être fixé) et se base uniquement sur les données d'entraînement. **Il peut être utilisé aussi bien pour la régression que pour la classification.** [17]

3.5.1 Principe de l'algorithme

L'idée est la suivante : à partir d'une base de données étiquetée, on peut estimer la classe d'une nouvelle donnée en regardant quelle est la classe majoritaire des k données voisines les plus proches. Le seul paramètre à fixer est k, k est le nombre de voisins à considérer (voir figure).

Pour appliquer cette méthode, les étapes à suivre sont les suivantes :

- On fixe le nombre de voisins k.
- On détecte les k-voisins les plus proches des nouvelles données d'entrée que l'on veut classer par une fonction de distance (dépends du type de donnée à classifier).
- On attribue les classes ou les valeurs correspondantes par vote majoritaire (pour ces voisins, l'algorithme se basera sur leurs variables de sortie ou output variable pour calculer la valeur de la variable de l'observation qu'on souhaite prédire).

Par ailleurs :

- **Si K-NN est utilisé pour la régression**, c'est la moyenne (ou la médiane) des variables des plus proches observations qui servira pour la prédiction

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{: Moyenne arithmétique} \tag{3.14}$$

- **Si K-NN est utilisé pour la classification**, c'est le mode ou la valeur dominante des variables des plus proches observations qui servira pour la prédiction

Pour effectuer une prédiction, l'algorithme K-NN ne va pas calculer un modèle prédictif à partir d'un *Training Set* comme c'est le cas pour la régression logistique ou la régression linéaire. En effet, K-NN **n'a pas besoin de construire un modèle prédictif.** Ainsi, pour K-NN il n'existe pas de phase d'apprentissage proprement dite. C'est pour cela qu'on le catégorise parfois dans le **Lazy Learning**. Pour pouvoir effectuer une prédiction, K-NN **seulement base sur le jeu de données** pour produire un résultat. [27] [28] [29] [30] [31]

Mais, comment choisit-on ce paramètre k lors de l'implémentation de l'algorithme ?

- On fait varier k (moins on utilisera de voisins K, plus on sera sujette au sous-apprentissage (underfitting). Par ailleurs, plus on utilise de voisins K, plus sera fiable dans notre prédiction. Toutefois, si on utilise nombre de voisins avec K =N et N étant le nombre d'observations, on risque d'avoir du overfitting.

  Pour chaque valeur de k, on calcule le taux d'erreur de l'ensemble de tests

$$taux\ erreur = \frac{|nombre\ total\ d'estimation - nombre\ total\ de\ valeur\ exact|}{nombre\ total\ de\ valeur\ exact} X100 \qquad (3.15)$$

- On garde le paramètre k qui minimise ce taux d'erreur test (varie en fonction du jeu de données). [27] [28] [29] [30] [31]

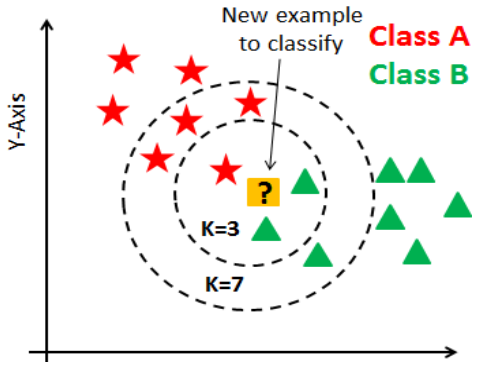


Figure 3.20 : *Représentation du fonctionnement de l'algorithme KNN*

3.5.1.1 calcule de similarité dans l'algorithme du KNN

K-NN a besoin d'une fonction de calcul de distance entre deux observations. Plus deux points sont proches l'un de l'autre, plus ils sont similaires et vice versa. Il existe plusieurs fonctions de calcul de distance selon le cas d'utilisation de l'algorithme, notamment : [27] [28] [29] [30] [31]

- **La distance euclidienne** pour les données quantitatives (exemple : reconnaissance faciale, reconnaissance d'émotion, etc.) et **du même type.** C'est la distance qui calcule la racine carrée de la somme des différences carrées entre les coordonnées de deux points :

$$\boldsymbol{D_e(x,y) = \sqrt{\sum_{j=1}^{n}(x_j - y_j)^2}} \qquad (3.16)$$

- **Distance Manhattan** est une bonne mesure à utiliser quand **les données d'entrée ne sont pas du même type** (exemple : estimation de l'âge, reconnaissance du genre, etc.). C'est le calcul la somme des valeurs absolues des différences entre les coordonnées de deux points :

$$\boldsymbol{D_m(x,y) = \sum_{i=1}^{k}|x_i - y_i|} \qquad (3.17)$$

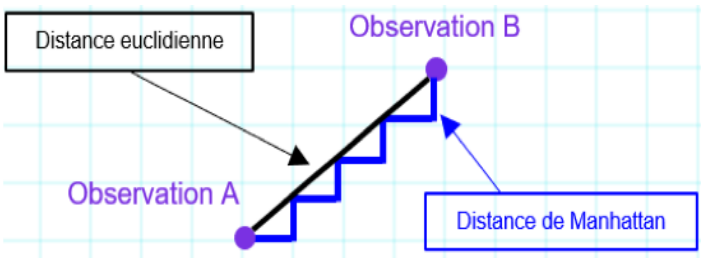


Figure 3.21 : *Représentation du fonctionnement*

3.6 Classificateur SVM

Les SVM (Support Vector Machine ou séparateur à vaste marge) sont des classificateurs qui permettent de traiter des problèmes non linéaires en les reformulant en problèmes d'optimisation quadratique (ils peuvent aussi être utilisés dans les problèmes de régression). Qui sont beaucoup plus faciles à résoudre : Ce sont un ensemble de techniques d'apprentissage supervisé qui ont pour objectif de trouver, dans un espace de dimension N>1, l'hyperplan qui divise au mieux un jeu de donné en deux. Les SVM sont des **séparateurs linéaires**, c'est-à-dire que la frontière séparant les classes est une droite, mais aussi utiliser dans les problèmes linéairement non séparables. [32]

3.6.1 Principe de l'algorithme SVM

Ils reposent sur deux idées clés :

- **La notion de marge maximale :** Sans trop rentrer dans les détails théoriques, la marge maximale est la frontière de séparation des données qui maximise la distance entre la frontière de séparation et les données les plus proches (c.-à-d. qui maximise la marge)
- **La notion de fonction noyau :** Sans trop rentrer dans les détails théoriques, en quelques mots également, une fonction noyau est une sorte d'alternative à un produit scalaire dans un espace à très grande dimension.

$$noyau\ polynomiale: (\boldsymbol{a}\ \boldsymbol{X}\ \boldsymbol{b} + \boldsymbol{r})^{\boldsymbol{d}} \tag{3.18}$$

$$noyau\ radial : \boldsymbol{e}^{-\boldsymbol{\gamma}(\boldsymbol{a}-\boldsymbol{b})^2} \tag{3.19}$$

a et **b** : valeurs des deux classes ; **r**: détermine le coefficient du polynôme ; **d** ou $\boldsymbol{\gamma}$: dégrée du polynôme

Fin de trouver cette fameuse frontière séparatrice :

- Il faut donner au SVM des données d'entrainement. On donne à l'algorithme un jeu de données dont on connait déjà les deux classes.
- On entre alors dans la phase d'entrainement. Le SVM va déterminer la frontière la plus plausible. Mais comment choisir la frontière alors qu'il y en a une infinité ?
- C'est là qu'intervient la première idée clé : la marge maximale.

- La frontière choisie doit maximiser sa distance avec les points les plus proches de la frontière. Les points d'entraînement les plus proches de la frontière sont d'ailleurs appelés vecteurs support. Ils sont appelés comme cela, car la frontière donnée par un SVM ne dépend que des vecteurs support. Ils sont donc le support que le SVM utilise pour construire la frontière
- Après la phase d'entrainement, le SVM a « appris ». Ainsi, après plusieurs phases d'entrainement, le SVM sait où placer la frontière pour de nouvelles données. [32] [33] [34]

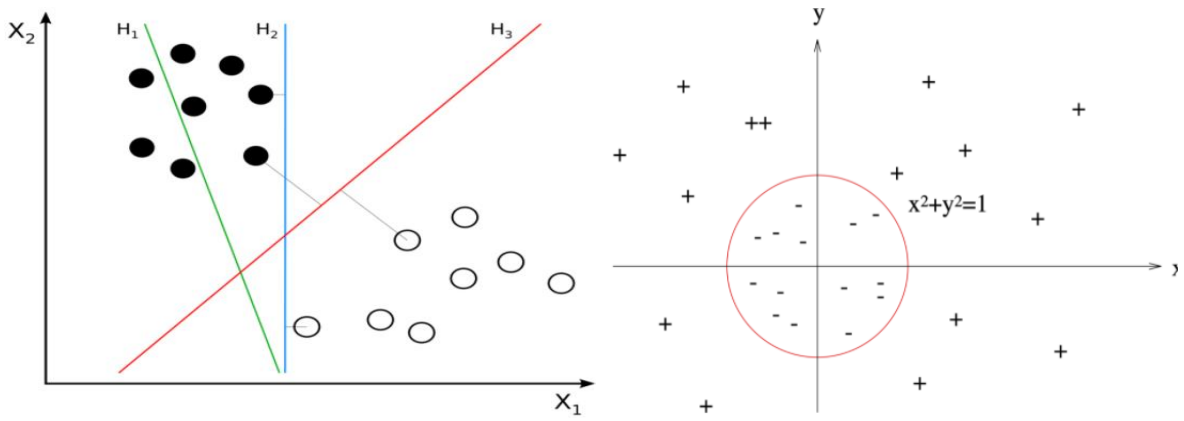


Figure 3.22 : *Exemple illustrant séparateur de deux classes*

Dans de la première figure, nous avons ci-dessus un exemple d'hyperplan séparateur pour N=2. H1 ne sépare pas correctement le jeu de donnée ; H2 le sépare bien, mais pas de façon optimale ; H3 sépare le jeu de donnée avec la marge maximale. (Utilise le noyau polynomial)

La seconde figure représente, un problème non linéairement séparable qui utilise la fonction noyau radiale

Remarque :

- Afin de contourner le problème non linéairement séparable, on plonge dans un espace de dimension supérieur (éventuellement infini) et on reconsidère le problème dans cet espace-là.

Cette manœuvre permet de passer d'un problème non linéairement séparable à un problème linéairement séparable. La fonction noyau joue un rôle primordial.

- Pour traiter les problèmes de classification à plusieurs classes, ils utilisent plusieurs solutions, comme le one vs all (cette approche consister à créer autant de SVM que de catégories présentes) ou même le KNN. [33] [34]

3.7 Conclusion

Pour conclure, dans ce chapitre nous avons vu les principales connaissances requises pour comprendre le deep-learning ou l'apprentissage profond : nous avons vu ce qu'est un réseau de neurones, son fonctionnement et comment il apprend. Nous avons aussi appris en profondeur le réseau de neurones à convolution concernant son fonctionnement, ces avantages, son architecture générale. Nous avons aussi parlé du fonctionnement et du principe du CNN en profondeur et du CNN séparable en profondeur.

On a aussi présenté l'algorithme des K-plus proches voisins et du SVM ou séparateur à vaste marge qui sont des méthodes de classification de donnée. L'implémentation des techniques de classification développer dans ce chapitre devient de nos jours de plus en plus facile, grâce à l'essor sans arrêt des ressources matérielles (processeurs graphiques) et donc plus de capacité de calcul.

Grâce à ce qu'on a vu dans ce chapitre nous avons appris que l'apprentissage profond englobe certains modèles de prédiction pour arriver à ses fins. Pour le traitement d'image, le réseau de neurones à convolution est le plus efficace et le plus adapter. De plus, combiner à la technique de CNN séparable en profondeur il devient encore plus performant. Il nous permet de faire la classification d'image : en extrayant d'abord ces caractéristiques après le passage dans plusieurs couches de convolution et de pooling. Puis grâce aux cartes de caractéristiques, un réseau Multi Perceptron Layer ou MPL va faire son travail d'apprentissage pour classifier les images.

Notre dernier chapitre va se focaliser sur l'architecture utilisée pour le fonctionnement et la réalisation de notre projet.

CHAPITRE 4

LA RÉALISATION DU PROJET

4.1 Introduction

Que savez-vous de votre visage ? Avez-vous déjà utilisé ces expressions "Rire du bout des dents", "Avoir le front de faire quelque chose", "Faire bonne figure", "S'en mordre les lèvres", "Avoir les paupières lourdes", "Rester bouche bée » ? Ces expressions sont largement utilisées par les peuples du monde pour faire passer une sensation, une émotion ou une idée par le biais des traits du visage. Le visage d'une personne nous permet tout d'abord de l'identifier, aussi d'estimer son âge et de savoir son sexe. De là, nous pouvons imaginer le rôle important du visage dans la communication et sa puissance de transmettre un message spécifique à un interlocuteur parce que le visage est l'un des canaux les plus puissants de la communication non verbale. C'est pourquoi nous allons concevoir un système intelligent pour la vidéo surveillance basée sur le visage humain.

4.2 Architecture générale d'un système de reconnaissance d'image faciale

Pour faire le traitement d'image, une image doit passer les blocs suivants à fin d'être reconnue par la machine dans la vision par ordinateur. Ce sont les blocs communs utiliser pour faire des reconnaissances basées sur l'image faciale. [34]

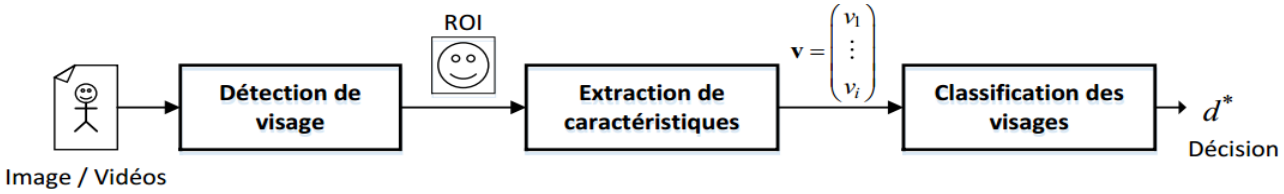


Figure 4.01 : *Schéma-blocs général d'un système de reconnaissance de visages*

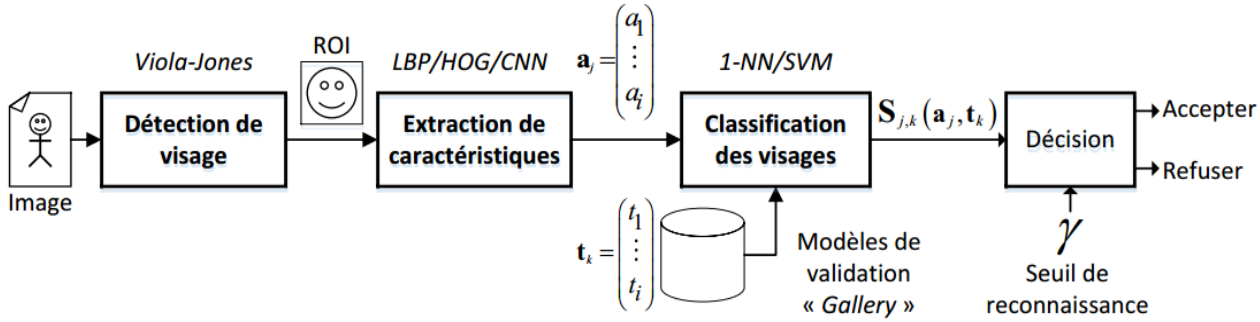


Figure 4.02 : *Schéma-blocs adapté au système de reconnaissance de visages selon les divers algorithmes employés*

Un système qui effectue une reconnaissance automatique dans le domaine du visage que ce soit pour l'identifier ou reconnaitre son expression ou estimer son âge ou identifier son sexe est généralement composé de **trois modules principaux**, comme illustrés dans la figure 4.01 et 4.02.

- Le premier module consiste à détecter et enregistrer la région du visage dans les images ou les séquences d'images d'entrée. Il peut s'agir d'un détecteur pour détecter le visage dans chaque image ou simplement détecter le visage dans la première image, puis suivre le visage dans le reste de la séquence vidéo.
- Le deuxième module consiste à extraire et représenter les cartes de caractéristique de chaque image.
- Le dernier module détermine une similarité entre l'ensemble des caractéristiques extraites et un ensemble de caractéristiques de référence pour faire la classification. D'autres filtres ou modules de prétraitement de données peuvent être utilisés entre ces modules principaux pour améliorer les résultats de détection, d'extraction de caractéristiques ou de classification.

Dans le module de classification, il peut s'agir de : classification pour la reconnaissance faciale, de reconnaissance d'expression faciale, d'estimation de l'âge, reconnaissance du genre ou tous en même temps. C'est ce que nous allons réaliser dans le projet de mémoire. [34] [35]

4.2.1 Acquisition du visage

Le problème de détection et d'enregistrement des visages implique l'identification de la présence de visages dans une image et la détermination des emplacements et des échelles des visages. La précision de la détection et l'enregistrement du visage est particulièrement importante dans des conditions réalistes, où la présence du visage dans une scène et sa localisation globale ne sont pas connues a priori. Un système complet de localisation du visage devrait faire face à certains défis décrits ci-dessous :

- **Pose :** Les traits du visage, y compris les yeux, le nez et la bouche, peuvent être partiellement invisibles ou déformés en raison de la pose relative du visage ou de la caméra.
- **Occlusion :** Les traits du visage peuvent être obstrués par une barbe, une moustache et des lunettes. De même, le maquillage peut provoquer l'apparition de régions artificielles sur le visage ou cacher les limites faciales normales.
- **Expression :** les traits du visage montrent de grands changements dans leur forme sous différentes expressions. Certaines caractéristiques peuvent devenir invisibles ou d'autres ne sont visibles que sous différentes expressions. [34] [35]

- **Conditions d'acquisition de l'image :** L'éclairage et les changements dans les caractéristiques de la caméra affectent de manière significative la chrominance des régions du visage. Certaines caractéristiques peuvent être masquées ou combinées avec des ombres ou des brillances sur le visage provoquant ainsi une perte d'informations sur les couleurs.

4.2.1.1 Détection du visage

La première étape d'un système d'analyse faciale entièrement automatique consiste à localiser la région du visage et ses limites. Le but de la détection du visage est de déterminer si un visage est présent ou non sur l'image et, le cas échéant, de localiser son emplacement (voir figure 4.03, le rectangle rouge englobe le visage détecté). Une étude exhaustive sur les algorithmes de détection de visages dans a regroupé les différentes méthodes en deux catégories :

- La première catégorie est basée sur des modèles rigides et comprend les variations de boosting. Les principaux algorithmes de cette catégorie comprennent l'algorithme de détection de visage Viola-Jones (VJ) et ses variations, les algorithmes basés sur des réseaux neuronaux convolutionnels (Convolutionnel Neural Network, CNN) et CNN profond (Deep CNN, DCNN), et les méthodes qui appliquent des stratégies inspirées de l'extraction d'images (image-retrieval) et la transformée généralisée de Hough.
- La deuxième catégorie est basée sur l'apprentissage profond et le transfert learning ou l'application d'un modèle hybride (ou combinaison de plusieurs méthodes pour avoir un système efficace). Ces méthodes peuvent également combiner détection de visage et localisation de la partie faciale et sa classification. Cette famille d'algorithmes s'articule principalement autour des extensions et des variations de la méthodologie générale de détection d'objets. **Pour notre projet nous avons choisi cette catégorie.** [34] [35]

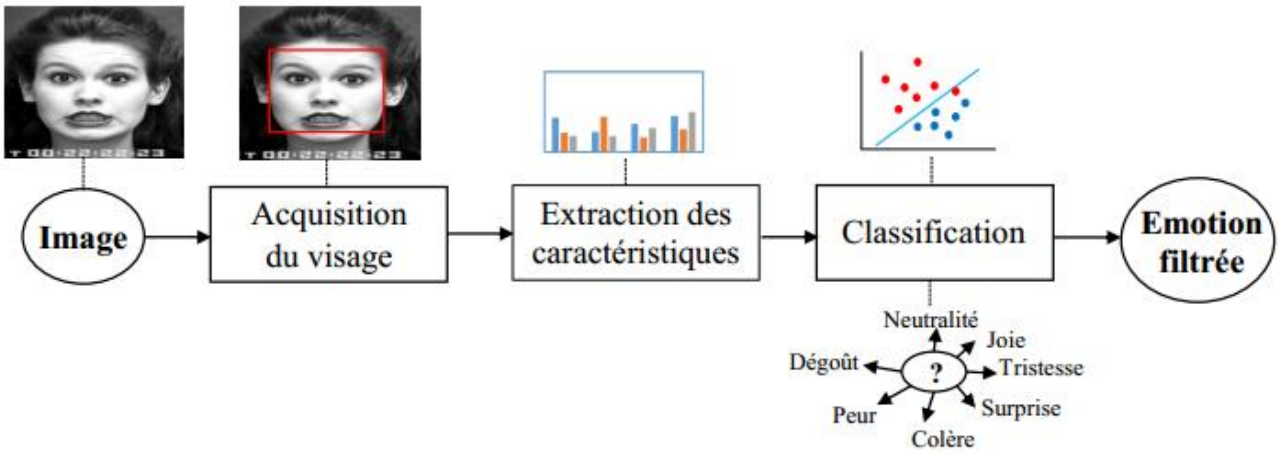


Figure 4.03 : *Modules du système d'analyse des expressions faciales*

4.2.1.2 Détection des points caractéristiques du visage

Les points caractéristiques du visage sont principalement situés autour des composants faciaux tels que **la bouche, le sourcil droit, le sourcil gauche, l'œil droit, l'œil gauche, le nez, la mâchoire.** La détection des points caractéristiques du visage commence habituellement à partir d'une boîte englobante rectangulaire renvoyée par un détecteur de visage qui localise ce dernier (voir Section précédente 4.2.1.1). Bien qu'optionnelle, cette étape de détection des points faciaux est importante, car elle facilite la décomposition du visage (voir Section 4.2.1.3), l'extraction de caractéristiques géométriques telles que les contours des composants faciaux, les distances faciales, etc., et fournit les emplacements où les caractéristiques d'apparence peuvent être calculées (voir Section 1.3.2.2). **Pour notre projet cette partie sera effectuée par un modèle basé sur les CNNs est ses variants.** [34] [35]

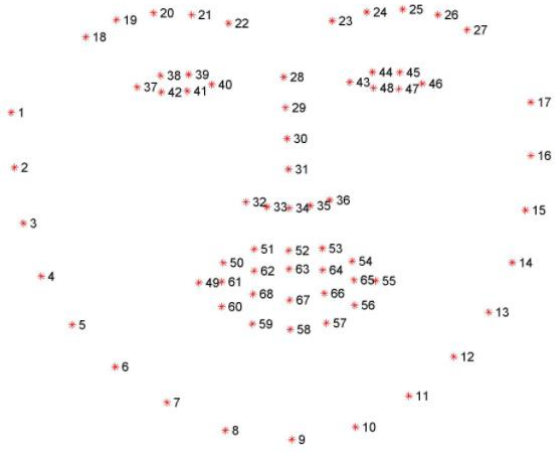


Figure 4.04 : *Visualisation des 68 points de repère du visage*

4.2.1.3 Traitement du visage : visage entier vs régions spécifiques

L'enregistrement du visage est une étape fondamentale pour son référencement. En général, il vise à trouver la région d'intérêt (Region Of Interest, ROI), principalement le visage entier, à partir de l'image d'entrée, et à normaliser par la ROI trouvée en détectant certains composants faciaux internes tels que les yeux. Par exemple, le visage peut être aligné et normalisé en fonction des emplacements des yeux détectés et de la distance entre eux, ce qui entraîne la suppression de la translation et la différence d'échelle **donc, la réduction des paramètres de calcule pour le prochain bloc de traitement (rapidité du prochain traitement) et d'avoir plus de précision sur la région à traiter**. Cependant, cette approche simple reste sensible aux rotations de la tête et à la variation des sujets. Mais grâce, à la méthode basée sur l'enregistrement de l'emplacement des points caractéristiques par

un modèle d'apprentissage profond. Il peut être effectué en utilisant les coordonnées des points faciaux estimés par un détecteur de points caractéristiques (voir Section 4.2.1.2). [34] [35]

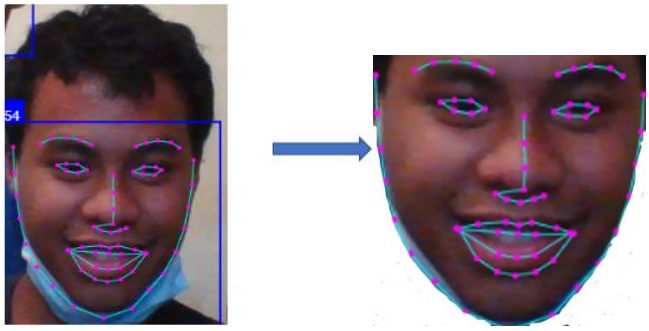


Figure 4.05 : *ROIs du visage obtenu grâce aux 68 points de repère faciaux*

Cette méthode est appliquée pour perfectionner la reconnaissance (dans le cas de la vue de profil ou d'autre vue différente de la vue de face) **et aussi surtout utile pour la reconnaissance d'expression faciale et de l'estimation de l'âge, mais aussi de la détection du genre** (qui se base tout sur l'apparence de ces 68 points de repère faciale) sur la totalité de la zone du visage détecté. En effet, certaines régions du visage sont totalement indépendantes de la production d'expression ou de l'estimation de l'âge. Elles peuvent être supprimées de la zone détectée sans influence sur la reconnaissance des expressions faciale et l'estimation de l'âge. Ainsi, la plupart des régions du visage ne participent pas à la production d'expressions faciales et de l'âge de la personne. Par conséquent, différents travaux ont sélectionné les sous-régions du visage qui subissent un changement durant ou lors d'une expression faciale.

Pour les anciennes méthodes, les caractères spécifiques du visage ont été segmentés en plusieurs ROIs, tels que : les sourcils, œil gauche, œil droit, autour des yeux et des lèvres, le nez, les joues, la mâchoire et la bouche, puisque ce sont les régions les plus représentatives des expressions faciales et de l'âge d'une personne (les cheveux ne le sont pas puis qu'ils peuvent être juste coloré). Cette région du visage détectée est insérée dans l'étape d'extraction **de ROIs qui en sortie donne une carte de caractéristique du visage qui permet la discrimination de chaque visage**.

Dans notre cas cette tâche est faite par un *CNN* spécialiser qui s'assure de ne pas faire aucune supposition sur l'importance des *features ou vecteurs caractéristiques,* et donc, **les vecteurs descripteurs obtenus deviennent théoriquement aussi discriminants que possible tout en devenant spécifiques au cas de reconnaissance dans l'image faciale**. [34] [35]

4.2.2 Extraction des caractéristiques du visage

Une fois l'enregistrement du visage effectué (zone spécifique du visage dans la section 4.2.1.3), l'étape suivante consiste à extraire et représenter les vecteurs caractéristiques du visage pour les diverses reconnaissances faciales (dans notre cas, le système utiliser est hybride). **L'obtention de caractéristiques efficaces d'expression faciale, à partir de ROI(s) extraite(s), est cruciale pour une reconnaissance d'image faciale réussie**.

Les expressions faciales et l'âge : sont définies principalement par la contraction des muscles faciaux qui produisent des changements dans l'apparence et la forme du visage. De ce fait, les méthodes d'extraction des caractéristiques pour l'analyse d'expression et de l'âge peuvent être séparées en quatre types d'approches :

- **Les caractéristiques géométriques** : représentent la forme et l'emplacement des composants du visage (y compris la bouche, les yeux, les sourcils et le nez). Les composants faciaux ou les traits faciaux sont extraits pour former un vecteur de caractéristiques représentant la géométrie du visage. Une méthode basée sur la géométrie est que les expressions faciales affectent la position relative et la taille des divers traits faciaux et que, en mesurant le mouvement de certains points faciaux, l'expression faciale sous-jacente peut être déterminée.
- **Les caractéristiques d'apparence :** représentent les changements d'apparence (texture de la peau) du visage tels que les rides et les sillons, ces caractéristiques d'apparence peuvent être extraites sur tout le visage ou sur des régions spécifiques du visage. Exemple l'ACP (Analyse en Composantes Principales)
- **Caractéristiques hybrides :** Il est connu que les caractéristiques basées sur la géométrie et l'apparence ont des avantages et limitations spécifiques respectifs. Par exemple, les caractéristiques géométriques sont efficaces dans le calcul alors qu'elles sont sensibles au bruit; en revanche, les caractéristiques basées sur l'apparence sont robustes au désalignement de l'image, mais cela prend beaucoup de temps de calcul.
- **Caractéristiques basées sur le deep learning :** L'apprentissage profond ou deep learning est un paradigme qui permet d'apprendre des représentations hiérarchiques multicouches à partir de données d'apprentissage. C'est la méthode la plus robuste et efficace pour l'extraction des caractéristiques d'image (basé sur le CNN). **Notre projet de se basera sur cette caractéristique d'extraction**.

L'extraction des caractéristiques du visage étant la dernière étape commune d'un système d'analyse faciale. [34] [35]

Figure 4.06 : *Représentation des cas de figure d'expression faciale*

4.2.3 Classification d'image faciale

La classification d'image faciale dépend étroitement du type de reconnaissance à faire. Chaque classificateur a sa spécificité selon le type de problème à résoudre :

- **La reconnaissance faciale est un problème de classification :** elle peut être résolue par les classificateurs suivants par exemple, le KNN et le SVM. Dans notre cas nous avons choisi le modèle de classification KNN ou K plus proche voisin, c'est le mode ou la valeur dominante des variables des plus proches observations qui servira pour la prédiction de l'image faciale. Ici on fait la mesure des distances euclidienne entre les K voisins et le voisin ayant la plus faible distance euclidienne sera la référence pour prédire le mode de la classification.
- **La reconnaissance d'expression faciale est aussi un problème de classification :** elle peut être résolue par les classificateurs suivants par exemple, le KNN et le SVM. Dans notre cas nous avons choisi le modèle de classification SVM ou séparateur à vaste marge qui se trouve plus performant et efficace pour classifier le vecteur caractéristique d'un visage parmi les différentes classes qui sont les expressions : neutre, heureux, surpris, triste, dégoûter, angoisser et énerver. C'est un problème non linéairement séparable qui utilise la fonction noyau radiale pour simplifier le problème de classification en augmentant la dimension des classes à prédire à fin d'utiliser **le one vs all** (Cette approche consister à créer autant de SVM que de catégories présentes).
- **L'estimation de l'âge est un problème de régression :** elle peut aussi être résolue par les classificateurs KNN et SVM. Dans notre cas nous avons choisi le modèle de classification KNN appliquer dans les problèmes de régression parce qu'il se trouve meilleurs que le SVM dans ce type de problème. Pour ce faire le modèle KNN prend les valeurs des K voisins définis et calcule la moyenne des valeurs de ces variables qui servira pour la prédiction de l'âge.

4.3 Architecture globale notre système de traitement d'image faciale

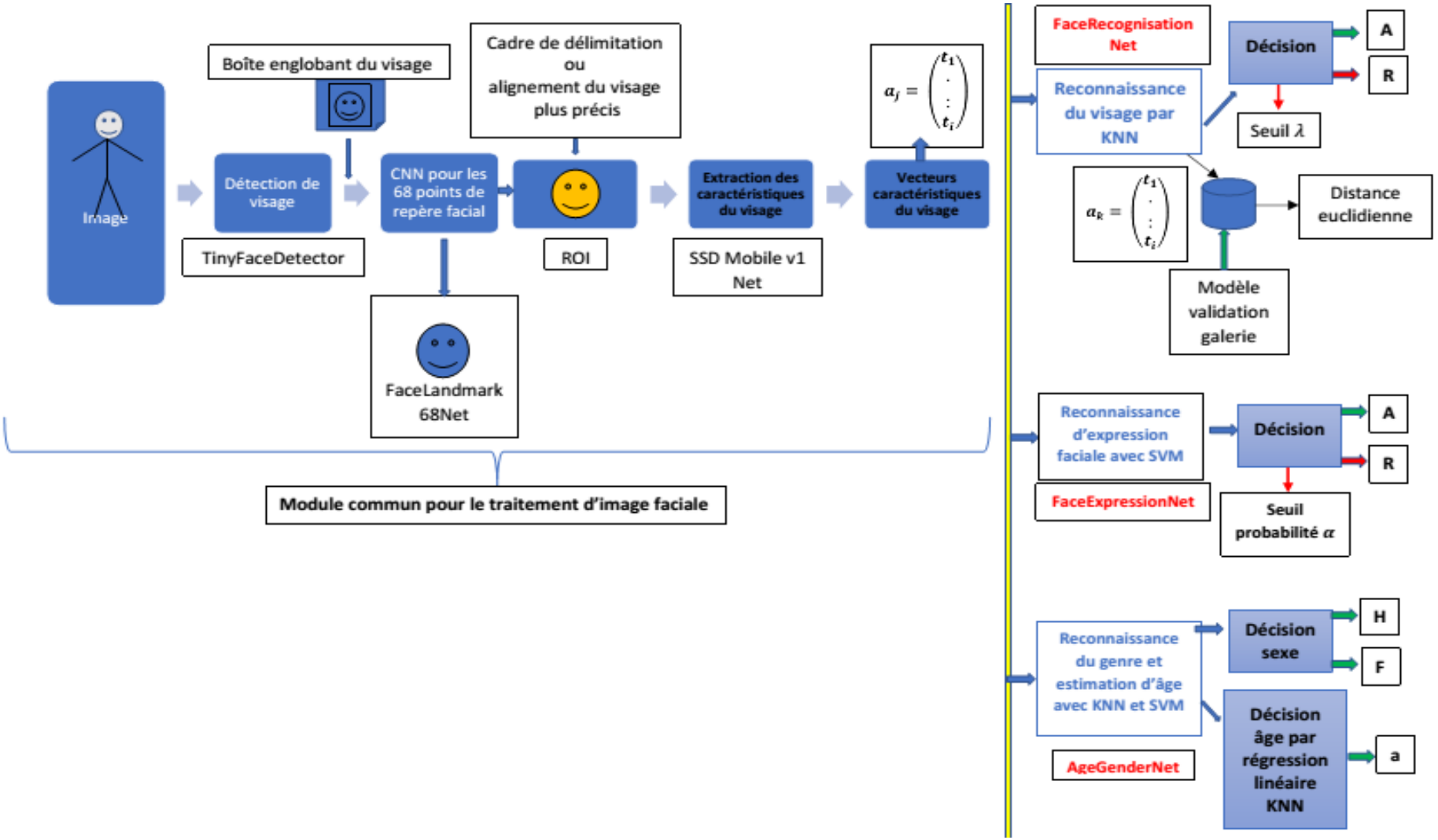


Figure 4.07 : *Architecture globale de notre modèle hybride de reconnaissance d'image faciale*

4.3.1 Fonctionnement de notre architecture hybride

4.3.1.1 Acquisition du visage

Pour l'acquisition ou la détection du visage on a utilisé le modèle de CNN Tiny Face Detector : c'est un détecteur de visage en temps réel très performant, qui est beaucoup plus rapide, plus petit et moins consommateur de ressources que le détecteur de visage SSD Mobilenet V1, en retour il fonctionne un peu moins bien pour détecter les petits visages. Ce modèle est extrêmement mobile et convivial pour le Web, il devrait donc être notre détecteur de visage GO-TO sur les appareils mobiles et les clients aux ressources limitées. La taille du modèle quantifié n'est que de 190 Ko.

Le détecteur de visage a été formé sur un ensemble de données personnalisé d'environ 14 000 images étiquetées avec des boîtes englobantes. De plus, le modèle a été formé pour prédire les boîtes englobantes, qui couvrent entièrement les points des caractéristiques faciales, il produit donc en général de meilleurs résultats en combinaison avec la détection ultérieure des points de repère du visage que SSD Mobilenet V1.

Ce modèle est essentiellement une version encore plus petite de Tiny Yolo V2 (You Only Look Once), remplaçant les convolutions régulières de Yolo par des convolutions séparables en profondeur. Yolo est entièrement convolutif, il peut donc facilement s'adapter à différentes tailles d'image d'entrée pour échanger la précision contre les performances (temps d'inférence).

4.3.1.2 Alignement du visage et détection des 68 points de repère faciale

Après avoir eu la boîte englobante du visage dans le bloc précédent. Cette image sera affinée avec précision à l'aide d'un détecteur de repère facial à 68 points très léger et rapide, mais précis. Le modèle par défaut a une taille de seulement 350 Ko (**face_landmark_68_model**) et le petit modèle ne fait que 80 Ko (**face_landmark_68_tiny_model**). Les deux modèles utilisent les idées de convolutions séparables en profondeur ainsi que de blocs densément connectés. Les modèles ont été formés sur un ensemble de données d'environ 35 000 images de visage étiquetées avec 68 points de repère de visage. **Cette étape permet d'avoir un cadre de délimitation de visage précis donc, de réduire les paramètres de calcul pour le prochain bloc de traitement (accélérer les calculs dans le prochain traitement) et d'avoir plus de précision sur la région à traiter.**

4.3.1.3 Extraction des caractéristiques du visage

SSD MobileNet (Single Shot MultiBox Detector) est un modèle d'architecture du réseau de neurones à convolution (CNN) qui se concentre explicitement sur la classification des images pour les applications mobiles. Plutôt que d'utiliser les couches de convolution standard, il utilise **des couches**

de convolution séparables en profondeur. Ce qui distingue ce modèle, c'est que son architecture réduit le coût de calcul et qu'une très faible puissance de calcul est nécessaire pour exécuter ou appliquer l'apprentissage par transfert.

Détecteur multibox à un seul coup : c'est aussi un détecteur multibox à prise unique est un algorithme qui ne prend qu'une seule prise de vue pour détecter de nombreux objets dans l'image à l'aide du multibox. Il utilise un seul réseau de neurones profonds pour y parvenir. Ce détecteur fonctionne à différentes échelles, il est donc capable de détecter des objets de différentes tailles/échelles dans l'image. Dans ce projet il reçoit en entrée une image du visage précis et bien délimité grâce au travail du bloc précédant, cela permettra au réseau de faire des opérations d'extractions de caractéristiques précises à partir de sa partie convolution fortement entrainer et rapide (le modèle de détection de visage a été formé sur l'ensemble de données WIDERFACE).

4.3.2 Fonctionnement de chaque bloc de classification ou de régression

4.3.2.1 Classificateur KNN pour la reconnaissance faciale

Face Recognisation Net est une architecture de type ResNet-34. Il est implémenté pour calculer un descripteur de visage à partir de n'importe quelle image de visage donnée, qui est utilisée pour décrire les caractéristiques du visage d'une personne. Il permet de déterminer la similarité de deux visages arbitraires en comparant leurs descripteurs de visage, par le calcule la distance euclidienne dans son classificateur KNN. Les poids ont été entraînés avec la base d'image davisking et le modèle atteint une précision de prédiction de **99,38 %** sur la référence LFW (Labeled Faces in the Wild) pour la reconnaissance faciale. La taille du modèle quantifié est d'environ 6,2 Mo. **Pour notre cas nous allons utiliser que la partie classificatrice MPL très performante du modèle Face Recognisation Net. Il reçoit en entrée les vecteurs caractéristiques calculés par le bloc précédent SSD Mobile net.** [36]

Pour le classificateur KNN ou K plus proches voisins, **c'est le mode ou la valeur dominante des K plus proches voisins qui servira pour la prédiction de l'identité faciale**. Ici on fait la mesure des distances euclidienne entre les K voisins et le voisin ayant la plus faible distance euclidienne sera la référence pour prédire le mode de la classification. Dans ce projet le nombre de K voisin sera paramétré automatiquement en fonction du nombre total d'images de référence de chaque personne à identifier dans la base d'image (le meilleur K d'après expérimentation). **Et la distance euclidienne**

de validation sera inférieure ou égale à 0.65 pour la reconnaissance éloignée et 0,5 pour la reconnaissance à courte distance. (Ces paramètres doivent être paramétrés manuellement)

4.3.2.2 Classificateur SVM pour la reconnaissance d'expression faciale

Face Expression Net : c'est un modèle léger, rapide et offre une précision raisonnable. Le modèle a une taille d'environ 310 Ko et il utilise des convolutions séparables en profondeur et des blocs densément connectés. Il a été formé sur une variété d'images provenant d'ensembles de données accessibles au public ainsi que sur des images récupérées sur le Web. Notez que le port de lunettes peut diminuer la précision des résultats de prédiction. Pour résoudre les problèmes de classification il utiliser le SVM. [36]

Le classificateur SVM ou séparateur à vaste marge qui se trouve plus performant et efficace pour problème de classification d'émotion. Il s'agit de classifier le vecteur caractéristique d'un visage parmi les différentes classes expressions faciales : neutre, heureux, surpris, triste, dégoûter, angoisser et énerver. C'est un problème non linéairement séparable qui utilise la fonction noyau radiale pour simplifier le problème de classification en augmentant la dimension des classes à prédire à fin d'utiliser **le one vs all** (Cette approche consister à créer 7 classes séparer non linéairement par le SVM dont chaque classe correspond à une émotion spécifique).

4.3.2.3 Classificateur SVM pour la reconnaissance du genre et KNN pour l'estimation de l'âge

Age Gender Net est un classificateur efficace et léger qui utilise le classificateur SVM ou séparateur à vaste marge : Il va chercher à trouver à l'aide de la fonction noyau le meilleur moyen pour représenter les deux classes (homme et femme) dans un hyperplan N = 2 en faisant recoure à la déclassification de certain point qui permettrait de faire une vraie séparation de classe pour faciliter la reconnaissance du genre. [36]

Cette étape faite, pour l'estimation de l'âge le SVM va résoudre ce problème en appliquant la solution de régression du KNN : dans le principe est de collecter les valeurs d'âges des points voisins K de notre point de référence et de calculer leurs moyennes (Cela toujours dans l'hyperplan N =2 de notre classificateur SVM).

4.4 Réalisation du projet

4.4.1 Langage et outils du développement

Pour la réalisation de ce projet, nous avons utilisé les outils suivants :

- Visual Studio (Editeur de texte)
- Javascript (langage de programmation)

- Html (Le HTML : HyperText Markup Language est le principal langage de programmation utilisé dans le WEB contrôlant la manière dont les pages web sont disposées et sont perçues. C'est donc en HTML qu'on écrit ce qui doit être affiché sur la page : du texte, des liens, des images).

- CSS (Cascading Style Sheets, aussi appelés Feuilles de style gère l'apparence de la page web : agencement, positionnement, décoration, couleurs, taille du texte…).

- Bootstrap (c'est un Framework ccs et JavaScript qui permet d'offrir du code déjà structuré et organisé, il contient aussi des plugins jQuery de qualité pour enrichir la page).

- Model Canevas (Canevas est un framework comme bootstrap)

- Typescript (Typescript est un langage de programmation fortement typé qui s'appuie sur le javascript. Le code Typescript est converti en JavaScript, qui s'exécute partout où JavaScript s'exécute**).**

- NodeJs (NodeJs est un environnement d'exécution permettant d'utiliser le Javascript côté serveur. Grâce à son fonctionnement non bloquant, il permet de concevoir des applications en réseau performantes, telles qu'un serveur web, une API)

- ReactJS (ReactJS permet aux développeurs de créer de grandes applications web qui peuvent modifier les données, sans avoir à recharger la page. L'objectif principal de React est d'être rapide, évolutif et simple. Il ne fonctionnement que sur les interfaces utilisateurs de l'application)

- Local Storage (l'interface locale Storage mémorise les données sans limite de durée de vie. Contrairement à la session Storage, elles ne sont pas effacées lors de la fermeture d'un onglet ou du navigateur. Son rôle est d'enregistrer des données de manière organisée afin d'aider à les retrouver facilement plus tard et de créer un site dynamique. Local Storage est plus pratique ici parce qu'il fait partie du navigateur donc, plus facile et rapide d'accès pour notre système en streaming).

- Navigateur (Google Chrome)

- Face-api.js (Face-api.js est un module de javascript, construit sur le dessus de **tensorflow.js noyau** , qui met en œuvre plusieurs **CNN**s **(**convolutifs Neural Networks) pour résoudre la

détection du visage, la détection des points de repère faciale et la reconnaissance d'image faciale, optimisée pour le Web et pour les appareils mobiles. **Dans ce projet nous allons utiliser des modèles déjà entrainés pour la détection du repère faciale, la reconnaissance faciale, la reconnaissance d'expression faciale, l'estimation de l'âge et la reconnaissance du genre** (voir la section 4.3 pour l'architecture de notre système).

- Webcam
- Live server (Serveur Web une extension installable sur Visual Studio).

4.4.2 Environnement de travail

La configuration de notre Pc pour la simulation de ce projet est les suivantes :

- Système d'exploitation : Windows 10
- Processeur : corie 5-7200U CPU @ 2.50 – 2.7 Ghz avec 4 Cœurs
- Mémoire RAM : 16 Gb
- Mémoire Graphique : AMD Radeon™ R5 M430 - 10 Gb
- Caméra HD 720 p Logitech

4.5 Présentation de l'application web

Dans ce projet de mémoire nous avons fait un système de VSI qui est la combinaison de sept (07) fonctionnalités. Nous avons pu inclure les basiques requis pour qu'un VSI soit considéré comme intelligent pour notre simulation : ce sont la détection de repère faciale et la détection des points de repère faciale, la reconnaissance faciale, la reconnaissance d'émotion, la reconnaissance du genre, l'estimation de l'âge. (Voir la section 1.3.2)

Les avantages que présentent ses fonctionnalités basiques dépendent du domaine de son application

Cependant nos fonctionnalités actuelles nous permettent les avantages spécifiques suivants :

- **Identifier une ou plusieurs personnes en même temps ou simultanément** et savoir leurs personnalités (nom, sexe, émotion, âge estimer) ;

- **Gagner du temps** grâce aux analyses intelligentes des images vidéo (divers types de

reconnaissance intelligent) destinées à assister les agents de sécurité ou les personnels : l'humain peut se concentrer sur des tâches à valeurs ajoutées pendant que son système de vidéosurveillance intelligent trie et analyse. Elle vient assister l'opérateur humain et le rend plus performant.

- **Influencer les concernés** : la simple mise en place d'une telle installation permet de dissuader tout individu mal intentionné, les employées ou simple individu dans un local privé ou public (écarter et réduire les risques et les dangers).

4.5.1 Structure de codage du programme

Pour la réalisation de notre travail, nous avons utilisé le modèle MVC (Modèle Vue Contrôleur). Le pattern MVC permet de bien organiser son code source. Ila va nous aider à savoir quels fichiers créer, mais surtout à définir leur rôle. Le but de MVC est justement de séparer la logique du code en trois parties que l'on retrouve dans des fichiers distincts. Elle permet de créer une application web pour bien gérer la structuration d'un projet en trois parties.

- **Modèle** :

  Cette partie gère les *données* de votre site. Son rôle est d'aller récupérer les informations « brutes » dans la base de données, de les organiser et de les assembler pour qu'elles puissent ensuite être traitées par le contrôleur. On y trouve donc entre autres les requêtes SQL.
- **Vue** :

  Cette partie se concentre sur l'*affichage.* Elle ne fait presque aucun calcul et se contente de récupérer des variables pour savoir ce qu'elle doit afficher. On y trouve essentiellement du code HTML mais aussi quelques boucles et conditions PHP très simples, pour afficher par exemple une liste de messages.
- **Contrôleur** :

  Cette partie gère la logique du code qui prend des décisions. C'est en quelque sorte l'intermédiaire entre le modèle et la vue : le contrôleur va demander au modèle les données, les analyser, prendre des décisions et renvoyer le texte à afficher à la vue. [37]

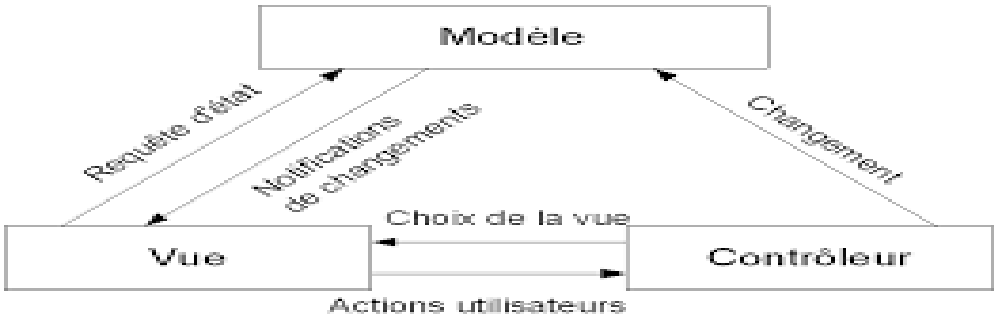


Figure 4.08 : *Architecture MVC*

4.5.2 Principe de la prise de photo pour la base de référence

Pour avoir un meilleur résultat lors de la reconnaissance faciale il est conseillé d'avoir au moins cinq images de référence pour chaque personne sous diverses poses comme l'exemple suivant (plus on a d'images de référence plus notre système est efficace) :

Figure 4.09 : *Les cinq différentes poses pour la base de référence*

Remarque : Pour augmenter encore plus la performance du système, on peut ajouter des visages avec les sept expressions faciales pour augmenter les caractéristiques discriminatoires de chaque personne.

4.5.3 Comment démarrer le projet

Pour démarrer le projet, il faut aller juste :

- Ouvrir le dossier contenant le projet dans visual studio code.
- Sélectionner le fichier « index reconnaissance facial.html ».
- Faire un clic droit et démarrer le serveur web Live server, ensuite une page web s'ouvre puis autoriser l'utilisation du Webcam dans le navigateur.
- Attendre le chargement des modèles et la construction de la base de référence d'image.
- Activer le bouton lecture et la simulation en streaming commence.

4.6 Test et évaluation du système

4.6.1 Extrait de test réaliser

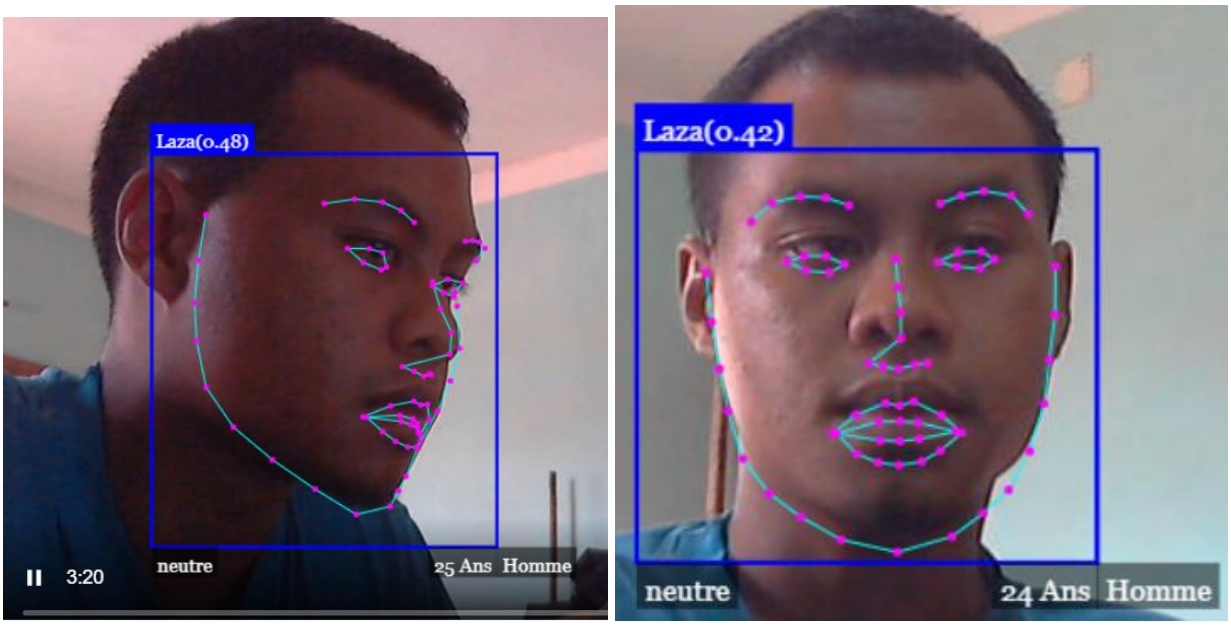


Figure 4.10 : *Reconnaissance vue de profil et vue de face avec exacte précision (âge réel 24 ans)*

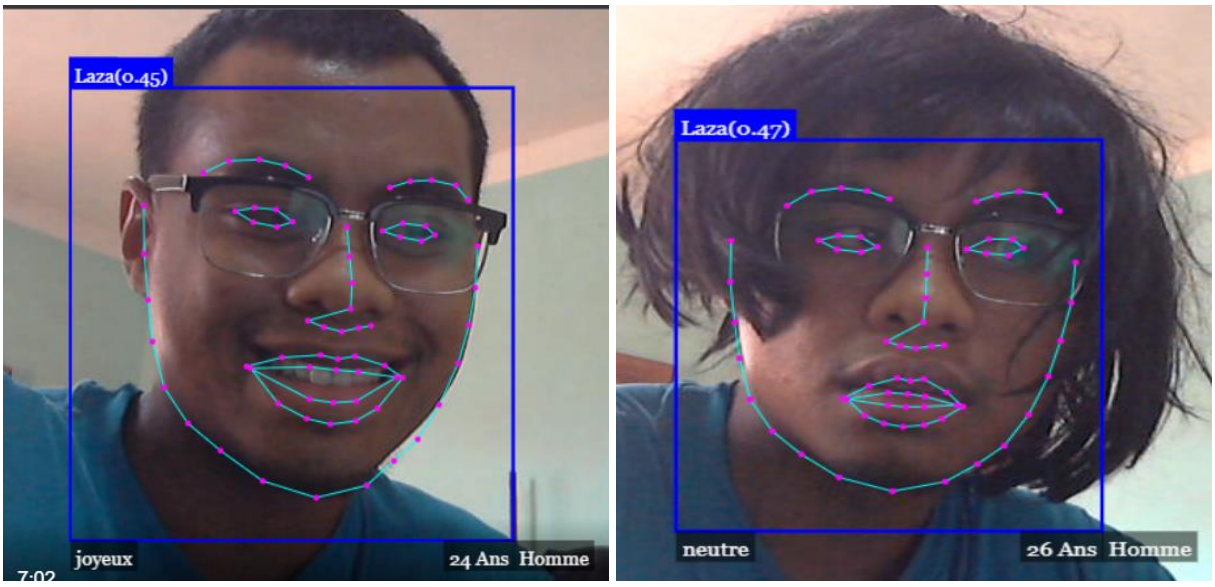


Figure 4.11 : *Reconnaissance vue de face avec lunette et faux cheveux réussis*

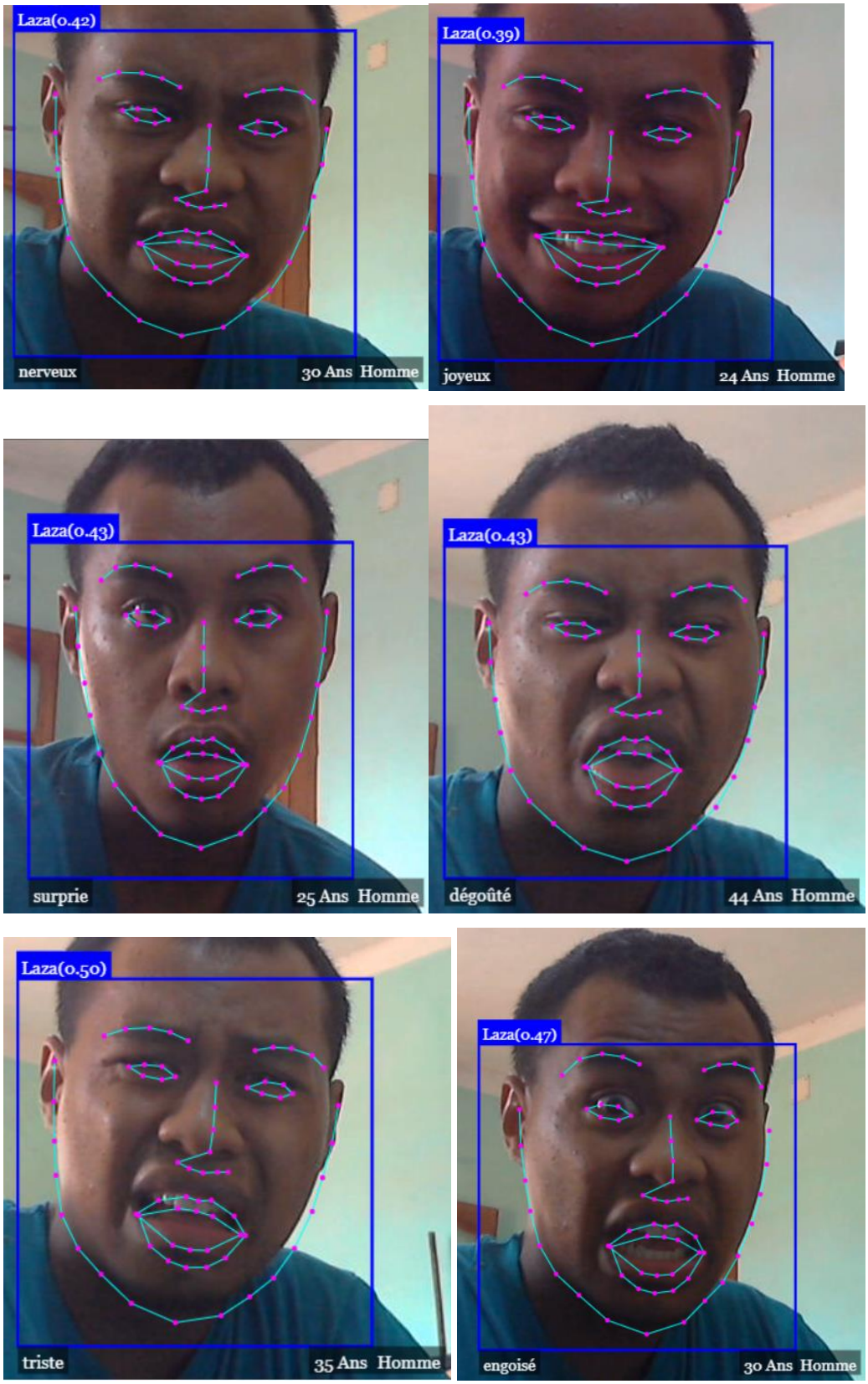


Figure 4.12 : *Reconnaissance des expressions faciales vues de face réussit*

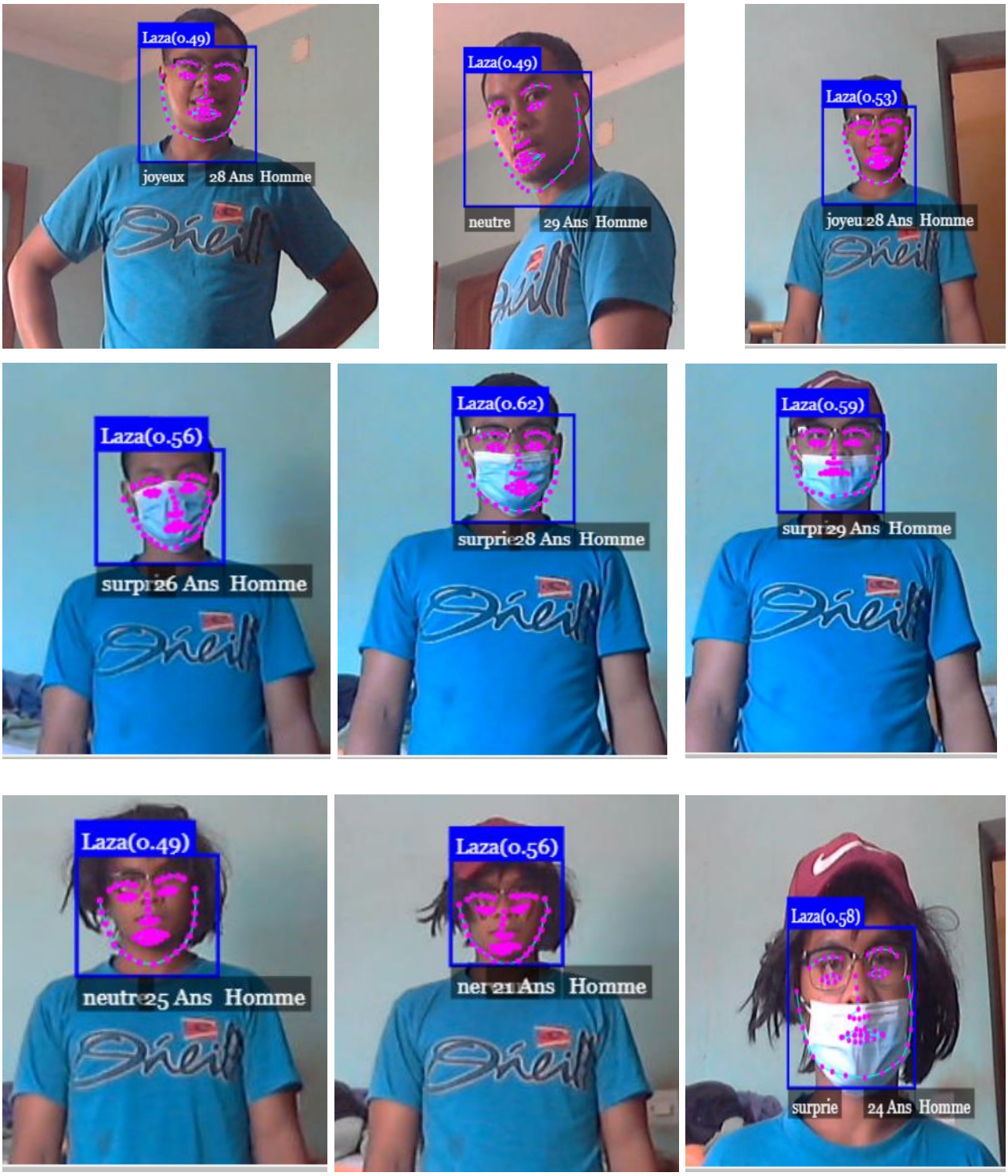


Figure 4.13 : *Reconnaissance d'image faciale avec échelle et accessoire pour déformer le visage vu de face réussie*

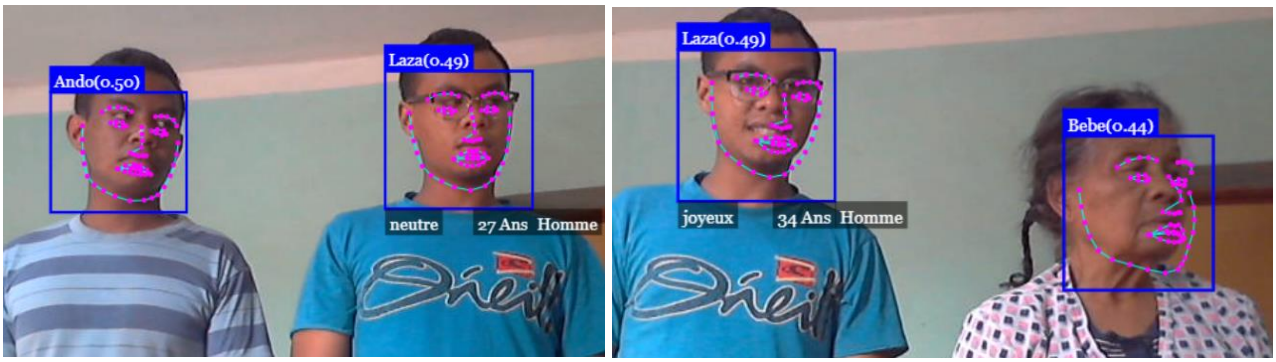


Figure 4.14 : *Reconnaissance d'image faciale de deux personnes avec échelle sous différentes vues réussies*

4.6.2 Résultats des taux de reconnaissance

La performance d'un système biométrique peut se mesurer principalement à l'aide de trois critères : sa précision, son efficacité (vitesse d'exécution) et le volume de données qui doivent être stockés pour chaque utilisateur et ces performances dépendent de plusieurs facteurs qui interviennent à plusieurs niveaux et qui peuvent limiter le degré de précision.

Cependant, il serait judicieux de s'intéresser à ces facteurs avant de mesurer la performance d'un système de reconnaissance. Nous citons ici les principaux facteurs :

- L'environnement au moment de l'acquisition.
- Les différentes positions des capteurs.
- La qualité des capteurs.
- La mauvaise interaction entre l'utilisateur et les capteurs [38]

Après plusieurs tests en streaming, on a eu les résultats suivants :

Distance entre la personne et la caméra	Reconnaissance faciale	Reconnaissance du genre	Estimation de l'âge	Reconnaissance d'émotion
Réduit (Environ 1 mètre)	99.38%	98%	70%	70%
Moyenne (Environ 3 mètres)	98%	97%	63%	60%
Long (Environ 5 mètres)	87%	85%	60%	60%
Très long (> 5 mètres)	45%	≈Aucune prédiction	≈Aucune prédiction	≈Aucune prédiction

Tableau 4.01 : *Résultats des taux de reconnaissance*

$$\text{Taux de reconnaissance} = \frac{\text{nombre de décision correctes}}{\text{le nombre de décision totale}} \; X \; 100 \qquad (4.01) \; [38]$$

4.6.3 Limites et atouts

Les pourcentages de réussites de chaque test sont très variables et dépendent de plusieurs éléments, comme : la distance, la luminosité, la netteté de l'image, la chrominance, les variations de poses et des expressions faciales.

Limite :

- Les contractions et les expressions du visage font varier de manière significative l'estimation de l'âge du système

- Le porté de la reconnaissance faciale est limité à environ 5 mètres de la position de la caméra

- Les expressions faciales de l'angoisse, de la tristesse et du dégoût s'avèrent difficiles à discriminer pour le système : du fait que les caractéristiques faciales de ces émotions se rapprochent.

- L'expression énerver est moyennement reconnue par le système, mais souvent confondue par l'expression de la tristesse et du dégoût.

- Il faut savoir que l'expression faciale et l'émotion sont différents (mais dans ce travail les deux sont utilisés l'une comme l'autre pour traiter l'expression faciale). L'expression faciale est l'ensemble des traits caractéristiques visibles sur le visage humain qui reflète l'émotion interne d'une personne, donc trucable. Alors que l'émotion interne est la sensation réelle de la personne. L'expression faciale fait partie seulement d'un des canaux de la communication de l'émotion humaine.

Atout :

- L'expression neutre, heureux et surpris sont facilement reconnue par le système.

- La reconnaissance faciale est très robuste : malgré les déformations faciales dues aux expressions faciales et les accessoires (lunette, cache bouche, casquette, faux cheveux) le système arrive à prédire presque toujours la bonne réponse de prédiction.

- À courte distance la prédiction d'émotion et d'estimation de l'âge sont assez efficaces

- Le système est rapide dans sa totalité et peut être qualifié de fiable

4.7 Conclusion

Dans ce chapitre, nous avons vu l'architecture générale d'un système de reconnaissance d'image faciale qui est composer des étapes suivantes : acquisition du visage, extraction des caractéristiques du visage et la classification d'image faciale. Ensuite, nous avons présenté d'architecture que nous avons adaptée pour la réalisation de ce travail. C'est un système hybride basé sur l'apprentissage profond et le transfert learning utilisant des réseaux de neurones à convolutions CNN et CDN ou Convolutional Deep Neuronal pour l'extraction des vecteurs caractéristique du visage. Chaque bloc du traitement correspond à un modèle spécifique et sont disposés en série pour travailler ensemble. Ainsi, la sortie d'un bloc devient l'entrée d'un autre. De cette façon le système sera efficace et rapide. Pour résoudre les problèmes de classification et de régression, les modèles utilisés font appel au classificateur KNN et SVM : qui nous permet de faire la reconnaissance faciale, la reconnaissance d'expression faciale, l'estimation de l'âge et la reconnaissance du genre. À la fin de ce chapitre, nous avons fait des tests d'évaluations de notre travail. Grâce à cette architecture nous avons pu faire un système intelligent qualifiable et performant pour la vidéo surveillance.

CONCLUSION GENERALE

Dans ce travail, notre objectif consiste à étudier la mise en place d'un système intelligent pour la vidéo surveillance. Nous avons, ainsi, fait des études approfondies sur le traitement automatique des images. Afin d'assurer au système de reconnaissance automatique d'image vidéo une bonne performance et une robustesse, les modèles que nous avons utilisés sont tous basés sur l'apprentissage profond. Nous avons utilisé un architecture hybride composé de plusieurs modèle pré-entrainer et efficace dans chacun de leurs domaines. Ces derniers ont prouvé leurs efficacités grâce à leurs architectures basées sur les réseaux de CNN et de DCN.

Nous avons, ainsi, réaliser notre système intelligent pour la vidéo surveillance. D'abord, la reconnaissance faciale permet d'identifier chaque personne dans la zone de couverture de la caméra. Ce qui est très important dans le domaine de la sécurité IT. Ensuit, connaitre l'expression faciale de chaque personne visible par la caméra permet de nous rassurer sur l'état d'esprit de chacun de ces personnes présentes. L'estimation de l'âge quant à lui nous permet de reconnaitre l'âge des différentes personnes qui fréquente nos lieux. Et bien sûre savoir le sexe de ces personnes nous aides encore plus à s'informer.

L'application que nous avons développée se présente être efficace et performant dans la reconnaissance faciale et du genre d'une part et d'autres part ayant quand même des limites sur certain de ces fonctionnalités. D'après les tests effectuer les pourcentages de réussite de chaque prédiction sont très variables et dépends de plusieurs éléments, comme : la distance, la luminosité, la netteté de l'image vidéo, la chrominance, les variations de poses et des expressions faciale et la qualité du capteur vidéo.

Evidemment l'application que nous avons créée peut-être largement améliorée et ajoutée de nouvelle fonctionnalité. Les recherche dans ce domaine ne cesse de se développer et il est possible d'étendre l'application à des nouvelles fonctionnalité, comme : la détection d'objet et la détection d'activité humaine et encore d'autres.

En gros, nous avons conçu une application permettant de faire la reconnaissance faciale, la reconnaissance d'expression faciale, la reconnaissance du genre et l'estimation de l'âge qui sont les bases principales d'un système de vidéosurveillance.

A la fin de ce travail, je peux dire que j'ai bien pu avoir une visibilité concrète sur un domaine bien spécifique qui est le traitement automatique de l'image. Cette partie importante de l'intelligence

artificielle. Ce travail m'a été profitable en termes d'acquérir un bénéfice intellectuel. Sa réalisation m'a apporté plusieurs connaissances sur le traitement automatique en générale et surtout sur le traitement automatique d'image, le domaine de la programmation dynamique et de la gestion des travails. Malgré les difficultés, je n'en ai tiré que des profits.

ANNEXE 1

EXTRAITS DE CODE SOURCE

A.1.1 Chargement des modèles et activations de la caméra

```
const video = document.getElementById("videoInput");
let predictedAges = [];
Promise.all([
        faceapi.nets.tinyFaceDetector.loadFromUri("../public/models"),
        faceapi.nets.faceLandmark68Net.loadFromUri("../public/models"),
        faceapi.nets.ssdMobilenetv1.loadFromUri("/public/models"),
        faceapi.nets.faceRecognitionNet.loadFromUri("../public/models"),
        faceapi.nets.faceExpressionNet.loadFromUri("/public/models"),
        faceapi.nets.ageGenderNet.loadFromUri("/public/models"),
  ]).then(startVideo);
function startVideo() {
        document.body.append('Models Loaded');
        navigator.getUserMedia(
        { video: {} },
         stream => (video.srcObject = stream),
         err => console.error(err)
         );
        console.log(" video Ajouter -> En marche !");
        recognizeFaces();
}
```

A.1.2 Paramètre seuil de la reconnaissance faciale

```
const labeledDescriptors = await loadLabeledImages();
console.log(labeledDescriptors);
// Seuille du distance euclidienne 0.65 distance euclidienne maximale
  const faceMatcher = new faceapi.FaceMatcher(labeledDescriptors, 0.65);
  console.log(faceMatcher);
```

A.1.3 Parties du code qui traite la reconnaissance faciale

```
const detections = await faceapi
        .detectAllFaces(video, new faceapi.TinyFaceDetectorOptions())
        .withFaceLandmarks()
        .withFaceDescriptors()
        .withFaceExpressions()
        .withAgeAndGender();
const resizedDetections = faceapi.resizeResults(detections, displaySize);
canvas.getContext("2d").clearRect(0, 0, canvas.width, canvas.height);
faceapi.draw.drawDetections(canvas, resizedDetections);
faceapi.draw.drawFaceLandmarks(canvas, resizedDetections);
// Code pour la reconnaissance faciale
const results = resizedDetections.map((d) => {
return faceMatcher.findBestMatch(d.descriptor);
});
// Fonction qui traite chaque face détecter pour la reconnaissance faciale
results.forEach( (result1, i) => {
const box = resizedDetections[i].detection.box;
      var verification =   results[i];
      var result = verification.label;
      var dist = verification.distance;
      var distanceToString = dist.toString();
      var distance = distanceToString.substring( 0,4) ;
      console.log( distance );
```

```
if(result != 'unknown' ){
        var resultat = result;
        console.log(typeof(resultat) );
    }else{
        var resultat = 'Inconnue' ;
        console.log( "objet contenant :" + resultat);
    };
```

A.1.4 chargement des vecteurs caractéristiques de la base de référence

```
function loadLabeledImages() {
        const labels =[ 'Neny','Laza','Bebe','Dada','Fanasina', 'Ando'];
        return Promise.all(
                labels.map(async (label)=>{
                const descriptions = [];
                // i nombre d'image de référence de chaque personne dans la base
                for(leti=1;i<=5;i++){const img = await
                faceapi.fetchImage(`../public/labeled_images/${label}/${i}.jpg`);
                const detections = await faceapi.detectSingleFace(img)
                .withFaceLandmarks()
                .withFaceDescriptor();
                console.log(label + i + JSON.stringify(detections));
                descriptions.push(detections.descriptor);
                console.log("Label Image Etudier : " + label)
        }
        // retourne le nombre de visage traiter
        document.body.append(label+' Faces Loaded | ');
        return new faceapi.LabeledFaceDescriptors(label, descriptions);
    })
  );
}
```

BIBLIOGRAPHIE

[1] Y.Asma, L.Abderrahim, « *Mise au point d'une application de télésurveillance* »,Université Abou Berk Belkaid, 18 juin 2017

[2] N.Boutadara, « *Proposition d'une approche intelligente pour la reconnaissance d'actions humaines à partir d'image de vidéosurveillance* », Université Ahmed Draia – Adrar, 2016

[3] S.Angot, « *Conception d'un système de vidéosurveillance intelligente pour l'IMT* », Central Marseille, 2022

[4] L.Beddiaf, « *Vidéosurveillance principe et technologies* », L'usine nouvelle, 2.019

[5] J. Ah-Pine, *« Apprentissage automatique »*, 2019-2020

[6] A. Géron, *« Machine Learning »*, Dunod : Paris, 2017.

[7] I. Teller, *« Apprentissage automatique pour le TAL »*, HAL, 02 Septembre 2010

[8] M. Taffar, *« Initiation à l'apprentissage automatique »*, Cours Master, Ment Informatique-Faculté des Sciences Exactes et de l'Informatique, 2018

[9] V. Bisson, *« Algorithme d'apprentissage pour la recommandation »*, Septembre 2012

[10] J. Ah-Pine, *« Méthodes avancées en apprentissage supervisé et non supervisé »*, Mars 2019

[11] A. Ravel, *« Apprentissage semi-supervisé »*, 25 Mai 2016

[12] J. B. Metomo, *« Machine Learning : Introduction à l'apprentissage automatique »*, https://www.supinfo.com/articles/single/ ,10 octobre 2017.

[13] I. Bellin, *« Apprentissage par renforcement »*, https://dataanalyticspost.com/, Février 2019.

[14] Y. LeCun, *« Les Enjeux de la Recherche en Intelligence Artificielle »*, Chaire Informatique et Sciences Numériques Collège de France, AU : 2015-2016

[15] C.Migneault,E.Granger *« Évaluation de méthodes de reconnaissance de visages pour l'identification d'individus à partir d'une image de référence »,* ResearchGate., Août 2016

[16] M. Zaffagni, *« Cette IA a appris à conduire une voiture autonome en 20 minutes »*, https://www.futura-sciences.com/tech/actualites/intelligence-artificielle-cette-ia-apprisconduire-voiture-autonome-20-minutes-71942/, 09 Juillet 2018.

[17] R.J.Manolo « *Conception d'un système de recherche par reconnaissance d'image et application dans un site-commerce »*, ESPA, 15 mai 2019

[18] B.Thomas, *« Réseau de neurones : définition et fonctionnement », https://datascientest.com/fonctionnement-des-reseauxneurones#rappel reseauxneurones/,* 15 juin 2020

[19] P.Guillaume, *« Algorithme de descente de gradient »,* https://datascientest.com /descente-de-gradient, 20 juillet 2020

[20] B.Gary, *«Convolutional neural network»,* https://datascientest.com/ convolutional-neural-network, 25 juin 2020

[21] *« Les réseaux de neurones convolutifs »*, https://www.natural-solutions. eu/blog/la-reconnaissance-dimage-avec-les-rseaux-de-neurones-convolutifs /Qu'est-ce-que-a-la-reconnaissance-d'image, 17 Avril 2018

[22] Quantblog, *« MobileNet, optimisation de la convolution pour les réseaux de neurones embarqués »*, https://www.quantmetry.com/blog/mobilenet-optimisation-de-la-convolution -pour-les-reseaux-de-neurones-embarques/, 03 Mars 2019

[23] ICHI.PRO, *« Comprendre les convolutions séparables en profondeur et l'efficacité des réseaux mobiles »* https://ichi.pro/fr/comprendre-les-convolutions-separables-en-profondeur-et-l-efficacite-des-reseaux-mobiles-120825327265027, 2022

[24] A.Lima, *« Réseaux de neurones convolutifs séparables en profondeur »* https://fr.acervolima.com/reseaux-de-neurones-convolutifs-separables-en-profondeur/, 2021

[25] ICHI.PRO, *« Une introduction aux convolutions séparables avec revue de la littérature »,* https://ichi.pro/fr/une-introduction-aux-convolutions-separables -avec-revue-de-la-litterature-226958156990252, 2021

[26] Y. Benzaki, *« Logistic Regression pour Machine Learning – Une Introduction Simple »*, https://mrmint.fr/logistic-regression-machine-learning-introduction-simple, 6 septembre 2017

[27] C.Yohan, *« Qu'est-ce que l'algorithme KNN ? »*, https://datascientest. com/knn, 19 novembre 2020

[28] P.Guillaume, *« Entraînez votre premier k-NN »*, https://openclassrooms.com/fr/courses/4011851-initiez-vous-au-machine-learning/4022441-entrainez-votre-premier-k-nn,

[29] Y.Benzaki, *« Introduction à l'algorithme K Nearst Neighbors (K-NN) »*, https://mrmint.fr/ introduction-k-nearest-neighbors, 2 Octobre 2018

[30] Équipe Data Science, *« Comprendre la méthode des "k-plus proches voisins" en 5 min »*, https://blog.ysance.com/algorithme-n5-comprendre-la-methode-des-k-plus-proches-voisins-en-5-min, 21 Septembre 2020

[31] K.Harifi, *« Bien comprendre l'algorithme des K-plus proches voisins (Fonctionnement et implémentation sur R et Python) »*, https://medium.com/@kenzaharifi/bien-comprendre-lalgorithme-des-k-plus-proches-voisins-fonctionnement-et-implémentation-sur-r-et-a66d2d372679, 21 Septembre 2019

[32] IBM, *« Fonctionnement de SVM »*, https://www.ibm.com/docs/fr/spss-modeler/SaaS ?topic=models-how-svm-works, 17 Août 2021

[33] S.Gadat, *« Algorithmes de Support Vector Machines »*, UMR CNRS-UPS, 2022

[34] C.Migneault, E.Granger *« Évaluation de méthodes de reconnaissance de visages pour l'identification d'individus à partir d'une image de référence »*, ResearchGate, Août 2016

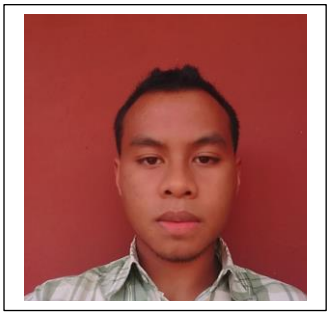
[35] K.Lekdioui, *« Reconnaissance d'états émotionnels par analyse visuelle du visage et apprentissage machine »*, HAL, 23 Mars 2019

[36] Kevin, *« face-api.js »*, https://justadudewhohacks.github.io/face-api.js/docs/index.html, 2022

[37] Openclassrooms, *« Comment fonctionne une architecture MVC ? »*, https://openclassrooms.com/fr/ courses/4670706-adoptez-une-architecture-mvc-en-php/4678736-comment fonctionne-une-architecture-mvc, 19 Octobre 2021

[38] A.Chiheb *« Performances d'un système de reconnaissances de visage »* https://www.memoireonline.com/02/13/6979/m_Reconnaissance-de-visages -par -Analyse-Discriminante-LineaireLDA-10.html, 2018

FICHE DE RENSEIGNEMENTS



Nom : RAKOTONARIVO

Prénom : Laza Manampisoa

Adresse de l'auteur : Lot AKT IE 10 Andohony - Antanety II Vontovorona

Antananarivo 102 – Madagascar

Tel : +261 34 94 143 95

E-mail : lazamanampisoa.rkt@gmail.com

Titre du mémoire :

« *CONCEPTION D'UN SYSTEME INTELLIGENT POUR LA VIDEO SURVEILLANCE* »

Nombre de pages : 85

Nombre de tableaux : 01

Nombre de figures : 48

Directeur de mémoire :

Nom : RAVONIMANANTSOA

Prénoms : Ndaohialy Manda-Vy

Tel : +261 34 11 358 00

Email : ravonimanantso@gmail.com

RÉSUMÉ

La reconnaissance automatique d'image vidéo est une technologie informatique permettant à un logiciel ou une machine d'interpréter les images vidéo. Elle permet à une machine d'extraire les informations que renferme un flux vidéo et de les interpréter. Dans ce travail un système intelligent pour la vidéo surveillance a été conçue. Grâce à la reconnaissance automatique d'image vidéo basée sur l'apprentissage profond et le transfert learning ou apprentissage par transfert. Ce système utilise sur une méthode hybride. Il consiste à faire la reconnaissance faciale, la reconnaissance d'expression faciale, la reconnaissance du genre et l'estimation de l'âge. En occurrence, la performance de notre système varie autour de 85%.

Mots clés : Apprentissage automatique, Intelligence artificielle, Image et vidéo numérique, traitement d'image, vidéo surveillance.

ABSTRACT

Automatic video image recognition is a computer technology that allows software or a machine to interpret video images. It allows a machine to extract the information contained in a video stream and to interpret it. In this work an intelligent system for video surveillance was designed. Thanks to automatic video image recognition based on deep learning and transfer learning. This system uses on a hybrid method. It consists of doing facial recognition, facial expression recognition, gender recognition and age estimation. In this case, the performance of our system varies around 85%.

Keywords : machine learning, artificial intelligence, digital image and video, image processing, video surveillance.