Khadraoui Djihene
Master 2 - Image et Vie Artificielle

9 juillet 2019

Table des matières

Table des figures

Liste des abréviations

SVM	Support Victor Machine
PDM	Points Distribution Model
HSV	Hue, Saturation, Value
RGB	Red, Green, Blue
Camera CCD	Charge Coupled Device Camera
FACS	Facial Action Coding System
MPEG-4	Motion Picture Experts Group Layer-4
AU	Units Actions
FDP	Facial Displacement Points
FAP	Facial Animation Parameters
FAPU	Facial Animation Parameters Units
HMM	Hidden Markov Model
NN	Neural Network
KNN	K-Nearest Neighbors
EEG	Electroencephalography
SNW	Sparse Network of Winnos
ENIAC	Electronic Numerical Integrator And Computer
OCR	Optical Character Recognition
EAR	Eye Aspect Ratio
MAR	Mouth Aspect Ratio

Introduction générale

Les humains ont inventé le langage, et ils en sont bien fiers. Ils parlent tant qu'ils ont fini par délaisser des méthodes de communication plus fondamentales comme : le toucher, les gestes, le contact visuel et les expressions faciales. Ces modes d'expression fondamentaux peuvent transmettre plus d'information, plus rapidement que la parole, mais ils peuvent aussi exprimer des messages qu'on ne peut communiquer par le langage. Même si personne ne peut le voir, le bébé transmet continuellement ses émotions par des expressions faciales et des gestes, les adultes font de même, et leurs visages se modifient juste à penser à quelque chose, même si personne n'en est témoin. On peut définir l'expression faciale comme un signe visible sur le visage qui indique ce que ressent une personne, l'expression du visage peut ainsi montrer la joie, la tristesse, la douleur, la fatigue, etc.

L'exploitation des algorithmes issus du domaine de la vision artificielle pour reconnaitre les différentes expressions de visage attire de plus en plus l'intention des spécialistes dans le domaine, et aussi le grand public. Ce genre de processus peut s'avérer très utile pour la sécurité, la médecine, la communication, et l'éducation, etc. Nous proposons ainsi un système basé sur des techniques de la vision artificielle permettant de reconnaitre certaines expressions faciales dans le but de détecter un éventuel état de fatigue chez les conducteurs.

En effet, face à la nécessité croissante d'utiliser les moyens de transport et aux accidents de la route grandissants pour de nombreuses raisons, notamment : vitesse excessive, somnolence ou fatigue, il est préférable de doter chaque conducteur des équipements nécessaires pour éviter de tels accidents. L'une des possibilités est de concevoir des systèmes peuvent alerter le conducteur et surveiller son niveau de vigilance pour le tenir éveillés, et par conséquent réduire le nombre d'accidents de la route. Ainsi, nous proposons un système basé sur des techniques de la vision artificielle permettant de reconnaitre certaines expressions faciales dans le but de détecter un éventuel état de fatigue chez les conducteurs.

Ce mémoire est donc organisé comme suit : le premier chapitre portera sur une révue de littérature qui donne un aperçu général sur les expressions faciales, les principales approches utilisées pour la reconnaissance des expressions faciales, et celles utilisée dans le cadre de détection de fatigue. Le deuxième chapitre introduit le concept des réseaux de neurones, l'apprentissage, qui seront utilisé dans notre système proposé.Le troisième chapitre décrit la conception du système proposé. Le dernier chapitre donne les détails d'implémentation de notre système.

Chapitre 1

Révue de littérature

Introduction

Le visage humain est impliqué dans une variété impressionnante d'activités différentes, puisqu'il contient la majorité de notre appareil sensoriel : yeux, oreilles, bouche, et nez, nous permettant de voir, écouter, goûter et sentir. À part ces fonctions biologiques, le visage humain fournit un nombre de signaux essentiels pour la communication interpersonnelle dans notre vie sociale. Le visage comporte plusieurs systèmes qui coopèrent pour produire un ensemble de signaux de communication comme : la parole, le regard, le positionnement et les mouvements de la tête, les expressions faciales, etc. Ces signaux étant produits essentiellement par le système de production de la parole et le système musculaire facial. Ils sont primordiaux pour déduire l'état affectif et les intentions d'une personne. D'autres informations comme, l'attraction, l'âge et le genre peuvent être aussi dérivées á partir du visage d'une personne. Automatiser les analyses des signaux faciaux, et en particulier les signaux faciaux rapides (actions des muscles faciaux), devrait être fortement bénéfique pour plusieurs champs d'intérêt comme la sécurité, la médecine, la communication, et l'éducation. Dans le contexte de la sécurité, les expressions faciales jouent un rôle crucial dans l'établissement ou l'évaluation de la crédibilité. En médecine, les expressions faciales fournissent une signification directe permettant d'identifier les processus mentaux spécifiques, par exemple, quand une personne est en état de sommeil. En éducation, les pupilles des auditeurs informent le professeur de la nécessité d'ajuster le message d'instruction. Les interfaces normales entre les êtres humains et les ordinateurs (ordinateurs personnels/ robots/ machines) sont aussi concernés, les expressions faciales fournissent une possibilité pour communiquer des informations de base sur des besoins et des demandes à une machine. [1]

En fait, les analyses automatiques des signaux faciaux rapides semblent avoir une place naturelle dans divers sous-ensembles de systèmes de vision, incluant les outils automatisés pour le regard et reliés aussi à la lecture des lèvres, traitement de la parole bimodal, visage / synthèse de la parole visuelle, et le traitement facial. Certains signaux faciaux (clignement de l'œil) peuvent être aussi associés à certaines commandes (clic de souris) offrant une alternative à des commandes du clavier et souris traditionnelles. Les possibilités humaines à entendre dans les environnements bruyants au moyen de la lecture sur les lèvres sont la base du traitement de la parole bimodal qui peut mener à la réalisation des interfaces de discours robustes. La capacité humaine pour lire les émotions à partir des expressions faciales de quelqu'un est la base de l'analyse des messages faciaux pouvant mener au développement d'interfaces d'extension avec la communication émotive et, alternativement, à obtenir une interaction plus flexible, plus adaptable, et normale entre les hommes et les machines. Bien que les êtres humains soient parfaitement capables d'estimer l'état affectif d'une personne à partir d'une image statique, il y a assurément plus d'informations sur le comportement facial contenu dans des observations dynamiques de la séquence vidéo d'un visage humain. Par conséquent, nous présentons plusieurs approches à travers ce chapitre qui visent à tirer bénéfice de cette information additionnelle.

1.1 Anatomie du visage humain

1.1.1 Muscles faciaux

Les muscles faciaux entrant dans la production des expressions sont nombreux. Ils ont certaines particularités , il semble par exemple qu'ils soient organisés en réseaux fibrés, ce qui permettrait de mettre en place un système de rétroaction. En effet, de telles fibres permettent au système nerveux central d'avoir connaissance de l'ètat actuel d'activation du muscle, même si le muscle a été activée involontairement. Ce mécanisme pourrait expliquer la manière dont sont " ressenties " les émotions. [2]

De plus, contrairement aux autres muscles liés au squelette, les muscles faciaux n'ont pas de muscles antagonistes correspondants. En effet, chaque muscle squelettique ne peut se contracter que si son muscle antagoniste est relâché. Les muscles faciaux reviennent à une position de repos quand il n'y a plus d'activation, puisque l'ensemble des tissus du visage oppose une réaction au mouvement. [2]

L'enchevêtrement des muscles au niveau du visage permet une grande mobilité. Les points d'attache des muscles aux os du crâne sont relativement peu nombreux et beaucoup sont liés directement les uns aux autres [2], voir la figure 1.1.

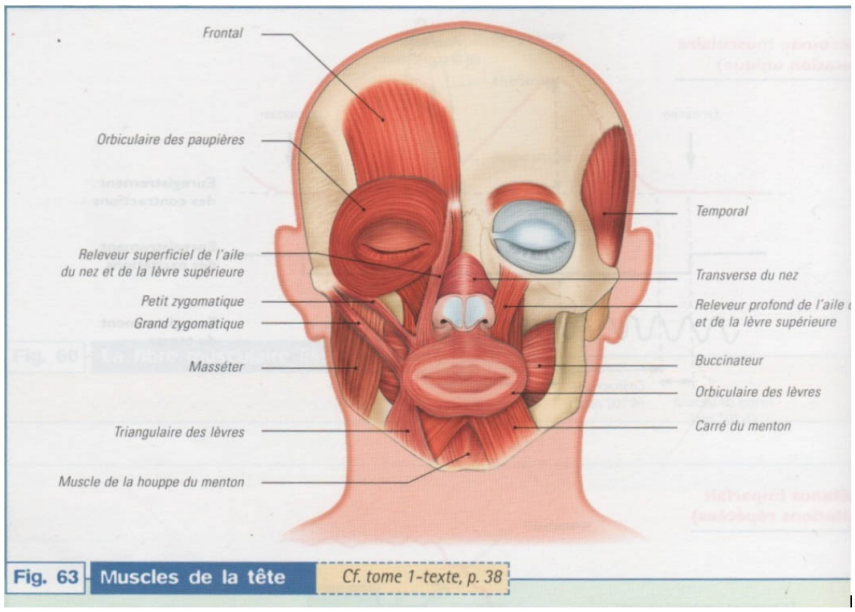


FIGURE 1.1 – Points d'attache des muscles faciaux (extrait de [2])

1.1.2 Informations portés par le visage

Le visage est une source d'information importante. On distingue principalement deux informations portées par le visage : l'identité et les expressions.

L'identité est déterminée quasi intégralement par la forme et la position des os du crâne. Ces caractéristiques, uniques pour chaque individu permettent de le distinguer des autres. Ainsi, il est possible d'avoir une bonne idée de l'identité d'un individu uniquement à partir de ses os. Cependant, l'identité à un instant donné est aussi déterminée par des caractéristiques comme la couleur et texture de la peau et la pilosité.

L'apparence statique d'un visage porte aussi des informations sur l'âge, le sexe et l'origine ethnique. Lâge peut modifier la texture des muscles, de la peau et modifier l'épaisseur de la couche graisseuse. La forme de la mâchoire, de l'arcade sourcilière et la protubérance occipitale diffèrent généralement entre l'homme et la femme. La couleur de la peau, déterminée par

la quantité et la nature des mélanines de la peau, la présence de carotène alimentaire et l'hémoglobine peut donner une information sur le sexe (les femmes ont un teint légèrement plus clair que les hommes). La couleur, texture de la peau et structure de la couche graisseuse sont des indices sur l'origine ethnique.

Les expressions sont déterminées par l'activation des muscles faciaux. De plus, certains processus émotionnels peuvent faire changer localement la couleur de la peau, en la colorant localement par un afflux sanguin plus important qu'à l'accoutumée. [2]

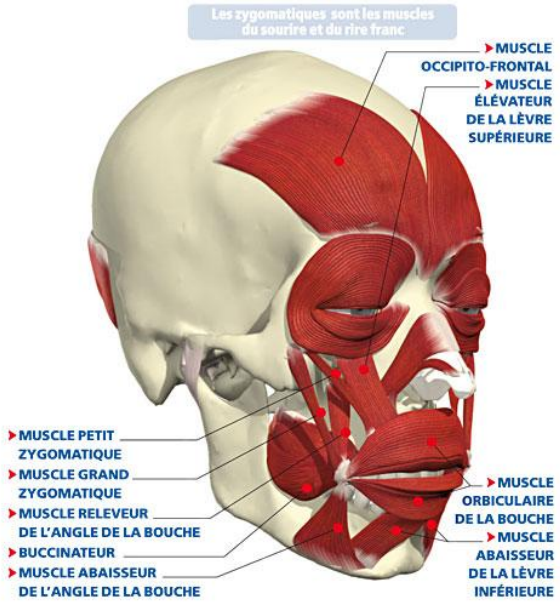


FIGURE 1.2 – Muscles faciaux (extrait de [2])

1.2 Analyse de visage

L'analyse du visage est une discipline dont les premiers travaux sont liés à l'essor de l'Intelligence Artificielle, discipline phare de l'informatique des années 1960. En effet, les premiers travaux sur l'analyse automatique du visage remontent aux travaux de SAKAI, NAGAO et FUJIBAYASHI [3] qui proposent un système permettant de détecter l'existence ou l'absence d'un visage dans une image. Suivent les travaux de KELLY [4] qui, à partir de trois images de chaque individu (une image du corps, une image de l'arrière-plan et une image du visage), propose une extraction des contours de la tête et une localisation des yeux, du nez et de la bouche. En 1973, TAKEO KANADE [5] présente un système de reconnaissance automatique de visage, basé sur une seule image. Dans ce qui suit, nous allons introduire les différents concepts liés à l'analyse de visage.

1.2.1 Détection de visage

La détection de visage consiste à déterminer la présence ou l'absence de visages dans une image. C'est une tâche préliminaire nécessaire à la plupart des techniques d'analyse du visage. Les techniques utilisées sont généralement issues du domaine de la reconnaissance des formes. En effet, le problème peut être vu comme la détection de caractéristiques communes à l'ensemble des visages humains : il s'agit de comparer une image à un modéle générique de visage et d'indiquer s'il y a ou non ressemblance. Les principales difficultés sont la robustesse aux différentes identités, poses du visage, expressions faciales et aux variations d'illumination. La sortie d'un détecteur de visage indique le nombre de visages présents dans l'image. De plus,

la plupart des détecteurs de visage actuels sont aussi des localisateurs de visages : ils renvoient une localisation des visages détectés (une boîte englobante par exemple).

1.2.2 Reconnaissance de visage

La reconnaissance de visage consiste á associer une identité á un visage après l'avoir détecté. On rencontre deux cas différents : l'identification où il s'agit de trouver dans une base de données de visages, le visage le plus ressemblant á celui étudie et l'authentification où il s'agit de vérifier que le visage étudié a bien l'identité qu'il prétend posséder. Ce deuxième cas relevant davantage de la biométrie, nous nous intéresserons plus particulièrement au premier cas.

Les systèmes d'identification de visages possèdent une base de données sur laquelle est effectuée un apprentissage. Cette base définit les différentes identités connues du système. Une nouvelle image est présentée au système et le but est de décider à quelle identité connue appartient ce visage ou s'il ne s'agit d'aucun des visages connus.

Les traitements d'un système d'identification de visages peuvent être séparés en deux étages distincts : une première pour trouver une représentation du visage qui permette de regrouper les visages de la même identité et de discriminer les différentes identités et une étape pour trouver la classe la plus vraisemblable, lors de la présentation d'un nouveau visage. Les principales difficultés d'un système de reconnaissance de visage sont la robustesse aux changements d'expressions, de pose, d'illumination, ainsi qu'aux changements morphologiques dus à l'âge et ceux dus à la présence d'artefacts visuels comme des lunettes.

La présentation du visage est basée sur l'extraction des caractéristiques peut consister en un ensemble de mesures discriminantes de l'identité : le visage est segmenté en composantes (nez, bouche, yeux, etc.) et certaines propriétés locales sont extraites.

1.2.3 Reconnaissance d'émotion

La reconnaissance d'émotions consiste à associer une émotion à une image de visage. Le but est donc de déterminer, d'aprés son visage, l'état émotionnel interne de la personne. L'ensemble considéré des émotions affichables par un visage est généralement de petite taille : il s'agit de l'ensemble des émotions universelles présenté par EKMAN [6].

Il s'agit d'un problème du même ordre que la reconnaissance de visage : le visage en entrée doit être classé parmi un ensemble fini de classes représentant les émotions. Cependant, ici les caractéristiques extraites doivent être indépendantes de l'identité (et de la pose, illumination, etc.) Les techniques utilisées sont donc très proches de celles utilisées pour la reconnaissance de visage, seules vont être changées les composantes faciales à prendre en compte pour la représentation d'un visage.

1.3 Méthodes d'analyse de visage

Il existe des différentes modélisations informatiques du visage utilisées dans la littérature. Il peut s'agir de modèles qui servent aussi bien à l'analyse de l'identité qu'à l'analyse des expressions. Nous avons distingué deux types de méthodes : celles basées sur une segmentation explicite du visage en composantes et une description des caractéristiques de ces composantes faciales et celles basées directement sur des caractéristiques de l'image dans sa globalité. De plus, on présent en tant que méthodes hybrides.

1.3.1 Approche globale sans segmentation

La modélisation la plus simple du visage, consiste à prendre en compte un ensemble de points du visage représentant l'état de certaines composantes. Ces points doivent correspondre

à des indices visuels qu'il est possible de mettre en correspondance sur toutes les observations de l'étude. Les points à analyser sont différents quand il s'agit d'analyser l'identité de quand il s'agit d'analyser l'expression.

Un visage peut âtre caractérisé par les coordonnées de chacun des points du modèle ainsi que par la valeur des pixels en leur voisinage, permettant de définir un descripteur de visage plus puissant que l'image brute. Certaines méthodes considèrent un traitement particulier en chacun des points d'intérêt : le résultat du traitement en chacun des points formant le vecteur d'entrée du système d'analyse [2]. Par exemple, la transformée par ondelettes de Gabor peut être utilisée sur une grille de points aussi bien pour la reconnaissance d'identité [7] que pour la reconnaissance d'émotions [8]. L'ensemble des réponses des filtres de Gabor forme un vecteur d'entrée à un système de reconnaissance. Dans [9], ces vecteurs d'entrée sont présentés à un réseau de neurones pour la reconnaissance d'émotions. Après un ensemble d'expériences sur les points à choisir et sur les paramètres du réseau de neurones, l'auteur conclut sur l'importance de chacun des points choisis : les points les plus importants pour la tâche de reconnaissance des émotions sont les points autour des yeux, de la bouche, des sourcils et du menton. [2]

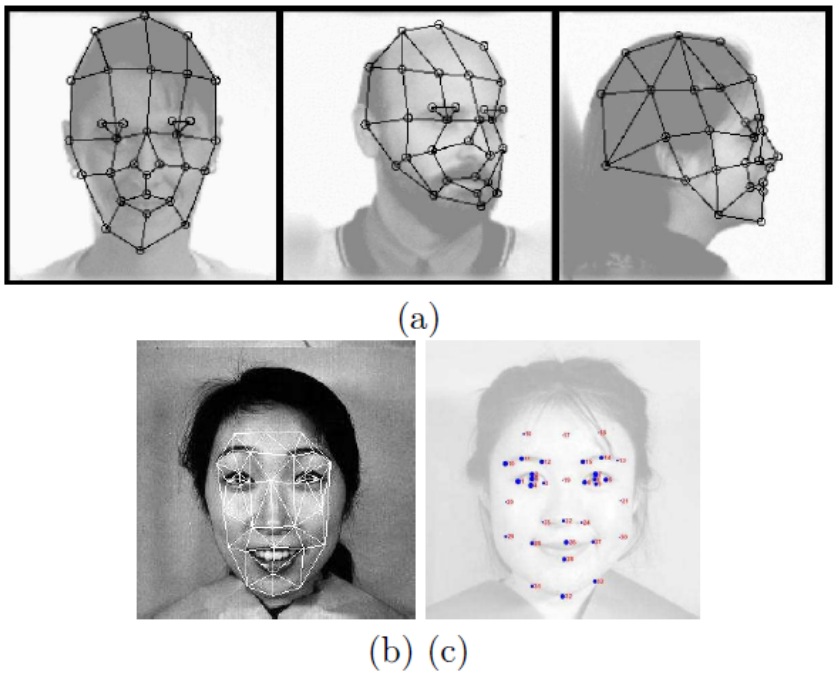


FIGURE 1.3 – Modèles de visage basé sur une grille de points. (a) modèle pour la reconnaissance d'identité [7]. (b) modèle pour la reconnaissance d'émotions [8]. (c) importance des points pour la reconnaissance démotions [9].

LITTLEWORT et al. [10] proposent un système d'analyse automatique des expressions faciales. Il s'agit d'un système de classification permettant de détecter la présence d'action units ainsi que leur intensité au cours d'une vidéo. L'extraction des données utiles à la classification se fait à partir de la texture du visage quasiment brute : le visage et les yeux sont détectés par des techniques proches de celles développées par VIOLA et JONES [11]. L'image du visage est alors normalisée en une fenêtre de 96x96 pixels où les yeux sont à une position fixe. Les caractéristiques extraites sont les réponses d'un ensemble de filtres de Gabor à différentes échelles et orientations en chacun des pixels de l'image (représentant au total plus de 650 000 filtres). Les réponses des filtres de Gabor sont alors données en entrée à des SVM (Un SVM est utilisé pour chaque action unit à détecter). Chaque SVM a été précédemment entraîné sur la base COHN-KANADE [12] : la présence de l'action unit est la réponse positive, et une réponse négative est renvoyée dans tous les autres cas.

Le système dans son ensemble permet un taux de reconnaissance de plus de 90%. De plus,

l'utilisation des SVM permet de mesurer l'intensité de l'action unit reconnue. Le principal avantage réside dans le fait que les techniques d'apprentissage et de classification sont génériques et il est donc aisé d'ajouter de nouveaux action units à détecter, à condition d'avoir la base d'apprentissage correspondante. Cependant, le système n'est capable que de traiter des visages vus de face et sans occultations. [2]

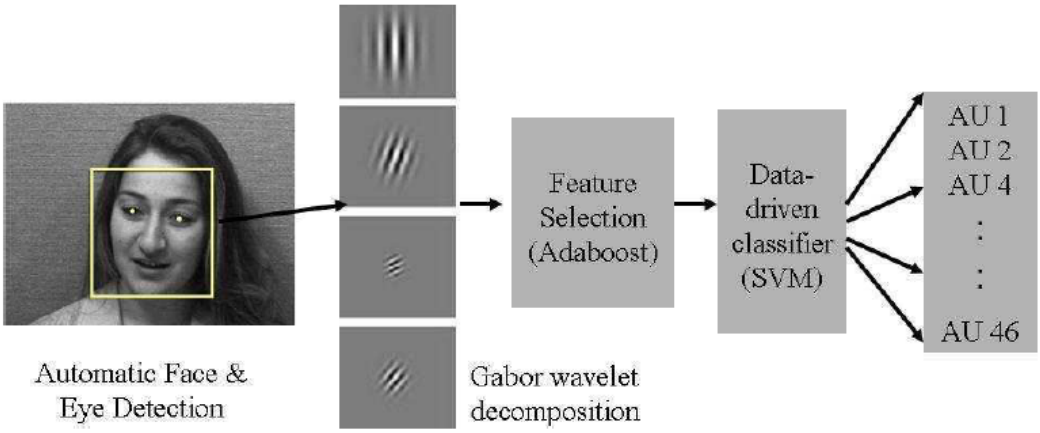


FIGURE 1.4 – Vue schématique du système de reconnaissance de [10]

1.3.2 Approche avec segmentation

L'approche basée sur l'analyse des composantes faciales consiste à employer une méthode particulière d'analyse pour chacune des composantes faciales. Dans cette catégorie, on trouve des techniques basées sur des modèles paramétriques : il s'agit d'un ensemble de points d'intétêt des composantes faciales liés entres eux par certaines contraintes. Il s'agit généralement de contraintes imposées sur la forme de la composante. Par exemple, les lèvres, qui sont des composantes très étudiées, ont des contours bien marqués et peuvent donc être modélisées par des polynômes ou bien encore par des formes plus libres. Les techniques d'analyse consistent alors à superposer les contours du modèle utilisé avec les contours réels ; il s'agit généralement de méthode d'optimisation, maximisant un critère de ressemblance (la répartition des points à fort contraste sur les contours du modèle par exemple). [2]

MORIYAMA et al. [13] présentent une méthode précise d'analyse des mouvements de l'œil. La méthode est basée sur l'utilisation d'un modèle gènèratif de l'œil humain : il s'agit d'un modèle 2D texturé, organisé en couches plus ou moins transparentes. L'œil est découpé en plusieurs sous composantes : paupières, sourcils, iris, etc. auxquelles est associé un ensemble de paramètres de forme et d'aspect (intensité lumineuse et couleur). Les paramètres permettent de faire évoluer le modèle selon une morphologie particulière (paramètres de structure) ou selon une expression particulière (paramètres de mouvement). Le but de l'algorithme est de trouver les paramètres du modèle à chaque image, afin que celui-ci ressemble le plus possible à l'image observèe. Une fois le visage normalisé en forme (en prenant en compte les rotations et en redressant l'image via l'utilisation d'une mèthode de suivi de demi cylindre 3D) et en luminosité, les auteurs utilisent l'algorithme de Lucas Kanade [2], modifié de façon à prendre en compte les déformations possibles du modèle.

Bien que la position du modèle soit initialisée manuellement sur la première image de chaque séquence à analyser, les résultats sont excellents en termes de robustesse et de précision. Cependant, ce type de modèle est très difficile à dèvelopper. En effet la forme de chacune des composantes doit être fidelement modélisée ainsi que ses variations possibles d'aspect. De

plus, les variations de forme et d'aspect doivent être différenciées selon qu'elles sont dues à des variations interpersonnelles (variations morphologiques) ou à des variations intra-personnelles (variations expressives). [2]

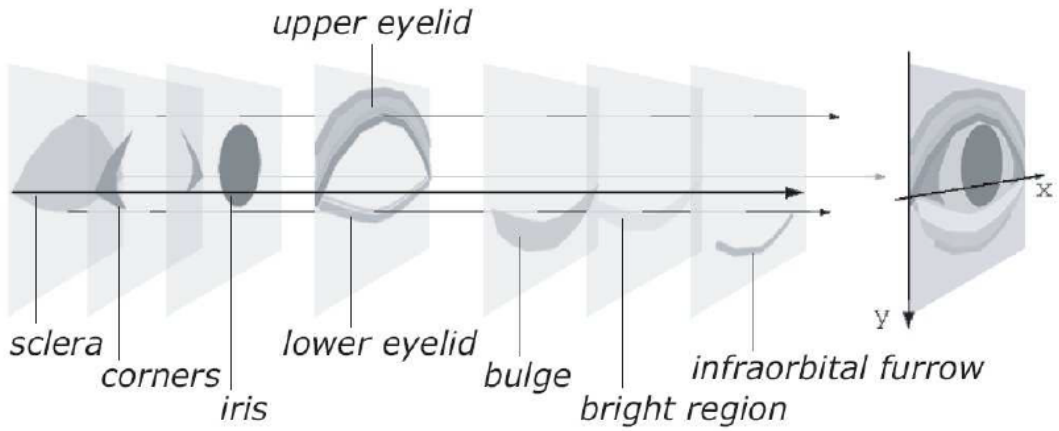


FIGURE 1.5 – Modèle génératif de l'œil utilisé dans le modèle de MORIYAMA [13]

TIAN et al. [14] proposent un système d'analyse automatique des expressions faciales, par reconnaissance d'action units. La classification est effectuée par des réseaux de neurones. Les données d'entrée des réseaux de neurones sont un ensemble de mesures effectuées sur des descripteurs principalement géométriques du visage, qui sont extraits par des méthodes ad hoc. Les sourcils et le haut des joues sont par exemple modélisés par deux segments de droite, la bouche et les yeux par des courbes paramétriques. Les modèles de la bouche et des yeux ont plusieurs états possibles : ouvert, semi ouvert et fermé. En plus de ces descripteurs, la présence de rides est détectée par une analyse de contour dans certaines zones (haut du nez par exemple) : l'opérateur de Canny appliqué à ces zones permet de déterminer s'il y a présence de ride en comparant le nombre de contours aux contours présents sur la première image de la séquence.

L'ensemble de ces données est fourni en entrée à des réseaux de neurones multicouches ayant une sortie par action unit. Le système est capable de reconnaître 15 actions units et certaines combinaisons avec un taux d'environ 90% et un taux de fausses alarmes d'environ 10%. Bien que le système offre de bonnes performances, les méthodes d'extraction des paramètres sont très spécifiques et construites empiriquement sur des indices de couleur, contours et mouvements. De plus, le système doit être initialisé manuellement sur la première image. Aucune information concernant la robustesse du système aux occultations manuelles et aux rotations du crâne n'est disponible. [2]

1.4 Les expressions faciales

1.4.1 Définition

Tout d'abord, il est important de faire la distinction entre la reconnaissance des expressions faciales et la reconnaissance d'émotions. Les émotions résultent de plusieurs facteurs et peuvent être révélées par la voix, la posture, les gestes, la direction de regard et les expressions faciales. Par contre, les émotions ne sont pas la seule origine des expressions faciales. En effet, celles-ci peuvent provenir de l'état d'esprit (ex : la réflexion), de l'activité physiologique (la douleur ou la fatigue) et de la communication non verbale (émotion simulée, clignotement de l'œil, froncement des sourcils). Néanmoins, sept émotions de base correspondent chacune à une expression faciale unique, et ce, quelles que soient l'ethnicité et la culture du sujet, ces émotions sont :

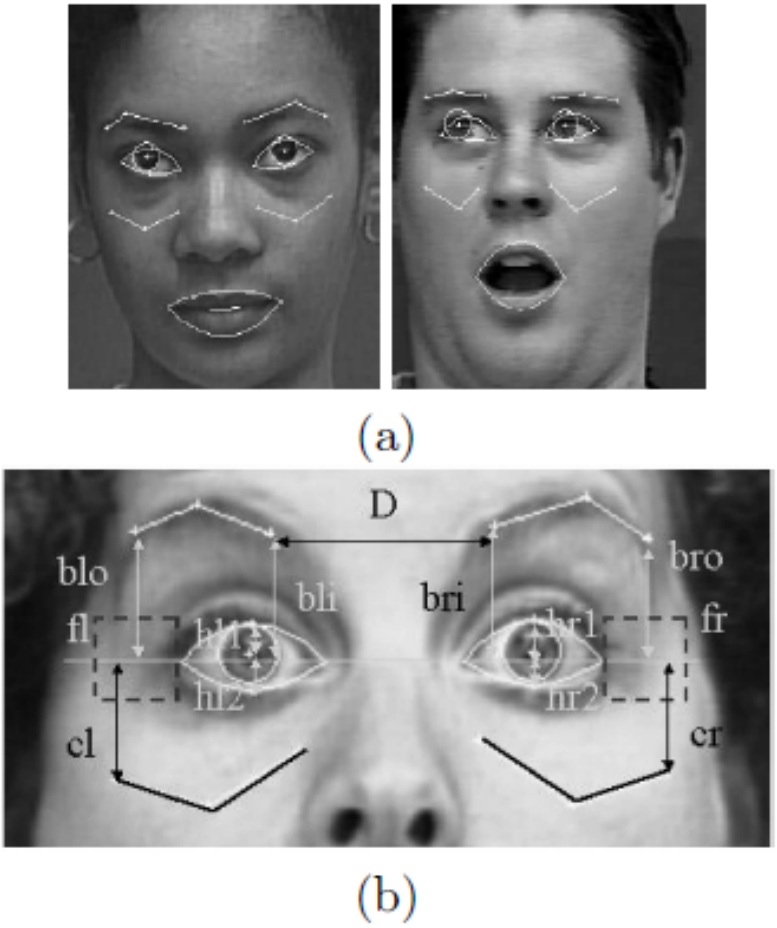


FIGURE 1.6 – : Modèle de visage basé sur des courbes paramétrées [14] (a) Modèle. (b) Mesures associées au modèle

la colère, le dégoût, l'étonnement, la joie, le mépris, la peur, et la tristesse. La reconnaissance des expressions faciales consiste à classer les déformations des structures faciales et les mouvements faciaux uniquement à partir des informations visuelles. La reconnaissance des émotions, quant à elle, est une tentative d'interprétation qui requiert une information contextuelle plus complète.

L'émotion est l'un des sujets les plus controversés de la psychologie, source de discussions intenses et de désaccords entre les premiers philosophes et autres penseurs jusqu'à nos jours.

L'émotion peut être décrite relativement à différents composants sur la base de facteurs physiologiques ou psychologiques, y compris des éliciteurs d'émotion, des processus neuraux d'émotion et des visages d'émotion. Il existe une longue histoire d'intérêt pour le problème de la reconnaissance des émotions liées à l'expression faciale dans plusieurs disciplines : philosophie (René Descartes), biologie (Charles Darwin) et psychologie (William James, Paul Ekman).

Depuis 1649, DESCARTES [15] a introduit les six " passions simples " : *merveille, amour, haine, envie, sourire* et *tristesse* et a supposé que tous les autres sont composés de certains de ces six. En 1872, DARWIN [16] affirma qu'il existe des émotions innées spécifiques et que chaque émotion comprend un schéma spécifique d'activation de l'expression du visage et du comportement. Inspiré des travaux de DARWIN [16], EKMAN [6], FRIESEN et ELLSWORTH [17] ont montré que les observateurs pouvaient s'accorder sur la manière de qualifier les expressions faciales posées et spontanées en termes de catégories émotionnelles ou de dimensions

émotionnelles à travers les cultures. EKMAN [6] montra des images d'expressions faciales à des habitants des Etats-Unis, du Japon, de l'Argentine, du Chili et du Brésil et découvrit qu'ils jugeaient ces expressions de la même manière. Des expressions faciales similaires ont tendance à se produire en réponse à des événements provoquant des émotions particulières. Mais cela n'a pas été décisif, car tous ces gens auraient pu apprendre le sens des expressions en regardant la télévision. L'expérience nécessite alors des personnes isolées visuellement, non exposées au monde moderne ni aux médias. Ekman [6] les a trouvés dans les hauts plateaux de Papouasie Nouvelle Guinée. Les résultats de l'expérience montrent que les sujets jugent les expressions proposées de la même manière. De plus, leur réponse à une émotion particulière correspond à la même expression. Sur la base de ces résultats, Ekman [6] a confirmé l'universalité des expressions émotionnelles et dressé une liste de six expressions émotionnelles de base, à savoir Surprise, Colère, Dégoût, Bonheur, Tristesse et Peur (voir la figure1.7). [18]

FIGURE 1.7 – Les six expressions émotionnelles universelles dans l'ordre suivant : bonheur, peur, dégoût, colére, tristesse et surprise [17]

Cependant, le terme "expression" implique l'existence de quelque chose qui est exprimé et les gens peuvent tromper leur sentiment interne en simulant une autre expression (comme les acteurs). Il existe donc une différence entre une expression faciale qui ne peut être reconnue que par l'analyse des traits du visage et une "émotion" qui correspond à une sensation interne et nécessite plus que des déformations de traits pour être reconnue. Ensuite, la reconnaissance d'une expression faciale permet d'obtenir des informations sur l'état émotionnel, mais ne suffit pas pour le confirmer et peut nécessiter le recours à d'autres modalitès (par exemple, la voix, le geste) [18].

1.4.2 Représentation des expressions faciales

Les expressions faciales représentent un important canal de communication non verbale. Même si l'être humain a acquis la capacité puissante d'un langage verbal, le rôle des expressions faciales dans les interactions interpersonnelles reste important et l'amélioration de ces compétences est souvent recherchée. Ils sont souvent à la base d'impressions significatives telles que la convivialité, la fiabilité ou le statut.
L'expression d'un visage donné à un moment donné est traduite par un composite de signaux provenant de plusieurs sources d'apparence faciale. Ces sources comprennent la forme générale, l'orientation (pose) et la position de la tête, les formes et positions des traits du visage (par exemple les yeux, la bouche) et la présence de rides et de leurs formes. Surtout, la source la plus importante est le comportement des traits du visage. Par exemple, la figure 1.8 en bas montre que le visage sans les traits permanents du visage ne transmet aucune expression ; tandis que chaque combinaison spécifique de forme des yeux et de la bouche conduit à une expression faciale spécifique (voir Figure 1.8 en bas). [18]

Cependant, les modifications de l'apparence des traits du visage sont le résultat de mouvements musculaires produits par une partie des muscles du visage. Les muscles faciaux sont

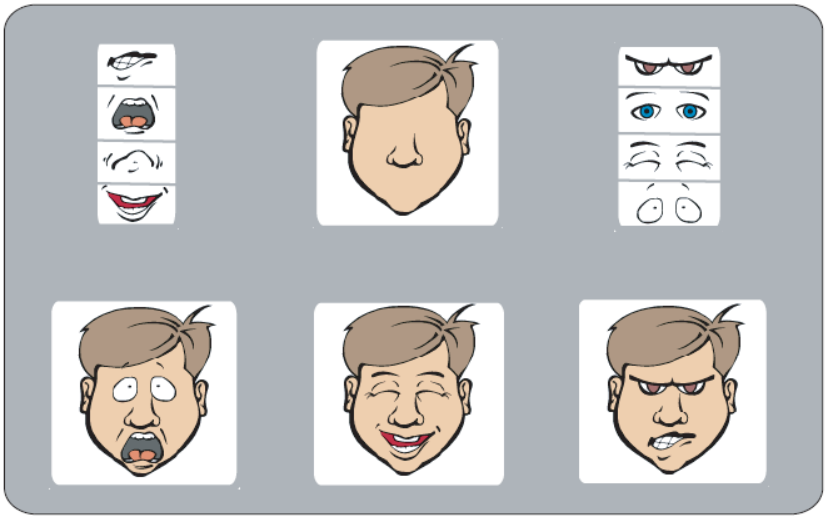


FIGURE 1.8 – En haut : informations nécessaires à l'interprétation de l'expression faciale ; en bas : trois exemples de configuration des traits du visage (yeux et bouche) conduisant à la peur, au sourire et à la colère. [18]

comme des feuilles élastiques étirées en couches sur le crâne, les os du visage, les ouvertures qu'ils forment, le cartilage, la graisse et d'autres tissus de la tête. La figure 1.9 montre une vue de 3/4 des muscles faciaux. Chaque expression faciale correspond à une combinaison de ces muscles faciaux. [18]

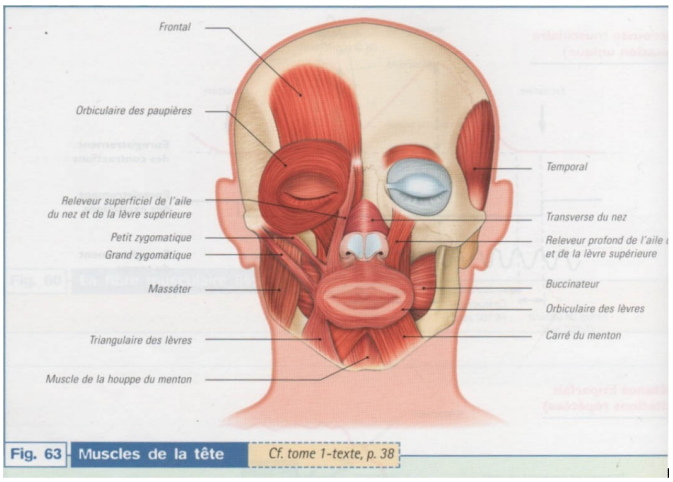


FIGURE 1.9 – Les muscles du visage impliqués pour produire des expressions faciales. [18]

1.4.3 Détection de visage et les expressions faciales

Le résultat attendu de tous les recherches et tous les projets est l'implémentation d'un système qui peut effectuer des analyses automatisées des expressions faciales. En général, deux étapes doivent être exécutées pour résoudre ce problème. Premièrement, avant qu'une expression faciale soit analysée, le visage doit d'abord être détecté dans une scène. Ensuite, les expressions faciales sont extraites à partir des séquences d'images. Ce processus correspond à l'extraction du visage et de ses caractéristiques dans une scène. Dans la plupart des travaux en analyse des expressions faciales, les conditions d'acquisition de chaque séquence d'images sont contrôlées. Habituellement, le visage est capté sous forme d'une vue frontale. Par conséquent, la présence d'un visage dans la scène est assurée, et la position du visage dans la scène est aussi

connue a priori. Cependant, la détermination de l'endroit exact du visage dans une image faciale digitalisée est un problème plus complexe. [1]

D'abord, la dimension et l'orientation du visage peuvent changer d'une image à une autre. Si les images sont captées avec une caméra fixe, les visages peuvent être captés dans les images à des tailles diverses et des orientations différentes dues aux mouvements de la personne observée. Ainsi, il est difficile de rechercher un modèle fixe dans l'image, aussi la présence du bruit et d'occlusions rend le problème bien plus difficile. On croit généralement que les images à deux niveaux de gris de 100 à 200 pixels forment une limite inférieure pour la détection d'un visage par un observateur humain. Une autre caractéristique du système visuel humain est qu'un visage est perçu dans son ensemble, pas comme une collection des caractéristiques faciales. La présence de ces caractéristiques et leur rapport géométrique réciproque semblent être plus importants que les détails de ces caractéristiques [1]. Ci-dessous y ont quelques exemples pour tout ça :

- Pour représenter le visage, C.L. HUANG et Y.M. HUANG [19] appliquent un modèle statistique de distribution des points (Points Distribution Model, PDM). Afin de réaliser un placement correct d'un PDM initial dans une image d'entrée, Huang et Huang utilisent un détecteur d'arêtes pour obtenir une évaluation grossière de la position du visage dans l'image. La vallée dans la fonction de luminance située entre les lèvres et les deux bordures verticales symétriques représentant les frontières verticales externes du visage découle d'une évaluation grossière de l'endroit de cette dernière. Le visage ne doit pas être couvert de cheveux et de lunettes. La tête doit être statique et les variations d'illumination doivent être linéaires pour que le système fonctionne correctement. [1]

- PANTIC et ROTHKRANTZ [20] détectent le visage comme une unité totale. Ils utilisent des paires d'images faciales en entrée : une de face et une de profil (Figure 1.10). Pour déterminer les frontières externes verticales et horizontales de la tête, ils analysent l'histogramme vertical et horizontal de l'image de vue frontale proposé dans [19]. Pour localiser le contour du visage, ils utilisent un algorithme de détection et de localisation du contour de visage basé sur le modèle de couleur HSV, qui est similaire à l'algorithme basé sur le modèle relatif RGB. Ainsi, ils utilisent des images en vue de profil pour faciliter la détection automatique des expressions faciales en cas de mouvement de la tête durant le traitement. Pour les images en vue de profil, ils appellent un algorithme de détection de profil, qui représente une approche spatiale pour prélever le contour de profil d'une image seuillée et pour le seuillage de l'image de profil en entrée, la valeur du seuil découlant du modèle de couleur HSV est exploitée. Les cheveux dans le visage et les lunettes ne sont pas permis non plus. [1]

- KOBAYASHI et HARA [21] ils ont utilisé une caméra CCD en mode monochrome pour obtenir une distribution en niveau de brillance du visage humain. Premièrement, la distribution des niveaux de brillance du visage de 10 sujets est obtenue. Ensuite, le système extrait la position des iris en utilisant une technique de corrélation, cette dernière tente de déterminer la position dans une image d'un visage à partir de modèles de base où la corrélation entre les modèles d'iris et des régions dans l'image du visage est maximale. Une fois que les iris sont identifiées, la position globale du visage est déterminée en employant la position relative des caractéristiques faciales dans le visage. Le sujet observé doit faire face à la caméra tout en se situant à la distance approximative d'un mètre devant elle. [1]

- YONEYAMA et al. [22] ils extraient les coins externes des yeux, la taille des yeux et la

taille de la bouche d'une manière automatique. Une fois que ces caractéristiques sont identifiées, la taille du domaine facial examiné est normalisée et une grille rectangulaire de 8 par 10 est placée au-dessus de l'image [20]. Il n'est pas énoncé quelle méthode a été appliquée et aucune limitation de la méthode utilisée n'a été rapportée par [22]. [1]

- KIMMURA et YACHIDA [23] proposent des méthodes automatiques pour extraire des points de caractéristiques du visage à partir des images de couleurs normales. La méthode proposée inclut la localisation du visage, la position des caractéristiques du visage, les contours de ces caractéristiques, et les séquences de points de ces dernières. Pour extraire robustement le contour d'une caractéristique de visage, ils ont proposé des modèles de contour actif, qui emploient n contours pour résoudre le problème des méthodes de contours originales une fois appliquées au problème contenant les contours et ils proposent une nouvelle fonction d'énergie pour ces méthodes. [1]

1.4.4 Analyse automatique des expressions faciales

À ce point, une distinction claire sera faite entre deux termes, nommés, caractéristiques faciales et modèle de caractéristiques faciales. Les caractéristiques faciales sont les sourcils, les yeux, le nez, la bouche, et le menton. Le modèle de caractéristiques faciales est présenté sous forme d'une combinaison de caractéristiques utilisées pour représenter le visage, qui peut se représenter de plusieurs façons, soient en unité totale (représentation holistique), sous forme d'un ensemble de caractéristiques (représentation analytique), ou en une combinaison de ces dernières (approche hybride). La dernière étape consiste à définir les catégories qu'on voudra utiliser pour la classification des expressions faciales et/ou l'interprétation des expressions faciales, et de diviser le mécanisme de catégorisation.

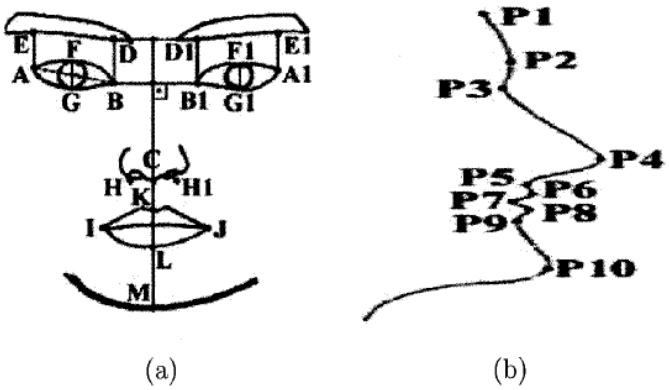
1.4.5 Extraction des expressions faciales

Après avoir localisé un visage dans une image, la prochaine étape est d'extraire les informations sur l'expression faciale produite d'une manière automatique. Un analyseur d'expressions faciales entièrement automatique doit alors être développé. La représentation du visage et le genre d'images d'entrée affectent le choix de l'approche utilisée pour l'extraction des informations sur les expressions faciales.

Un des objectifs fondamentaux de l'analyse des expressions faciales est la représentation de l'information visuelle qu'un visage testé peut contenir [24]. Les résultats de JOHANSSON [25] ont donné un indice de l'importance de ce problème. Les expériences de l'extraction des informations sur les expressions faciales suggèrent que les propriétés du visage, concernant les expressions faciales pourraient être obtenues en décrivant les mouvements des points de contrôle associés aux caractéristiques faciales (sourcils, yeux, et bouche) et en analysant les rapports entre ces mouvements. Ceci a poussé les chercheurs sur l'analyse de visages à faire différentes tentatives pour définir les propriétés visuelles des ensembles de points faciaux permettant la modélisation des expressions faciales (par exemples figure 1.11 [21]). [1]

La figure 1.10 représente une vue frontale et une vue de côté de l'ensemble des points du visage (approche analytique).

La figure 1.11 illustre les points caractéristiques du visage selon H. KOBAYASHI et F. HARA [21], (approche analytique).



(a) (b)

FIGURE 1.10 – Points faciaux [20]. (a) Modèle de visage à vue frontale. (b) Modèle de visage en vue de côté.

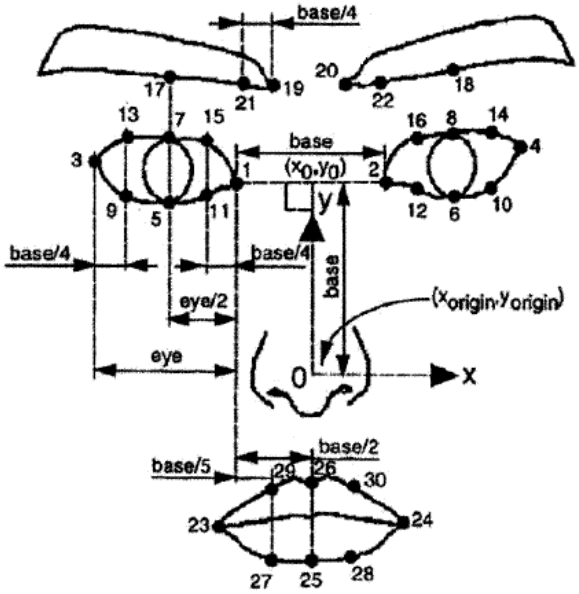


FIGURE 1.11 – Points caractéristiques du visage selon [21]

La figure 1.12 représente le modèle planaire pour représenter des mouvements rigides du visage et le modèle affine de courbure pour représenter des mouvements non rigides de visage (approche holistique).

Indépendamment du genre d'images d'entrée, les images faciales ou images arbitraires, la détection de la position du visage d'une image observée ou d'une séquence d'images a été approchée de deux manières :

* Dans l'approche holistique, le visage est considéré comme une unité totale.
* Dans l'approche analytique, le visage est localisé premièrement par la détection de certaines caractéristiques importantes du visage (ex : yeux, bouche, front).

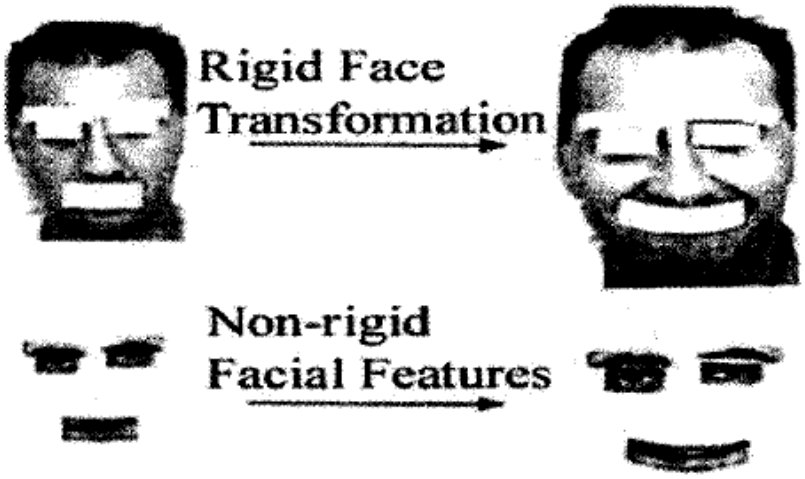


FIGURE 1.12 – Modèle planaire pour représenter les mouvements faciaux rigide et le modèle affine de courbure pour représenter les mouvements faciaux non rigides [26]

Le visage peut aussi être modélisé en utilisant une approche hybride, ce qui caractérise une combinaison de l'approche analytique et holistique pour la représentation du visage. Dans cette approche, un ensemble de points faciaux sont habituellement utilisés pour déterminer la position initiale d'un modulateur du visage. Le système qui utilise cette approche est proposé par KIMURA et YACHIDA [23]. Ils utilisent la fonction d'énergie pour adapter cette fonction à une image faciale normale. Ils calculent d'abord le contour de l'image en appliquant un filtre différentiel. Ensuite, afin d'extraire la force externe, qui correspond au gradient de contour dans l'image, ils appliquent un filtre gaussien. L'image filtrée est désignée sous le nom d'un champ énergétique. [1]

La figure 1.13 représente la fonction d'énergie et le champ énergétique correspondant (Approche hybride).

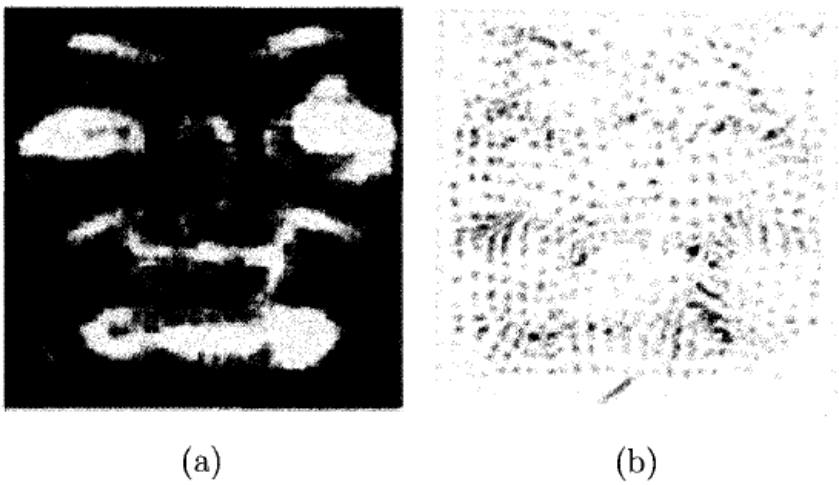


FIGURE 1.13 – La fonction d'énergie et son champ énergétique correspondant [23]. (a) La fonction d'énergie. (b) Le champ énergétique.

Indépendamment du genre de visage, le modèle est appliqué, des tentatives doivent être faites pour modeler et ensuite extraire les informations sur l'expression faciale montrée en perdant peu ou beaucoup de cette information. Plusieurs facteurs rendent cette tâche complexe. Le

premier est la présence des cheveux, les lunettes, etc., ce qui cache les expressions faciales. Un autre problème est la variation au niveau de la taille et l'orientation du visage lors de la capture d'une séquence d'images. Ceci rend difficile la recherche des modèles fixes dans les images. Finalement, le bruit et l'occlusion sont toujours présents dans une certaine mesure.

1.4.6 Classification des expressions faciales

Après que le visage et son aspect ont été perçus, la prochaine étape d'un analyseur automatisé d'expressions faciales est de classifier (identifier) cette expression donnée par le visage. Une problématique fondamentale de la classification des expressions faciales est de définir un ensemble de catégories que nous voulons traiter. Une autre problématique qui en découle est de concevoir des mécanismes de catégorisation. Des expressions faciales peuvent être classifiées de plusieurs façons en termes des actions faciales qui causent une expression, en termes de certaines expressions sans prototype (ex. : fronts augmentés) ou en termes de certaines expressions à prototype (expressions émotives). Il y a plusieurs approches pour la reconnaissance des changements du visage humain linguistiquement universelles basées sur l'activité musculaire du visage. [1]

Parmi ces approches, on trouve le système de codage d'actions faciales (FACS) proposé par EKMAN et al [6], qui est la meilleure approche connue et utilisée. C'est un système désigné pour les observateurs humains pour décrire les changements dans l'expression faciale en termes d'activations visuellement observables des muscles du visage humain. En plus des FACS, il existe un autre modèle tel que le standard MPEG4 [27]. MPEG4 est une norme de compression multimédia basée sur les objets, qui permet le codage de différents objets audiovisuels dans la scène de manière indépendante.

Système de Codage des Actions Faciales :

Les signaux rapides du visage humain sont les mouvements des muscles du visage qui tirent la peau, entraînant une déformation provisoire de la forme des caractéristiques faciales (yeux, bouche, nez, front) et de l'aspect des plis, des sillons, et des bombements de la peau. La terminologie commune pour décrire les signaux rapides du visage se réfère l'un ou l'autre aux limites linguistiques culturellement dépendantes indiquant un changement spécifique dans l'apparaissance d'une caractéristique faciale particulière (sourire, sourire affecter, froncement des sourcils, ricanement, etc.). Le système de codage des actions faciales (FACS) est probablement l'étude la plus connue sur l'activité faciale. C'est un système qui a été développé pour faciliter la mesure objective de l'activité faciale pour des investigations comportementales de la science sur le visage. FACS est conçu pour les observateurs humains pour détecter les changements subtils indépendants de l'aspect facial provoqué par des contractions des muscles faciaux. Les changements au niveau de l'expression faciale sont décrits avec les FACS en termes de 46 unités d'action différentes, qui sont anatomiquement liées à la contraction d'un muscle spécifique du visage ou d'un ensemble de muscles. [1]

La figure 1.14 illustre quelques images des unités d'actions faciales ainsi que la description de chaque unité d'action faciale.

Avec la définition des différentes unités d'action (AUs), les codeurs FACS fournissent aussi les règles permettant la détection visuelle des unités d'action et leurs segments temporels (début, milieu, et fin de l'unité d'action) d'une image faciale. En utilisant ces règles, un codeur FACS décompose une expression faciale fournie en plusieurs unités

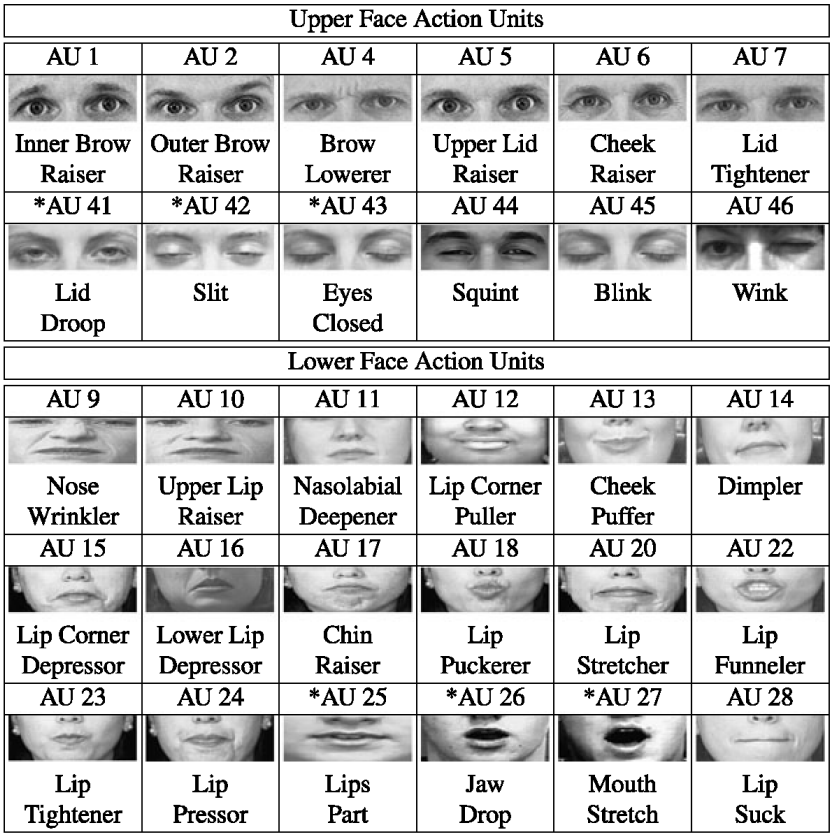


FIGURE 1.14 – Quelques exemples des unités d'actions faciales présentés par le FACS [6]

d'action qui produisent l'expression faciale. Bien que les FACS fournissent une bonne fondation pour le codage des unités d'actions des images faciales par les observateurs humains, réaliser la reconnaissance des unités d'actions par un ordinateur n'est pas une tâche triviale. Le problème majeur de cette méthode est que les unités d'action peuvent se produire dans plus de 7000 combinaisons complexes différentes entraînant des bombements et divers mouvements d'entrée et de sortie des images planes de caractéristiques faciales permanents qui sont difficiles à détecter dans les images faciales à deux dimensions. [1]

MPRG4 :

Elle spécifie un modèle de visage dans son état neutre avec un ensemble de points caractéristiques : le jeu de paramètres de définition faciale (FDP) (Figure 1.15 (c) et (d)). L'objectif principal de ces points de fonctionnalité est de fournir le jeu de paramètres d'animation faciale (FAP). Les FAP représentent un ensemble d'actions faciales de base (mouvements de la langue, des yeux et des sourcils) et permettent la représentation des expressions faciales. Les paramètres de mouvement en translation sont exprimés en termes d'unités de paramètres d'animation faciale (FAPU). Les FAPU sont illustrées à la figure 1.15 (a) et correspondent à des fractions de distances entre certaines caractéristiques faciales clés. Ces unités sont définies en fonction des distances dans une expression neutre afin de permettre l'interprétation des FAP. Ensuite, l'utilisation de la norme MPEG4 consiste à définir les expressions faciales comme un ensemble de mesures (FDP) et de transformations (FAP). [18]

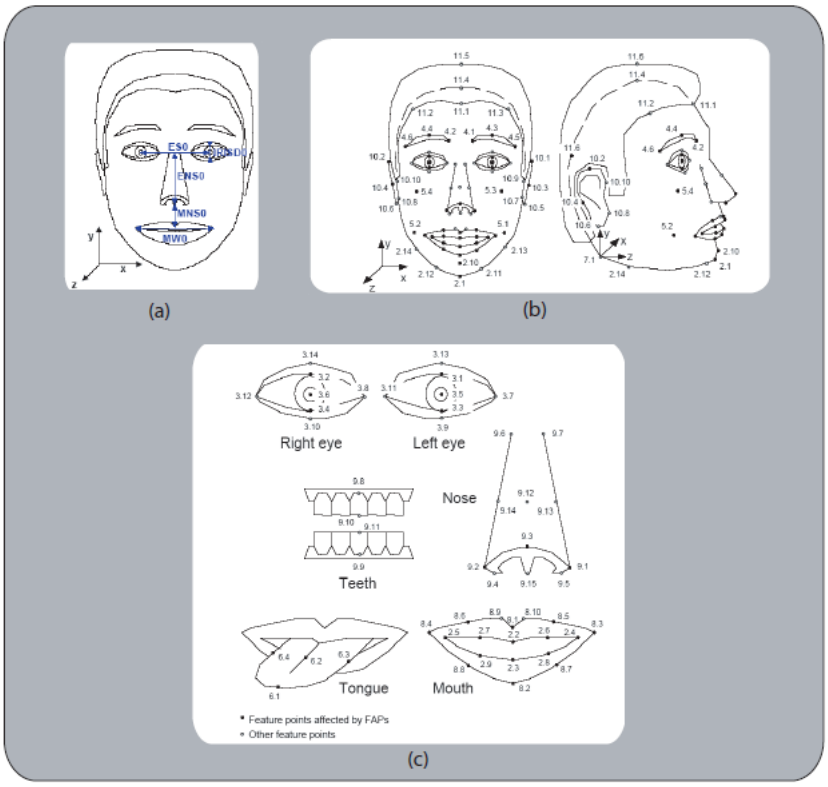


FIGURE 1.15 – (a) Un modèle de visage dans son état neutre et les unités de paramètres d'animation faciale ; (b) et (c) Paramètres de définition faciale utilisés pour la définition de l'animation faciale [27]

En résumé, différents codages des comportements des caractéristiques faciales peuvent être choisis en fonction de leur pertinence par rapport au contexte de l'application. Ainsi, une analyse automatique de l'expression émotionnelle d'un visage humain nécessite un certain nombre d'étapes de prétraitement pour extraire des informations du visage. Ces informations correspondent à la détection et au suivi du visage ; la localisation d'un ensemble de traits du visage (tels que les yeux, la bouche, le nez et les rides) ; leur représentation (par exemple FACS ou MPEG4) ; leur interprétation et enfin la reconnaissance de l'expression.

1.4.7 Reconnaissance des expressions faciales basée sur la vision d'ordinateur

Il y a plusieurs moyens pour faire une étape de reconnaissance des expressions faciales, par exemple l'utilisation des capteurs, mais nous nous intéressons aux méthodes de la vision par ordinateur.

En vision par ordinateur, de nombreuses recherches sur la classification des expressions faciales ont conduit de nombreux systèmes à adopter des approches différentes. Une description détaillée peut être trouvée dans PANTIC et al. [20] et FASEL et al. [29]. Il existe trois approches principales : l'analyse de flux optique à partir d'actions faciales, les techniques basées sur des modèles et les méthodes basées sur des points de repère. [18] Nous présentons ci-dessous les principes de ces méthodes :

Approches basées sur le flux optique :

Des informations de mouvement précises peuvent être obtenues en calculant le flux op-

tique, qui représente la direction et la magnitude du mouvement. Plusieurs tentatives de reconnaissance d'expressions faciales se sont concentrées sur l'analyse de flux optique à partir d'une action du visage, le flux optique étant utilisé pour modéliser les activités musculaires ou estimer les déplacements de points caractéristiques.

- YACOOB et DAVIS [30] ont proposé une représentation du mouvement du visage basée sur le flux optique afin de reconnaître les six expressions faciales universelles. Cette approche est divisée en trois étapes : premièrement, les régions rectangulaires entourant les traits permanents du visage (yeux, sourcils, nez et bouche) sont supposées être données dans la premiére image et suivies dans les images restantes de la séquence (voir Figure 1.16) ; deuxièmement, une estimation du flux optique sur ces caractéristiques définit la représentation de niveau moyen qui décrit les modifications faciales observées à chaque image en fonction de la première image (mouvements rigides et non rigides) ; troisièmement, cette représentation de niveau moyen est classèe dans l'une des six expressions faciales à l'aide d'un système basé sur des règles combinant les actions de base des composants de fonction [31] et les règles de repères de mouvement décrites dans [32]. [18]

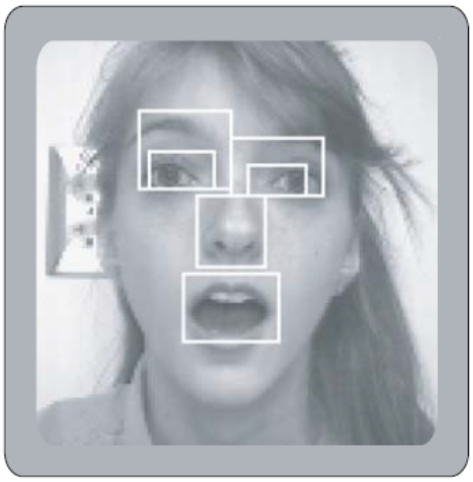


FIGURE 1.16 – Exemple de rectangles entourant les régions de visage d'intérêt [30]

- ROSENBLUM et al. [33] a étendu le système ci-dessus avec un réseau de neurones à fonction de base radiale pour chacune des six expressions faciales universelles, ce qui permet d'apprendre les corrélations entre les mouvements du visage et les expressions faciales. Cependant, les performances de la classification proposée ne sont testées que pour les expressions : Smile et Surprise. Pour améliorer la précision du modèle et être robuste au mouvement de la tête, BLACK et YACOOB [26] ont présenté une approche avec un modèle paramétré local du mouvement de l'image pour l'analyse de l'expression faciale. Les mouvements de tête rigides sont représentés par un modèle planaire permettant de récupérer les informations sur le mouvement de la tête (voir figure 1.17). Le mouvement des traits du visage est déterminé relativement au visage. Les mouvements non rigides des traits du visage (yeux, sourcils et bouche) sont représentés par un modèle affine plus courbure. Un ensemble de paramètres estimés à partir des modèles à l'aide d'un schéma de régression [34] basé sur l'hypothèse de la constance de la luminosité est utilisé pour définir des prédicats de niveau moyen décrivant le mouvement des caractéristiques faciales. Un ensemble de règles est ensuite défini pour combiner les prédicats de niveau intermédiaire entre le début et la fin des expressions afin de le reconnaître (la description détaillée des

règles se trouve dans [26]). Dans cette approche, les régions initiales pour la tête et les caractéristiques faciales sont sélectionnées manuellement et sont automatiquement suivies dans les images restantes de la séquence. De plus, les seuils utilisés pour détecter les prédicats de niveau moyen dépendent de la taille du visage. [18]

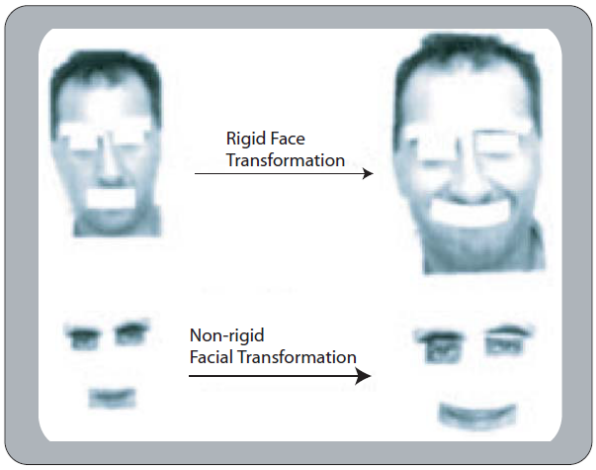


FIGURE 1.17 – Modèle plan pour la représentation des mouvements du visage rigide et modèle affine-plus-courbure pour la représentation des mouvements du visage non rigides [33]

- ESSA et PENTLAND [35] ont proposé la combinaison d'un modèle physique dynamique et de l'énergie du mouvement pour la classification des expressions faciales. Le mouvement est estimé à partir d'un flux optique et est affiné par le modèle physique dans une estimation récursive et un cadre de contrôle. Un modèle de visage physique est appliqué pour modéliser l'activation des muscles faciaux et un mouvement 2D idéal est calculé pour les cinq expressions étudiées (Colère, Dégoût, Bonheur, Surprise et sourcils levés (les autres expressions sont difficiles à simuler par leurs sujets) (voir Figure 1.18). Chaque modèle a été délimité en faisant la moyenne des motifs de mouvement générés par deux sujets pour chaque expression. La classification des expressions faciales est basée sur la distance euclidienne entre le modèle appris d'énergie de mouvement et l'énergie de mouvement de l'image observée. [18]

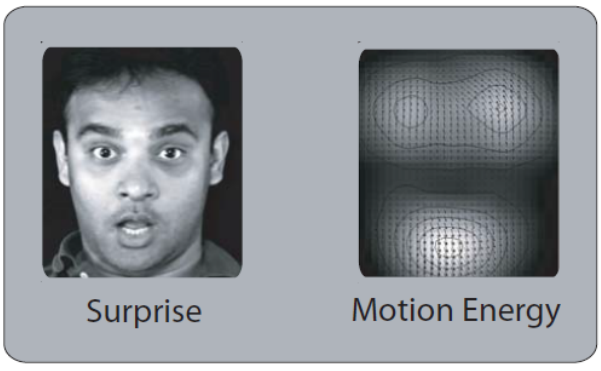


FIGURE 1.18 – Représentation d'énergie de mouvement spatio-temporelle du mouvement facial pour une surprise [35]

- COHN et al [36] proposent un système de reconnaissance automatique basé sur la modélisation des UA. Le déplacement de 36 points caractéristiques situés autour des yeux, des sourcils, du nez et de la bouche (voir Figure 1.19) est estimé à l'aide du flux optique.

Des groupes séparés (matrices de variance-covariance) ont ét'e utilisés pour la classification des UA. Ils ont utilisé deux fonctions discriminantes pour trois UA de la région des sourcils, deux fonctions discriminantes pour trois UA de la région des yeux et cinq fonctions discriminantes pour neuf UA des régions du nez et de la bouche. Cependant, le flux optique est calculé au niveau de chaque pixel dans une région d'intérêt spécifiée. Cette approche ne permet pas toujours de distinguer le flux optique causé par le mouvement des caractéristiques faciales de celui provoqué par un autre bruit non lié, ce qui conduit à des résultats de détection erronés. Par exemple, dans le cas de Surprise, la détection d'une zone de la bouche correspondant à un état ouvert peut être due à l'erreur d'estimation du flux optique provoquée par la variation de luminance. De plus, les estimations de flux optiques sont facilement perturbées par des mouvements non rigides. Ils sont également sensibles à l'inexactitude de l'enregistrement des images et aux discontinuitès de mouvement. [18]

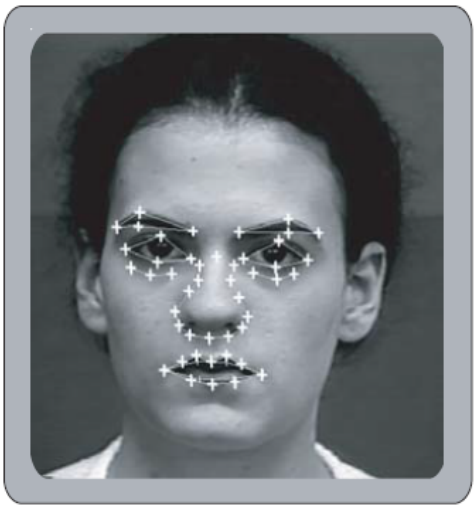


FIGURE 1.19 – Suivi des points caractéristiques [36]

Approches basées sur un modèle :

Plusieurs systèmes de reconnaissance d'expression faciale ont utilisé des techniques basées sur des modèles. Certaines d'entre elles appliquent un processus de déformation d'image pour mapper des images de visage sur un modèle géométrique. D'autres réalisent une analyse locale où des noyaux spatialement localisés sont utilisés pour filtrer les caractéristiques faciales de l'extrait. Un certain nombre de travaux ont appliqu{e une analyse holistique des niveaux de gris basée sur une analyse en composantes principales, une analyse en ondelettes de Gabor ou une approche à surface propre et FisherFace. [18]

- HUANG et HUANG [19] calculent d'abord 10 paramètres d'action AP (voir la figure 1.20) en fonction de la différence entre les paramètres de caractéristique de modèle d'un visage neutre et ceux de l'expression faciale examinée de la même personne. Les deux premiers termes des valeurs propres sont utilisés pour représenter les variations des points d'accès. Ensuite, un classificateur de distance minimale a été utilisé pour regrouper les deux paramètres d'action principaux de 90 échantillons d'image d'apprentissage en six groupes (les six expressions émotionnelles de base). Etant donné que la distribution en composantes principales de chaque expression

est superposée à la distribution d'au moins deux autres expressions, trois meilleures correspondances sont sélectionn'ees. Le score le plus élevé des trois corrélations détermine la classification finale de l'expression examinée. Cependant, les tests étant réalisés sur le même sujet, on ne sait pas comment la méthode se comportera pour un sujet inconnu. [18]

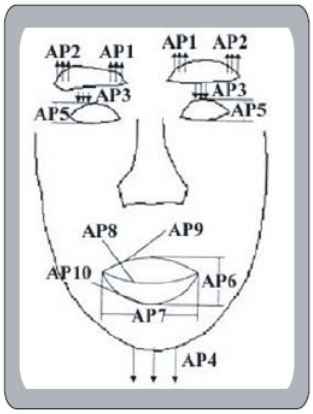


FIGURE 1.20 – Paramèmtres d'Action [19]

- Sur la base de la reprśentation en ondelettes de Gabor, LYONS [8] a présenté une méthode de classification des six expressions universelles plus l'expression neutre. Une grille de 34 points caractéristiques du visage est initialisée manuellement sur le visage (Figure 1.21). Les coefficients d'ondelettes de Gabor de chaque point caractŕistique de la grille sont calculś et combinś en un seul vecteur. Les principales composantes des vecteurs caractŕistiques des images d'apprentissage sont calculés. Ensuite, une analyse discriminante linaire est utilisé afin d'agrger les vecteurs rsultants en grappes ayant diffŕents attributs faciaux. Enfin, la classification a été rálisé en projetant le vecteur d'entrée d'une image test le long des vecteurs discriminants. Le vecteur d'entré est affecté au cluster le plus proche. [18]

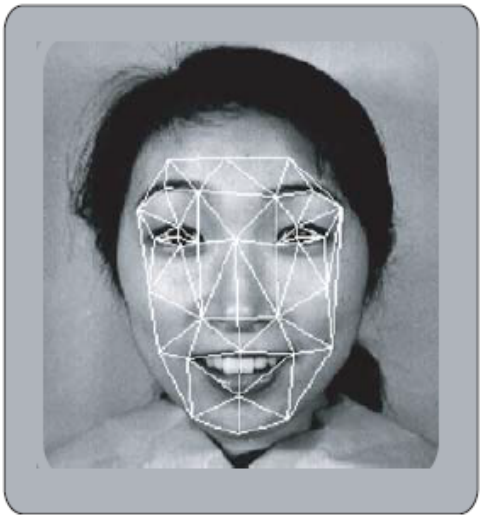


FIGURE 1.21 – Représentation graphique élastique marquée par Gabor de l'image faciale [8]

- OLIVER et al. [37] ont appliqué un modèle de Markov caché (HMM) à la reconnaissance de l'expression faciale, basé sur la déformation de formes de la bouche suivies en temps réel. La forme de la bouche est caractérisée par son aire, ses valeurs propres spatiales (par exemple, la largeur et la hauteur) et son cadre de sélection. La figure

1.22 décrit le vecteur de caractéristiques de la bouche extrait. Basées uniquement sur la forme de la bouche, les expressions étudiées sont bouche ouverte, tristesse, sourire et bouche ouverte. Chacun dès les expressions basées sur la bouche sont associées à un HMM formé sur le vecteur de caractéristiques de la bouche. L'expression faciale est identifiée en calculant la probabilité maximale de la séquence d'entrée pour tous les HMM formés. Cependant, seule une partie des expressions faciales présente le motif caractéristique contenu dans la forme de la bouche. [18]

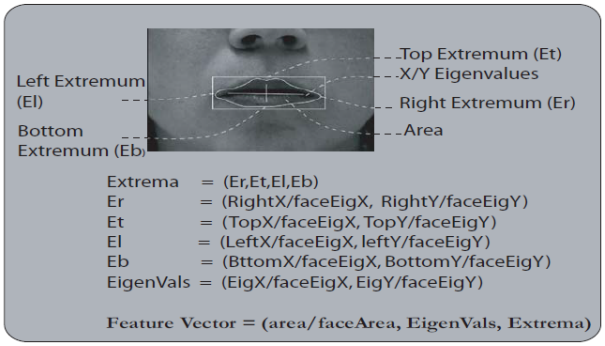


FIGURE 1.22 – Extraction du vecteur de caractéristiques de la bouche [37]

- ANDERSON [38] a proposé un système automatisé à plusieurs étapes pour la reconnaissance en temps réel de l'expression faciale. Il utilise le mouvement du visage pour caractériser les vues frontales monochromes d'expressions faciales et est capable de fonctionner efficacement dans des scènes encombrées et dynamiques. Les visages sont localisés à l'aide d'un algorithme de suivi de modèle de ratio spatial. Le flux optique du visage est ensuite déterminé à l'aide d'une implémentation en temps réel d'un modèle de gradient. Le système de reconnaissance d'expression effectue ensuite la moyenne des informations de vitesse faciale sur des régions identifiées du visage et annule les mouvements de tête rigides en prenant des ratios de ce mouvement moyenné. Les signatures de mouvement produites sont ensuite classifiées à l'aide de machines à vecteurs de support comme non expressives ou comme l'une des six expressions universelles. Cependant, le système est spécifique à un seul utilisateur à une distance et un aspect relativement fixe de la caméra. [18]

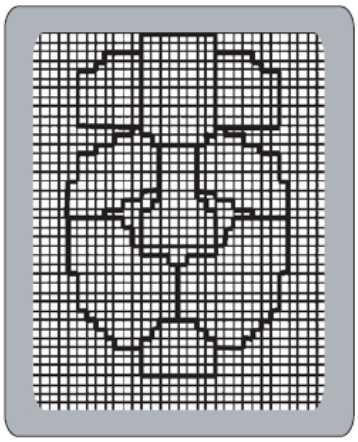


FIGURE 1.23 – Régions pour la moyenne de mouvement [38]

Il est difficile de concevoir un modèle physique déterministe qui représente avec précision les propriétés géométriques du visage et les activités musculaires. L'approche holistique implique généralement une phase d'entraînement intensive. Le modèle formé est souvent

peu fiable pour des utilisations pratiques en raison des variations interpersonnelles, du fait que les expressions sont agies, de la variabilité de l'éclairage et des difficultés à faire face aux séquences émotionnelles dynamiques. [18]

Approches basées sur des points de repère :

Les dernières années ont vu l'utilisation croissante de l'analyse de caractéristiques géométriques pour représenter des informations faciales. Dans ces approches, les mouvements du visage sont quantifiés en mesurant les déplacements géométriques des points caractéristiques du visage entre la trame actuelle et la trame initiale.

- LIEN et al. [39] proposent une méthode hybride basée sur : premièrement, le suivi des points caractéristiques (points autour des contours des yeux, des sourcils, du nez et de la bouche détectés manuellement dans le premier cadre), deuxièmement, le flux optique et le troisième, détection de sillon pour extraire les informations d'expression. La classification des expressions est basée sur les unités d'action du système de codage des actions facialesFACS [32]. Les HMM sont utilisés pour la discrimination entre chaque UA ou combinaison d'UA en fonction de la configuration des mouvements de caractéristiques. Un lien dirigé entre les états du HMM représente la transition inhérente possible d'un état du visage à un autre. Une UA est identifiée si son HMM associé a la probabilité la plus élevée parmi tous les HMM, à partir d'un vecteur de caractéristiques faciales. [18]

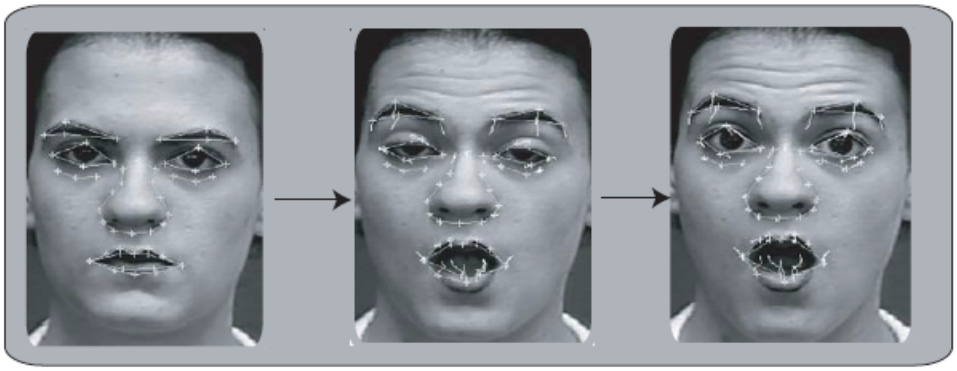


FIGURE 1.24 – Séquence de suivi des points caractéristiques du visage [39].

- TIAN ET AL. [14] utilisent deux approches distinctes basées sur des réseaux de neurones pour reconnaître 6 UA supérieures et 10 inférieures en fonction des caractéristiques faciales permanentes (yeux, sourcils et bouche) et des caractéristiques faciales transitoires (approfondissement des sillons faciaux). Les caractéristiques faciales ont été regroupées dans des collections distinctes de paramètres car les actions faciales des faces supérieure et inférieure sont relativement indépendantes pour la reconnaissance UA. Les entrées des NN pour l'entraînement et la classification sont les descriptions paramétriques des traits du visage permanents et transitoires. Les traits du visage sont initialisés manuellement dans la première image et suivis dans les images restantes de la séquence. La reconnaissance de l'expression faciale est réalisée par la combinaison des UAs du haut et du bas du visage. [18]

- PANTIC et ROTHKRANTZ [20] utilisent des modèles de visage composés de points de vue doubles pour la classification des expressions faciales : la vue de face et la vue de côté. Après la segmentation automatique des traits du visage (yeux, sourcils et bouche), ils codent plusieurs points caractéristiques (tels que les coins des yeux, des coins de la bouche, etc.) en UA en utilisant un ensemble de règles. Ensuite, le FACS [32] est utilisé pour reconnaître les six expressions faciales universelles. La classification est effectuée en comparant la description codée par l'UA des expressions faciales de l'expression observée par rapport aux descripteurs de règle FACS. [18]

- PARDAS et al. [40] et TSAPATSOULIS et al. [41] proposent une description des six expressions faciales universelles utilisant l'ensemble de paramètres de définition faciale MPEG-4 (FDP) [27], TSAPATSOULIS et al. [41] utilisent tous les FAP (définis dans [27]) et proposent une classification basée sur un système d'inférence floue. Basé uniquement sur les sourcils et la bouche segmentation. PARDAS et al. [40] ont utilisé les FAP correspondants (8 pour les sourcils et 10 pour la bouche) pour le processus de classification. Celui-ci est basé sur un HMM système qui attribue à l'entrée l'expression dont la probabilité est la plus élevée. [18]

- COHEN et al. [12] ont développé un système basé sur un algorithme de suivi de visage non rigide [42] pour extraire les caractéristiques de mouvement locales. Ces caractéristiques de mouvement sont les entrées d'un classifieur de réseau bayésien utilisé pour reconnaître les six expressions faciales universelles. La représentation basée sur les caractéristiques nécessite une détection et un suivi précis et fiables des caractéristiques faciales afin de faire face aux variations d'éclairage, aux mouvements importants de la tête et à la rotation, ainsi que le changement de fonctionnalité non rigide. [18]

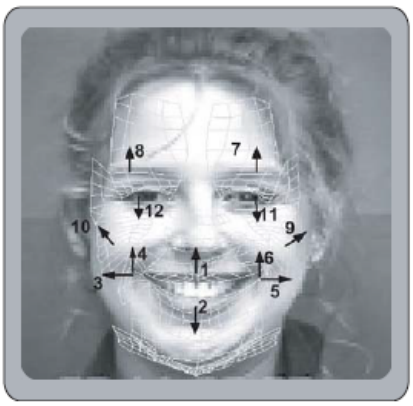


FIGURE 1.25 – Les mesures de mouvement du visage [12]

1.5 Utilisation des expressions faciales pour la détection de la fatigue

A côté du mouvement de la tête et des yeux, les expressions faciales sont l'une des plus importantes sélections visuelles, parmi les sélections visuelles existantes, on trouve le mouvement de paupières, la détection de la direction du regard, le mouvement de la tête, et les expressions

faciales. Ainsi les expressions faciales d'une personne en état de fatigue ou bien d'une personne au début de la fatigue sont souvent caractérisées par le traînement des muscles et le bâillement. Le développement des systèmes actifs pour alerter un conducteur et surveiller son niveau de vigilance est très important pour tenir les conducteurs réveillés et par conséquent réduire le nombre d'accidents. Certains efforts ont été rapportés dans la littérature sur le développement des systèmes actifs pour surveiller la fatigue en temps réel [43], mais la majorité de ces systèmes emploient une seule sélection visuelle ce qui est insuffisant.

On peut utiliser différentes techniques pour analyser l'épuisement du conducteur :

- Un ensemble de techniques place des capteurs sur des composants de véhicule standard, par exemple un volant, une pédale d'accélérateur, et analyse les signaux envoyés par ces capteurs pour détecter la somnolence.

- Un deuxième ensemble de techniques se concentre sur la mesure de signaux physiologiques tels que la fréquence cardiaque, le pouls et l'électroencéphalographie (EEG). Des chercheurs ont rapporté que plus le niveau de vigilance diminuait, plus la puissance EEG des bandes alpha et thêta augmentait. Donnant ainsi des indicateurs de somnolence. Cependant, cette méthode présente des inconvénients sur le plan pratique car elle oblige une personne à porter une casquette EEG pendant la conduite. [43]

- Un troisième ensemble de solutions est axé sur les systèmes de vision par ordinateur capables de détecter et de reconnaître le mouvement du visage et les changements d'apparence se produisant pendant la somnolence. Nous nous intéressons à cette catégorie et voilà quelques travaux connexes :

- Pour améliorer le confort du conducteur, des systèmes de mesure sans contact utilisant des techniques de vision par ordinateur ont été étudiés. En particulier, de nombreuses études ont porté sur les changements liés aux yeux. HIROSHI et al. (1994) [45], CHU et al.(2004) [46], D'ORAZIO et al. (2007) [47], AYUMI et al. (2009) [48], MARCO et al. (2010) [49], ARTEM et JONG (2012) [50] et GARCIA et al (2012) [51] ont estimé la somnolence sur la base de modifications de l'apparence des yeux. Cependant, les modifications induites par la somnolence des traits du visage ne concernent pas uniquement les yeux. De plus, l'apparence des yeux ne change que lorsqu'une personne est extrêmement somnolente, ce qui est le cas exact juste avant un accident. [44]

- VIDYAGOURI et UMAKANT (2013) [52], MOHAMMAD et MOHAMMAD (2011) [53], et PING et LIN (2012) [54] ont présenté plusieurs études prenant en compte d'autres traits du visage. Ils ont détecté la somnolence en utilisant des techniques de vision pour faire clignoter et bailler. Ils ont déterminé qu'un conducteur somnolait lorsqu'il bâillait ou que son clignement des yeux ralentissait, ce qui est très simple et reste discutable. [44]

- ESRA et al. (2007, 2010) [55] ont examiné diverses expressions faciales en détectant des unités d'action faciale. Cependant, leur objectif était de détecter les expressions du visage juste une minute avant un accident, qui ne sont clairement d{etectables qu'en utilisant des fonctions oculaires. Par conséquent, l'efficacit{e des traits d'expression n'est pas claire. De plus, il est trop tard pour avertir le conducteur une minute avant l'accident. Un système qui prévient au début de la somnolence est nécessaire. Après avoir analysé une vidéo d'expressions faciales capturées, SATORI et al. (2010) [56] et KENJI et al. (2010) [57] ont

mis l'accent sur le mouvement des points caractéristiques du visage sur les yeux, les sourcils et la bouche. [44]

- Récemment, GU & JI [43] a présenté l'une des premières études sur la fatigue intégrant certaines expressions faciales autres que le clignement des yeux. Leur étude fournit des informations sur les unit{es d'action en tant qu'entrée dans un réseau bayésien dynamique. Le réseau a été formé sur des sujets présentant un état de fatigue. Les segments vidéo ont été classés en deux étapes : inattention, bâillement ou endormissement. Pour prédire l'endormissement, on a utilisé des hochements de tête, des clignements de yeux, des rides du nez et des tenseurs de paupières. [44]

1.5.1 La performance du conducteur

La mesure objective de la manière dont le conducteur contrôle le véhicule est un élément clé de ses performances. Ces mesures constituent le moyen le moins invasif de détecter létat du conducteur car il n'y a pas d'interaction directe avec celui-ci. Mais d'autre part, ces mesures résultent directement de l'intervention du conducteur dans le contrôle du véhicule, telles que la direction, l'accélérateur et les freins. Les véhicules sont progressivement équipés de systèmes de détection des métriques du conducteur ; ces mesures sont donc particulièrement appropriées. Les mesures du conducteur se composent de la position de la voie, de l'avance et de l'angle du volant [58]. Les performances du conducteur peuvent être influencées par de nombreux facteurs, tels que l'expérience, les distractions et les conditions de conduite ; par conséquent, les performances de conduite ne sont pas nécessairement étroitement liées à l'état du conducteur.Cependant, les chercheurs développent encore leurs études sur l'impact de l'état du conducteur sur certains accidents.

1.5.2 L'état du conducteur

La capacité du conducteur à conduire peut-être déterminée par la façon dont il / elle se comporte au volant. Les comportements révélateurs de fatigue ou d'autres situations de conduite dangereuse, telles que la distraction, se manifestent sous forme de bâillements, cligner des yeux, fermer les yeux, bouger la tête, utiliser un appareil mobile ou avoir une vue dégagée. La première étape vers la détection de la somnolence en fonction des caractéristiques comportementales consiste à détecter le visage du conducteur. Dans ce cas, la zone de recherche de tout trait du visage sera réduite à la région du visage. Il existe de nombreuses techniques pour le traitement de détection de visage ; des images contenant des visages ont été développées dans différentes catégories de recherche telles que la reconnaissance des visages, le suivi des visages, l'estimation de la pose et la reconnaissance de l'expression. Pour construire un système capable d'analyser les informations incluses dans les images de visage, un algorithme de détection de visage robuste et efficace est requis. Et reste l'objectif de la détection de visage est de reconnaître toutes les régions d'image contenant un visage sans tenir compte de sa position, de son orientation et des conditions d'éclairage. Et pour notre recherche, nous nous concentrons sur ces trois détections :

Détection faciale

Pour la détection de visage elle-même, plusieurs approches ont été utilisées dans la littérature connexe. Les méthodes basées sur la connaissance [59] tentent de coder la connaissance humaine sur les caractéristiques d'un visage typique, telles que les relations entre les traits du visage, et de les utiliser comme moyen de détecter des visages dans une image. Le but des

approches invariantes des caractéristiques [60] est de trouver des caractéristiques structurelles du visage, telles que les sourcils, les yeux, le nez, la bouche et les cheveux, qui persistent sous diverses poses, points de vue ou éclairages et de les utiliser pour détecter les visages. Ces caractéristiques sont principalement extraites à l'aide de détecteurs de bord. Par exemple :

- SIROHEY[61] a proposé une méthode d'identification du visage à partir d'un fond encombré basé sur la segmentation. Le détecteur de bords Canny et les méthodes heuristiques servent de mappage de bords pour supprimer et regrouper des bords. Ensuite, l'ellipse est ajustée à la limite entre la région de la tête et l'arrière-plan et le visage sera détecté. Et une autre méthode de détection des visages basée sur la localisation des traits du visage est développée par GRAF et al. [62]. Dans cette méthode, les opérations morphologiques seront appliquées pour trouver les zones de forte intensité avec certaines formes. Sur la base de la valeur de crête importante de l'histogramme de l'image, le seuil adaptatif sera choisi pour créer des images binarisées. Ensuite, le composant connecté dans les images binarisées sera évalué en tant que candidats aux caractéristiques faciales afin de déterminer l'emplacement du visage. [58]

La texture des visages humains [63] ou la couleur de leur peau [64] se sont également révélées être des caractéristiques efficaces pouvant être utilisées pour la détection des visages.

- La méthode proposée par YING et al. [65], ils ont considéré la couleur de la peau comme la caractéristique la plus importante pouvant être séparée des autres parties de l'arrière-plan en utilisant le seuil de variance maximum des variétés. SAXE et FOULDS [66] ont développé un système de détection de visage qui utilise l'intersection d'histogrammes dans l'espace colorimétrique HSV pour mettre en évidence la région de la peau [64]. Dans leur méthode, un patch initial de couleur de peau sera utilisé pour lancer l'algorithme itératif. Afin de détecter la couleur de la peau, la méthode pr{esente un histogramme de contrôle, qui sera appliqué sur différents patchs de l'image et l'histogramme actuel à des fins de comparaison. Ensuite, la valeur de seuil sera assignée pour être comparée au résultat de la comparaison d'histogramme afin d'analyser la région de peau. [58]

Les méthodes d'appariement de modèles [67] stockent plusieurs motifs standard de visages différents pour décrire séparément le visage ou les traits du visage, et calculent les corrélations entre une image d'entrée et les motifs stockés afin de déterminer le degré de similitude de le motif à un visage.

- CRAW et al. [68] ont proposé la méthode suivante : la face frontale est détectée sur la base d'une correspondance de gabarit. Les arêtes extraites du filtrage Sobel seront regroupées pour localiser la face. Ensuite, la même procédure sera répétée pour trouver d'autres traits faciaux tels que les yeux, la bouche et le nez dans le visage candidat. Une autre méthode de détection de visage est décrite par A. SAMAL et al. [69] utilisant des silhouettes comme modèles pour la localisation des visages. L'analyse en composantes principales est utilisée pour collecter un ensemble de silhouettes de visage, représentées par un tableau de bits. Ensuite, la transformation de Hough et les silhouettes propres seront utilisées pour la localisation du visage. [58]

Dans les méthodes basées sur l'apparence, les modèles de visage sont appris à partir d'un ensemble d'images d'apprentissage, qui incluent la variabilité représentative de l'apparence du visage. Ces méthodes peuvent tirer parti de 15 réseaux de neurones, appliqués à de nombreux problèmes de reconnaissance de modèle, de machines à vecteurs de support, de classificateurs

Nive Bayes ou de modèles de Markov cachés, en tant qu'outils permettant d'évaluer l'appariement du modèle à la base de données de formation.

- EL-KHAMY et al. [70] dcrivent une méthode de reconnaissance du visage humain utilisant un algorithme de réseau de neurones et l'extraction de caractéristiques statistiques. Le bord de l'image du visage est détecté en appliquant un filtre Sobel lors de l'étape de prétraitement. Ensuite, l'image en noir et blanc à deux dimensions sera transformée en un vecteur à une dimension. Enfin, sept caractéeristiques seront extraites sur la base de l'analyse statistique. L'algorithme de propagation rapide en retour sera utilisée dans l'étape de reconnaissance. [58]

Les systèmes de détection de visage à la pointe de la technologie actuels reposent principalement sur l'utilisation de classificateurs. Le système de détection de visage le plus connu et le plus utilisé de cette catégorie est l'algorithme de détection de visage VIOLA-JONES [11]. Il est capable de détecter efficacement les faces frontales neutres car il a ét'e entrîané avec une grande base de données de visages.

- ERDEM et al. [71] combinaison de deux méthodes de détection de visage pour des résultats plus précis et fiables. La première méthode est le détecteur de visage basé sur les caractéristiques Haar, développé par VIOLA-JONES [11] pour les images en niveaux de gris, et la seconde méthode est un filtre de couleur de peau, qui fournit des informations complémentaires dans les images en couleur. Dans leur procédé, l'image passe par un détecteur de visages basé sur les caractéristiques de Haar, qui présente un nombre élevé de fausses détections et un faible nombre de visages manqués. Ensuite, la méthode de post-filtrage couleur de peau est utilisée pour éliminer bon nombre de ces fausses détections. [58]

Détection des yeux

Différentes méthodes de détection de la fatigue du conducteur sont mises œuvre par d'autres chercheurs qui se concentrent sur les changements et les mouvements oculaires. Ces techniques analysent les changements dans la direction du regard du conducteur, la fermeture des yeux et la fréquence de clignotement.

- À mesure que les gens deviennent somnolents, leurs schémas clignotants changent. SIGARY [72] a proposé une méthode de détection d'hypovigilance par traitement de la région de l'œil et sans étape de détection de l'œil explicite. Pour extraire les symptômes de fatigue et de distraction, une projection horizontale du demi-segment supérieur de l'image faciale est nécessaire. Pour déterminer la somnolence, le pourcentage de fermeture des yeux et de distance des paupières change avec le temps. [58]

- Une autre méthode de détection de la somnolence basée sur le mouvement des paupières a été proposée par LIU et al. [73] Dans leur méthode basée sur les changements de paupière à partir d'une image de différences temporelles, la situation de fatigue sera analysée. Le nombre de pixels blancs peut être utilisé pour le critère de jugement de fatigue dans la première étape. Ensuite, le nombre de pixels avec un changement positif dans l'image de différence de trois niveaux et le nombre de pixels avec un changement négatif entre l'image actuelle et l'image précédente représenteront le mouvement de la paupière d'ouvert à fermer, ce qui sera utile en tant qu'indicateur de somnolence.[58]

- OMIDYEGANEH et al. [74] ont utilisé une méthode de détection de la fatigue en appliquant la mesure de similarité structurelle pour trouver l'emplacement de l'œil. Dans leur méthode, la valeur de mesure de similarité structurelle sera évaluée entre -1 et 1. Lorsque deux images sont identiques, la valeur maximale gagnée sera 1 et lorsqu'il y aura quelques différences, le résultat sera -1. Ensuite, les projections horizontale et verticale seront appliquées sur la région des yeux pour déterminer le degré de fermeture des yeux et aligner la région des yeux détectée. [58]

- TABRIZI et ZOROOFI [75] ont proposé un moyen simple et non intrusif de détecter la fatigue en déterminant si l'œil est ouvert ou fermé. Dans leur algorithme, les trois étapes ont ét'e analysées, telles que la détermination des régions de l'œil par carte oculaire et la localisation du centre de la pupille par le centre de gravité de l'image de la région de l'œil. La dernière étape consiste à affiner le centre de la pupille et à détecter la limite de l'iris. Afin d'analyser l'état de l'œil afin de déterminer le stade de somnolence, un algorithme basé sur la chromatique a été utilisé, qui offre un meilleur taux de détection pour les yeux fermés. [58]

Détection des bâillements

- Afin de déterminer l'état de somnolence du conducteur, YUFENG et al. [76] ont proposé une méthode centrée sur la recherche du visage dans la première étape. Cette étape peut être déterminée en utilisant la différence en images entre deux images dans une séquence d'images. La méthode de seuil adaptatif peut être utilisée pour segmenter la zone en mouvement dans laquelle se trouvent le contour du visage et de la tête pour cette localisation. L'emplacement du menton et des narines est déterminé à l'étape suivante en fonction de l'emplacement du menton dans la moitié inférieure de la région du visage. La projection intégrale directionnelle sera utilisée pour trouver le milieu des narines. L'état de bâillement est déterminé en fonction du calcul de la distance entre le menton et l'emplacement du point médian des narines. [58]

- La méthode de détection de visage robuste et fiable basée sur la théorie de VIOLA-JONES [11] a été utilisée par WANG et SHI [77] pour limiter la zone de recherche de la bouche à la région du visage. La région de la bouche sera localisée sur la base d'une binarisation á plusieurs seuils dans l'espace d'intensité et du modèle gaussien dans l'espace colorimétrique RGB. Le coin de la lèvre sera trouvé en calculant la projection intégrale de la bouche dans la direction verticale. Les deux lignes traversant les limites des lèvres inférieure et supérieure résultant de la projection intégrale représentent l'ouverture de la bouche. Dans cette méthode, le stade de bâillement sera déterminé en déterminant le degré d'ouverture de la bouche en fonction du rapport de format du rectangle englobant la bouche. Une grande ouverture de la bouche au-dessus d'un seuil prédéfini pour un nombre continu d'images signifie que le conducteur est somnolent. [58]

- Selon la méthode d'ALIOUA et al.[78], les conditions de somnolence et de fatigue peuvent être déterminées par des détections au microsommeil et au bâillement, respectivement. La transformation de quantification moyenne locale successive est utilisée au début pour détecter l'emplacement du visage. Ensuite, le visage est divisé en fonction du classificateur Sparse Network of Winnows (SNW). La transformation circulaire de Hough sera appliquée sur les régions extraites des yeux et de la bouche afin de déterminer la situation de bâillement. La condition de bâillement est détectée si la grande zone sombre avec une forme circulaire qui montre la bouche est largement ouverte. [58]

- Le système de détection de fatigue développé par NAROLE et al. [79] repose sur les yeux et la bouche du conducteur. Après avoir trouvé une région du visage par segmentation de la couleur de la peau, la zone des yeux et de la bouche peut être détectée par un processus de seuillage et de segmentation. À cette fin, les pixels des lèvres peuvent être identifiés à l'aide du rapport Rouge / Vert, qui a des valeurs différentes pour la peau et pour les lèvres. À la fin, le réseau de neurones et l'algorithme génétique sont utilisés pour détecter la somnolence du conducteur. [58]

Conclusion

Cette revue de littérature nous a permis d'expliquer succinctement comment détecter automatiquement les expressions faciales d'une personne, en présentant les différentes approches qui ont été proposée.

En effet, l'analyse des expressions faciales est un problème intrigant que les humains résolvent avec une assez grande facilité. Nous avons identifié trois aspects importants de ce problème : la détection du visage, l'extraction des informations associées à chaque expression faciale, et la classification d'expressions faciales. Elle devrait servir comme un point de référence à n'importe quel système de vision automatique essayant de réaliser les mêmes fonctionnalités. Noter que la plus part des algorithmes proposés utilisent des seuils appliqués aux caractéristiques extraits en étudiant les différentes expressions. Ces valeurs dans la plupart des cas ne permettent pas forcément de différencier les différentes expressions, comme par exemple l'ouverture de la bouche pour sourire et son ouverture quand une personne baille.

Dans notre travail, nous proposons un système basé sur l'utilisation d'une base d'exemple qui donne la possibilité de déterminer les bons valeurs permettant de différencier au mieux les différentes expressions, par conséquent déterminer si une personne est en état de fatigue ou autres.

Chapitre 2

Réseaux de neurones

Introduction

L'informatique est la science du traitement automatique de l'information. Son développement est souvent confondu avec celui des machines de traitement : les ordinateurs.

Depuis les débuts (ENIAC 1946) jusqu'à aujourd'hui, les ordinateurs sont devenus de plus en plus puissants. Cependant, cette augmentation de puissance ne permet pas toujours de résoudre les tous les problèmes. En effet, les algorithmes utilisés peuvent être classés en deux classes : algorithmes classique, et algorithme basés sur la connaissance. Les algorithmes classique execute un certain nombre d'instruction prédéfinit par l'utilisateur. Par contre, les algorithmes basés sur la connaissance, dits algorithmes d'apprentissage, permettent eux même de trouver un modèle de prédiction en se basant sur des exemples, et cela dit qu'ils peuvent en quelque sorte s'adapter aux différentes situation (il sont plus intéligents). Elles sont largement utilisées dans la reconnaissance de formes (images ou signaux), le diagnostic, le contrôle qualité, la traduction automatique, de la compréhension du langage,etc.

Dans ce chapitre nous intéréssons à la deuxième catégorie d'algorithme. Nous allons donc décrit un certain type d'algorithmes d'apprentissage qui est les réseaux de neurones artificels.

2.1 Les neurones

La figure 2.1 reprend l'hypothèse proposée par de nombreux biologistes : pour recréer le comportement intelligent du cerveau, il faut s'appuyer sur son architecture, en fait, tenter de l'imiter. [80]

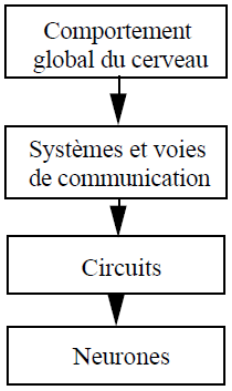


FIGURE 2.1 – Hypothèse biologique de génération d'un comportement intelligent [80]

Définition :

« Les réseaux de neurones artificiels sont des réseaux fortement connectés de processeurs élémentaires fonctionnant en parallèle. Chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit. Toute structure hiérarchique de réseaux est évidemment un réseau » [80].

Les réseaux de neurones formels sont à l'origine une tentative de modélisation mathématique du cerveau humain. Les premiers travaux datent de 1943 et sont l'œuvre de MM. MAC CULLOCH et PITTS (1943) [81]. Ils présentent un modèle assez simple pour les neurones et explorent les possibilités de ce modèle, tel que l'idée principale des réseaux de neurones " modernes " est la suivante :

» » On se donne une unité simple, un neurone, qui est capable de réaliser quelques calculs élémentaires. On relie ensuite entre elles un nombre important de ces unités et on essaye de d{eterminer la puissance de calcul du réseau ainsi obtenu. Il est important de noter que ces neurones manipulent des données numériques et non pas symboliques.

2.1.1 Neurone nature

Le cerveau se compose d'environ 1012 (mille milliards) de neurones interconnectés, avec 1000 à 10000 synapses (connexions) par neurone (Figure 2.2).
Les neurones ne sont pas tous identiques, leur forme et certaines caractéristiques permettent de les répartir en quelques grandes classes : [80]

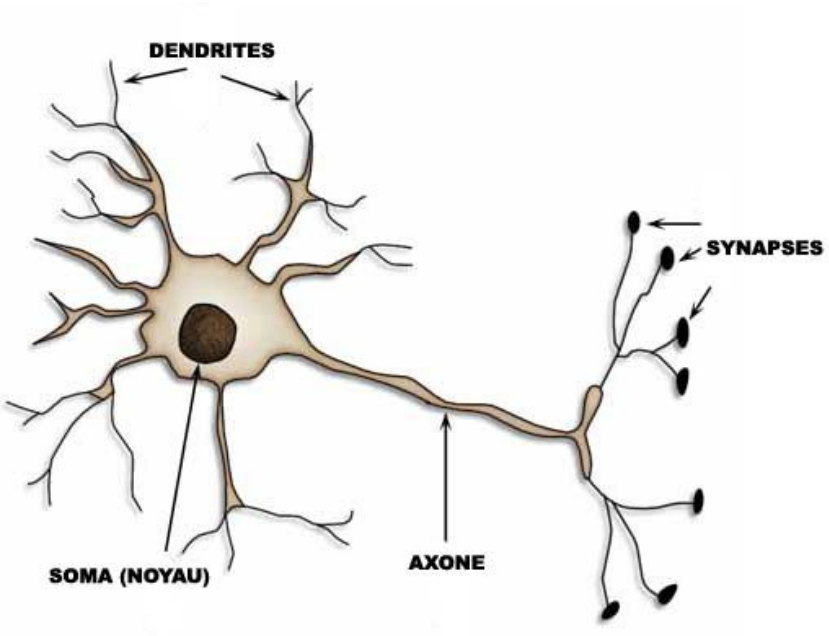


FIGURE 2.2 – Neurone naturel [82]

- **Le corps cellulaire** contient le noyau du neurone et effectue les transformations biochimiques nécessaires à la synthèse des enzymes et des autres molécules qui assurent la vie de la cellule. Ce corps cellulaire ayant une forme sphérique ou pyramidale, sa taille est de quelques microns de diamètre. [80]

- **Les dendrites** sont de fines extensions tubulaires qui se ramifient autour du neurone et forment une sorte de vaste arborescence. Les signaux envoyés au neurone sont capt{es

par les dendrites. Leur taille est de quelques dizaines de microns de longueur. [80]

- **L'axone** est la fibre nerveuse, sert de moyen de transport pour les signaux émis par le neurone. Il est plus long que les dendrites, et se ramifie à son extrémité où il se connecte aux dendrites des autres neurones. Les connexions entre deux neurones se font en des endroits appelés synapses où ils sont séparés par un peu espace synaptique de l'ordre d'un centième de microns. [80]

Chaque neurone est une unité autonome au sein du cerveau. Le neurone reçoit en continu des entrées. Le corps cellulaire du neurone est le centre de contrôle. C'est là que les informations reçues sont interprétées. La réponse, unique, à ces signaux est envoyée au travers de l'axone. L'axone fait synapse sur d'autres neurones (un millier). Le signal transmis peut avoir un effet excitateur ou inhibiteur. [80]

2.1.2 Neurone artificielle

Un neurone est considéré comme une unité de traitement élémentaire. Il reçoit les entrées et produit un résultat à la sortie. Le premier modèle du neurone artificiel est prposé par le neuropsychiatre MCCULLOCH et l'informaticien PITTS [81], ils représentent une abstraction du neurone physiologique (La figure 2.3 illustre ce modèle.).

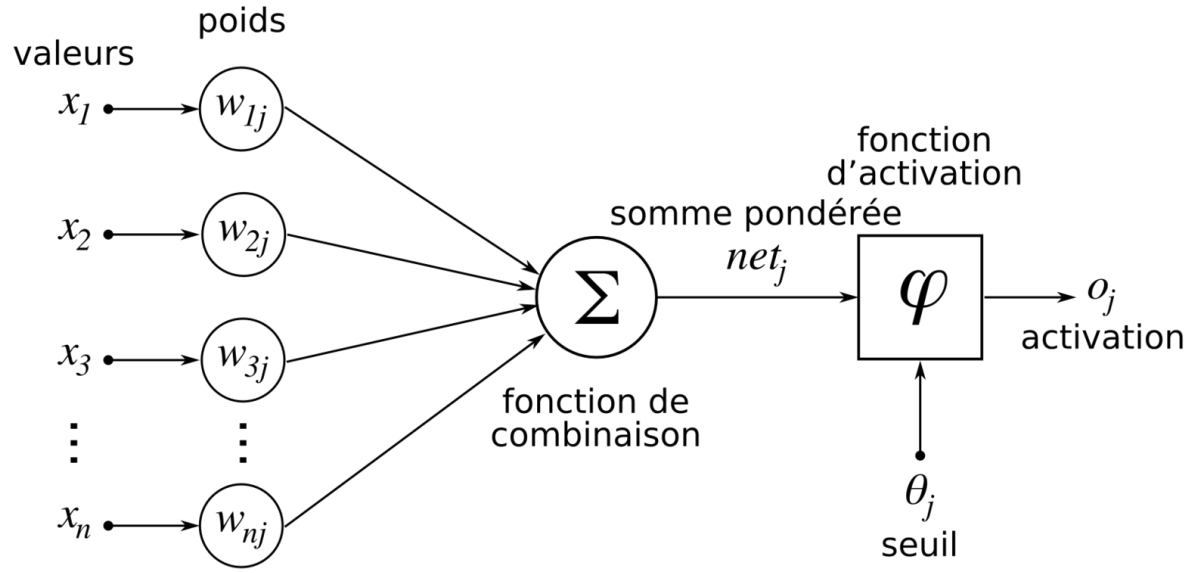


FIGURE 2.3 – Neurone formel de MCCULLOCH et PITTS [81]

Tels que :

- ($x1, x2, ..., xn$) : sont les entrées du neurone (signaux qui lui parviennent).

- ($w1j, w2j, ..., wnj$): les poids associés à chaque connexion.

- Σ : la somme pondérée des entrées (potentiel d'activation).

- θj : le seuil d'activation.

- φ: fonction d'activation qui est une fonction binaire.

$$\varphi = \sum_{i=1}^{n} wjxi - w0$$

- oj : la sortie du neurone (réponse du neurone « activé y=1, ou non activé y=0 »).

$$oj \begin{cases} 1 \; si \; \theta \geq 0 \\ 0 \; si \; \theta < 0 \end{cases}$$

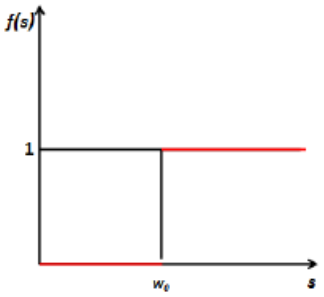
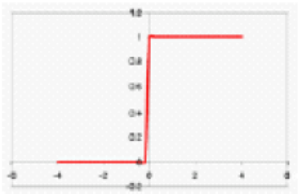
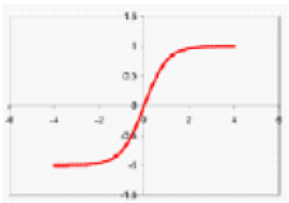


FIGURE 2.4 – Fonction d'activation d'un neurone formel [81]

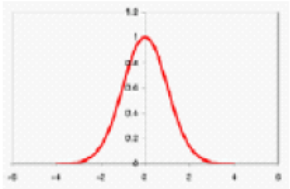
A partir de ce modèle ont été définis divers modèles de neurones et avec d'autres fonctions d'activations.



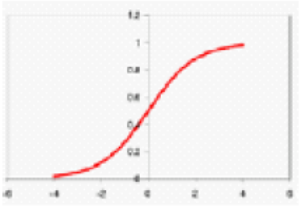
Fonction seuil



La tangente hyperbolique



Fonction Gaussienne



Le sigmoïde standard
(Fonction logistique)

FIGURE 2.5 – Exemples des fonctions d'activations [80]

2.1.3 Réseaux de neurones

Un réseau de neurone est un ensemble de neurones formels interconnectés, associés en couches et fonctionnant en parallèle. L'information donnée au réseau se propage couche par couche, de la couche d'entrée à la couche de sortie, en passant soit par aucune couche, par une couche ou par plusieurs couches intermédiaires. La capacité de traitement d'un réseau de neurones est stockée sous forme de poids d'interconnexions obtenus par un processus d'apprentissage. Les connexions entre les neurones qui composent le réseau décrivent la topologie du modèle. Elle peut être quelconque, mais le plus souvent il est possible de distinguer une certaine régularité.

Réseau multicouche (au singulier) : les neurones sont arrangés par couche. Il n'y a pas de connexion entre neurones d'une même couche et les connexions ne se font qu'avec les neurones des couches avales (figure 2.6). Habituellement, chaque neurone d'une couche est connecté à tous les neurones de la couche suivante et celle-ci seulement. Ceci nous permet d'introduire la notion de sens de parcours de l'information (de l'activation) au sein d'un réseau et donc définir les concepts de neurone d'entrée, neurone de sortie. Par extension, on appelle couche d'entrée l'ensemble des neurones d'entrée, couche de sortie l'ensemble des neurones de sortie. Les couches intermédiaires n'ayant aucun contact avec l'extérieur sont appelés couches cachées.

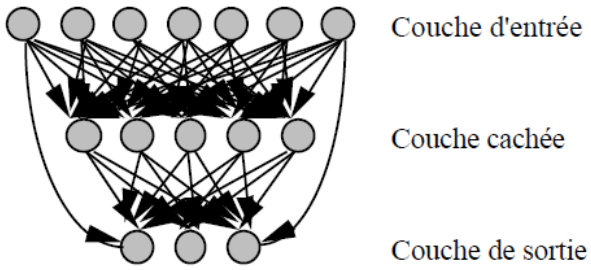


FIGURE 2.6 – Définition des couches d'un réseau multicouche [80]

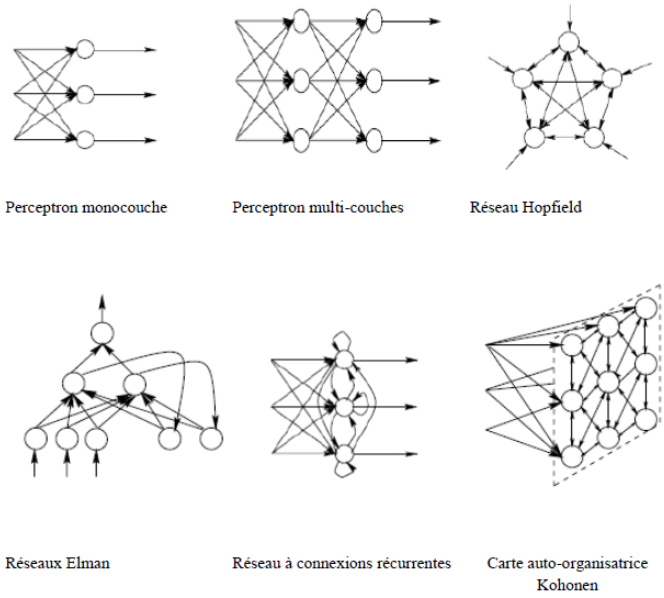


FIGURE 2.7 – Modèles des réseaux de neurones [80]

2.2 L'apprentissage en utilisant les réseaux de neurones

2.2.1 Définition

L'apprentissage automatique (en anglais *Machine Learning*) est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches statistiques pour donner aux ordinateurs la capacité d " apprendre " à partir de données, c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune. Plus largement, cela concerne la conception, l'analyse, le développement et l'implémentation de telles méthodes [83]. En général, l'objectif de l'apprentissage automatique est de comprendre la structure des données et de les intégrer dans des modèles qui peuvent être compris et utilisés par les tout le monde. L'apprentissage automatique comporte généralement deux phases. La première consiste à calculer un modèle à partir de données, appelées observations, qui sont disponibles et en nombre fini, lors de la phase de conception du système. L'estimation du modèle consiste à résoudre une tâche pratique, telle que traduire un discours, estimer une densité de probabilité, reconnaître la présence d'un chat dans une photographie ou participer à la conduite d'un véhicule autonome. Cette phase dite " d'apprentissage " ou " d'entraînement " est généralement réalisée préalablement à l'utilisation pratique du modèle. La seconde phase correspond à la mise en production : le modèle étant déterminé, de nouvelles données peuvent alors être soumises afin d'obtenir le résultat correspondant à la tâche souhaitée (prédiction).

Bien que l'apprentissage automatique soit un domaine de l'informatique, il diffère des approches informatiques traditionnelles. En effet, les algorithmes classiques sont des ensembles d'instructions explicitement programmées utilisées par les ordinateurs pour calculer ou résoudre des problèmes. Par contre, les algorithmes d'apprentissage automatique permettent aux ordinateurs de s'entraîner sur les entrées de données et utilisent l'analyse statistique pour produire des valeurs qui se situent dans une plage spécifique. Pour cette raison, l'apprentissage automatique facilite l'utilisation des ordinateurs dans la construction de modèles à partir de données d'échantillonnage afin d'automatiser les processus de prise de décision en fonction des données saisies.

L'utilisation des algorithmes d'apprentissage automatique sont utilisés de plus en plus. La technologie de reconnaissance faciale par exemple permet aux plateformes de médias sociaux d'aider les utilisateurs à marquer et partager des photos d'amis. La technologie de reconnaissance optique des caractères (OCR) convertit les images du texte en caractères. Les moteurs de recommandation, alimentés par l'apprentissage automatique, suggèrent les films ou émissions de télévision à regarder en fonction des préférences de l'utilisateur. Les voitures autonomes qui utiliseront l'apprentissage automatique pour naviguer seront bientôt disponibles pour les consommateurs, et nous trouvons que l'apprentissage est vraisemblablement la propriété la plus intéressante des réseaux neuronaux. Elle ne concerne cependant pas tous les modèles, mais les plus utilisés [84] comme on peut dire :

« L'apprentissage est une phase du développement d'un réseau de neurones durant laquelle le comportement du réseau est modifié jusqu'à l'obtention du comportement désiré. L'apprentissage neuronal fait appel à des exemples de comportement. » [80]

Dans le cas des réseaux de neurones artificiels, on ajoute souvent à la description du modèle l'algorithme d'apprentissage. Le modèle sans apprentissage présente en effet peu d'intérêt. Dans la majorité des algorithmes actuels, les variables modifiées pendant l'apprentissage sont les poids des connexions. L'apprentissage est la modification des poids du réseau dans l'optique d'accorder la réponse du réseau aux exemples et à l'expérience. Il est souvent impossible

de décider à priori des valeurs des poids des connexions d'un réseau pour une application donnée. A l'issu de l'apprentissage, les poids sont fixés : c'est alors la phase d'utilisation. Certains modèles de réseaux sont improprement dénommés à apprentissage permanent. Dans ce cas il est vrai que l'apprentissage ne s'arrête jamais, cependant on peut toujours distinguer une phase d'apprentissage (en fait de remise à jour du comportement) et une phase d'utilisation. Cette technique permet de conserver au réseau un comportement adapté malgré les fluctuations dans les données d'entrées. [80]

2.2.2 Types d'apprentissage

Dans l'apprentissage automatique, les tâches sont généralement classées en grandes catégories. Ces catégories sont basées sur la façon dont l'apprentissage est reçu ou comment le feedback sur l'apprentissage est donné au système développé. Deux des méthodes d'apprentissage automatique les plus largement adoptées sont l'apprentissage **supervisé** qui forme des algorithmes basés sur des données d'entrée et de sortie étiquetées par l'homme et l'apprentissage **non supervisé** qui ne fournit pas à l'algorithme des données étiquetées pour lui permettre de trouver une structure et de découvrir une logique dans données entrées. Explorons donc ces méthodes plus en détail.

Apprentissage supervisé (rétro-propagation)

Dans l'apprentissage supervisé (*Supervised Learning* en anglais), les exemples d'entrées sont étiquetés (sorties connnues). Dans ce type de méthode,les algorithmes puisse « apprendre » en comparant sa sortie réelle avec les sorties « enseignées » pour trouver des erreurs et modifier le modèle en conséquence. L'apprentissage supervisé utilise donc des modèles pour prédire les valeurs d'étiquettes sur des données non étiquetées supplémentaires. [83]

Les exemples annotés constituent une base d'apprentissage, et la fonction de prédiction apprise peut aussi être appelée « hypothèse » ou « modèle ». On suppose cette base d'apprentissage représentative d'une population d'échantillons plus large et le but des méthodes d'apprentissage supervisé est de bien généraliser, c'est-à-dire d'apprendre une fonction qui fasse des prédictions correctes sur des données non présentes dans l'ensemble d'apprentissage. [84]

Dans ce type apprentissage, l'environnement fournit au réseau des couples entrées/sorties qui vont former un jeu d'entraînement. Le réseau va mettre à jour ses poids en utilisant la différence entre le résultat qu'il a calculé, en fonction des entrées fournies, et la réponse attendue en sortie (la réponse donnée). Ainsi, le réseau va se modifier jusqu'à ce qu'il trouve la bonne sortie. On l'appelle le mode supervisé car l'environnement doit fournir la sortie correcte pour chaque jeu d'entrées, jouant ainsi le rôle du superviseur, par exemples : SVM,KNN, etc.

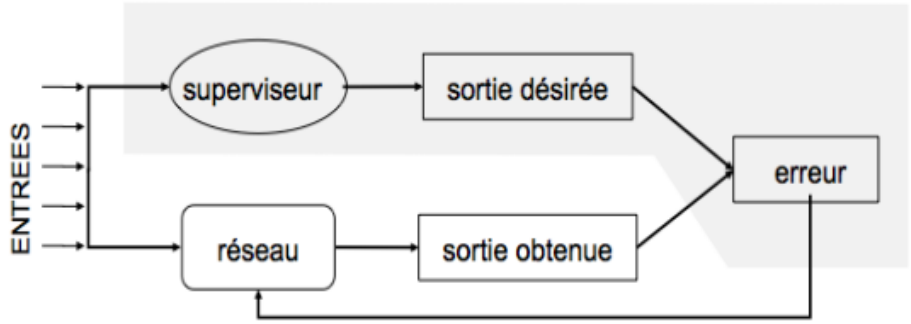


FIGURE 2.8 – L'apprentissage supervisé [85]

Apprentissage non supervisé (auto-organisationnel)

L'apprentissage non supervisé (appelé en anglais *Unsupervised Lerarning*) est un autre type d'apprentissage automatique. Il s'agit, pour un logiciel, de trouver des structures sous-jacentes à partir de données non étiquetées. Puisque les données ne sont pas étiquetées, il n'est pas possible d'affecter au résultat de l'algorithme utilisé un score d'adéquation. Cette absence d'étiquetage (ou d'annotation) est ce qui distingue les tâches d'apprentissage non-supervisé des tâches d'apprentissage supervisé [83], par exemple le K-means, etc.

Dans l'apprentissage non supervisé, les données sont non étiquetées, de sorte que l'algorithme d'apprentissage trouve tout seul des points communs parmi ses données d'entrée. Les données non étiquetées étant plus abondantes que les données étiquetées, les méthodes d'apprentissage automatique qui facilitent l'apprentissage non supervisé sont particulièrement utiles. L'objectif de l'apprentissage non supervisé peut être aussi simple que de découvrir des modèles cachés dans un ensemble de données, mais il peut aussi avoir un objectif d'apprentissage des caractéristiques, qui permet à la machine intelligente de découvrir automatiquement les représentations nécessaires pour classer les données brutes. [84]

Sans une réponse « correcte », les méthodes d'apprentissage non supervisées peuvent examiner des données complexes, plus expansives et apparemment sans point commun, afin de les organiser de manière potentiellement significative. L'apprentissage non supervisé est souvent utilisé pour la détection d'anomalies, y compris pour les achats frauduleux de cartes de crédit et les systèmes de recommandation qui conseille sur les produits à acheter ensuite. [84]

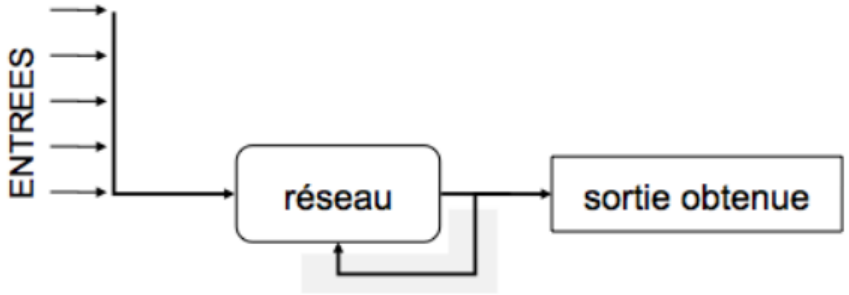


FIGURE 2.9 – L'apprentissage non supervisé [85]

2.2.3 Stratégie d'apprentissage

L'apprentissage au sein des différentes architectures dépond de l'architecture du réseau et de l'environnement du problème. Les deux règles d'apprentissage pour mettre (faire) corriger les poids d'un neurone ne concernent qu'un neurone seul. Ces règles peuvent servir pour mettre à jour les poids d'un neurone, de certains de neurones, mais ne peuvent être généralisées et s'appliquer à n'importe quelle architecture. Chaque architecture possède ses spécificités et nécessite une règle d'adaptation des poids qui lui est propre. L'apprentissage n'est pas modélisable dans le cadre de la logique déductive : celle-ci en effet procède à partir de connaissances déjà établis dont on tire des connaissances dérivées. Or il s'agit ici de la démarche inverse : par observations limitées tirer des généralisations plausibles.

- **La mémorisation** : le fait d'assimiler sous une forme dense des exemples éventuellement nombreux.

- **La généralisation** : le fait d'être capable, grâce aux exemples appris, de traiter des exemples distincts, encore non rencontrés, mais similaires. [85]

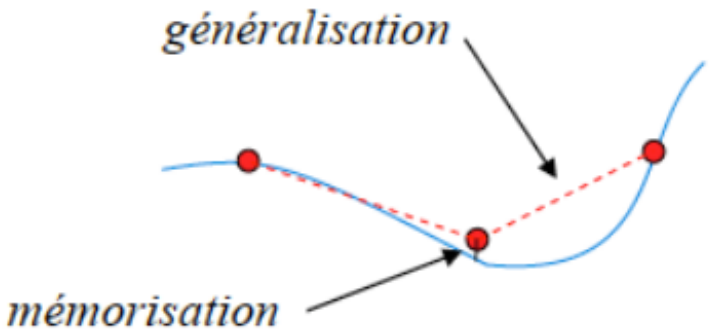


FIGURE 2.10 – La mémorisation et la généralisation [85]

Le concept de la mémorisation et celui de la généralisation sont partiellement en opposition. Si on privilégie l'un, on élaborera un système qui ne traitera pas forcement de façon très efficace l'autre. Il faut trouver un compromis en choisissant un coefficient d'apprentissage satisfaisant (par essai-erreur). [85]

Conclusion

Dans ce chapitre, nous avons présenté les concepts de bases liés aux réseau de neurones.

Dans nos jours, l'utilisation de ce genre d'algorithmes est en pleine expansion, et cela dû aux résultats qu'il fournit. Nous allons faire de même, et utiliser ce concept dans ce travail pour concevoir un système de détection de fatigue. Nous donnons les détails de conception de ce système dans le chapitre suivant.

Chapitre 3

Conception de système

Introduction

Dans les chapitres précédents, nous avons présenté une étude théorique sur l'analyse des expressions faciales ainsi que son possible utilisation pour détection de la fatigue chez les conducteurs. Dans ce chapitre, nous allons mettre en pratique ce concept et proposer un système de détection de fatigue qui a comme but de surveiller le conducteur pour détecter un éventuel état de fatigue. Il est organisé comme suit :

Dans la première partie, nous décrivons les grandes lignes de notre modèle proposé. Dans la deuxième partie, nous décrivons en détaille la conception du modèle propose en donnant les détails de chaque module de la conception. Nous définissons par la suite les paramètres et les détails techniques relatifs à l'analyse des expressions ainsi que la méthode d'apprentissage utilisée.

3.1 Conception générale de notre système proposé

L'objectif de cette application est d'analyser les expressions faciales d'un être humain, dont le but de détecter un possible état de fatigue. Comme on peut le savoir, chaque individu a une manière d'exprimer la fatigue via des expressions faciales, comme par exemple, la fermeture des yeux pendant quelques secondes ou involontairement, le bâillement répétitif avec des périodes espacées ou convergentes, ou alors le suivi de l'état des pupilles des deux yeux durant un moment. Dans ce travail, nous cherchons à proposer une méthode capable de détecter le visage d'une personne, d'extraire les points caractéristiques à partir de ce dernier, et puis par la suite interpréter et reconnaitre les expressions permettant de détecter un état fatigue ou pas. Pour cela, nous allons utiliser un système d'apprentissage en utilisant une base d'exemples. Les exemples sont les frames extraits à partir des différentes vidéos contenant dans la base de données utilisée. La base de données contient une centaine de vidéos, qui sont des séquences montrant un conducteur avec deux état possible : fatigué, normal. Pour une nouvelle vidéo (temps réel ou séquences enregistrée) non classé, nous utilisons le modèle d'apprentissage construit pour prédire l'état du conducteur parmi les deux états possibles.

La figure 3.1 décrit le processus d'exécution de notre système :

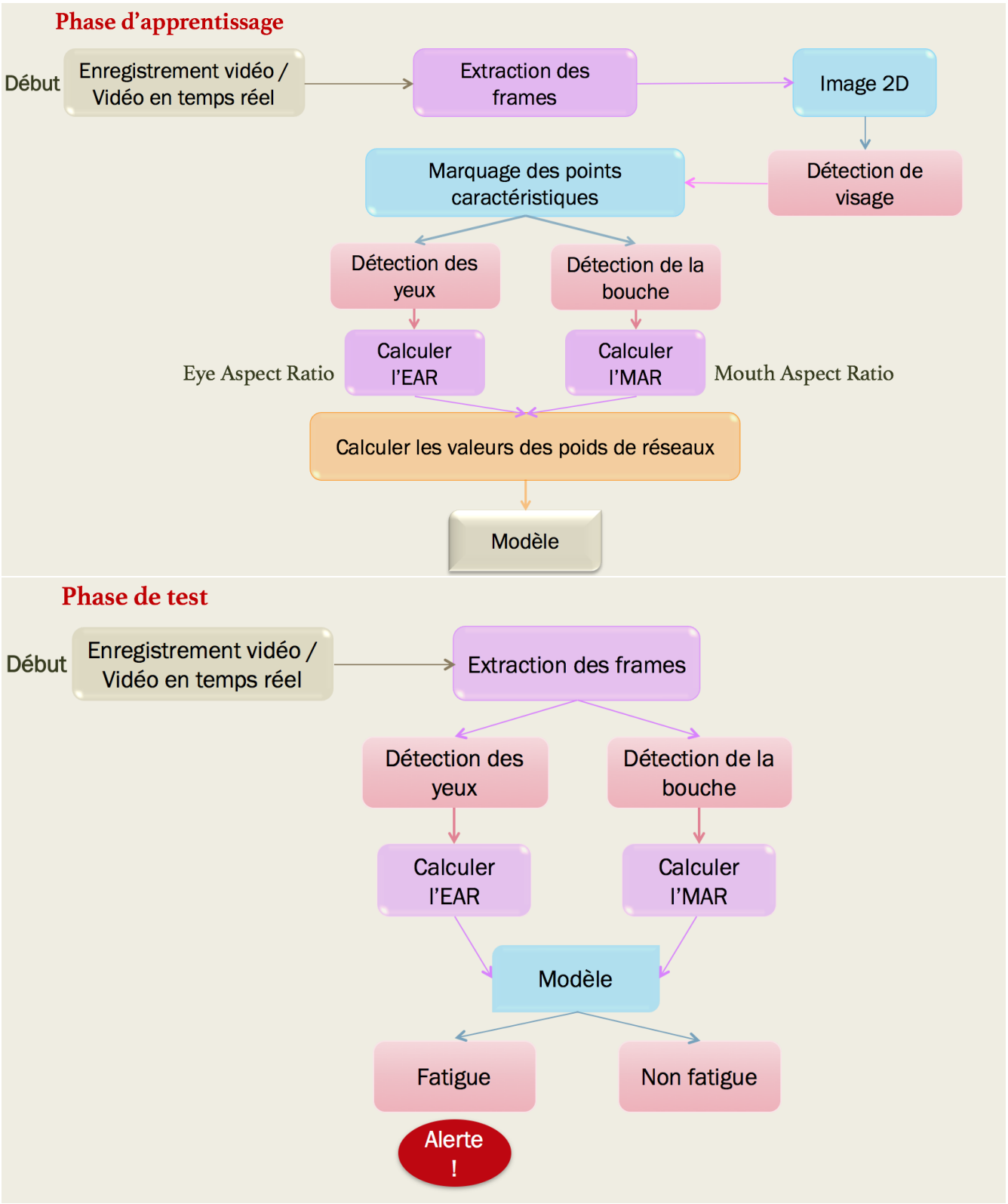


FIGURE 3.1 – Schéma globale du système

3.2 Conception détaillé de système

Dans ce qui suit, nous allons détailler chacune des étapes décrites ci-dessus.

3.2.1 Acquisition des données

L'acquisition des données est la première étape de notre système. Elle permet d'acquérir des données à partir du monde réel, et de concevoir une base d'exemples (apprentissage). Autrement dit, cette étape sert à transformer des séquences d'observations (vidéo), décrivant l'état des personnes en des images (notre base d'exemples), qui sera utilisées dans les étapes suivantes.

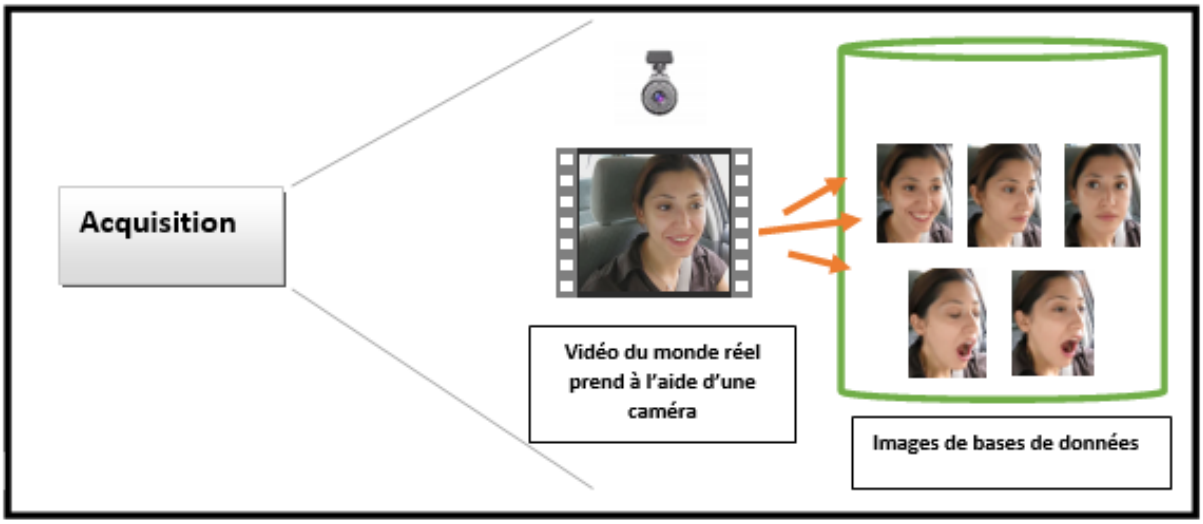


FIGURE 3.2 – Processus d'acquisition des images servant comme une base d'exemples.

3.2.2 Détection du visage

L'étape de détection du visage est une étape primordiale permettant de chercher la présence d'un visage dans une image. Une fois détecté, on doit extraire et récupérer les points caractéristiques représentant les yeux, et la bouche qui seront utilisés ultérieurement.

L'objectif de la détection de visage est d'identifier toutes les régions d'image comprenant un visage, quelles que soient sa position, son orientation et les conditions d'éclairage. Un tel problème pose des problèmes car les visages ne sont pas rigides et présentent une grande variabilité en taille, forme, couleur et texture. On suppose en principe que la caméra est installée à l'intérieur du véhicule face au conducteur selon un angle fixe. Par conséquent, le problème de la pose relative du visage de la caméra est moins difficile dans notre cas, alors que la position de la tête peut toujours varier d'un conducteur à l'autre. Il existe également une grande variabilité entre les visages, notamment leur forme, leur couleur et leur taille. La présence de traits du visage tels que barbes, moustaches et lunettes peut également faire toute la différence. L'autre facteur important est constitué par les conditions d'éclairage. Celles-ci sont principalement affectées par la lumière ambiante, qui peut changer en fonction des conditions météorologiques et de l'heure de prise des captures vidéo.

Étant donné que nous traitons des séquences vidéo, la phase de détection implique implicitement la phase de suivi du visage dans la scène, puisque nous traitons la séquence vidéo image par image. Cette étape se décomposée de trois tâches à savoir, la détection du visage à l'aide de la méthode « VIOLA & JONES [11] », l'extraction des traits faciaux ou bien les points caractéristiques (yeux, bouche) et le suivi des déplacements du visage dans la scène.

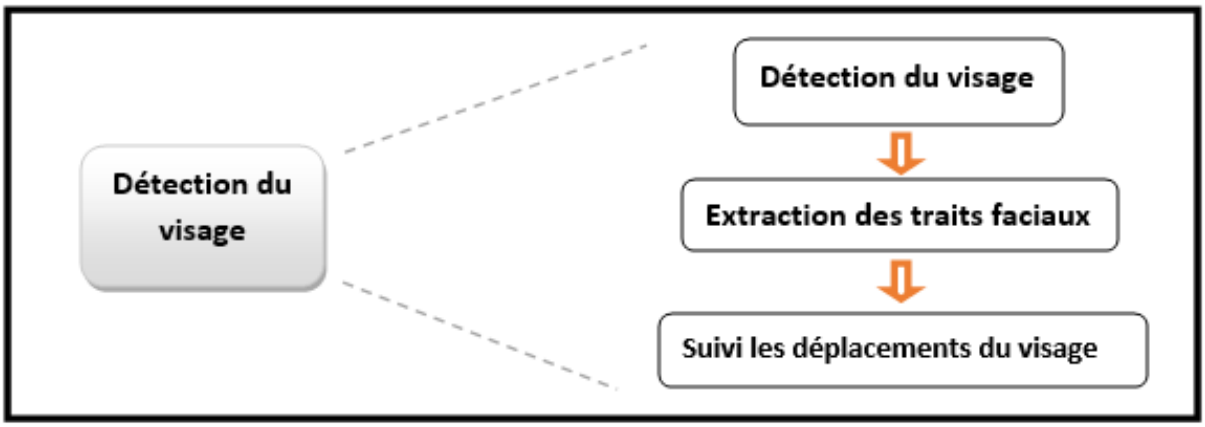


FIGURE 3.3 – Processus de détection du visage.

Méthode de Viola & Jones [11]

La méthode « VIOLA & JONES [11] » a été proposée au départ pour la détection de visages dans une image numérique ou séquence vidéo puis utilisée pour détecter d'autres objets comme les voitures... La bibliothèque *OpenCV* présente une implémentation de cette méthode sous le nom « détecteur en cascades de **Haar** ». Le point fort de cette méthode est la rapidité de détection ce qui la rend capable de s'exécuter en temps réel et de répondre aux exigences du traitement vidéo. Toutefois, elle présente quelques limites telles que la difficulté de détection simultanée du plusieurs vues de même objet, la durée nécessaire à la phase d'apprentissage des cascades est relativement assez grande et le nombre d'échantillons d'apprentissage est important.

3.2.3 La détection des points caractéristiques dans un visage

L'étape de caractérisation consiste à dégager les caractéristiques pertinentes de l'expression faciale étudiée et d'éliminer les informations redondantes.

Puisque nous intéressons à la fatigue, nous cherchons à extraire des points caractéristiques spécifiques dans le visage : ce sont ceux des yeux et la bouche. La détection de ces points est implémentée dans la librairie **dlib** [86] utilisé sous le langage *Python*. Elle permet de produit 68 point 2D de coordonnées (x, y) qui cartographient des structures faciales spécifiques. Ces points sont stockés dans un tableau indexé. Voici donc les indices de chaque point parmi les 68 points (figure 3.4) :

FIGURE 3.4 – Le résultat de détection des points caractéristiques à partir du visage en utilisant **dlib** [86]

Dans notre implémentation, nous nous intéressons à œil droit, œil gauche et à la bouche.

La première tâche de cette étape est la détection de ces caractéristiques par trouver les points caractéristiques (Landmarks) des yeux et de la bouche dans le visage.

La deuxième tâche est le maintien de ces points propres générés qui disposent de l'information caractéristique de visages. Et la dernière tâche est de sauvegardé ces repères pour le calcul de la matrice des poids dans l'espace engendré par les visages propres retenus.

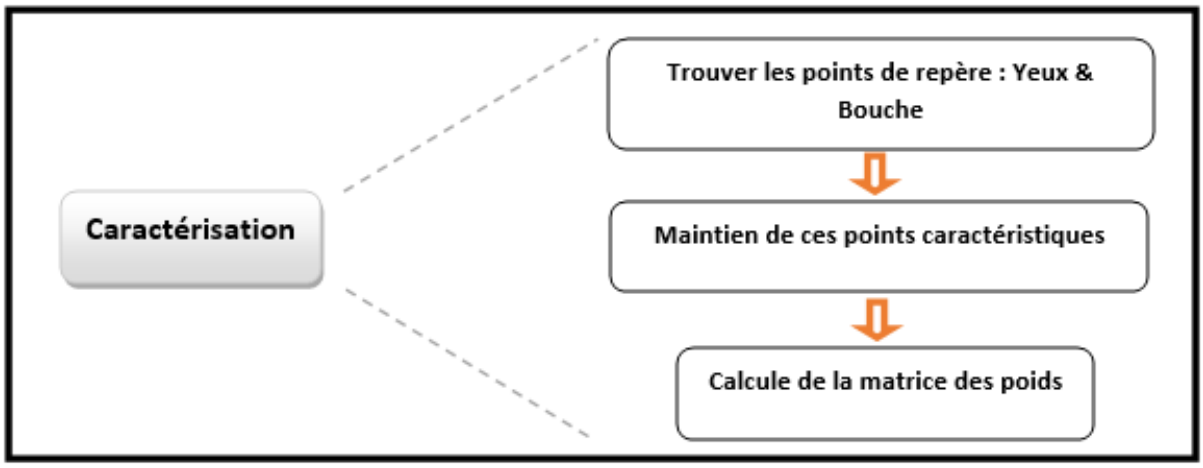


FIGURE 3.5 – Détail de l'étape de Caractérisation

Détection des yeux fermés

Comme on peut le savoir, la fermeture des yeux représente un signe principal des symptômes de la somnolence. Dans **dlib** [86], chaque œil est représenté par 6 points de coordonnées (x, y) comme indiqué dans la figure 3.6.

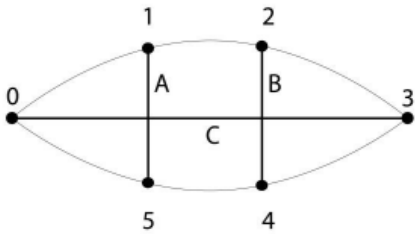


FIGURE 3.6 – Région d'un œil représentée par les points caractéristiques

Pour décrire et représenter les points caractéristiques décrivant les yeux, nous allons utiliser une mesure appelée Rapport d'Aspect Oculaire, en anglais « Eye Aspect Ratio (EAR) ». Cette mesure représente le rapport entre la largeur et la hauteur de l'œil. La valeur de cette mesure va nous permettre de définir et de mesurer l'ouverture de l'œil, plus elle grande, plus l'œil est ouvert. Elle va donc nous permettre plus tard de caractériser la base des exemples (images) pour trouver un modèle d'apprentissage permettant de prédire l'état d'une nouvelle instance. La valeur d'EAR est calculé comme suit :

$$Ae = d(eye[1], eye[5]) \tag{3.1}$$

$$Be = d(eye[2], eye[4]) \tag{3.2}$$

$$Ce = d(eye[0], eye[3]) \tag{3.3}$$

Tel que *Ae* et *Be* mesurent respectivement la distance verticale de l'œil et *Ce* calcule les dimensions horizontales de l'œil, alors :

$$EAR = (Ae + Be)/(2.0 * Ce) \tag{3.4}$$

- *Seuil oculaire (Te)*
  C'est la valeur qui permet de définir l'état de l'œil. Si la valeur du EAR est en dessous de cette valeur, l'œil est considéré comme fermé.

$$EyeState = Closed, EAR < Te$$
$$= Open, EAR \geq Te$$

  Dans la littérature, la plupart des travaux essaie de trouver cette valeur empiriquement [87]. La valeur de *Te* peut être déterminée par essais et erreurs, en recherchant les différentes valeurs de *Te*, de sorte que le système puisse correctement classer les différentes instances la valeur la plus représentative est [87] : **Te = 0,3**. Dans ce travail nous allons essayer de déterminer et calculer cette valeur autrement. Pour cela nous allons utiliser un algorithme d'apprentissage permettant de calculer les différentes valeurs du EAR et pour chaque état parmi les deux états possibles, puis comparer les résultats.

Détection de la bouche et du bâillement

Après avoir détecté l'œil et calculé la valeur d'EAR, la prochaine étape est la détection des bâillements qui consiste à localiser la bouche et les lèvres. Pour ce faire, la zone de la bouche marquée par des point caractéristiques sont calculés aussi par **dlib** [86], ils sont indiqués dans la figure 3.7.

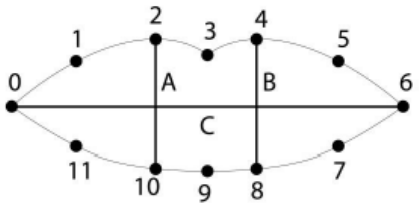


FIGURE 3.7 – Région de la bouche représentée par les points caractéristiques.

De la même façon, nous calculons une mesure appeleé Rapport d'Aspect de la Bouche, en anglais « Mouth Aspect Ration (MAR) », qui définie le rapport entre la hauteur et la largeur de la bouche. Sur la base des valeurs du MAR nous pouvons savoir si la bouche est ouverte (dans un état de bâillement), ou fermée. Cette valeur est calculée comme suit :

$$Am = d(mouth[2], mouth[10]) \tag{3.5}$$

$$Bm = d(mouth[4], mouth[8]) \tag{3.6}$$

$$Cm = d(mouth[0], mouth[6]) \tag{3.7}$$

Am et *Bm* mesurent l'ouverture verticale de la bouche et *Cm* calcule la largeur de la bouche, alors :

$$MAR = (Am + Bm)/(2.0 * Cm) \tag{3.8}$$

- *Seuil de la bouche (Tm)*
  C'est la valeur seuil de MAR qui permet de définir si la bouche est ouverte en situation de baillement ou autres. Si la valeur d'MAR dépasse cette valeur, la bouche est considérée comme étant dans un état de bâillement.

$$MouthState = Yawn, MAR > Tm$$
$$= Closed/Talking, MAR \leq Tm$$

Dans la littérature, la valeur de *Tm* est déterminée par essais et erreurs, en recherchant différentes valeurs de *Tm*, de sorte que le système puisse correctement classifier une instance de bâillement et de bouche fermée. Lors de la détermination du seuil optimal pour la bouche béante, un problème de détection de faux positifs a été rencontré, dans lequel le conducteur pouvait avoir la bouche ouverte tout en parlant. Cela ne doit pas être identifié comme un état de bâillement. Étant donné que la bouche humaine s'ouvre beaucoup plus largement sur un certain nombre d'images consécutives en bâillant que de parler, le seuil est sélectionné de manière à être suffisamment élevé pour ignorer la bouche ouverte lorsque vous parlez dans des conditions normales [87]. Par cette méthode nous obtenons une valeur optimale de Tm égale à : **Tm = 0,9**. De la même façon pour le seuil du MAR, nous allons essayer de la calculée en utilisant une méthode d'apprentissage sur la base des valeurs du MAR connus.

3.2.4 Algorithme de détection de fatigue

Notre système proposé commence par capturer la vidéo du conducteur, et la traitée image par image. Dans chaque image le visage du conducteur est détecté. La détection des visages est réalisée à l'aide d'un classifier en cascade de **Haar** formé à reconnaître les visages humains [11]. Lors de l'identification de la région rectangle du visage, l'étape suivante consiste à déterminer l'emplacement des principaux points caractéristiques du visage. Cette détection est réalisée en utilisant un ensemble d'arbres de régression [88] formés pour estimer la position de repères clés (les points caractéristiques), tels que la région des yeux et de la bouche pour notre système. L'ensemble est formé à l'aide d'un ensemble de données étiquetées d'images qui spécifient les coordonnées des régions entourant chaque trait du visage. Ce prédicateur de référence permet une détection rapide et précise en temps réel. La prochaine étape de la détermination du niveau de fatigue du conducteur consiste à analyser l'état des points caractéristiques. Pour analyser l'état de l'œil, considérons la forme convexe formée autour de la région de l'œil et calculons son rapport d'aspect, comme illustré dans la section précédente. Si l'EAR tombe en dessous de *Te*, l'œil est considéré comme fermé. De même pour la bouche, le rapport d'aspect bouche est calculé. Si ce MAR dépasse *Tm*, la bouche est considérée ouverte en bâillant. En fonction de la fréquence des bâillements, le système détermine le niveau de fatigue et émet les alertes appropriées.

3.2.5 Apprentissage

Imaginons que le réseau de neurones soit utilisé pour reconnaître les photos qui comportent au moins quelques signes de fatigue. Pour pouvoir identifier les conducteurs fatigues sur les photos, l'algorithme doit être en mesure de distinguer les différents types de signes, et de reconnaître un conducteur fatigue de manière précise quel que soit l'angle sous lequel il est photographié. Afin de déterminer le seuil pour les valeurs d'EAR et du MAR, notre système exige une phase d'apprentissage. A partir d'une base d'exemples (valeurs EAR, MAR), on déduit le bon couplage des valeurs EAR-MAR permettant de classifier l'état du conducteur.

Afin d'y parvenir, notre réseau de neurones doit être entraîné. Pour ce faire, il est nécessaire de compiler un ensemble d'images d'entraînement pour pratiquer le Deep Learning. Cet ensemble va regrouper des milliers de photos de conducteurs différents, mélangés avec des images de conducteurs qui ne sont pas fatigues. Ces images sont ensuite converties en données et transférées sur le réseau. Les neurones artificiels assignent ensuite un poids aux différents éléments. La couche finale de neurones va alors rassembler les différentes informations pour déduire s'il s'agit ou non d'un conducteur fatigue.

FIGURE 3.8 – Détail de l'étape d'Apprentissage.

3.2.6 Reconnaissance

Au cours de la phase de reconnaissance, il y aura une comparaison entre les caractéristiques acquises de l'expression requête avec celles issues de l'apprentissage. C'est une comparaison du vecteur de poids test par rapport aux informations sauvegardées dans la matrice des poids.

L'objectif de cette étape est la détermination qu'une expression de fatigue existe bien dans notre référence. Pour le cas de notre approche, nous avons recourt à estimation comme étant comparateur entre le vecteur requête et la référence. Avoir une valeur minimale entre le vecteur test et une parmi les colonnes de la matrice de poids indique que l'image correspondante à ce dernier vecteur est la plus proche de l'image test et sa valeur est celle de la valeur de test.

Le réseau de neurones va ensuite comparer cette réponse aux bonnes réponses indiquées par les humains. Si les réponses correspondent, le réseau garde cette réussite en mémoire et s'en servira plus tard pour reconnaître les conducteurs fatigues. Dans le cas contraire, le réseau prend note de son erreur et ajuste le poids placé sur les différents neurones pour corriger son erreur. Le processus est répété des milliers de fois jusqu'à ce que le réseau soit capable de reconnaître un conducteur fatigue sur un frame dans toutes les circonstances et c'est la technique d'apprentissage qu'est appelée « Supervised Learning » ou apprentissage supervisé.

FIGURE 3.9 – Détail de l'étape de Reconnaissance.

Conclusion

Ce travail met l'accent sur l'étude des expressions faciales qui est relié à la fatigue chez les conducteurs. L'étude présentée dans ce chapitre constitue l'analyse des expressions faciales à partir de la détection de visage et l'extraction ses points caractéristiques qui signent à la fatigue c.-à-d. les yeux et la bouche à but d'identifier si le conducteur est en état de sommeille ou il est bien réveillé, en utilisant une stratégie d'apprentissage pour facilite la réponse et la prédiction de leur état au futur.

Les détails de l'implémentation notre application ainsi que les résultats obtenus seront présentés dans le chapitre suivant.

Chapitre 4

Implémentation et résultats

Introduction

Dans ce chapitre, nous allons décrire la mise en œuvre des différentes étapes de notre système proposé pour la détection de fatigue. Dans un premier temps, nous allons commençons par présenter le langage de programmation qui a été utilisé dans le développement de notre application, ensuite nous détaillons les structures de données utilisées, et enfin nous présentons notre l'algorithme utilisé, en expliquant le principe de détection de visage réaliser en temps réel, ainsi la détection des points caractéristiques pour détecter l'état de fatigue.

La dernière partie concerne les résultats obtenus, et une petite discussion est donnée à la fin de ce chapitre.

4.1 Outils utilisés

* **Dispositifs matériels**

Notre configuration matérielle inclut les dispositifs suivants :

- Modèle : MSI GT70 - Dragon
- Processeur : Intel ^{R} Core ^{TM} i7-3630QM CPU $^{@}$2.40 GHz
- Carte graphique : Nvidia GeForce GTX 680M
- Mémoire : 10 Go
- Ecran : 17.3"
- Système d'exploitation : Windows 10, 64 bits
- Capteur Webcam

* **Outils logiciels**

- ***Présentation de langage utilisé et motivation du choix***

Python est un langage de programmation généraliste lancé par GUIDO VAN ROSSUM qui est devenu très populaire très rapidement, principalement pour sa simplicité et sa lisibilité du code. Il permet au programmeur d'exprimer des idées en moins de lignes de code sans réduire la lisibilité.

Comparé à des langages tels que *C / C ++*, *Python* est plus lent. Cela dit, *Python* peut facilement être étendu avec *C / C ++*, ce qui nous permet d'écrire du code exigeant en calculs intensifs en *C / C ++* et de créer un environnement *Python* utilisable en tant que modules *Python*. Cela nous donne deux avantages : premièrement, le code est aussi rapide que le code original *C / C ++* (puisqu'il s'agit du code *C ++* réel travaillant en arrière-plan) et deuxièmement, il est plus facile de coder en *Python* que *C / C ++*. *OpenCV-Python* est une enveloppe *Python* pour l'implémentation d'origine *OpenCV C ++* et nous allons utilisé cette bibliothèque pour réaliser notre travail.

- ***La bibliothèque OpenCV et les bibliothèques de Python***

OpenCV « Open Source Computer Vision Library » est une bibliothèque écrit en *C++* conçue pour résoudre les problèmes de vision par ordinateur. *OpenCV* a été développé à l'origine en 1999 par **Intel**, mais il a ensuite été pris en charge par WILLOW GARAGE. *OpenCV* est utilisable avec une grande variété de langages de programmation tels que *C ++, Python, Java, etc*. Il peut aussi être executé sous plusieurs plates-formes, notamment Windows, Linux et MacOS.

OpenCV-Python utilise *NumPy* « Numeric Python », qu'est une bibliothèque hautement optimisée pour les opérations numériques avec une syntaxe proche de celle *MATLAB*. Toutes les structures de tableau *OpenCV* sont converties vers et à partir de tableaux *NumPy*. Cela facilite également l'intégration avec d'autres bibliothèques utilisant *NumPy*, telles que *SciPy* « Scientific library » et *Matplotlib* « Plotting library ».

- ***Deep Learning et Keras***

A présent, nous connaissons déjà l'apprentissage automatique, une branche de l'informatique qui étudie la conception d'algorithmes pouvant apprendre. Ici, nous allons nous concentrer sur l'apprentissage en profondeur, un sous-champ de l'apprentissage automatique constitué d'un ensemble d'algorithmes inspiré de la structure et des fonctions du cerveau. Ces algorithmes sont généralement appelés réseaux de neurones artificiels (RNA). L'apprentissage en profondeur est l'un des domaines les plus appliqué pour l'analyse des données. De nombreuses études de cas ont donné des résultats impressionnante dans les domaines de la robotique, de la reconnaissance d'images et de l'intelligence artificielle.

Keras est l'une des bibliothèques *Python* les plus puissantes et les plus faciles à utiliser pour développer et évaluer des modèles d'apprentissage en profondeur. Il enveloppe les bibliothèques de calcul numériques efficaces, en l'occurence, *Theano* et *TensorFlow*. L'avantage son utilisation est principalement de pouvoir d'utiliser les réseaux de neurones de manière simple et amusante.

4.2 Implémentation

4.2.1 Description des processus de notre système

* **Méthodologie de détection de visage**

L'étape de détection permet de décider exclusivement sur l'existence ou non d'un visage dans une image. Nous avons donc besoin d'un détecteur (classifieur) de visages permettant de chercher les caractéristiques relatives à ce dernier. La bibliothèque *OpenCV* contient la méthode de détection de visages « VIOLA & JONES » [11] qui donne comme résultat une liste de fichiers « **.xml** » dits classifieurs en cascade de **Haar**. Nous avons aussi utilisé les fonctions de cette même bibliothèque pour détecter le visage dans chaque frame de la séquence vidéo. Les coordonnées du visage détecté sont passées à une fonction chargée d'englober ce dernier dans un rectangle tout en traitant les séquences vidéo image par image. Cette fonction permet de suivre implicitement le mouvement du visage détecté. Nous avons aussi synchronisé le dessin de rectangles englobants avec la détection afin qu'ils prennent en charge toute modification de position et de dimension du visage (figure 4.1).

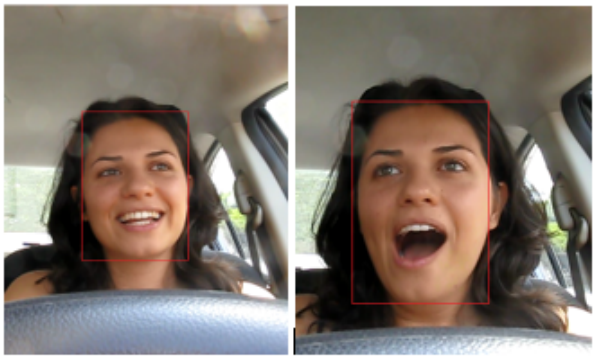
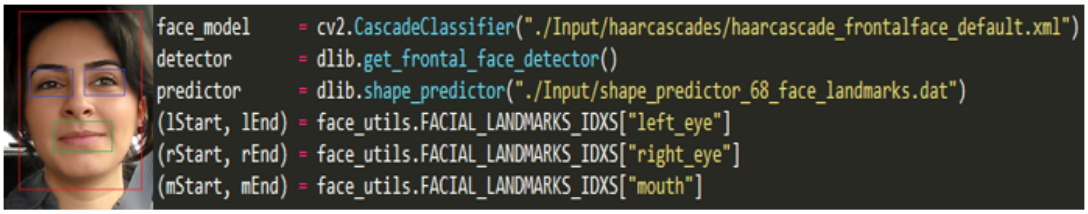


FIGURE 4.1 – Détection de visage et dessin de rectangle englobant dans chaque frame.

* **Définition d'une segmentation spatiale du visage**

Après avoir détecté le visage, nous avons utilisé la bibliothèque **dlib** [86] pour localiser les régions contenant les éléments significatifs du visage tels que la bouche, les yeux (les points caractéristiques). C'est la méthode qui exploite les connaissances a priori relatives à la répartition spatiale du visage humain. Autrement dit, la géométrie d'un visage humain ici est toujours formée de bas vers le haut d'un front, menton, bouche, nez, yeux et sourcils. Dans la figure 4.2, nous montrons comment nous avons traité la détection des yeux et de la bouche selon le visage et déterminer les coordonnées du rectangle englobants la bouche et yeux et ce, après avoir récupéré les coordonnées du visage.



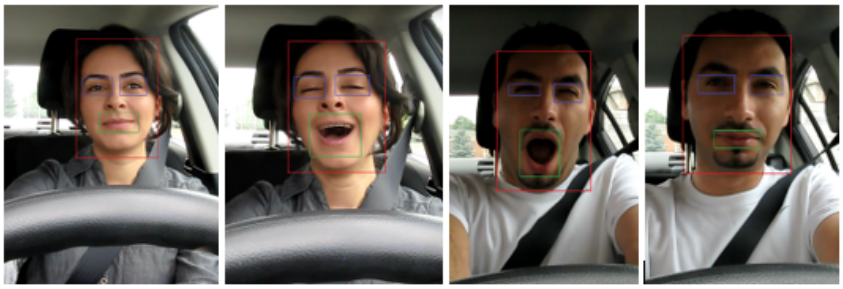


FIGURE 4.2 – Segmentation spatiale du visage.

* **Calcule les valeurs d'EAR et du MAR**

À travers cette détection, et après avoir récupéré les coordonnées du visage, nous avons la possibilité de calculer le EAR pour chaque œil et le MAR pour la bouche, en se basant sur les leurs points caractéristiques. La figure 4.3, montre comment calculer les valeurs EAR et MAR en utilisant *Python*.

```python
def eye_aspect_ratio(eye):
    A = dist.euclidean(eye[1], eye[5])
    B = dist.euclidean(eye[2], eye[4])
    C = dist.euclidean(eye[0], eye[3])
    #Compute eye aspect ratio
    ear = (A+B)/(2.0*C)

    return ear
```

```python
def mouth_aspect_ratio(mouth):
    A = dist.euclidean(mouth[5],mouth[8])
    B = dist.euclidean(mouth[1],mouth[11])
    C = dist.euclidean(mouth[0],mouth[6])
    #Compute eye aspect ratio
    mar = (A+B)/(2.0*C)

    return mar
```

FIGURE 4.3 – Calcule des valeurs EAR et MAR.

Après avoir calculer ces deux valeurs pour tous les examples, nous enregistrons dans un fichier .csv (l'un des formats de fichiers utilisé pour stocker des caractéristiques) pour l'utilisera comme entré de notre algorithme d'apprentissage. Dans cet ensemble de données, il existe 3 attributs (à savoir 3 colonnes dans le fichier **.csv**) décrivant la valeur EAR, la valeur MAR et la classe soit 0 ou 1 auquel appartiennent les deux valeurs (à savoir une seule ligne) et un total de X instances (à savoir le nombre total de lignes ou le nombre de frames extraites), voir Figure 4.4.

CalculeEAR_MAR.csv

```
EAR,MAR,CLASS
0.34498121820233907,0.4123582169027884,1

0.4647854279665712,0.31385400469869607,1

0.4487775015352331,0.363566600945335,1

0.2999732029784743,0.8258308643916009,0

0.37337584367412147,0.734041989130465,1

0.10229935593909842,0.2392896971193136,0

0.1941394659350138,0.16230990366393222,0

0.16774943638443116,0.16718586998475482,0
```

FIGURE 4.4 – Vue globale à notre fichier .csv.

* **Notre solution pour la reconnaissance**

Le processus de reconnaissance s'appuie sur les deux phases précédentes : caractérisation et apprentissage. Il implique l'ensemble des images de visages décrivant des expressions. Pour déterminer si l'expression de la personne présente en face de la caméra est une expression de fatigue ou non, nous avons téléchargé utilisé une base de données contenant des vidéos [91]. Elle contient des conducteurs masculin et féminin de différents âges, ethnique, sans ou avec des lunettes.

Nous avons alors choisi de créer notre propre base d'expressions (images) à partir de ces vidéos [91] pour optimiser la classification ; les images représentant un état fatigue nommées : **drowsy***i* et les autres nommées : **normal***j*. La figure 4.5 ci-dessous montre des exemples des expressions faciales d'extraits à partir de notre base.

FIGURE 4.5 – Des exemples des expressions faciales d'extraits à partir de notre base d'apprentissage.

Une fois les visages détectés et les points caractéristiques localisés, nous passons à la phase d'apprentissage, qui utilise les résultats de calcule des valeurs d'EAR et du MAR sauvegardés sous forme de fichier « .csv » qui représente les entrées au notre réseau. À partir des frames qui sont enregistrées dans notre base et classées manuellement (**drowsy***i* et **normal***j*) et leur valeur EAR, MAR ainsi que la classe, le système d'apprentissage va apprendre à partir de cette base est créer son modèle. Une fois le modèle trouvé et enregistrer, nous pouvons trouver la classe d'une nouvelle instance (phase de prédiction), en utilisant les ses valeurs EAR, MAR.

À chaque étape, les « mauvaises » réponses sont éliminées et renvoyées vers les niveaux en amont pour ajuster le modèle mathématique. Au fur et à mesure, le programme réorganise les informations en blocs plus complexes. Lorsque ce modèle est appliqué à d'autres cas, il est normalement capable de reconnaître un conducteur fatigue sans que personne ne lui ait jamais indiqué comment le faire. Les données de départ sont essentielles : plus

le système accumule des expériences différentes, plus il sera performant. On peut faire un rappele à la définition de l'apprentissage au profond :

« Le Deep Learning (apprentissage au profond) est un ensemble de techniques d'apprentissage automatiques ou la machine apprend à reconnaître des motifs en s'entraînant sur des modèles de données qualifiées. Les algorithmes qui permettent de réaliser ces opérations utilisent des concepts mathématiques reposant essentiellement sur des transformations non linéaires. Le Deep Learning apprend à un modèle informatique comment réaliser des tâches de classification directement à partir d'images, de textes ou d'audio. Les modèles de Deep Learning peuvent atteindre un niveau de précision exceptionnel, parfois supérieur aux performances humaines. L'entraînement des modèles s'effectue via un vaste ensemble de données labellisées et d'architectures de réseaux de neurones qui contiennent de nombreuses couches. »

- *Diagramme de Keras pour le Deep Learning*

  L'apprentissage supervisé signifie que nous avons un ensemble de données libélés où les résultats sont connus (classes ou valeurs réelles). Nous formons un modèle d'apprentissage approfondi avec les données d'apprentissage afin qu'il puisse prédire le résultat (classe ou valeur réelle) des futures données invisibles (ou des données de test).
  Tout problème complexe lié à l'apprentissage supervisé peut être résolu à l'aide de cet organigramme. Nous allons passer en revue chaque étape une par une en détail.

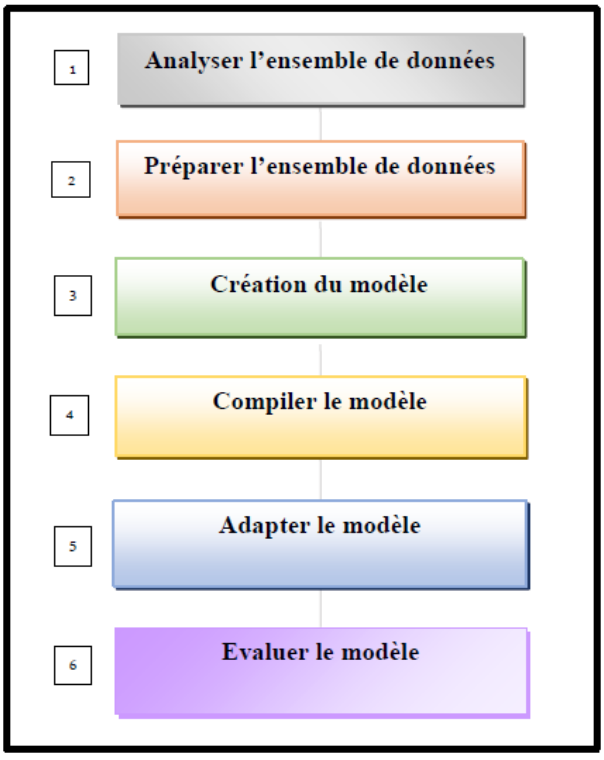


FIGURE 4.6 – Diagramme de *Keras* pour le Deep Learning.

1. La première étape de tout problème d'apprentissage en profondeur (Deep Learning) consiste à collecter plus de données, à les analyser et à en comprendre les différents paramètres / attributs. Nous avons déjà notre propre jeu de données, qui est une collection d'exemples de visages présentant différents statuts de conducteurs, représnetés par les valeurs EAR, MAR.

2. La deuxième étape est la structuration les données pour que les ordinateurs puissent les comprendre. Représenter nos données analysées est la prochaine étape de Deep Learning. Les données seront représentées sous la forme d'une matrice à **n** dimensions dans la plupart des cas (qu'il s'agisse de données numériques, d'images ou de vidéos). Nous allons utiliser le format de fichier **csv** dans notre travail, comme nous l'avons indiqué ci-dessus.
3. Dans *Keras*, un modèle est créé à l'aide du modèle Sequential. Nous voudrons peut-être rappeler que les réseaux de neurones contiennent un grand nombre de neurones résidant dans plusieurs couches séquentielles.

   Nous allons créer un modèle comportant des couches entièrement connectées, ce qui signifie que tous les neurones sont connectés d'une couche à l'autre.

   Nous allons utiliser l'architecture de réseau de neurones profonds ci-dessus qui comporte une seule couche d'entrée, 2 couches cachées et une seule couche de sortie.

   Les données d'entrée de taille 2 sont envoyées à la première couche cachée ayant initialisé de manière aléatoire 8 neurones. C'est une approche très utile, si nous n'avons aucune idée du nombre de neurones à spécifier dès la première tentative. À partir de là, nous pouvons facilement effectuer une procédure d'essai et d'erreur pour renforcer l'architecture du réseau et produire de bons résultats. La couche cachée suivante a 6 neurones et la couche de sortie finale a 1 neurone renovi en résultat l'état du conducteur : fatigué ou non.

   La figure 4.7 illustre notre réseau de neurones profonds utlisé.

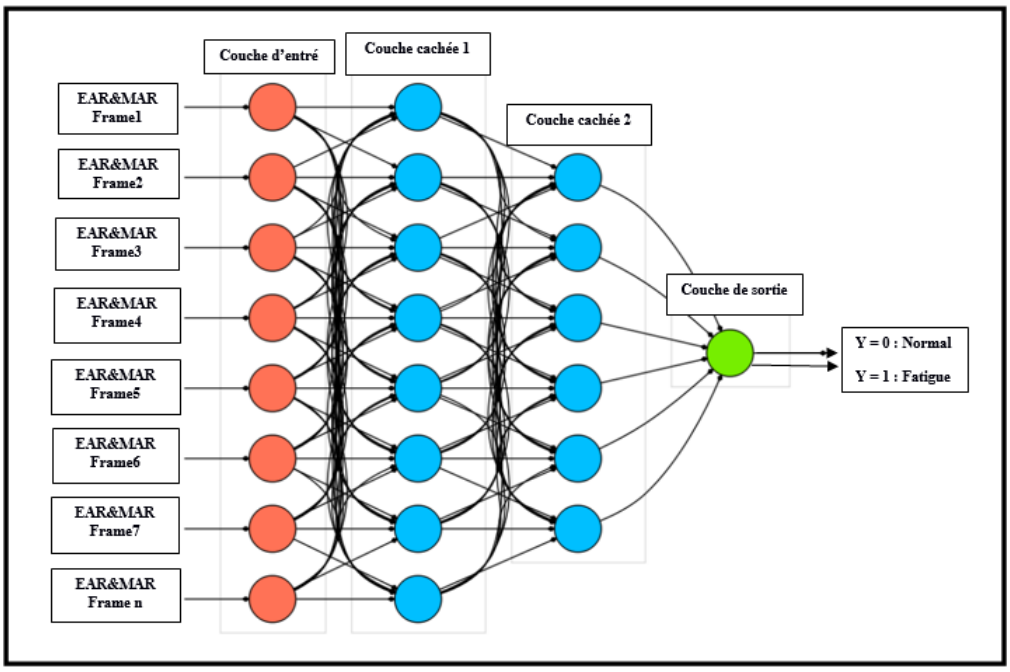


FIGURE 4.7 – Notre réseaux de neurones profonds (Perceptron multicouche).

4. Après avoir créé le modèle, nous devons compiler le modèle dans *Keras*. Les paramètres doivent être ajustés en fonction du problème car notre modèle nécessite une optimisation en arrière-plan (prise en charge par *Theano* ou *TensorFlow*) afin qu'il tire des leçons des données à chaque époque (ce qui signifie une réduction de l'erreur entre la sortie réelle et la valeur prédite).
5. Une fois le modèle compilé, l'ensemble de données doit être adapté au modèle.

6. Une fois le jeu de données adapté au modèle, celui-ci doit être évalué. L'évaluation du modèle formé avec un jeu de données de test non vu montre comment notre modèle prédit la sortie sur des données non vues.

Ensuite après cette étape d'apprentissage, nous sauvegardons le model pour l'utiliser avec une nouvelle vidéo (phase de test).

Pour tester une nouvelle vidéo, pour chaque frame nous calculons le EAR et le MAR, et nous donnons ces résultats au réseau, en utilisant cette fonction :

$$Ynew = model.predictclasses(Xnew) \tag{4.1}$$

Si elle renvoi 1 c.à.d. état de fatigue, sinon normal.

```
Xnew = np.array([[ear, mouthEAR]])
ynew = loaded_model.predict_classes(Xnew)
```

FIGURE 4.8 – Fonction de prédiction d'une nouvelle instance.

4.2.2 Description de l'application

Nous commençons notre expérimentation par l'entrainement de système et l'initialisation des données, pour l'entrainer. Lorsque nous aurons le modèle, l'étape suivante consiste à reconnaître et classifier une nouvelle instance (vidéo). Nous avons travaillé sur deux types d'entrées différentes : vidéo enregistré et vidéo en temps réel en utilisant une WebCam. La vidéo au cours d'exécution contient un conducteur qui fait des gestes (mouvements, signes de fatigue et autre normal comme le rire ou parler), et travers notre système nous lancer des alertes quand un état de fatigue est détecté .

4.3 Résultats obtenus

4.3.1 Détection de visage et les points caractéristiques

La première étape dans le processus de détection de la somnolence est basée sur la segmentation du visage. Cette étape est divisée en cinq phases : détection du visage, détection des yeux, détection de la bouche, détection du cligne des yeux et du bâillement. Comme indiqué dans la section précédente, le visage est détecté en utilisant la méthode de « VIOLA & JONES» [11]. L'implémentation de cette détection nous permet de tracer des rectangles colorés autour du visage, les yeux et la bouche.

La figure 4.9 montre les résultats obtenus en utilisant cette méthode.

Après avoir localisé ces trois parties de visage, la prochaine étape consiste à déterminer l'état des yeux et de la bouche (yeux fermés, bâillement). Cela est réaliser en appliquant l'algorithme d'apprentissage décrit précédemment, appliqué aux valeurs du EAR et de MAR. En effet, un conducteur est considéré comme fatigué s'il ferme les yeux (valeur EAR trouvé par le réseau de neurone) un certain nombre de fois, ou s'il baille. Par si les yeux seront fermés (et/ou) la bouche largement ouverte, ce qui augmentera le temps en comptant le nombre de fois que cela est produit. Si le nombre de fois dépasse une certaine valeur, le conducteur est considéré comme fatigué.

FIGURE 4.9 – Détection de visage et les points caractéristiques.

4.3.2 Le modèle d'apprentissage

Dans l'étape suivante, les coordonnées des repères extraites des images agiront comme entrée dans l'algorithme d'apprentissage en profondeur, basé sur le classifieur multicouche de perceptron avec deux couches cachées. Au cours de cette étape, un processus de formation s'ensuivra où diverses prédictions seront établies à partir desquelles un modèle sera formé ; des corrections sont apportées au modèle si les prédictions tournent mal. La formation sera traitée jusqu'à ce que le niveau de précision souhaité soit atteint.

4.3.3 Phase de prédiction

C'est la dernière étape dans le processus de détection de fatigue. Il s'agit de déterminer pour une nouvelle instance (frames), et selon les valeurs EAR, MAR, ainsi que le nombre de fois où le conducteur baille ou ferme les yeux, l'état de conducteur.

La figure 4.10 montre le résultat de l'algorithme de détection du visage, des yeux et de la bouche dans différents scénarios de conduite.

Nous avons appliqué nos tests sur différents vidéos de conducteur, et selon différents angles de vues. Les résultats obtenus ont montré que pour les vidéos où la caméra est installée sous le rétroviseur, le taux de détection du visage est de 85% environ, mais le taux de détection des yeux et de la bouche, suivi de la fermeture des yeux et du bâillement, est d'environ 40%. Dans le cas où la caméra est installée sur le tableau de bord, les régions du visage, des yeux et de la bouche sont détectées correctement et en temps réel. Et le taux de réussite de la détection du bâillement par la fermeture des yeux est de 95%, car la caméra capte une vue frontale du visage du conducteur, ce qui facilite la détection. La détection des yeux et de la bouche dans ce scénario est de 85% et la détection de la fermeture des yeux et des bâillements est de 60%.

FIGURE 4.10 – Détection de visage et les points caractéristiques.

Comme le montre la figure 4.11 la région du visage est encadrée en vert, la région des yeux est entourée en bleu, tandis que la région de la bouche est entourée en rouge. Lorsque le système détermine que les yeux sont en position de fermeture et / ou que la bouche est en position de bâillement, un message signalant le début de la détection de la somnolence du conducteur. Par conséquent, en comptant le nombre total de fermetures oculaires et le nombre total de bâillements, le système est en mesure de déterminer si le conducteur bâille ou non.

Le système de détection de fermeture des yeux / de bâillement mis en œuvre est rapide, et il est également fiable et précis pour déterminer la fatigue du conducteur en comptant le nombre de yeux fermés et de bâillements sur une courte période. Lorsque le système détecte plus de 4 fermetures de yeux et /ou bâillements de quelques secondes, une alarme est déclencher.

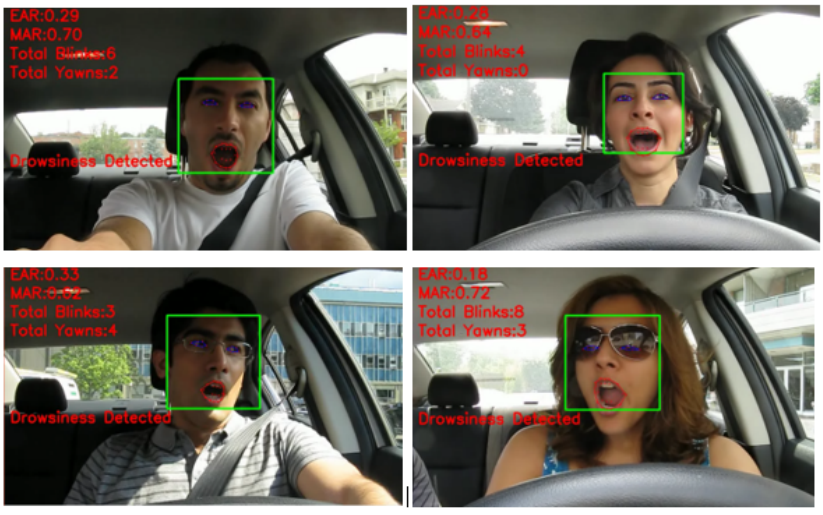


FIGURE 4.11 – Alerte de somnolence lorsque le seuil passe la 4ème fois des détections.

Notre modèle était capable de classer correctement une séquence d'images consécutives à partir de vidéos. Il détectait une personne somnolente avec un niveau de confiance de 95,25% dans la plupart de nos tests. Pour visualiser la fonction de perte, nous avons exécuté plus de 200 époques, ce qui a donné le graphique de la figure 4.12 (en haut) , tandis que la perte a une confiance inférieure à 20%, comme le montre la figure 4.12 (en bas) .

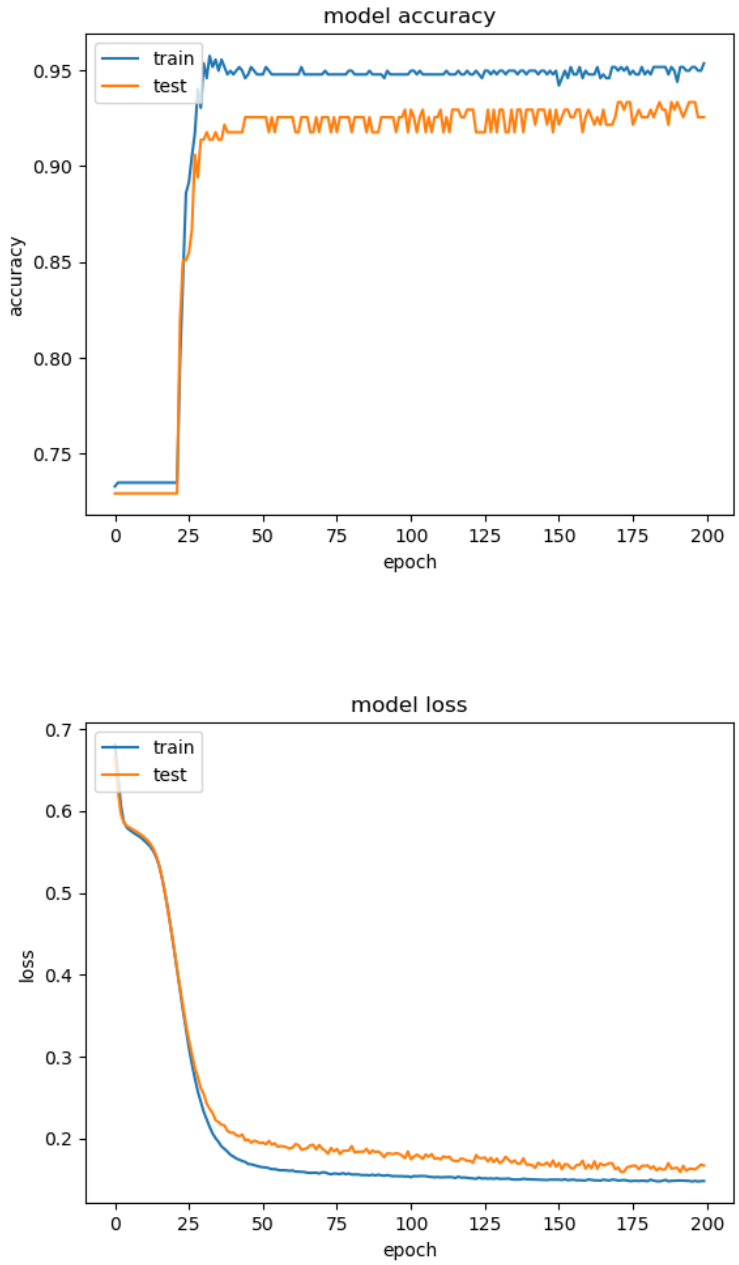


FIGURE 4.12 – Traces d'entraînement et précision de la validation (en haut) et perte (en bas) lors de la formation de notre réseau de neurones pour la détection de la somnolence.

Conclusion

Lors de ce chapitre nous nous avons introduit les détails de l'applémentation et de l'utilisation de notre système proposé. Nous avons donc présenté le modèle du réseau de neurones utilisé pour construire un modèle d'apprentissage en se basant sur des exemples de valeurs des descripteurs extraits à partir des yeux et de la bouche.

Conclusion générale

Nous avons à travers ce mémoire présenté une nouvelle contribution au domaine de la vision artificielle. Elle consiste en une nouvelle méthode basée sur la description de certaines expressions faciales pour la détection de fatigue chez les conducteurs.

Le système proposé utilise des mesures appliquées à des caractéristiques extraites à partir des yeux et de la bouche pour déterminer l'état du conducteur. L'ouverture de la bouche à un certain degré et la fermeture ou clignes des yeux permet de prévoir un éventuel état de fatigue. Pour déterminer les valeurs de ces mesures permettant de différencier des expressions de fatigues parmi d'autres, nous avons utilisé un algorithme d'apprentissage approfondi (Deep Learning). Cet algorithme permet de transformer ces mesures calculées pour une base d'exemples (frames) en un robuste modèle permettant de prédire la valeur d'une nouvelle instance (état de conducteur).

Le processus de détection de fatigue se déroule en trois phases : phase de prétraitement, phase d'apprentissage, et phase de test. Dans la première phase, les exemples (vidéo) sont transformés en frames, puis chacune en points caractéristiques extraits à partir du visage. Nous calculons ensuite pour chaque exemple de données deux mesures caractérisant les yeux et la bouche. Leurs valeurs vont être par la suite utilisées comme entrée de l'algorithme d'apprentissage. Dans la deuxième phase, nous utilisons un algorithme d'apprentissage approfondie qui permet de trouver un modèle déterminant les couplets des deux mesures qui différencier une expression de fatigue parmi les autres expressions possibles. Dans la dernière phase, et pour une nouvelle instance (vidéo) représentant un conducteur, nous utilisons le modèle trouvé dans la phase précédente pour déterminer son état.

La méthode proposée montre une grande fiabilité, et une grande invariance aux changements : âge, genre et ethnicité du conducteur. Il est aussi important de signaler que contrairement aux méthodes existantes, notre méthode ne nécessite aucune utilisation des seuils, ni de choisir les valeurs des mesures utilisées empiriquement.

Dans les futurs travaux, nous comptons améliorer notre méthode pour qu'elle utilise plus de mesures, avec moins de données d'exemples, et avec autant d'efficacité. On peut aussi envisager de l'implémenter sur d'autres plateformes pour qu'elle soit par exemple utilisée avec des systèmes embarqués.