UNIVERSITE DES ANTILLES
Laboratoire de Mathématiques, Informatique et Applications



# THÈSE

en vue de l'obtention du grade de Docteur

**Spécialité : Informatique Affective & Intelligence Artificielle**

**Mention : Informatique**

présentée et soutenue publiquement par

**Stéphane CHOLET**

le 17 juin 2019

---

# Évaluation automatique des états affectifs et dépressifs : vers un système de prévention des risques psychosociaux

---

Dirigée par le Professeur Hélène PAUGAM-MOISY

| | |
|---|---|
| Examinateurs : | Mme Suzy GAUCHER-CAZALIS |
| | M. Guillaume LEPAGE |
| | M. Didier PUZENAT |
| Rapporteurs : | M. Frédéric ALEXANDRE |
| | M. Renaud SEGUIER |
| Encadrants : | Mme Hélène PAUGAM-MOISY |
| | M. Lionel PREVOST |
| | M. Sébastien REGIS |

# RÉSUMÉ

Les risques psychosociaux sont un enjeu de santé publique majeur, en particulier à cause des troubles qu'ils peuvent engendrer : stress, changements d'humeurs, burn-out, etc. Bien que le diagnostic de ces troubles doive être réalisé par un professionel, l'*Affective Computing* peut apporter une contribution en améliorant la compréhension des phénomènes. L'*Affective Computing* (ou Informatique Affective) est un domaine pluridisciplinaire, faisant intervenir des concepts d'Intelligence Artificielle, de psychologie et de psychiatrie, notamment. Dans ce travail de recherche, on s'intéresse à deux éléments pouvant faire l'objet de troubles : l'état émotionnel et l'état dépressif des individus.

Le concept d'émotion couvre un très large champ de définitions et de modélisations, pour la plupart issues de travaux en psychiatrie ou en psychologie. C'est le cas, par exemple, du circumplex de Russell, qui définit une émotion comme étant la combinaison de deux dimensions affectives, nommées *valence* et *arousal*. La *valence* dénote le caractère triste ou joyeux d'un individu, alors que l'*arousal* qualifie son caractère passif ou actif. L'évaluation automatique des états émotionnels a suscité, dans la dernière décénie, un regain d'intérêt notable. Des méthodes issues de l'Intelligence Artificielle permettent d'atteindre des performances intéressantes, à partir de données capturées de manière non-invasive, comme des vidéos. Cependant, il demeure un aspect peu étudié : celui des intensités émotionnelles, et de la possibilité de les reconnaître. Dans cette thèse, nous avons exploré cet aspect au moyen de méthodes de visualisation et de classification pour montrer que l'usage de classes d'intensités émotionnelles, plutôt que de valeurs continues, bénéficie à la fois à la reconnaissance automatique et à l'interprétation des états.

Le concept de dépression connaît un cadre plus strict, dans la mesure où c'est une maladie reconnue en tant que telle. Elle atteint les individus sans distinction d'âge, de genre ou de métier, mais varie en intensité ou en nature des symptômes. Pour cette raison, son étude tant au niveau de la détection que du suivi, présente un intérêt majeur pour la prévention des risques psychosociaux. Toutefois, son diagnostic est rendu difficile par le caractère parfois anodin des symptômes et par la démarche souvent délicate de consulter un spécialiste. L'échelle de Beck et le score associé permettent, au moyen d'un questionnaire, d'évaluer la sévérité de l'état dépressif d'un individu. Le système que nous avons développé est capable de reconnaître automatiquement le score dépressif d'un individu à partir de vidéos.

Il comprend, d'une part, un descripteur visuel spatio-temporel bas niveau qui quantifie les micro et les macro-mouvements faciaux et, d'autre part, des méthodes neuronales issues des sciences cognitives. Sa rapidité autorise des applications de reconnaissance des états dépressifs en temps réel, et ses performances sont intéressantes au regard de l'état de l'art. La fusion des modalités visuelles et auditives a également fait l'objet d'une étude, qui montre que l'utilisation de ces deux canaux sensoriels bénéficie à la reconnaissance des états dépressifs.

Au delà des performances et de son originalité, l'un des points forts de ce travail de thèse est l'interprétabilité des méthodes. En effet, dans un contexte pluridisciplinaire tel que celui posé par l'*Affective Computing*, l'amélioration des connaissances et la compréhension des phénomènes étudiés sont des aspects majeurs que les méthodes informatiques sous forme de "boîte noire" ont souvent du mal à appréhender.

# Abstract

Psychosocial risks are a major public health issue, because of the disorders they can trigger : stress, mood swings, burn-outs, etc. Although propoer diagnosis can only be made by a healthcare professionnel, Affective Computing can make a contribution by improving the understanding of the phenomena. Affective Computing is a multidisciplinary field involving concepts of Artificial Intelligence, psychology and psychiatry, among others. In this research, we are interested in two elements that can be subject to disorders : the emotional state and the depressive state of individuals.

The concept of emotion covers a wide range of definitions and models, most of which are based on work in psychiatry or psychology. A famous example is Russell's circumplex, which defines an emotion as the combination of two emotional dimensions, called valence and arousal. Valence denotes an individual's sad or joyful character, while arousal denotes his passive or active character. The automatic evaluation of emotional states has generated a significant revival of interest in the last decade. Methods from Artificial Intelligence allow to achieve interesting performances, from data captured in a non-invasive manner, such as videos. However, there is one aspect that has not been studied much : that of emotional intensities and the possibility of recognizing them. In this thesis, we have explored this aspect using visualization and classification methods to show that the use of emotional intensity classes, rather than continuous values, benefits both automatic recognition and state interpretation.

The concept of depression is more strict, as it is a recognized disease as such. It affects individuals regardless of age, gender or occupation, but varies in intensity or nature of symptoms. For this reason, its study, both at the level of detection and monitoring, is of major interest for the prevention of psychosocial risks. However, his diagnosis is made difficult by the sometimes innocuous nature of the symptoms and by the often delicate process of consulting a specialist. The Beck's scale and the associated score allow, by means of a questionnaire, to evaluate the severity of an individual's state of depression. The system we have developed is able to automatically recognize an individual's depressive score from videos. It includes, on the one hand, a low-level visual spatio-temporal descriptor that quantifies micro and macro facial movements and, on the other hand, neural methods from the cognitive sciences. Its speed allows applications for real-time recognition of

depressive states, and its performance is interesting with regard to the state of the art. The fusion of visual and auditory modalities has also been studied, showing that the use of these two sensory channels benefits the recognition of depressive states.

Beyond performance and originality, one of the strong points of this thesis is the interpretability of the methods. Indeed, in a multidisciplinary context such as that of Affective Computing, improving knowledge and understanding of the studied phenomena is a key point that usual computer methods implemeted as "black boxes" can’t deal with.

# Remerciements

L'exercice de la thèse fut au bas mot édifiant, fatiguant, long et captivant. J'ai eu, pendant ce temps, l'occassion de travailler sur mes propres états émotionnels et dépressifs. Je profite de ce passage pour remercier celles et ceux qui m'ont apporté leur aide et sans qui l'expérience aurait été radicalement différente.

Je remercie Hélène PAUGAM-MOISY pour sa patience et son incroyable rigueur, qui ont su m'inspirer et me guider pendant qu'elle dirigeait ma thèse. Sa capacité à apprendre de nouveaux sujets et à transmettre ceux qu'elle connaît m'ont particulièrement marqués.

Je remercie Lionel PREVOST, à qui je dois de m'être lancé dans cette aventure, et qui a dirigé ma thèse pendant la première année. Merci également à Sébastien REGIS pour avoir participé à l'encadrement de la thèse.

Je remercie Frédéric ALEXANDRE et Renaud SEGUIER pour avoir accepté de rapporter ma thèse, et Didier PUZENAT et Suzy GAUCHER-CAZALIS pour avoir accepté d'en être les examinateurs.

Merci à Raphaël PASQUIER, ingénieur de recherche au Centre Commun de Calcul Intensif (C3i, Université des Antilles) pour sa disponibilité et pour la qualité de ses explications.

Je remercie Léa et Ténissia, dont j'ai partagé le bureau, pour leur présence, leur discrétion et leur amitié, qui a fait de ce temps de travail un moment agréable et vivable. Merci aux membres et aux encadrants du LAMIA pour leur accueil et leur confiance. Merci à Larissa pour son intarissable bonne humeur, à Jimmy pour sa discrétion redoutablement efficace, à Maëva pour sa sympathie, à Audrey pour avoir dévoilé ma passion pour l'électronique.

Mon investissement en tant qu'élu étudiant puis vice-président étudiant m'a permis d'acquérir des compétences que mon parcours universitaire initial ne prévoyait pas. J'ai été amené à collaborer avec des personnes que je respecte profondément

et que je remercie pour leur confiance et pour leur dévotion envers l'Université.

Merci aux encadrants du SUAPS et du service culturel : Raymond, Simone, J.P., Marinette, Thierry ; à mes coéquipiers et coéquipières de l'ESM Volley ; à mes camarades de jeu. Ils ont aiguisé et entretenu mon goût pour la pratique sportive et culturelle, jusqu'à la transformer en besoin.

Je ne sais pas précisément qui remercier pour ces neuf années de service (qui n'ont cessé de s'améliorer), mais merci au restaurant universitaire du CROUS pour avoir été là lorsque le temps se faisait si rare que manger devenait facultatif.

Merci à mes parents et à mon frère pour leurs encouragements et leur compréhension. Merci à Cécilia pour tous ces moments passés et à venir. Merci à Emmanuel pour son amitié, et à la Team Montagnards pour ces expériences inoubliables que je pourrai très difficement réitérer.

Merci à Andrée pour son accueil (même si au début ce n'était pas gagné), pour m'en avoir tant appris, et pour avoir été à l'origine de rencontres et d'événements qu'elle même ne soupçonne pas. Merci à Mme CYRILLE pour son caractère légendaire, qui ne laisse jamais personne indifférent.

Merci à cette bonne étoile qui veillait sur moi toutes ces fois où j'ai quitté les lieux au petit matin, avec un seul œil ouvert.

# SOMMAIRE

# LISTE DES TABLEAUX

# TABLE DES FIGURES

CHAPITRE 1

# INTRODUCTION

## 1.1 L'Informatique Affective

Quotidiennement, la routine soumet les individus au stress, sans distinction de genre, d'âge ou de métier. Les changements de comportement induits peuvent avoir des conséquences aussi bien sur le plan personnel que familial ou professionnel. Sur ce dernier plan, on parle de risques psychosociaux. A moins de consulter un spécialiste, les individus ne sont que rarement conscients du fait qu'ils sont exposés ou atteints de troubles qui, dans une grande majorité des cas, peuvent se résoudre grâce à un suivi psychologique et/ou à la prescription de médicaments adaptés [24].

Toutefois, l'Intelligence Artificielle peut apporter une meilleure compréhension des désordres émotionnels et une assistance tant aux individus qu'aux spécialistes. Dans ses travaux de 1995, Rosalind Picard se pose en véritable précurseur du domaine, en définissant l'Informatique Affective (*Affective Computing*) comme "*l'informatique qui se réfère à, découle de ou influence délibérément les émotions*" [92]. A une époque où l'on ne peut pas encore parler d'automatisation de l'analyse comportementale, Picard dégage deux aspects clefs :

— L'analyse automatique des comportements affectifs humains : c'est une problématique désormais incontournable dans la mise au point de méthodes efficaces dans des domaines variés comme l'assistance individuelle, le divertissement, l'amélioration de l'expérience utilisateur, etc.

— Le raisonnement affectif : on considère qu'une décision efficace ne peut pas être prise sans un aspect affectif, chez l'humain. En ce sens, les machines doivent également raisonner affectivement pour implémenter des systèmes efficaces d'Intelligence Artificielle.

Cette thèse s'intéresse à l'analyse des comportements affectifs humains, et en particulier à la reconnaissance des états émotionnels et des états dépressifs.

## 1.2 Objectifs et plan de la thèse

D'un côté, l'évaluation des états affectifs est un domaine déjà bien avancé, et pour lequel les méthodes sont nombreuses et performantes. Des applications permettent déjà de déterminer l'état affectif d'un individu en temps réel dans le domaine discret ou continu. Toutefois, s'agissant du domaine continu, l'état de l'art ne s'intéresse que rarement aux singularités des intensités affectives ou à la difficulté de les discriminer entre elles.

D'un autre côté, l'évaluation des états dépressifs connaît un essor moins rapide, car elle est plus complexe à appréhender. Elle fait appel à des notions médicales réglementées (notamment en psychiatrie et en psychologie) dont il faut tenir compte en informatique. En ce sens, toute approche interprétable d'automatisation de l'évaluation des états dépressifs peut apporter une aide non seulement aux patients, dans un processus de suivi personnel, mais aussi aux professionnels, dans l'optique d'apporter de nouveaux vecteurs d'information sur l'état des sujets.

Le chapitre 2 présente les principaux modèles de classifieurs qui ont été utilisés pour la prédiction des états émotionnels et dépressifs. En particulier, en dehors des SVM (*Support Vector Machine*), les modèles que nous avons utilisés n'ont jamais, à notre connaissance, été appliqués à l'Informatique Affective. En conséquence, l'étude du classifieur incrémental à base de protytpes et de la triple mémoire associative multimodale sont approfondies. Les principales mesures de performance sont également exposées.

Le chapitre 3 propose une revue des représentations des émotions et des représentations de la dépression. Plusieurs jeux de données sont présentés et comparés, avec une étude plus détaillée du jeu de données qui va servir de support aux expériences tout au long de cette thèse. Des défis majeurs, liés aux jeux de données, sont aussi mis en avant ainsi des pistes pouvant conduire à leur résolution.

Le chapitre 4 dresse l'état de l'art de la reconnaissance des états émotionnels et de la reconnaissance des états dépressifs. Dans ce chapitre, la diversité des usages est mise en avant, car elle caractérise la difficulté à unifier ou à comparer les approches existantes.

Le chapitre 5 détaille notre contribution à la reconnaissance des intensités émotionnelles dans le domaine discret. Les résultats de notre première approche nous ont motivé à analyser la séparabilité des données, puis à opter pour une nouvelle approche plus efficace.

Le chapitre 6 s'intéresse à la classification des états dépressifs, en associant automatiquement à une vidéo un score BDI-II qui évalue la sévérité des états. Dans ce chapitre, on étudie entre autres la dépendance aux sujets des données, et on montre que la modalité auditive est plus adaptée que la modalité visuelle pour la reconnaissance des états dépressifs.

Le chapitre 7 explore plusieurs méthodes pour la fusion des modalités en vue de la classification des états dépressifs. Deux approches traditionnelles sont expérimentées, puis un *framework* pour la fusion abstraite des modalités est proposé.

Enfin, une conclusion synthétise nos contributions et propose des pistes pour leur amélioration et la poursuite des travaux de recherche.

CHAPITRE 2

# MODÈLES DE CLASSIFIEURS

## 2.1 Généralités

### 2.1.1 Vocabulaire

En Intelligence Artificielle, le *Machine Learning* fait référence à un ensemble de concepts permettant à un système d'apprendre à prédire une information à partir d'observations. En particulier, on distingue parmi ces concepts la classification et la régression.

On désigne par classification le processus qui consiste à associer automatiquement une catégorie à un élément. Par opposition, la régression consiste à associer automatiquement une valeur réelle à un élément. Une observation, appelé exemple ou motif, est représenté par un vecteur $\mathbf{x} = (x_1, \ldots, x_m)$ de dimension $m$. Un ensemble de $n$ exemples $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ est appelé jeu (ou base) de données. En classification, une catégorie, appelée classe, est représentée par une valeur discrète $y \in \{k_1, \ldots, k_c\}$ où $c$ est le nombre de classes possibles. En régression, les valeurs réelles à associer aux motifs sont définies dans le domaine continu.

L'apprentissage est supervisé quand la classe de chaque exemple est connue d'avance. On note $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ un ensemble de données annotées. L'objectif est alors d'entraîner un modèle de classifieur en apprenant sur les données afin de trouver une fonction $f : \mathbf{x} \in \mathbb{R}^m \mapsto y' \in \{k_1, \ldots, k_c\}$ qui associe à un exemple $\mathbf{x}$

une classe calculée $y'$. Si la classe (ou sortie) calculée $y'$ est égale à la classe $y$ de l'élément (sortie désirée), alors l'exemple est dit bien classé. Dans le cas contraire, il est dit mal classé. Des mesures de performance existent et permettent d'évaluer la capacité d'un classifieur à reconnaître un motif connu (performances en apprentissage) ou inconnu (performances en généralisation). Plusieurs de ces mesures sont présentées dans la section 2.6.2.

Si $c = 2$, alors le problème de classification est dit binaire. Dans ce cas, on parle souvent d'exemples et de contre-exemples pour désigner les exemples pour lesquels $y = k_1$ des exemples pour lesquels $y = k_2$. Si $c > 2$, le problème est dit multiclasse.

En classification non-supervisée, le concept de classe n'est pas prédéfini. On cherche alors à regrouper des exemples proches entre eux dans un même groupe, dit *cluster*, et on parle de méthodes de *clustering* (ou catégorisation). La notion de proximité se réfère à une mesure de similarité ou à une distance. Dans le cadre de cette thèse, nous n'utilisons pas de méthodes de classification non-supervisée.

### 2.1.2 Notations

Par convention, les vecteurs sont spécifiés en caractères gras et les valeurs scalaires en caractères normaux. Les matrices sont spécifiées par des lettres majuscules en caractères gras. Les symboles suivants sont utilisés dans le manuscrit :

| Symbole | Signification |
|---|---|
| $m$ | Dimension d'un vecteur ou nombre de colonnes d'une matrice |
| $n$ | Nombre de vecteurs ou nombre de lignes d'une matrice |
| $i$ | Utilisé pour l'indexation d'un vecteur ou des lignes d'une matrice |
| $j$ | Utilisé pour l'indexation des colonnes d'une matrice |
| $c$ | Nombre de classes |
| $k$ | Nom d'une classe |
| $\mathbf{x}$ | Vecteur de données, avec $\mathbf{x} = (x_1, \ldots, x_m)$ et $\forall i \in \{1, \ldots, m\}$, $x_i \in \mathbb{R}$ |
| $\mathbf{X}$ | Matrice des données, avec $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ et $\forall j \in \{1, \ldots, n\}$, $\mathbf{x}_j \in \mathbb{R}^m$ |
| $\mathbf{y}$ | Vecteur d'annotations, avec $\mathbf{y} = (y_1, \ldots, y_n)$ et $y = \{k_1, \ldots, k_c\}$ |
| $S$ | Ensemble de données, tel que $S = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$ |
| $\mathbf{x}^\top$ | Transposée d'un vecteur |
| $K$ | Fonction noyau |

| | |
|---|---|
| $\mathbf{G}$ | Matrice de Gram |
| $h$ | Fonction de transfert (réseaux de neurones) |
| $\mathscr{H}$ | Espace de Hilbert |
| $\phi$ | Fonction de *mapping* (SVM) |
| $\mu$ | Moyenne |
| $\sigma$ | Ecart-type |
| $\sigma^2$ | Variance |
| $cov$ | Covariance |
| $\rho$ | Coefficient de corrélation linéaire de Pearson |
| $\rho_S$ | Coefficient de corrélation linéaire de Spearman |
| CCC | Coefficient de concordance des corrélations |
| Classe() | Primitive qui renvoie la classe d'un élément passé en paramètre |

## 2.2 Modèles de l'état de l'art

### 2.2.1 Réseaux de neurones

Les réseaux de neurones artificiels, souvent abrégés Réseaux de Neurones (RN), ou en anglais *Artificial Neural Networks* (ANN), sont une famille de méthodes de *Machine Learning* qui s'inspirent du fonctionnement du cerveau humain. Un tel réseau met en œuvre plusieurs cellules (ou neurones ou nœuds) de traitement, le plus souvent organisés en couches. L'origine des neurones artificiels qui composent ces réseaux remonte aux années 1940 avec la modélisationdu neurone formel par McCulloch et Pitts [80].

#### 2.2.1.1 Du neurone au réseau de neurones multicouches

**Généralités** Un neurone (Fig. 2.1) est caractérisé par ses poids $w_1,\ldots,w_m$, auxquels s'ajoute un biais $w_0$. Les entrées et les poids sont traités par une fonction d'activation $f(\mathbf{x},\mathbf{w})$, dont le résultat $a$ est transformé en sortie par une fonction de transfert $h$. Pour le neurone formel, la fonction d'activation est la somme pondérée des entrées (Eq. 2.1) et la fonction de transfert est la fonction de Heaviside ou la fonction signe (Eq. 2.2) :

(2.1) $$f(\mathbf{x},\mathbf{w}) = a = w_0 + \sum_{i=1}^{m} x_i w_i, \text{où } w_0 \text{ est l'opposé du seuil}$$

(2.2) $$H(a) = \begin{cases} 0 & \text{si } a < 0 \\ 1 & \text{sinon} \end{cases} \quad \text{ou} \quad sign(a) = \begin{cases} -1 & \text{si } a < 0 \\ 1 & \text{sinon} \end{cases}$$

Le processus de traitement du neurone formel est schématisé par la figure 2.1.

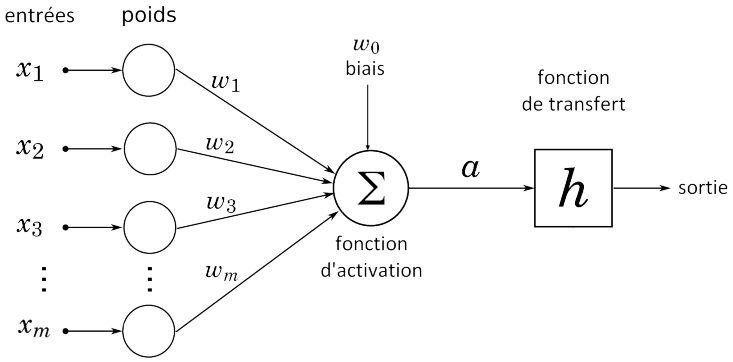



FIGURE 2.1 – Modèle d'un neurone artificiel.

Dans un réseau de neurones multicouche (*MultiLayer Perceptron*, MLP), un neurone donné a pour entrées l'ensemble des descripteurs $x_1,\ldots,x_m$ ou la sortie des neurones de la couche précédente, et sa propre sortie est proposée en entrée des neurones de la couche suivante. Le réseau est dit *feed-forward* lorsque que son graphe de connexions est acyclique, et récurrent si son graphe de connexions comporte des cycles.

**Classification supervisée avec des réseaux de neurones multicouches**

- La première couche, dite couche d'entrée, est de taille $m$, traditionnellement adaptée à la taille des entrées $\mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jm})$, $j \in \{1,\ldots,n\}$.
- Dans un problème à $c$ classes, la dernière couche, dite couche de sortie, est de taille $c$, où $c$ est le nombre de classes à prédire.
- Les couches intermédiaires, dites couches cachées, peuvent avoir des tailles différentes. Pour clarifier les explications, on considère un réseau multicouche à une couche cachée de taille $\bar{n}$.

On note $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{im})$ les poids entre la couche d'entrée et l'$i^{ème}$ neurone de la couche cachée, $i \in (1,\ldots,\bar{n})$. $\boldsymbol{\beta} = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{ic})$ est le vecteur de poids

entre l'$i^{ème}$ neurone caché et les $c$ neurones de sortie, et $\mathbf{y}'_j = (y'_{j1}, y'_{j2}, \ldots, y'_{jc})$ est le vecteur des états de la couche de sortie associés à l'exemple $\mathbf{x}_j$.

Fonctionnellement, on peut définir le réseau comme l'ensemble des activations associées à chaque exemple (Eq. 2.3). Il s'agit alors de trouver les poids et seuils du réseau en minimisant une fonction de coût $L$ (*loss function*), qui calcule l'erreur. La fonction de coût peut être l'erreur quadratique moyenne sur l'ensemble d'apprentissage (Eq. 2.4) :

$$\sum_{i=1}^{\bar{n}} \beta_i f(\mathbf{x}_j, \mathbf{w}_i) = \mathbf{y}'_j \qquad j = 1, \ldots, n \tag{2.3}$$

$$\text{en minimisant} \quad L = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (\mathbf{y}_j - \mathbf{y'}_j)^2} \tag{2.4}$$

L'apprentissage est achevé en présentant plusieurs fois le jeu de données au réseau, et en adaptant les poids des connexions avec pour objectif de faire converger la fonction de coût vers son minimum global.

Dans la littérature, on est amené à confondre les fonctions d'activation et de transfert pour ne plus parler que de fonction d'activation. La fonction d'activation cherche à modéliser le potentiel d'action des neurones, et plusieurs familles de fonctions peuvent être utilisées, dont les plus courantes sont les fonctions linéaires, à seuil ou sigmoïdes.

Dans l'exemple explicatif pris ci-dessus, on considère des fonctions d'activation linéaires, et la classe prédite pour un exemple $\mathbf{x}_j$ correspond à l'indice du maximum du vecteur $\mathbf{y}'_j$ (règle du *winner-takes-all*).

#### 2.2.1.2 Extreme Learning Machines

Les *Extreme Learning Machines* (ELM) sont des réseaux de neurones à une couche cachée.

Les auteurs ont montré que le système d'équations 2.3 peut être réécrit de manière compacte sous la forme d'une équation matricielle, tel que $\mathbf{H}\boldsymbol{\beta} = \mathbf{Y'}$, où $\mathbf{H}$ est la matrice des fonctions de décision de la couche cachée (Eq. 2.5) :

$$H(\mathbf{x},\mathbf{w}) = \begin{bmatrix} f(\mathbf{x}_1,\mathbf{w}_1) & \dots & f(\mathbf{x}_1,\mathbf{w}_{\bar{n}}) \\ \vdots & \ddots & \vdots \\ f(\mathbf{x}_n,\mathbf{w}_1) & \dots & f(\mathbf{x}_n,\mathbf{w}_{\bar{n}}) \end{bmatrix}_{n \times \bar{n}}, \tag{2.5}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1^{\mathsf{T}} \\ \vdots \\ \boldsymbol{\beta}_{\bar{n}}^{\mathsf{T}} \end{bmatrix}_{\bar{n} \times c} \quad \text{et } \mathbf{Y'} = \begin{bmatrix} \mathbf{y}_1'^{\mathsf{T}} \\ \vdots \\ \mathbf{y}_n'^{\mathsf{T}} \end{bmatrix}_{n \times c}$$

où $\bar{n}$ est le nombre de neurones de la couche cachée.

Les poids de la couche cachée sont déterminés de manière aléatoire, et les poids de la couche de sortie sont calculés de manière analytique. Les poids $(\beta_1^q, \dots, \beta_m^q)$ entre la couche cachée et un neurone de sortie $q$ sont calculés en appliquant la méthode des moindres carrés au système $\boldsymbol{\beta} = \mathbf{H}^{\mathsf{MP}}\mathbf{Y'}$, où $\mathbf{H}^{\mathsf{MP}}$ est l'inverse de Moore-Penrose de la matrice $\mathbf{H}$.

#### 2.2.1.3 Long Short-Term Memory Recurrent Neural Networks

Les LSTM-RNN [48] (*Long Short Term Memory – Recurrent Neural Networks*, Fig. 2.2) sont des réseaux de neurones récurrents capables de "se souvenir" et "d'oublier" certaines configurations apprises dans un certain intervalle de temps. Cette capacité les rend particulièrement adaptés aux problèmes où le temps joue un rôle essentiel.

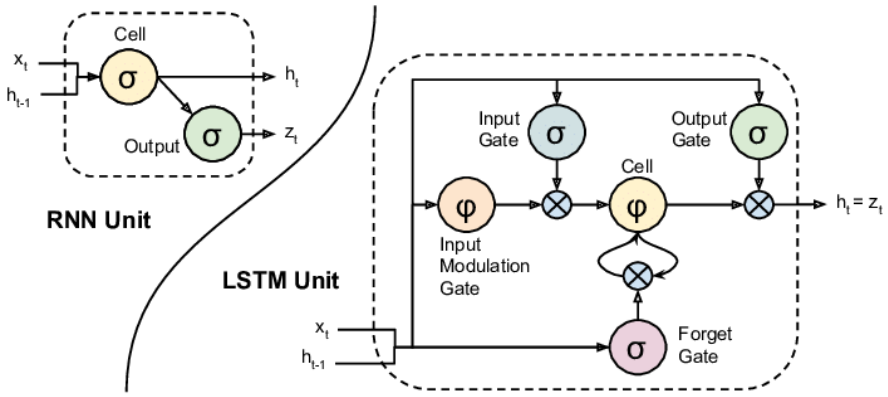



FIGURE 2.2 – Fonctionnement de cellules de Recurrent Neural Network et Long Short Term Memory, représenté par Donahue *et al.* [26].

#### 2.2.1.4 Réseaux de neurones profonds

Un réseau de neurones est dit profond lorsque le nombre de couches cachées est élevé et que l'apprentissage se fait couche par couche. En anglais, on parle de *deep neural network* (DNN) [39].

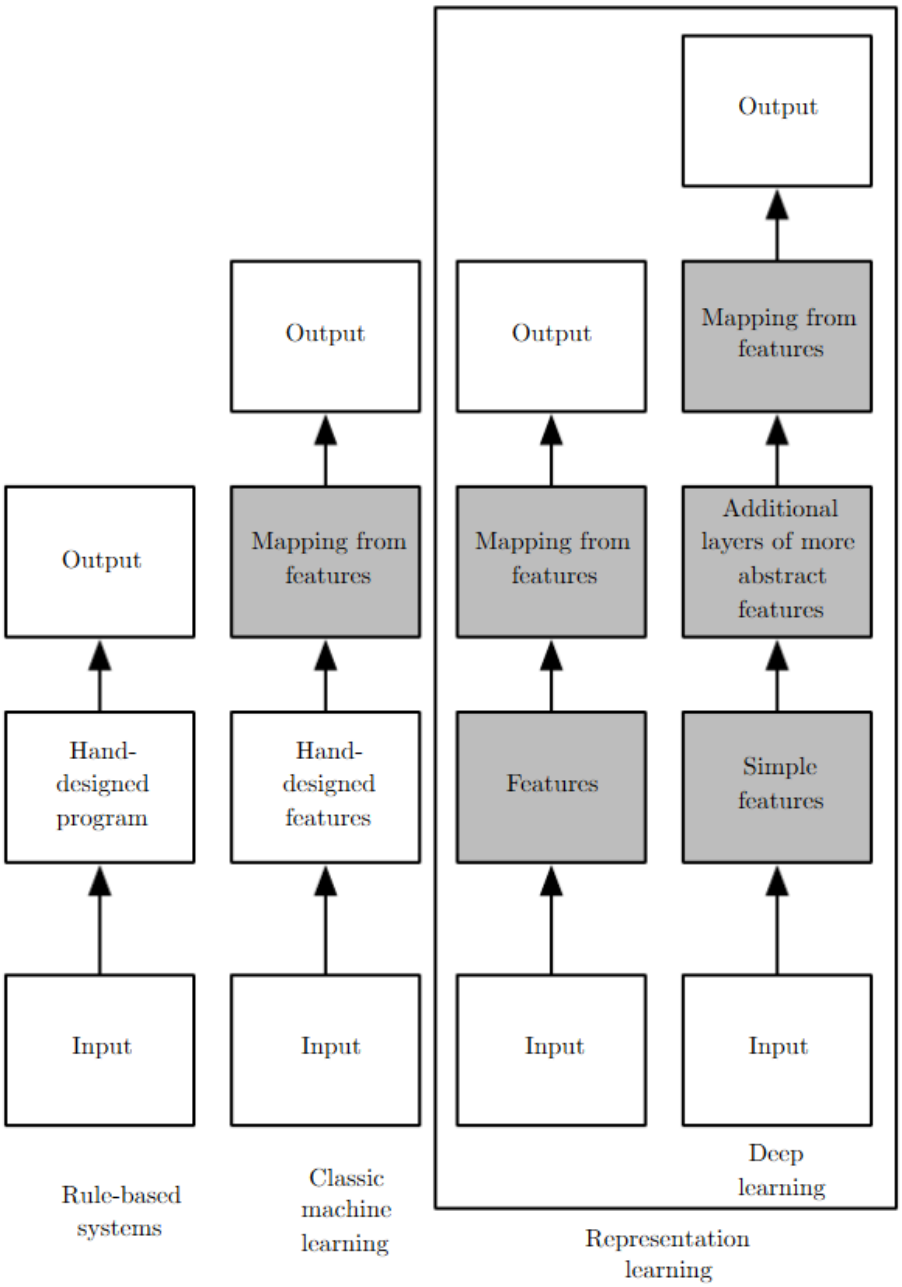



FIGURE 2.3 – Différences entre plusieurs méthodes d'intelligence artificielle. Les composants gris sont capables d'apprendre à partir de données, d'après [39].

**Motivation et différences avec les autres classifieurs** Les réseaux profonds, dont l'architecture se veut compacte, transforment les données en entrée en une succession de *features* de plus en plus abstraites (Fig. 2.3).

**Convolutional Neural Networks** En 1990, LeCun [71] pose les bases des réseaux de neurones convolutionnels avec le tout premier réseau du type, baptisé LeNet. Le concept a évolué jusqu'à s'imposer dans le domaine de la reconnaissance de formes, de scènes, et plus globalement pour la reconnaissance automatisée à partir d'images. Un réseau convolutionnel (ConvNet ou CNN) reçoit en entrée une image (non prétraitée) et produit un ensemble de *features* de haut niveau, que l'on utilise pour apprendre un classifieur. Un CNN est composé de deux types de couches principales, dites de convolution et de pooling. D'autres types de couches (de correction, totalement connectée ou de perte) peuvent également être utilisées. Dans un CNN, une couche peut être perçue comme un réseau de neurones en trois dimensions, avec une épaisseur liée au nombre de filtres utilisés.

**Couche de convolution - CONV** Une couche CONV est composée de filtres, chacun étant appliqué à une petite partie (appelée champ récepteur) de l'image en entrée. La dimension des filtres et le pas de chevauchement définissent la dimension de la couche de convolution.

**Couche de pooling - POOL** Le pooling peut être assimilé à un sous-échantillonage. On applique sur une "tuile" de quelques neurones adjacents un opérateur comme le maximum (max-pooling) ou la moyenne (*mean-pooling*) afin de créer une seule sortie, dont la valeur est définie par l'opérateur.

**Couche de correction - RECT** Cette couche joue le rôle de fonction d'activation sur les signaux issus des filtres. En particulier, la fonction ReLU (*Rectifier Linear Unit*, $f(u) = max(0,u)$) force les neurones à retourner des valeurs positives.

**Couche totalement connectée - FC** Une couche totalement connectée est construite à partir de tous les étages de la couche précédente, et est complètement connectée à la suivante. Cette couche (vectorielle, sans épaisseur) peut être consituée de neurones classiques.

**Couche de perte - LOSS** La couche LOSS estime l'erreur entre sortie désirée et sortie calculée via une fonction de coût, par exemple la fonction *softmax* qui transforme la valeur de chaque neurone de sortie en une probabilité d'appartenance à une classe.

Toutes ces couches peuvent être empilées, en plusieurs exemplaires, et permettre ainsi de construire des CNN profonds.

**Exemples de CNN** Typiquemment, un CNN est un enchaînement de couches dont l'ordre a un sens pour traiter des données en entrée (INPUT). L'architecture de base peut être la suivante : INPUT ↦ CONV ↦ ReLU ↦ FC. Des architectures plus complexes, comme VGG-Face [89], ResNet [45] (Fig. 2.4) ou SoundNet [5] ont été proposées pour solutionner des problématiques de reconnaissance faciale ou de traitement de données auditives.

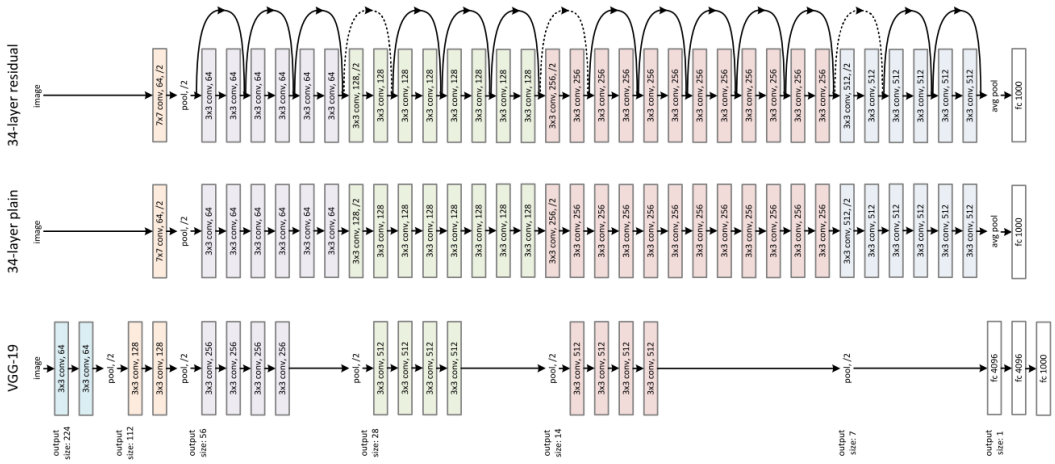



FIGURE 2.4 – Réseau convolutionnel ResNet (152 couches) [45].

## 2.2.2 Régresseurs linéaires

Les régresseurs linéaires sont une famille de régresseurs qui permettent de modéliser la relation entre un exemple $\mathbf{x}$, dont les composantes dont distribuées suivant une loi normale, et une valeur réelle $y$, par le biais de fonctions linéaires. Le modèle linéaire de base considère le système d'équations linéaires suivant (Eq. 2.6) :

$$y_i = p_0 + p_1 x_{i1} + \ldots + p_m x_{im} + \epsilon_i \quad i = 1, \ldots, n \tag{2.6}$$
$$y_i = \mathbf{x}^\top \boldsymbol{p} + \epsilon_i$$

avec $\boldsymbol{p}$ les paramètres du système et $\boldsymbol{\epsilon}$ une variable d'erreur.

Dans l'état de l'art, on retrouve des usages des moindres carrés partiels [123] (*Partial Least Square*, PLS) et du modèle linéaire généralisé (*Generalized Linear Model*, GLM).

La PLS consiste à modéliser la relation entre deux matrices **X** et **X'** en les projetant dans un espace où leur variance résiduelle est minimale. Pour ce faire, les matrices sont décomposées comme suit (Eq. 2.7 et Eq. 2.8) :

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{2.7}$$

$$\mathbf{X'} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \tag{2.8}$$

avec **T** et **U** les projections de **X** et **X'**, respectivement ; **P** et **Q** les coefficients de corrélation de la décomposition ; et **E** et **F** les erreurs résiduelles.

Le modèle GLM peut être vu comme une généralisation du modèle linéaire de base, où les données n'ont pas besoin de suivre une distribution normale. Pour ce faire, le modèle fait usage d'une fonction de lien (*link function*) afin d'exprimer la relation linéaire existant entre un prédicteur linéaire (Eq. 2.6) et la distribution des données.

## 2.3 La *Support Vector Machine*

### 2.3.1 Définition formelle

La *SupportVector Machine* (SVM), [18] est une méthode très documentée et très utilisée pour les cas de classification binaire et multiclasse.

Dans le cas binaire, on considère l'ensemble d'apprentissage $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, où les $y_i$ peuvent prendre la valeur +1 pour les exemples ou la valeur -1 pour les contre-exemples. L'action de la SVM consiste à discriminer les données $\mathbf{x} \in \mathbb{R}^m$ pour lesquelles $y = +1$ de celles pour lesquelles $y = -1$ en trouvant un hyperplan séparateur $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ qui vérifie la fonction de décision suivante (Eq. 2.9) :

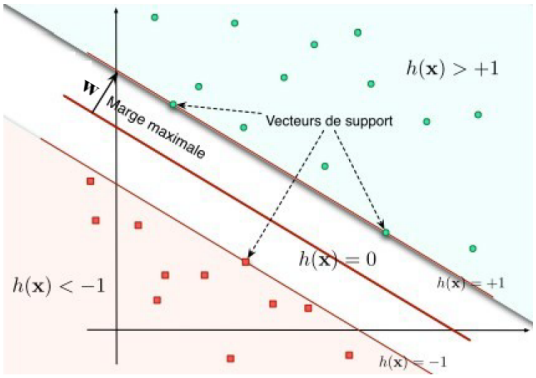



FIGURE 2.5 – Hyperplan optimal d'un problème binaire d'après [17]

$$\begin{cases} \mathbf{w}^\top x_i + b \geq 1, & \text{si } y_i = +1 \\ \mathbf{w}^\top x_i + b \leq 1, & \text{si } y_i = -1 \end{cases} \tag{2.9}$$

Les deux conditions (2.9) peuvent se résumer en $y_i\left(\mathbf{w}^\top x_i + b\right) \geq 1$. Dans la littérature, on montre que, pour que la marge $2/||\mathbf{w}||$ entre les vecteurs de support et l'hyperplan soit maximale (ce qui améliore la capacité de généralisation), il faut minimiser $||\mathbf{w}||$, ce qui donne l'expression primale du problème d'optimisation (Eq. 2.10) :

$$\begin{aligned} & min \quad \frac{1}{2}\mathbf{w}^\top \mathbf{w} \\ \text{sous les contraintes} \quad & y_i\left(\mathbf{w}^\top x_i + b\right) \geq 1, \qquad i = 1, \ldots, n \end{aligned} \tag{2.10}$$

Dans le cas où les données sont linéairement séparables, il est possible de trouver un hyperplan optimal qui vérifie ces contraintes. Dans le cas contraire, on tente de les rendre séparables en les projetant dans un espace $\mathscr{H}$ de plus grande dimension, grâce à une fonction de *mapping* $\phi : \mathbf{x} \in \mathbb{R}^m \mapsto \mathscr{H}$ (voir Astuce du noyau, 2.3.3). Comme le plongement dans un espace de plus grande dimension ne suffit pas toujours à rendre les données séparables, on introduit le paramètre $C$, dit de porosité, qui autorise la construction d'hyperplans avec des erreurs de classification résiduelles $\xi$. Le problème d'optimisation peut alors être reformulé ainsi (Eq. 2.11) :

$$\text{(2.11)} \qquad min \quad \frac{1}{2}\mathbf{w}^\top\mathbf{w} + C\sum_{i=1}^{n}\xi_i, \qquad C \in \mathbb{R}^+$$
$$\text{tel que} \quad \begin{cases} y_i\left(\mathbf{w}^\top\phi(x_i) + b\right) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

avec $i = 1, \ldots, n$.

Une formulation probabiliste de la SVM est la RVM [117] (*Relevance Vector Machine*). Le principe de la SVM peut aussi être appliqué à des problèmes de régression (et non de classification). On parle alors de SVR [27].

### 2.3.2 Extension multiclasse

On peut décomposer un problème multiclasse en un nombre fini de problèmes binaires. Cette approche est utilisée afin de résoudre ces problèmes à l'aide de SVM binaires. Par abus de langage, ces SVM sont qualifiées de multiclasses, bien que les SVM multiclasses [129] fassent référence à une modélisation différente (qui ne sera pas évoquée ici). Deux approches de décomposition se démarquent dans la littérature : l'approche un-contre-tous et l'approche un-contre-un.

L'approche un-contre-tous (*one-vs-rest*, OVR) consiste à construire $c$ classifieurs binaires, chacun apprenant à discriminer une classe de l'ensemble des autres. La résolution du problème d'optimisation (Eq. 2.11) permet de définir $c$ fonctions de décision (Eq. 2.9), chacune sur l'un des classifieurs. Un point $\mathbf{x}$ donné appartient à la classe dont la fonction de décision a la valeur maximale (Eq. 2.12) :

$$\text{(2.12)} \qquad \text{Classe}(\mathbf{x}) = \arg\max_c \left(\left(\mathbf{w}^c\right)^\top x_i + b^c\right)$$

où $c$ dénote le $c^{\text{ième}}$ classifieur et non un exposant.

L'approche un-contre-un (*one-vs-one*, OVO) consiste, elle, à construire autant de classifieurs qu'il y a de paires de classes. Dans ce cas, on applique traditionnellement un vote majoritaire : la classe qui aura récolté le plus de votes sera celle attribuée à l'exemple.

Hsu et Lin [50] ont montré que les SVM multiclasses un-contre-un se démarquaient de l'autre approche par leurs performances et leur praticité d'usage sur les jeux de données de grande dimension.

### 2.3.3 Astuce du noyau

L'usage d'une fonction noyau (*kernel function*) permet de projeter un point de dimension $m$ dans un espace de dimension $g$ supérieure ($g \gg m$). Pour ce faire, on utilise une fonction de *mapping* $\phi : \mathbf{x} \in \mathrm{R}^m \mapsto \mathscr{H}$ tel que $K(\mathbf{x},\mathbf{x}') = \left(\phi(\mathbf{x}),\phi(\mathbf{x}')\right)$. On appelle matrice de Gram la matrice semi-définie positive $\mathbf{G}$ du noyau $K(\mathbf{x},\mathbf{x}')$ où $\mathbf{G}_{ij} = K(\mathbf{x}_i,\mathbf{x}'_j)$.

Le choix du noyau détermine la forme des frontières qui pourront être définies dans l'espace initial. Pour cette raison, il est nécessaire de le choisir en accord avec la nature des données. Par exemple, un noyau linéaire définira des hyperplans, alors qu'un noyau gaussien (dit RBF, *radial basis function*) définira des hypersphères. D'autres noyaux, comme le noyau intersection d'histogrammes (HI) sont spécifiques aux données sous forme d'histogrammes et sont couramment utilisés pour des données images. Une liste de plusieurs noyaux et de leurs applications a été proposée par Deng *et al.* [23].

#### 2.3.3.1 Quelques noyaux

**Noyau linéaire** Le noyau linéaire est le produit scalaire des points et est défini par (Eq. 2.13) :

$$K\left(\mathbf{x},\mathbf{x}'\right) = \mathbf{x}^{\mathsf{T}}\mathbf{x} \tag{2.13}$$

**Noyau RBF** Egalement très courant dans la littérature, le noyau RBF est considéré comme une première approche nécessaire, indépendamment de la nature des données (Eq. 2.14) :

$$K\left(\mathbf{x},\mathbf{x}'\right) = exp\left(-\gamma||\mathbf{x}-\mathbf{x}'||^2\right) \tag{2.14}$$

où $\gamma$ est un coefficient réel.

**Noyau intersection d'histogrammes** Le noyau intersection d'histogrammes implémente une mesure de similarité entre deux histogrammes (Eq. 2.15) :

$$K\left(\mathbf{x},\mathbf{x}'\right) = \sum_{i=1}^{m} min\left(x_i,x'_i\right) \tag{2.15}$$

## 2.4 Classifieur neuronal incrémental à base de prototypes

Les classifieurs à base de prototypes regroupent des méthodes de classification supervisée et non supervisée, particulièrement adaptées à l'analyse de données complexes et de grande dimension [11]. L'utilisation d'un réseau de neurones comme support à un classifieur à base de prototypes tire son inspiration des cartes auto-organisatrices de Kohonen. Dans un classifieur neuronal à base de prototypes (*Incremental Neural Network*, INN), un prototype est assimilable aux poids d'un neurone caché, qui peut calculer une similarité entre le prototype stocké et une observation. Les prototypes sont représentables dans l'espace des données via leur vecteur de poids, ce qui les rend interprétables notamment pour des applications de recherche pluridisciplinaires.

L'INN est inspiré du modèle ART de Grossberg [41] et est fondé sur des principes issus des sciences cognitives. Il a été proposé par Azcarraga et Giacometti [6] puis repris et adapté à des tâches de classification supervisées par Puzenat [94].

Le classifieur neuronal incrémental utilisé ici est un réseau de neurones *feed-forward* à trois couches. En entrée, le réseau reçoit sur la première couche un vecteur $\mathbf{x}$ de dimension $m$. La seconde couche est la couche des prototypes. Sa taille peut augmenter au cours de l'apprentissage, et chacune de ses cellules est connectée à toutes les cellules en entrée et à une seule cellule en sortie. La dernière couche permet de lire les sorties, i.e. les classes. Une cellule de la dernière couche est connectée à tous prototypes correspondant à sa classe. Il n'y a pas d'apprentissage vers la dernière couche du réseau.

### 2.4.1 Contrôle et apprentissage

On considère un ensemble de données et leurs observations $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ tel que $\mathbf{x} \in \mathbb{R}^m$ et $y \in \{k_1, \ldots, k_c\}$ une étiquette de classe.

On présente successivement les exemples accompagnés de leur étiquette de classe à l'INN. A chaque fois que l'algorithme rencontre un nouvelle classe, il crée un nouveau prototype avec le premier exemple présenté pour cette classe.

Pour chaque exemple $\mathbf{x}$ présenté, l'algorithme cherche le prototype le plus proche $\mathsf{P}_{\text{meilleur}}$, conformément à une mesure de similarité ou à une distance $\delta$. Si

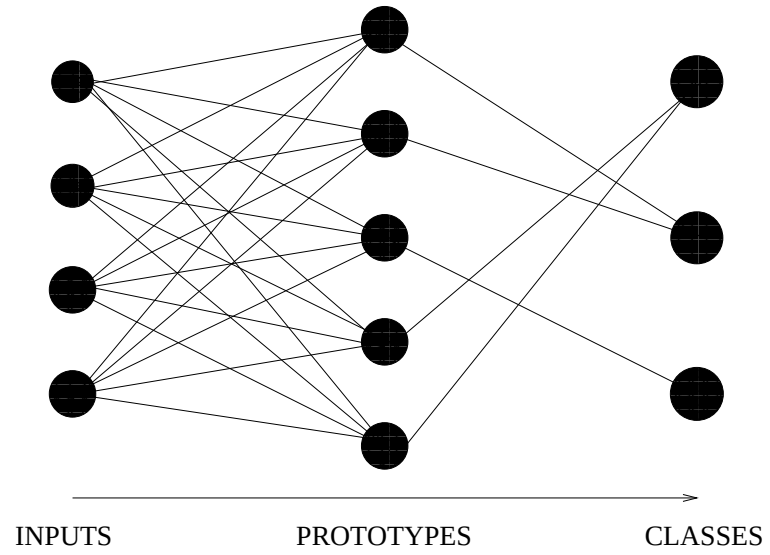



FIGURE 2.6 – Illustration schématique d'un INN [94].

$\mathsf{Classe}(\mathbf{x}) \neq \mathsf{Classe}(\mathsf{P}_{\mathsf{meilleur}})$, un nouveau prototype est créé. Sinon, on vérifie deux conditions :

1. $\delta(\mathsf{P}_{\mathsf{meilleur}}, \mathbf{x}) \leq \mathsf{s}_{\mathsf{inf}}$, avec $\mathsf{s}_{\mathsf{inf}} \in \mathbb{R}^+$ le seuil d'influence. On dit que l'exemple doit être sous l'influence de son meilleur prototype.
2. $\delta(\mathsf{P}_{\mathsf{meilleur}}, \mathbf{x}) - \delta(\mathsf{P}_{\mathsf{second}}, \mathbf{x}) > \mathsf{s}_{\mathsf{conf}}$, avec $\mathsf{s}_{\mathsf{conf}} \in \mathbb{R}^+$ le seul de confusion et $\mathsf{P}_{\mathsf{second}}$ le second meilleur prototype de $\mathbf{x}$ tel que $\mathsf{Classe}(\mathsf{P}_{\mathsf{second}}) \neq \mathsf{Classe}(\mathsf{P}_{\mathsf{meilleur}})$. On dit qu'il ne doit pas y avoir de confusion.

Si l'une des deux conditions n'est pas vérifiée, alors un nouveau prototype est créé. Sinon, les poids du prototype gagnant $\mathsf{P}_{\mathsf{meilleur}}$ sont modifiés afin de le rapprocher de l'exemple (Eq. 2.18).

L'algorithme exprimé ici est un algorithme par compétition (dit *winner-takes-all*) : le prototype gagnant est le seul à être activé et modifié.

#### 2.4.1.1 Contrôle

L'INN a trois hyperparamètres de contrôle de l'apprentissage ; le seuil d'influence $\mathsf{s}_{\mathsf{inf}}$, le seuil de confusion $\mathsf{s}_{\mathsf{conf}}$ et le coefficient de rapprochement $\alpha$. Ces hyperparamètres de contrôle permettent de répondre au dilemme de stabilité-plasticité, i.e. de tenir compte des nouveaux éléments à apprendre sans oublier ceux déjà mémorisés.

**Seuil d'influence** Le seuil d'influence $s_{inf}$ du classifieur caractérise la distance $\delta$ maximale acceptable entre un exemple $\mathbf{x}$ et son meilleur prototype $P_{meilleur}$ pour que $\mathbf{x}$ soit réputé sous l'influence de $P_{meilleur}$. Si cette condition (Eq. 2.16) n'est pas satisfaite en apprentissage, alors on crée un nouveau prototype.

$$\delta(P_{meilleur}, \mathbf{x}) \leq s_{inf} \tag{2.16}$$

**Seuil de confusion** Le seuil de confusion $s_{conf}$ du classifieur aide à lever les ambiguïtés de classification dans les zones frontières entre classes distinctes. Il caractérise l'écart entre, d'une part, la distance entre un exemple $\mathbf{x}$ et son meilleur prototype $P_{meilleur}$, et, d'autre part, la distance entre cet exemple et son second meilleur prototype $P_{second}$, avec $Classe(P_{second}) \neq Classe(P_{meilleur})$. Si cette condition (Eq. 2.17) n'est pas remplie en apprentissage, on crée un nouveau prototype.

$$\delta(P_{meilleur}, \mathbf{x}) - \delta(P_{second}, \mathbf{x}) > s_{conf} \tag{2.17}$$

Les valeurs optimales pour les seuils d'influence et de confusion dépendent fortement des données. La normalisation des données, ainsi que le calcul de leur variance sont des étapes permettant de faciliter la recherche en grille des hyperparamètres.

**Coefficient de rapprochement** Le coefficient de rapprochement $\alpha \in ]0,1]$ intervient dans les situations où les conditions 2.16 et 2.17 sont satisfaites, i.e. quand $\mathbf{x}$ est sous l'influence de son meilleur prototype et qu'il n'y a pas de risque de confusion. Les poids du neurone $P_{meilleur}$ sont alors modifiés afin de le rapprocher de $\mathbf{x}$. Si on note $\mathbf{w}$ le vecteur de poids de $P_{meilleur}$, le rapprochement du meilleur prototype s'exprime par :

$$\forall i \in \{1, \ldots, m\}, w_i \leftarrow w_i + \alpha * (x_i - w_i) \tag{2.18}$$

Une grande valeur de $\alpha$ implique un fort rapprochement des prototypes vers les exemples et par conséquent, un accroissement du nombre de prototypes. *A contrario*, une valeur faible limite le déplacement des prototypes et par conséquent leur nombre.

**Mesures de similarité** Le calcul de similarité entre les prototypes et les exemples se fait via une mesure de similarité ou via une distance, typiquement choisie en accord avec le problème à traiter. Cette possibilité rend les classifieurs incrémentaux flexibles et adaptables. La famille des distances de Minkowski (Eq. 2.19) sont des distances de l'état de l'art, dont le cas $p = 2$ revient à calculer une distance euclidienne.

$$d_{mink_p}(\mathbf{x},\mathbf{y}) = \left(\sum_{i=1}^{n} |\, x_i - y_i \,|^p\right)^{1/p} \tag{2.19}$$

#### 2.4.1.2 Algorithme pour l'apprentissage

Compte-tenu des possibilités de contrôle énoncées précédemment, il est possible de formuler un algorithme pour l'apprentissage de nouveaux motifs. Les hyperparamètres $s_{\text{inf}}$, $s_{\text{conf}}$ et $\alpha$ sont réputés trouvés (via une recherche en grille, par exemple).

1. **Présenter** un motif $\mathbf{x} = (x_1, \ldots, x_m)$ et son étiquette $y$
2. **Rechercher** $\mathsf{P}_{\text{meilleur}}$ tel que $\delta(\mathsf{P}_{\text{meilleur}}, \mathbf{x})$ soit minimale
3. **Rechercher** $\mathsf{P}_{\text{second}}$ tel que $\delta(\mathsf{P}_{\text{second}}, \mathbf{x})$ soit minimale, sous la contrainte $\mathsf{Classe}(\mathsf{P}_{\text{second}}) \neq \mathsf{Classe}(\mathsf{P}_{\text{meilleur}})$
4. **Si**

$$\begin{cases} \delta(\mathsf{P}_{\text{meilleur}}, \mathbf{x}) \leq s_{inf} \\ \delta(\mathsf{P}_{\text{meilleur}}, \mathbf{x}) - d(\mathsf{P}_{\text{second}}), \mathbf{x} > s_{conf} \\ \mathsf{Classe}(\mathsf{P}_{\text{meilleur}}) = \mathsf{Classe}(\mathbf{x}) \end{cases}$$

**Alors** rapprocher $\mathsf{P}_{\text{meilleur}}$ de $\mathbf{x}$ :

$$\forall i, w_i \leftarrow w_i + \alpha(x_i - w_i) \quad \text{avec } \mathsf{P}_{\text{meilleur}} = \{w_1, \ldots, w_m\} \text{ et } \mathbf{x} = \{x_1, \ldots, x_m\}$$

**Sinon** créer un nouveau prototype $\mathsf{P}$ :

$$\begin{array}{ll} \forall i \in \{1, \ldots, m\}, w'_i \leftarrow x_i & \text{avec } \mathsf{P} = \{w'_1, \ldots, w'_m\} \\ \text{et définir} & \mathsf{Classe}(\mathsf{P}) = y \end{array}$$

A chaque fois que l'algorithme rencontre un nouvelle classe, il crée un nouveau prototype avec le premier exemple présenté pour cette classe. Une fois tous les

motifs présentés, le modèle est composé de l'ensemble des prototypes appris par l'algorithme et peut être soit utilisé en généralisation, soit réappris (en lui présentant d'autres motifs).

### 2.4.2 Généralisation et interprétations

#### 2.4.2.1 Algorithme pour la généralisation

La généralisation respecte les mêmes contraintes que l'apprentissage. Toutefois, il n'y a plus ajout ni modification de prototypes. L'algorithme ci-dessous permet la classification de motifs n'ayant pas participé à l'apprentissage.

1. **Présenter** un motif $\mathbf{x}$
2. **Rechercher** $\mathsf{P}_{\mathsf{meilleur}}$ tel que $\delta(\mathsf{P}_{\mathsf{meilleur}}, \mathbf{x})$ soit minimale
3. **Rechercher** $\mathsf{P}_{\mathsf{second}}$ tel que $\delta(\mathsf{P}_{\mathsf{second}}, \mathbf{x})$ sous la contrainte $\mathsf{Classe}(\mathsf{P}_{\mathsf{second}}) \neq \mathsf{Classe}(\mathsf{P}_{\mathsf{meilleur}})$
4. **Si** $\mathsf{P}_{\mathsf{meilleur}}$ est trop éloigné de $\mathbf{x}$ ou s'il y a risque de confusion, i.e. si :

$$\begin{cases} \delta(\mathsf{P}_{\mathsf{meilleur}}, \mathbf{x}) > \mathsf{s}_{\mathsf{inf}} \\ \text{ou} \\ \delta(\mathsf{P}_{\mathsf{meilleur}}, \mathbf{x}) - d(\mathsf{P}_{\mathsf{second}}, \mathbf{x}) \leq s_{conf} \end{cases}$$

**Alors** rejeter $\mathbf{x}$ (non-réponse)
**Sinon Si** $\mathsf{Classe}(\mathsf{P}_{\mathsf{meilleur}}) = \mathsf{Classe}(\mathbf{x})$
**Alors** $\mathbf{x}$ est reconnu (bonne réponse)
**Sinon** $\mathbf{x}$ n'est pas reconnu (mauvaise réponse)

Les concepts de *non-réponse*, *bonne réponse* et *mauvaise réponse* sont discutés dans le paragraphe 2.4.2.2.

**Implémentation de l'algorithme** La recherche du meilleur prototype pour un exemple donné est basée sur une mesure de similarité entre les deux entités. Si on dispose d'un modèle de $p$ prototypes, cela implique de calculer, pour chaque élément en entrée, $p$ distances pour le classifier. Avec un nombre "raisonnable" de prototypes, le calcul (qui est de complexité $\mathcal{O}(p)$) est faisable. Toutefois, dans le pire des cas, la complexité est de $\mathcal{O}(n^2)$ (avec n le nombre d'exemples).

Afin de diminuer la complexité, on utilise un arbre *k-d* (*k-dimensional tree*) pour stocker les prototypes. De cette manière, les prototypes sont représentés dans un

espace où la relation d'ordre de grandeur entre eux est connue et implémentée. La complexité de la recherche du meilleur prototype pour $n$ exemples devient alors $\mathcal{O}(n * log(p))$. L'utilisation de *k-d trees* augmente considérablement la vitesse de classification au moyen d'un INN.

#### 2.4.2.2 Réponses du classifieur et interprétations

En sortie, le classifieur peut fournir trois types de réponses. Les bonnes réponses correspondent au cas où toutes les contraintes sont respectées, et où l'étiquette du motif inconnu est celle du prototype gagnant (on considère qu'on est en phase d'évaluation, et que les étiquettes sont disponibles). Intuitivement, plus le nombre de bonnes réponses est élevé, plus la qualité de la classification est bonne. Les mauvaises réponses, de manière analogue, correspondent aux cas où toutes les contraintes ont été respectées, mais où l'étiquette du motif inconnu n'était pas celle du prototype gagnant.

Les non-réponses sont rendues dans tous les autres cas, i.e. lorsque le prototype gagnant est trop éloigné du motif, ou que le risque de confusion entre le prototype gagnant et le second meilleur prototype est trop élevé. Une non-réponse est comparable avec le jugement humain face à une situation inconnue, et correspond à la réponse "je ne sais pas", marquant l'absence de décision plutôt qu'une décision bonne ou mauvaise. Le nombre de non-réponses peut être considéré comme un indicateur de la fiabilité des prédictions du système : un grand nombre de non-réponses correspond à un système peu fiable, bien que les bonnes réponses seront, elles, plus fiables [15].

## 2.5 La mémoire associative multimodale

La mémoire associative multimodale (*Bidirectional Associative Memory*, BAM) est un réseau de neurones récurrent capable de stocker et de rappeler des paires de vecteurs grâce à un processus de réverbération entre ses couches. Historiquement, la BAM est une adaptation du réseau de Hopfield pour la réalisation d'associations hétérogènes. Proposée par Kosko [66], la BAM est composée de couches complètement connectées dans les deux sens : l'activation suit un chemin qui boucle sur lui-même. Les couches, dont le nombre de cellules est indépendant, reçoivent en entrée des paires hétérogènes de vecteurs, par exemple un nom et un numéro de

téléphone. Le modèle ainsi entrainé (phase de stockage) est capable de se "souvenir" (phase de rappel) du membre manquant d'une paire à partir de l'autre membre, ou encore de débruiter des paires proposées au modèle.

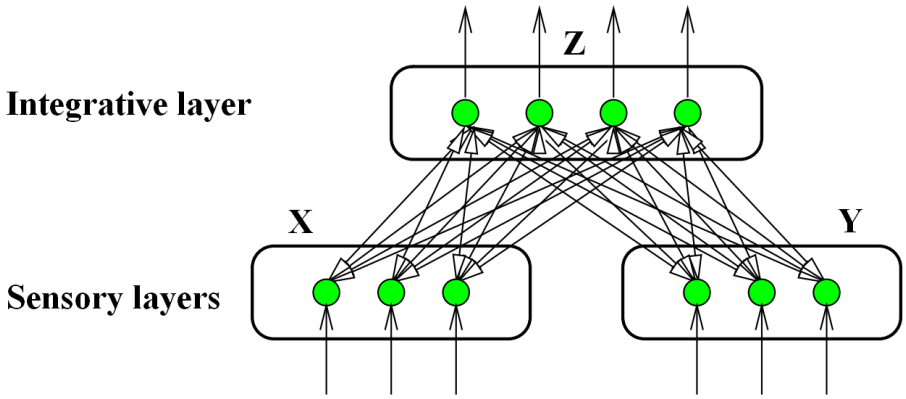



FIGURE 2.7 – Schéma de la m-BAM, d'après [98]

### 2.5.1 Apprentissage

L'apprentissage de la BAM implique la mise à jour des poids jusqu'à ce que ce ces derniers soient appris. La notion d'état stable est une notion fondamentale dans ce type de réseau, et qualifie un état dans lequel les poids du réseau n'évoluent plus en apprentissage. Un état stable est entouré d'un "bassin d'attraction", qui peut être vu comme la portion de l'espace d'entrée qui va conduire le réseau à évoluer vers l'état stable, c'est à dire une configuration favorable à la stabilité du réseau. L'algorithme d'apprentissage initial proposé par Kosko implique une matrice de poids symétrique. Kosko et d'autres auteurs [127] [110] ont montré que cette méthode limitait la capacité de la BAM (i.e. le nombre de paires qu'elle peut apprendre), estimée à la taille de la couche la plus petite.

L'algorithme PRLAB (*Pseudo Relaxation Learning Algorithm for BAM*) a été proposé par Oh et Kothari [85] pour faciliter l'apprentissage des paires, sans toutefois solutionner le problème de capacité limitée.

La BAM étudiée ici est issue des travaux de Reynaud [97], qui a proposé la multiple-BAM (m-BAM) afin de modéliser l'intégration multisensorielle telle qu'elle a lieu dans le cerveau humain. La m-BAM, ici présentée dans le cas d'une triple BAM, est composée de trois couches (Fig. 2.7). Les couches X et Y, dites basses ou sensorielles, reçoivent chacune en entrée un des deux vecteurs de la paire à

apprendre. La couche Z, dite haute ou d'intégration, permet la fusion des couches basses. Dans la m-BAM, les couches sensorielles ne sont pas connectées entre elles et sont complètement connectées à la couche d'intégration. Dans la définition générale, il peut y avoir un nombre quelconque (supérieur à deux) de couches basses.

#### 2.5.1.1 Algorithme PRLAB modifié

Reynaud [97] a modifié l'algorithme PRLAB afin de réaliser l'apprentissage d'un réseau à trois couches avec des matrices de poids non-symétriques. La non-symétrie des matrices de poids augmente considérablement la capacité du réseau.

On note $\mathbf{W}^{X\to Z}$, $\mathbf{W}^{Y\to Z}$, $\mathbf{W}^{Z\to X}$ et $\mathbf{W}^{Z\to Y}$ les matrices de poids entre les couches $X\to Z$, $Y\to Z$, $Z\to X$ et $Z\to Y$, respectivement. Les couches $X$, $Y$ et $Z$ sont de dimension $dimX$, $dimY$ et $dimZ$, respectivement. On note $\left(X^{(p)},Y^{(p)}\right)$ une paire hétérogène et $Z^{(p)}$ le résultat de son intégration. L'activation $S_{X_i}^{(p)}$ entre les couches $X\to Z$ fait intervenir les poids associés, notés $w_{ik}^{X\to Z}$, et le seuil du neurone $i$ de la couche $X$, noté $\theta_{X_i}$. Les hyperparamètres $\xi$ et $\lambda$ sont décrits dans le paragraphe 2.5.1.2. Les activations sont définies de manière similaire pour les couches $X$, $Y$ et $Z$ (Eq. 2.20) :

$$
\begin{aligned}
S_{X_i}^{(p)} &= \sum_{dimZ}^{k=1} w_{ik}^{X\to Z} Z_k^{(p)} - \theta_{X_i} \\
S_{Y_j}^{(p)} &= \sum_{dimZ}^{k=1} w_{jk}^{Y\to Z} Z_k^{(p)} - \theta_{Y_j} \\
S_{Z_k}^{(p)} &= \sum_{dimX}^{i=1} w_{ki}^{Z\to X} X_i^{(p)} + \sum_{dimY}^{j=1} w_{kj}^{Z\to Y} Y_j^{(p)} - \theta_{Z_k}
\end{aligned}
\tag{2.20}
$$

L'apprentissage avec PRLAB repose sur la résolution d'un ensemble d'inéquations linéaires (Eq. 2.21) :

$$
\begin{aligned}
S_{X_i}^{(p)}\, X_i^{(p)} - \xi \geq 0, &\qquad i = 1,\ldots,dimX \\
S_{Y_j}^{(p)}\, Y_j^{(p)} - \xi \geq 0, &\qquad j = 1,\ldots,dimY \\
S_{Z_k}^{(p)}\, Z_k^{(p)} - \xi \geq 0, &\qquad z = 1,\ldots,dimZ
\end{aligned}
\tag{2.21}
$$

L'algorithme d'apprentissage est le suivant :

1. Les poids $w_{ik}^{X\to Z}$, $w_{jk}^{Y\to Z}$, $w_{ki}^{Z\to X}$ et $w_{kj}^{Z\to Y}$ et les seuils $\theta_{X_i}$, $\theta_{Y_j}$ et $\theta_{Z_k}$ sont initialisés aléatoirement entre -1 et 1 (valeurs réelles) ;
2. Pour toutes les associations à apprendre entre$\left(X^{(p)},Y^{(p)}\right)$ et $Z^{(p)}$ :
   — pour les neurones de la couche $X$, si $S_{X_i}^{(p)}.X_i^{(p)} \leq 0$, mettre à jour poids et seuils selon :

$$\Delta w_{ik}^{X\to Z} = -\frac{\lambda}{1+dimZ}\left(S_{X_i}^{(p)} - \xi X_i^{(p)}\right)Z_k^{(p)}$$
$$\Delta\theta_{X_i} = \frac{\lambda}{1+dimZ}\left(S_{X_i}^{(p)} - \xi X_i^{(p)}\right)$$

   — pour les neurones de la couche $Y$, si $S_{Y_j}^{(p)}.Y_j^{(p)} \leq 0$, mettre à jour poids et seuils selon :

$$\Delta w_{jk}^{Y\to Z} = -\frac{\lambda}{1+dimZ}\left(S_{Y_j}^{(p)} - \xi Y_j^{(p)}\right)Z_k^{(p)}$$
$$\Delta\theta_{Y_j} = \frac{\lambda}{1+dimZ}\left(S_{Y_j}^{(p)} - \xi Y_j^{(p)}\right)$$

   — pour les neurones de la couche $Z$, si $S_{Z_k}^{(p)}.Z_k^{(p)} \leq 0$, mettre à jour poids et seuils selon :

$$\Delta w_{ki}^{Z\to X} = -\frac{\lambda}{1+dimX}\left(S_{Z_k}^{(p)} - \xi Z_k^{(p)}\right)$$
$$\Delta w_{kj}^{Z\to Y} = -\frac{\lambda}{1+dimY}\left(S_{Z_k}^{(p)} - \xi Z_k^{(p)}\right)$$
$$\Delta\theta_{Z_k} = \frac{1}{2}\left(S_{Z_k}^{(p)} - \xi Z_k^{(p)}\right)\left(\frac{\lambda}{1+dimX} + \frac{\lambda}{1+dimY}\right)$$

3. Répéter l'étape 2. jusqu'à stabilisation du réseau, i.e. jusqu'à ce que les poids et seuils n'évoluent plus.

#### 2.5.1.2 Hyperparamètres de la BAM

La BAM possède deux hyperparamètres :

— Un paramètre influant sur la taille des bassins d'attraction, $\lambda \in [0;2]$.

— Le facteur de relaxation (ou pas de modification des poids), $\xi$.

Les résultats théoriques de [97] montrent que les apprentissages les plus rapides sont produits par des valeurs de $\lambda \leq 1$ et des valeurs de $\xi \leq 20$. Par ailleurs, la

capacité de la m-BAM est estimée à $1.68t - 12$, avec $t$ la taille de la couche la plus petite.

La base d'apprentissage est présentée plusieurs fois (une présentation de la base complète est nommée "époque"), et à chaque présentation d'un exemple, les poids et seuils sont mis à jour. Lorsque plus aucune modification n'est nécessaire, i.e. lorsque les poids sont stables pour toute la base, l'algorithme s'arrête. Le nombre d'époques est un indicateur de la vitesse d'apprentissage.

L'algorithme présenté n'est valable que si les états du réseau sont bipolaires, i.e. si l'état d'un neurone est -1 ou 1. Bien que cela puisse être vu comme un frein à son application, Kosko[66] a montré que cela permettait au réseau de converger plus rapidement vers des états stables.

### 2.5.2 Rappel des paires apprises

Le rappel des paires apprises consiste à présenter au réseau des exemples proches (ou bruités) des paires apprises. A cette étape, il n'y a plus d'évolution des poids ni des seuils, mais les états des cellules sont modifiés pour "reconstituer" les état appris. Par exemple, pour la couche X, l'activation arrivant à une cellule $X_i$ à l'itération $t$ est comparée à son seuil $\theta_{X_i}$. L'état de la cellule $X_i$ est alors mis à jour suivant :

$$X_i(t+1) = \begin{cases} 1 & si \quad \sum_{k=1}^{dimZ} w_{ik} Z_k(t) > \theta_{X_i} \\ X_i(t) & si \quad \sum_{k=1}^{dimZ} w_{ik} Z_k(t) = \theta_{X_i} \\ -1 & si \quad \sum_{k=1}^{dimZ} w_{ik} Z_k(t) < \theta_{X_i} \end{cases} \tag{2.22}$$

Le nombre de modifications d'états pour un même exemple, aussi appellé nombre de "réverbérations", est un indicateur de la qualité de rappel du réseau.

## 2.6 Stratégies de validation et mesures de performance

En classification supervisée, la validation fait référence aux processus d'apprentissage et de généralisation (en test) et a un impact sur la qualité du modèle construit. L'évaluation de la qualité d'un classifieur ou d'un régresseur est réalisée grâce à une mesure adaptée, dite mesure de performance. Dans cette section, on

présente les principales méthodes de validation et mesures utilisées dans l'état de l'art et dans nos travaux.

### 2.6.1 Techniques de validation

L'apprentissage et la généralisation d'un modèle de classifieur peuvent se faire suivant plusieurs stratégies. Dans la suite, on retient deux techniques de validations : la validation *hold-out* (ou *leave-many-out*) et la validation croisée *leave-one-subject-out*.

La validation *hold-out*, souvent qualifiée de classique, est caractérisée par l'usage de deux jeux de données distincts, l'un consacré à l'apprentissage du modèle, et l'autre à la mesure de sa généralisation. Le modèle final est celui appris sur les données d'apprentissage, et ses performances sont celles estimées sur l'ensemble de généralisation. Cette technique est notamment adaptée lorsque les données sont nombreuses.

En validation croisée (*k-fold cross-validation*), on considère un seul jeu de données que l'on partitionne en $k$ groupes, de manière arbitraire. L'apprentissage est réalisé sur $k-1$ groupes, le groupe non utilisé servant pour la généralisation. Le processus est réitéré jusqu'à ce qu'on ait généralisé sur tous les groupes, séparément. La validation croisée est une technique intéressante, qui permet d'apprendre sur des données peu nombreuses ou bien d'estimer des performances sur l'ensemble des données à disposition. Les performances du modèle final (i.e. celui appris sur tous les groupes en même temps) correspondent aux performances moyennes obtenues en généralisation sur chaque groupe.

Il existe une variante de la validation croisée adaptée aux cas où il y a des groupes sémantiques dans les données, par exemple des groupes d'individus, appelés sujets. On parle de validation croisée *leave-one-subject-out* ou LOSO (Fig. 2.8) lorsque le partitionnement du jeu de données n'est plus arbitraire, mais qu'il est défini de manière à ce qu'un groupe contienne les données relatives à un même sujet.

En validation croisée *leave-one-subject-out*, les performances du modèle sont nécessairement *subject-independent* : à chaque itération, la généralisation a lieu sur un sujet qui n'a pas été vu en apprentissage. En validation *hold-out*, on ne peut parler d'approche *subject-independent* que si les sujets en apprentissage sont distincts de ceux en généralisation. Ces deux méthodes sont par conséquent

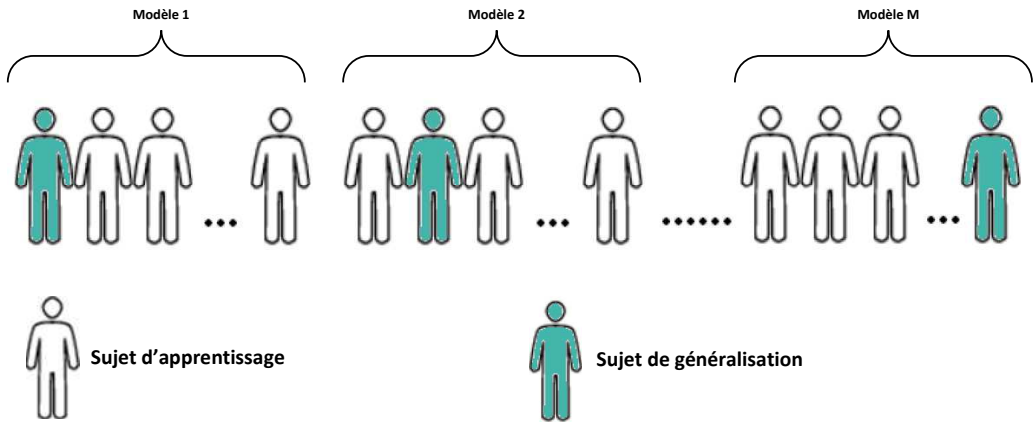



FIGURE 2.8 – Schématisation du processus de validation croisée *leave-one-subject-out*

adaptées à la construction et à l'évaluation de systèmes qui doivent classifier des éléments indépendamment des sujets.

### 2.6.2 Principaux indicateurs de performance

On note $y$ et $y'$ les sorties désirées et calculées d'un classifieur, respectivement, pour un exemple $\mathbf{x}$. L'exemple est dit bien classé lorsque la sortie calculée par le classifieur est la sortie désirée, i.e. lorsque $y = y'$. Dans le cas contraire, l'exemple est dit mal classé. Les mesures suivantes sont utilisées en classification supervisée.

La matrice de confusion CM est une représentation synthétique du nombre d'exemples bien classés et mal classés. C'est une matrice carrée où la cellule $j \in 1,\ldots,m$ de la ligne $i$ comporte le nombre d'exemples dont la sortie désirée est $i$ et dont la sortie calculée est $j$. Une classification est idéale si toutes les cellules ont la valeur zéro, sauf la première diagonale (qui comporte le nombre d'exemples bien classés). Quel que soit le nombre de classes, la matrice de confusion est le point de départ pour le calcul des indicateurs suivants :

— VP, le nombre d'exemples appartenant à une classe et qui y ont bien été assignés. On parle de prédictions positives correctes, ou de vrais positifs.
— VN, le nombre d'exemples n'appartenant pas à une classe et qui n'y ont pas été assignés. On parle de prédictions négatives correctes, ou vrais négatifs.
— FP, le nombre d'exemples n'appartenant pas à une classe mais qui y ont été assignés. On parle de faux positifs.

— FN, le nombre d'exemples appartenant à une classe et qui n'y ont pas été assignés. On parle de faux négatifs.

Le taux de succès (*accuracy*, Acc ou *Correct Classification Rate*, CCR) est le taux d'exemples bien classés (Eq. 2.23). La précision (*precision*) évalue seulement le taux de prédictions positives correctes (Eq.2.24).

$$\text{Acc} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}} \tag{2.23}$$

$$\text{Précision} = \frac{\text{VP}}{\text{VP} + \text{FP}} \tag{2.24}$$

Ensemble, ces deux indicateurs donnent une première interprétation du comportement du classifieur. Par exemple, un taux de succès élevé est synonyme d'une bonne classification. Toutefois, s'il est accompagné d'une précision faible, cela signifie que les exemples de la classe considérée sont peu nombreux (le classifieur aura d'avantage reconnu les exemples n'appartenant pas à la classe).

La sensitivité ou rappel (*sensitivity* ou *recall*), et la spécificité (*specificity*) fournissent une interprétation plus fine. Le rappel (Eq. 2.25) estime le nombre d'éléments appartenant à une classe et qui y ont été assignés, alors que la spécificité (Eq. 2.26) estime le nombre d'éléments n'appartenant pas à une classe et qui n'y ont pas été assignés.

$$\text{Rappel} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{2.25}$$

$$\text{Specificite} = \frac{\text{VN}}{\text{VN} + \text{FP}} \tag{2.26}$$

Le score F1 est la moyenne harmonique du rappel et de la précision (Eq. 2.27).

$$\text{F1} = \frac{2 * \text{Rappel} * \text{Précision}}{\text{Rappel} + \text{Précision}} \tag{2.27}$$

Ces principaux indicateurs sont pertinents lorsque l'on souhaite évaluer la capacité d'un classifieur à prédire exactement la classe d'un exemple. Dans certains cas, on souhaite évaluer la capacité d'un classifieur à prédire une classe "proche" de celle de l'exemple en entrée. Dans ce cas, on utilise des mesures d'erreur ou de corrélation.

### 2.6.3 Mesures d'erreur

Les mesures d'erreur estiment les écarts entre les sorties calculées et les sorties désirées. En plus de donner une indication sur le comportement des classifieurs, elles peuvent également être utilisées comme fonctions de coût. Ces mesures sont utilisées à la fois en classification et en régression.

L'erreur moyenne MSE rend compte de la distance moyenne entre deux variables (Eq. 2.28). L'erreur quadratique RMSE (Eq. 2.29) est la racine carrée de l'erreur moyenne. :

$$\mathsf{MSE} = \frac{1}{N}\sum_{i=1}^{n}(\mathbf{y}-\mathbf{y'})^2 \tag{2.28}$$

$$\mathsf{RMSE} = \sqrt{\mathsf{MSE}} \tag{2.29}$$

Si ces deux mesures sont conceptuellement très proches, la RMSE accorde la même importance à tous les écarts, alors que la MSE donne plus d'importance aux grands écarts. Les deux mesures sont définies dans $\mathbb{R}^+$ et sont à l'échelle des variables mesurées. Les valeurs proches de 0 indiquent une erreur faible, i.e. une bonne performance.

### 2.6.4 Mesures de corrélation

Les mesures de corrélation sont adaptées à l'évaluation des régresseurs. Toutefois, en Informatique Affective, on en retrouve l'usage aussi bien en classification qu'en régression.

Le coefficient de corrélation de Pearson (Eq. 2.30) est considéré comme le coefficient de corrélation linéaire par défaut. A ce titre, on le désigne souvent comme étant simplement le coefficient de corrélation. Il mesure le rapport entre la covariance $cov$ et le produit des écart-types $\sigma$ des variables.

$$\rho = \frac{cov(\mathbf{y},\mathbf{y'})}{\sigma_{\mathbf{y}}\sigma_{\mathbf{y'}}} \tag{2.30}$$

Il est apprécié pour son interprétation directe (Fig. 2.9). Il est borné entre -1 et 1. Un coefficient à 0 marque l'absence de corrélation, alors qu'un coefficient à -1 ou 1 marque une corrélation forte et négative ou forte et positive, respectivement.

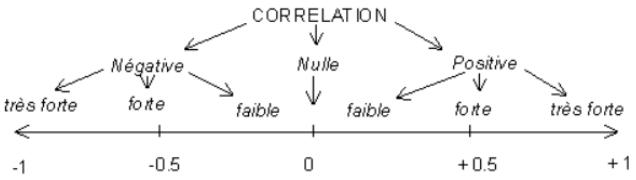



FIGURE 2.9 – Interprétation du coefficient de Pearson

Le *Concordance Correlation Coefficient* de Lin, noté CCC, mesure l'homogénéité de deux variables (Eq. 2.31) :

$$\mathrm{CCC} = \frac{2\rho\sigma_{\mathbf{y}}\sigma_{\mathbf{y}'}}{\sigma_{\mathbf{y}}^2 + \sigma_{\mathbf{y}'}^2 + \left(\mu_{\mathbf{y}} - \mu_{\mathbf{y}'}\right)^2} \tag{2.31}$$

avec $\mu$ la moyenne et $\sigma^2$ la variance. En Informatique Affective, le CCC est utilisé comme mesure de corrélation pour évaluer la sortie d'un classifieur, ou pour évaluer la concordance entre plusieurs annotateurs. Dans cette dernière utilisation, McBride [79] propose qu'un CCC inférieur à 0.90 indique une concordance faible.

Le coefficient de Spearman ($\rho_S$) tient compte de l'écart entre la classe prédite et la vérité terrain, plutôt que des correspondances entre ces deux éléments (Eq. 2.32) :

$$\rho_S = \frac{\sum_{i=1}^{n}\left(y_i - \mu_y\right)\left(y_i' - \mu_{y'}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i - \mu_y\right)^2 \sum_{i=1}^{n}\left(y_i' - \mu_{y'}\right)^2}} \tag{2.32}$$

## 2.7 Outils logiciels et matériels

La mise en œuvre de classifieurs nécessite l'usage d'implémentations spécifiques, autant pour les prétraitements, la classification, l'évaluation ou la production d'illustrations. Dans cette section, on fait un résumé des développements et implémentations réalisés ou utilisés pour la thèse.

Plusieurs tâches de prétraitements et de classification de l'état émotionnel ont été réalisées à l'aide de Matlab R2015a (licence académique) et R2017a (version d'évaluation). Toutefois, ces tâches ont été portées par la suite en Python. La plupart des supports d'illustration présents dans cette thèse sont issus de Matlab.

La plupart des expérimentations ont été codées en Python (> 2.7) et utilisent des packages issus de l'écosystème SciPy [57], dont Scikit-learn [91], NumPy [87],

Pandas [82] et Matplotlib [53]. Des éléments de la librairie Dlib [64] et OpenCV [12], notamment pour la détection de visages et de points faciaux, ont été employés.

Des modèles de classifieurs dont aucune implémentation publique n'existe ont été implémentés *from scratch* à partir d'articles de l'état de l'art. C'est le cas du classifieur incrémental neuronal (INN) et de la triple mémoire associative bidirectionnelle (m-BAM), que nous avons codés en Python.

Le calcul des matrices de Gram, très chronophage, a été implémenté pour s'exécuter sur une carte graphique. Cette implémentation a permis de réduire considérablement les temps de calculs (Fig. 2.10). En dehors de l'exécution sur un processeur graphique, cette approche a l'originalité de tenir compte de la taille souvent limitée de la mémoire sur les noeuds graphiques : le calcul est réalisé par *chunks* dont la taille est adaptée à la mémoire graphique disponible.

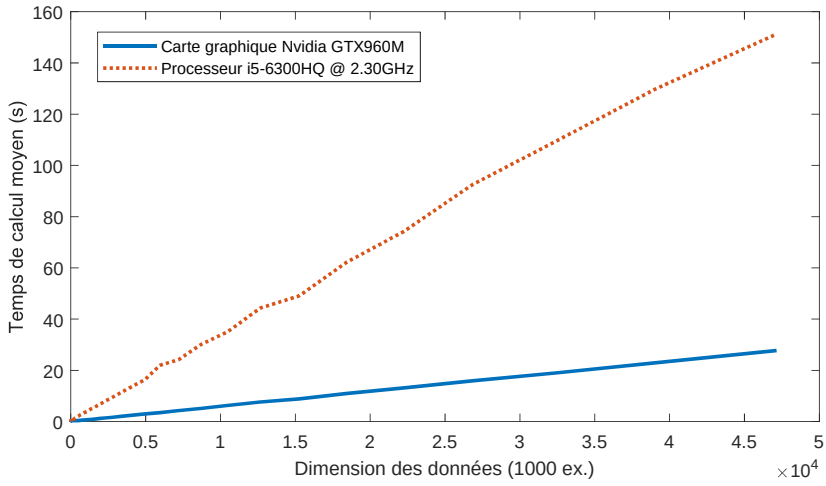



FIGURE 2.10 – Comparaison des temps de calcul moyens d'un noyau RBF en fonction du nombre d'exemples, sur un processeur (calcul parallèle sur les quatre cœurs) et sur une carte graphique.

Trois environnements matériels ont été utilisés :

— Un ordinateur de bureau, sous Linux Mint Sarah, équipé d'un processeur Intel(R) Core(TM) i7-4790S cadencé à 3.20GHz, pour la quasi-totalité des tâches réalisées.

— Le cluster de calcul du C3i (Centre Commun de Calcul Intensif de l'Université des Antilles), sous CentOS/SGE, et notamment un nœud graphique (Nvidia Quadro P5000, 16Gio GDDR5) pour le calcul rapide de matrices de Gram.

— Un ordinateur portable (sous Windows 10) équipé d'un processeur Intel(R) Core(TM) i5-6300HQ (4 cœurs physiques) cadencé à 2.30GHz et d'une carte graphique Nvidia GTX960M (2Gio GDDR5) pour l'impression des supports d'illustration et le calcul rapide de matrices de Gram.

## 2.8 Conclusion

Dans *Affective Computing*, il y a deux aspects, et c'est en cela que la thèse présente un caractère pluridisciplinaire. Ce chapitre constitue une partie purement *computing* des travaux.

Le choix d'un modèle de classifieur pour la résolution d'un problème n'est pas une tâche aisée et nécessite son étude théorique en amont de son utilisation. Le nombre d'hyperparamètres, la quantité de ressources informatiques requise ou encore l'existence d'implémentations pour un modèle sont autant d'éléments à considérer, tout particulièrement lorsqu'on doit traiter de grandes masses de données. Dans notre cas, plusieurs modèles ont dû être implémentés *from scratch*, en prenant en compte la puissance de calcul à disposition, qui est en grande part limitée à du matériel tout public. Parmi les modèles présentés, le classifieur incrémental neuronal à base de prototypes (INN) et la mémoire associative multiple (m-BAM) retiennent l'attention. Ces deux modèles s'inspirent du fonctionnement du cerveau et ont déjà été utilisés en sciences cognitives pour la modélisation de phénomènes perceptifs. Des études récentes ont rappelé le caractère adapté des systèmes à base de prototypes face à des données complexes, alors que la mémoire associative continue de susciter l'intérêt pour les problèmes d'intégration multimodale.

Dans le chapitre suivant, on se consacre à une partie purement *affective*, où les différentes représentations des émotions et de la dépression sont étudiées. En particulier, la complexité des données dites affectives pose plusieurs défis, dont la résolution présume du succès des modèles de classifieurs utilisés pour l'analyse des comportements émotionels et dépressifs.

CHAPITRE 3

# REPRÉSENTATIONS DES ETATS EMOTIONNELS ET DÉPRESSIFS

> Sous l'empire d'une colère médiocrement intense, l'action du cœur se surexcite légèrement, le visage se colore et les yeux brillent. Dans une vive colère, les yeux sont farouches, les sourcils sont abaissés et énergiquement froncés, les lèvres sont serrées.
>
> L'expression des émotions chez l'homme et les animaux [22], Darwin, 1872.

## 3.1 Représentation des émotions

Les nombreuses représentations des états émotionnels sont, pour la plupart, issues des travaux de psychiatres et de psychologues.

Des études entières sont consacrées à la seule définition du terme *émotion* [65].

En 2001, Plutchik [93] estime à plus de quatre-vingt-dix le nombre de définitions du terme *émotion*, qu'il qualifie pour sa part de *"hautement personnelle voire déroutante"*. Il émet des réserves sur le caractère évolutionnaire des émotions proposé par Darwin [22], selon lequel les émotions sont issues, dans leur expression, de comportements qui ont premis la survie de l'espèce humaine. Tomkins [104], un pionnier de la théorie de l'affect, considère que les émotions ont une origine purement biologique, et qu'elles se réfèrent à des mécanismes génétiquement

prévus chez tout un chacun. Ainsi, en Informatique Affective, la définition que l'on retient pour l'émotion est notamment motivée par l'application que l'on souhaite en faire. Selon Picard [92], l'émotion a *"cela de dérangeant pour la science de ne pas être expliquée de manière rationnelle, logique, ou de ne pas pouvoir donner lieu à des hypothèses vérifiables et à des tests que l'on peut répéter"*.

Dans la littérature et y compris dans cette thèse, les termes *état affectif* et *état émotionnel* sont synonymes.

### 3.1.1 Modélisation discrète

En 1872, Darwin décrit avec précision plusieurs combinaisons d'expressions faciales, qu'il regroupe sous le terme d'émotion. Il postule alors que les émotions sont universelles et innées. Dans l'ouvrage *The Expression of the Emotion in Man and Animals* [22], Darwin étudie avec soin le comportement et les mouvements expressifs chez plusieurs animaux (chiens, chats, chevaux, etc.) et chez l'homme. Ce faisant, il associe des comportements à des situations particulières, comme par exemple le hérissement des poils chez les chimpanzés lorsqu'ils sont effrayés brusquement. Plusieurs chapitres sont aussi consacrés aux mouvements faciaux chez l'être humain dans plusieurs contextes émotionnels : l'anxiété, le chagrin, la joie, la haine, etc.

#### 3.1.1.1 Le *Facial Action Coding System*

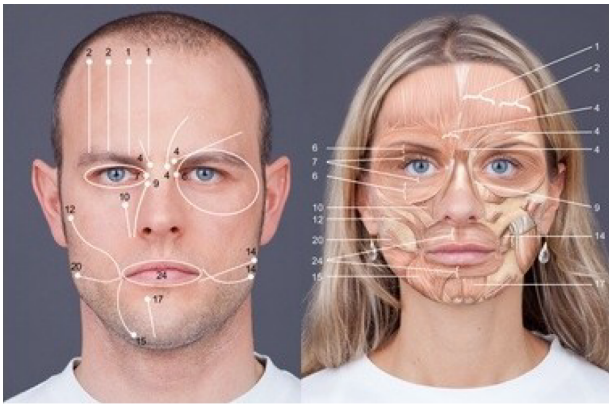


FIGURE 3.1 – Muscles peauciers intervenant dans les mouvements du visage d'après [28]

En 1978, Ekman et Friesen [30] proposent le *Facial Action Coding System* (FACS), que l'on peut traduire par "système d'encodage des mouvements faciaux". Le visage humain est composé d'environ quarante muscles peauciers, intervenant dans les mouvements des différentes parties du visage (Fig. 3.1). Un mouvement singulier, c'est-à-dire une micro-expression faciale, peut en conséquence être initié par un ou plusieurs muscles. Le FACS étend les travaux de l'anatomiste Hjortsjö [47] sur le langage facial humain et propose une nomenclature complète permettant l'encodage de quasiment toutes les micro-expressions. La méthode décompose les expressions en *action unit* (AU). Chaque AU décrit un micro-mouvement facial par le résultat visible et les muscles entrant en jeu (Fig. 3.2).

**Encodage des micro-mouvements** Pour encoder un micro-mouvement, le FACS dispose de vingt-huit AU principales, caractérisant chacune une action singulière. Par exemple, l'AU 6 correspond à la contraction des joues (*cheek raiser*) et l'AU 14 au retroussement des lèvres (*dimpler*). La combinaison AU 6 + AU 14 peut être assimilée à un sourire forcé. Sur le même principe, Ekman et Friesen [30] établissent les combinaisons entrant en jeu pour chacune des émotions de base, dites *basic-six* (Tab. 3.1). Depuis sa création, le FACS a subi plusieurs réactualisations, et permet d'encoder aussi bien les expressions faciales que leur symétrie (si une expression implique les deux côtés du visage ou un seul) et leur intensité (sur une échelle de 1 à 5).

TABLE 3.1 – AU entrant en jeu dans les émotions de base, d'après [30]

| Emotion de base | Action Units |
|---|---|
| joie | 6, 12, 25 |
| tristesse | 1, 4, 6, 11, 15 ,17 |
| peur | 1, 2, 4, 5, 20, 25, 26, 27 |
| surprise | 1, 2, 5, 26, 27 |
| colère | 4, 5, 7, 10, 17, 22, 23, 24, 25, 26 |
| dégoût | 9, 10, 16, 17, 25, 26 |

Il convient de distinguer la modélisation des émotions de la modélisation des expressions faciales, qui sont toutes deux rendues possibles par le FACS. Dans le premier cas, on est rapidement limité : la méthode ne permet de couvrir nativement qu'un ensemble d'émotions "prototypes". Dans le second, le FACS permet

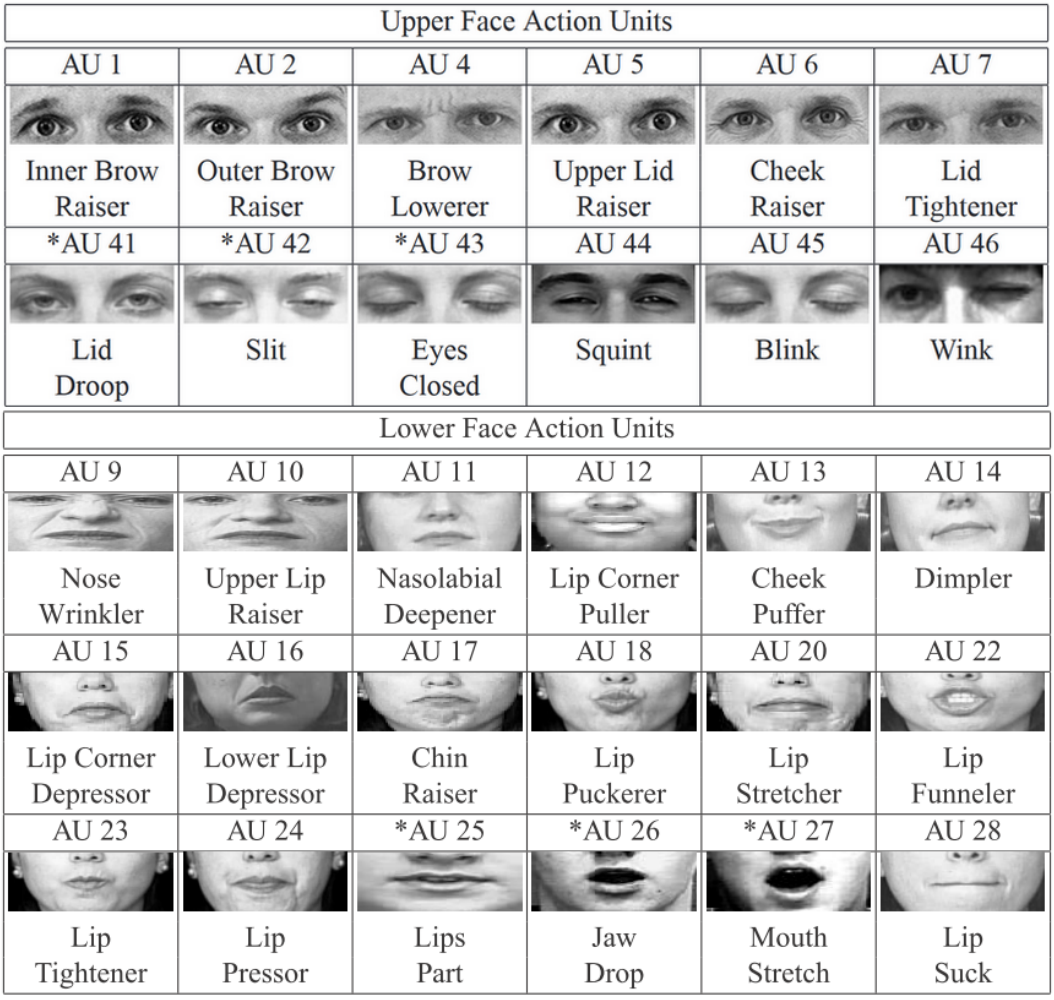



FIGURE 3.2 – Extrait du FACS. Action Units relatives à la partie inférieure du visage, d'après [116]. Les * indiquent la possibilité d'exprimer l'AU avec une certaine intensité.

un encodage objectif des micro-mouvements faciaux pour un usage ultérieur avec d'autres modélisations des émotions, qu'elles soient discrètes ou continues.

**Temporalité et spontanéité des expressions faciales** Le FACS introduit la notion de segmentation temporelle des expressions faciales, en postulant que toute expression se produit en quatre phases ordonnées :

— Le *neutral* est l'origine de toute expression : le visage est "au repos" et n'est par conséquent descriptible par aucune AU.

— L'*onset* est une phase plus ou moins rapide, coïncidant avec la contraction

des muscles mis en jeu dans l'expression qui sera réellement observée lors de la phase suivante.

— L'*apex* est souvent assimilée à un plateau, où l'expression reste à son maximum pendant un court laps de temps. C'est à l'*apex* que le FACS permet d'évaluer une expression faciale : les muscles sont alors au maximum de leur contraction.

— Enfin, l'*offset* est la phase de relâchement des muscles jusqu'à un retour progressif vers le *neutral*.

La phase de l'*apex* est réputée spontanée : contrairement aux autres phases, elle identifie une situation à son apogée. Par extension, les émotions ainsi exprimées sont également qualifiées de spontanées. Ces considérations sont cohérentes avec les méthodes de reconnaissance des émotions à partir de clichés.

L'utilisation du FACS implique que l'hypothèse où les expressions interviennent dans la segmentation temporelle ci-dessus est vérifiée. Par extension, cela implique aussi que chaque expression doit être séparée par une phase *neutral*, ce qui n'est pas toujours le cas selon Tian *et al.* [116], qui parlent de "simplification" du processus. Ils postulent alors que la reconnaissance des AU est aussi importante que la reconnaissance des transitions entre AU, ces transitions étant variables d'un individu à l'autre.

#### 3.1.1.2 La roue des émotions de Plutchik

Une autre approche, néanmoins plus subjective, est celle proposée par Plutchik [93] via sa roue des émotions (Fig. 3.3). Dans son approche, le psychologue considère l'émotion comme étant "psycho-évolutionnaire" : elle n'est pas un simple état d'esprit mais une chaîne complexe et connectée d'événements, provoquée par un *stimulus* et impliquant des ressentis, des changements physiologiques, des actions et des comportements spécifiques.

La représentation utilise des couleurs afin d'illustrer les états émotionnels intermédiaires ou proches. Elle est à tort appelée roue, puisqu'elle se présente sous la forme d'un cône en trois dimensions. On y considère huit émotions de base, s'opposant par paires : la joie vs. la tristesse, la confiance vs. le dégoût, la peur vs. la colère et la surprise vs. l'anticipation. Des états intermédiaires y sont également représentés. Dans cette représentation, des émotions proches sur le plan psychologique sont proches sur la roue, alors que des émotions opposées s'y

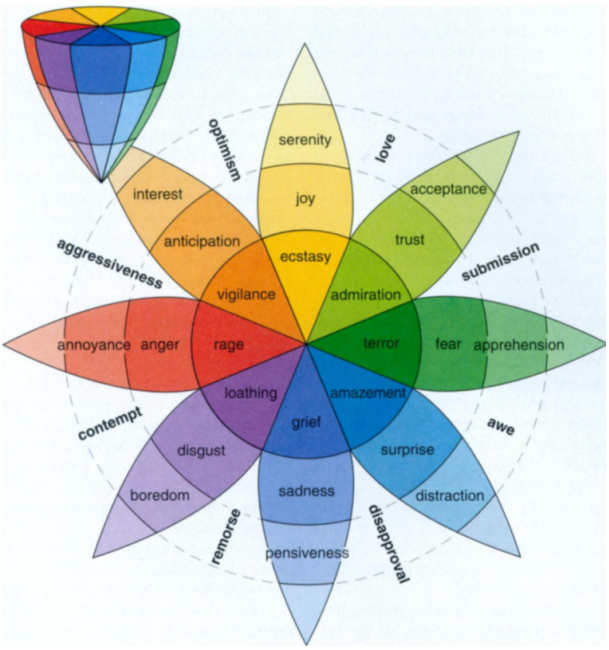



FIGURE 3.3 – Roue des émotions de Plutchik, d'après [93]

font face. La troisième dimension permet de modéliser l'intensité des émotions ressenties.

### 3.1.2 Modélisation continue

Les modélisations continues font appel à un paradigme différent de celui considéré pour les modélisations discrètes. Elles considèrent que toute émotion est nécessairement la combinaison de plusieurs "états d'esprits" en lien avec l'environnement du sujet. Les différences entre modélisations de ce type sont notamment fondées sur les axes choisis pour représenter les émotions.

#### 3.1.2.1 La modélisation tridimensionnelle de Schlosberg

En 1954, le psychologue Schlosberg [107] propose une modélisation tridimensionnelle de l'émotion. Les émotions ne sont alors plus considérées comme des états. Cette proposition tire profit de la théorie de l'excitation (*activation theory*), dont la pierre angulaire est le niveau d'excitation d'un l'individu. Un bas niveau correspond à des émotions comme le dégoût ou le mépris, alors qu'un haut niveau correspond plutôt à la surprise ou à la colère (Fig. 3.4a). Ainsi, trois axes sont définis : *sleep-tension*, *pleasant-unpleasant* et *attention-reject*. Ekman [29] critiquera

cette modélisation, jugeant les axes choisis trop intercorrélés pour exprimer les émotions sans biais.

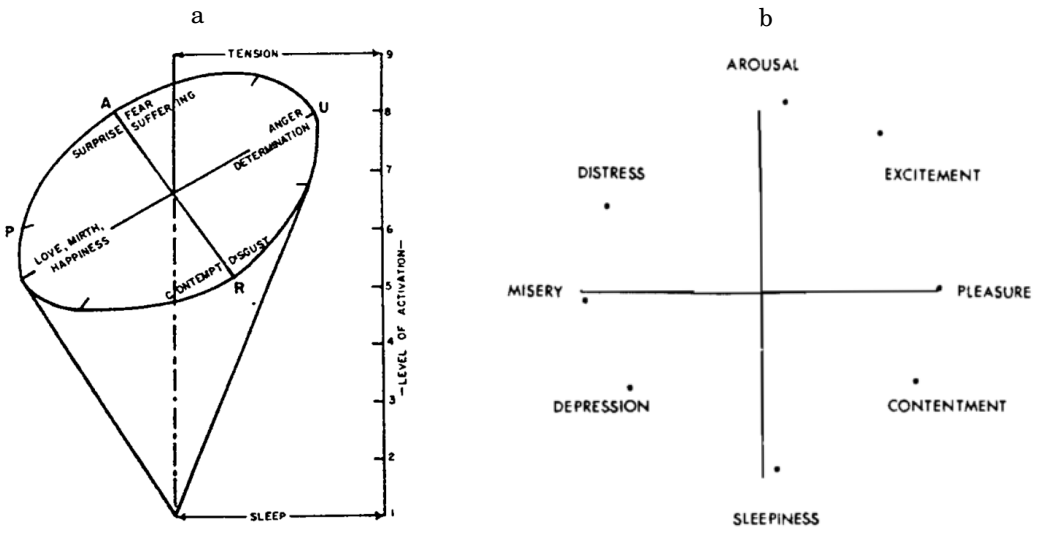



FIGURE 3.4 – (a) Modélisation tridimensionnelle de Schlosberg, d'après [107]. (b) Circumplex de Russell, d'après [103].

#### 3.1.2.2 Le circumplex de Russell

Le circumplex de Russell (Fig. 3.4b) est une modélisation qui a donné lieu à de nombreuses dérivations. En 1980, le psychologue Russell [103] place les émotions selon deux axes : l'*arousal* et la *valence*. L'*arousal* exprime l'état d'activité du sujet, entre la léthargie et l'éveil ; la *valence* exprime l'état d'esprit du sujet, entre la tristesse et la joie. Cette modélisation est très utilisée en Informatique Affective. Elle est en effet facile à visualiser et possède de solides fondements en psychologie, considérant que les états affectifs sont inter-connectés par nature. Parfois, une troisième dimension est utilisée, la *dominance*, qui exprime le degré de contrôle qu'estime avoir le sujet vis-à-vis de la situation d'interaction. Toutefois, Fontaine *et al.* [37] estiment que le monde des émotions n'est pas bi-dimensionel et ne peut pas être représenté efficacement par les axes *valence-arousal*. Ils préconisent alors quatre axes : les axes *evaluation-pleasantness*, *potency-control*, *activation-arousal* et *unpredictability*.

### 3.1.3 Discussion

Intuitivement, la représentation discrète des émotions l'emporte sur la modélisation continue : c'est la modélisation que l'on retrouve dans le langage de tous les jours. Il est plus commun de dire "je suis content" que de se situer sur les axes *valence-arousal*. Les modélisations discrètes sont plus aptes à l'aide au diagnostic, qu'il soit autoréalisé ou assisté par un tiers. Ces modélisations mettent en avant les liens et les transitions qui peuvent exister entre les émotions. Les modélisations continues nécessitent des outils adaptés pour leur évaluation, mais permettent d'étudier aussi bien les transitions que l'intensité des émotions.

Les modélisations discrètes et continues ont toutes deux des fondements en psychologie. Malgré la préférence notoire de l'Informatique Affective pour certaines d'entre elles, notamment le FACS ou le circumplex de Russell, il est erroné de croire que ces représentations ont été bâties pour l'informatique. Pour la prédiction des états émotionels, les modélisations discrètes servent généralement de support à des méthodes de régression, alors que les modélisation continues servent de support à des méthodes de classification (voir section 2.2 sur les modèles de l'état de l'art).

Le circumplex de Russell bénéficie d'avantages inhérents à sa conception même, comme celle de mettre en lien les émotions entre elles ou de les exprimer avec une intensité. Grâce à cette représentation, il devient possible d'étudier des aspects liés aux transitions ou aux intensités émotionnelles. En particulier, la représentation de Russell est adaptée à l'étude des intensités émotionnelles et de leur reconnaissance.

## 3.2 Représentations de la dépression

D'après l'Organisation Mondiale de la Santé [124], plus de trois cents millions de personnes souffrent de troubles dépressifs dans le monde. Ces troubles mentaux se caracérisent par de la tristesse, de la perte d'intérêt ou de plaisir, des sentiments de culpabilité ou de dévalorisation de soi, un sommeil ou un appétit perturbé, une certaine fatigue et des problèmes de concentration. En France, on estime à plus de 10 000 par an le nombre de suicides liés à un état dépressif, dont la majorité sont des hommes. Au niveau mondial, les femmes sont plus touchées par la maladie.

La maladie se décline en plusieurs termes, souvent liés au contexte d'occurrence et aux symptômes. L'un des plus courants est le trouble dépressif majeur (en anglais, *Major Depressive Disorder - MDD*), généralement appelé *dépression* dans

l'acception courante. D'autres troubles, comme le *burn-out*, le trouble dépressif persistant, le trouble bipolaire, la dépression *postpartum* (qui se produit après la grossesse) ou la dépression saisonière sont également des troubles dépressifs. Pour être diagnostiqué dépressif, un patient doit manifester les symptômes recherchés pendant au moins deux semaines [83]. A moins de consulter un spécialiste, les malades ne sont pas toujours en mesure de réaliser qu'ils sont atteints d'un trouble qui, dans une grande majorité des cas, peut se guérir grâce à un suivi psychologique et/ou à la prescription de médicaments adaptés [24].

### 3.2.1 Beck Depression Inventory

Le test d'évaluation de l'état dépressif *Beck Depression Inventory* (BDI) a été créé par le psychiatre Beck en 1961. L'outil met l'accent sur le ressenti du patient vis-à-vis de sa situation, et non sur les enjeux psychologiques motivant son comportement [9]. La version initiale de 1961 a été mise à jour à deux reprises, et la plus récente (BDI-II) date de 1996. Les modifications ont notamment visé à aligner l'outil avec les critères de diagnostic du *Diagnostic and Statistical Manual of Mental Disorders* [4] (DSM-VI) paru en 1994, qui fait autorité dans le domaine des troubles mentaux.

TABLE 3.2 – Interprétation standard du score au test BDI-II

| Score obtenu | Sévérité de la dépression |
|---|---|
| 0-13 | Minimale |
| 14-19 | Modérée |
| 20-28 | Moyenne |
| 29-63 | Sévère |

Le BDI-II [10] consiste en un questionnaire auto-administré à choix multiple de vingt-et-une questions, chacune évaluant un ou plusieurs symptômes dépressifs. Parmi celles-ci, quinze sont liées aux émotions, quatre aux changements de comportement et six aux symtpômes somatiques. Chaque question rapporte entre 0 et 3 points, et le total des points est un score pouvant aller de zéro à soixante-trois. Un haut score est synonyme de dépression sévère. L'outil s'accompagne d'une table d'interprétation des scores, bien qu'il en existe plusieurs (Tab. 3.2 pour la plus courante).

Le *Beck Depression Invetory II* est approprié à l'usage par des individus ne présentant pas de troubles mentaux ou psychosociaux (individus dits "normaux"). Sa réalisation prend entre cinq et dix minutes, et le test n'a pas vocation à être utilisé comme moyen de diagnostic, mais plutôt de dépistage et d'évaluation de la sévérité de la dépression.

### 3.2.2 Hamilton Rating Scale

La *Hamilton Rating Scale for Depression* [43] (HRSD) a été proposée par Hamilton en 1960. Elle permet d'évaluer la sévérité d'un trouble dépressif. Le psychiatre motive son travail par la nécessité d'une échelle fiable, critiquant assez sévèrement les tests auto-administrés. Dans ce test, l'accent est mis sur l'évaluation fiable d'un certain nombre de symptômes caractéristiques tels que les problèmes digestifs, les tendances suicidaires ou la perte de poids.

La HRSD consiste en l'évaluation de vingt-et-un symtpômes par un ou deux professionnels de santé [44]. Bien qu'une certaine latitude soit laissée aux évaluateurs, des instructions précises quant au déroulement de l'interview existent [8] telles que la durée (qui doit être de trente minutes), ainsi que des recommandations telles que le questionnement de personnes proches du patient. Le score final varie entre zéro et cinquante-deux, et ne considère que dix-sept des vingt-et-un symptômes évalués. Un score élevé indique une dépression sévère. Dans ses travaux, Hamilton n'a pas spécifié d'intervalles d'interprétation, mais on considère en général les intervalles proposés dans la table 3.3.

TABLE 3.3 – Interprétation standard du score au test HRSD

| Score obtenu | Sévérité de la dépression |
|---|---|
| 0-7 | Absence |
| 8-17 | Modérée |
| 18-24 | Moyenne |
| 25-52 | Sévère |

La *Hamilton Rating Scale for Depression* est adaptée à un usage sur des individus présentant des troubles mentaux ou psychosociaux. Son usage est recommandé lorsque de tels troubles ont déjà été diagnostiqués chez le patient.

### 3.2.3 Patient Health Questionnaire - 8

Le Patient Health Questionnaire - 8 (PHQ-8) est un outil de mesure de la sévérité dépressive, proposé en 2008 et composé de huit des neuf questions du PHQ-9 dont il est issu. Le PHQ a été introduit en 1999 par Spitzer *et al.* [112] comme une adaptation de l'outil PRIME-MD destiné à détecter et évaluer les troubles mentaux dans le contexte des soins primaires (i.e. chez le médecin généraliste, à la pharmacie, etc.). L'élaboration du PHQ-9 était alors motivée par la nécessité d'une version utilisable de manière plus large que le PRIME-MD, via un test plus condensé. La question figurant dans le PHQ-9 et qui a été otée du PHQ-8 visait à évaluer le penchant suicidaire du patient. Plusieurs études (Huang *et al.* [51], Lee *et al.* [72]) ont montré le rôle facultatif de cette dernière.

TABLE 3.4 – Interprétation du score au test PHQ-8

| Score obtenu | Sévérité de la dépression |
|---|---|
| 0-4 | Absence |
| 5-9 | Modérée |
| 10-14 | Moyenne |
| 15-19 | Moyennement sévère |
| 20-24 | Sévère |

Le PHQ-8 consiste en l'évaluation de huit critères dépressifs par un questionnaire auto-administré. Contrairement aux deux autres tests présentés dans cette partie, le PHQ-8 accorde une très grande importance à la durée des symptômes. Chaque réponse est attendue sous la forme du nombre de jours pendant lesquels les symptômes se sont manifestés durant les deux dernières semaines. La motivation du PHQ-8 est d'être utilisé comme outil de mesure de masse de l'état dépressif, par exemple pour des populations entières [69]. Le total des points peut atteindre vingt-quatre, et le diagnostic est réalisé en accord avec les seuils d'interprétation proposés dans la table 3.4.

### 3.2.4 Points de comparaison

Les tests d'évaluation de la sévérité d'une dépression présentés sont tous les trois très utilisés dans le domaine médical. Ils sont aussi accompagnés d'une table d'interprétation qui fournit une explication de la sévérité du score obtenu.

Le HRSD n'est pas auto-administré et est sujet à un protocole plutôt rigoureux, ce qui le rend fiable mais moins approprié que les autres aux études "de masse". Le PHQ-8 en revanche a déjà fait l'objet d'utilisations pour la mesure de la sévérité dépressive sur des populations. Le BDI-II bénéficie à la fois de la fiabilité du HRSD (étant positivement corrélé avec ce dernier) et d'une facilité d'administration. Ce dernier test met également l'accent sur le ressenti du patient et couvre plus de symptômes que le PHQ-8, ce qui en fait un choix intéressant pour l'analyse automatique des états dépressifs.

## 3.3 Jeux de données affectives

L'Informatique Affective a à sa disposition un nombre conséquent et croissant de bases de données, répondant chacune à une ou plusieurs problématiques spécifiques, telles que la détection d'*action unit*, la classification de l'état affectif ou de l'état dépressif. La constitution d'une base de données est un processus long et complexe, notamment lorsqu'elle est annotée. C'est pourtant un élément fondamental pour l'avancement de la recherche : sans données fiables et en nombre suffisant, il est très difficile de développer de nouvelles méthodes ou de modéliser un fonctionnement. Historiquement, les modalités physiologiques comme la conductance cutanée ou l'électrocardiographie (ECG) étaient très utilisées. Dorénavant, les modalités visuelles et auditives sont plébiscitées pour la faible invasivité de leur processus d'enregistrement, d'autant que le traitement automatique de grands volumes de données audio-visuelles est devenu possible.

Dans cette section, nous nous intéressons aux méthodes et aux subtilités liées à l'acquisition, à la description et à l'annotation de la modalité visuelle, puis nous réalisons un tour d'horizon de quelques bases de données de la littérature. La collecte et le traitement de ces données posent de véritables défis, dont certains sont discutés.

### 3.3.1 Généralités sur l'acquisition

Des bases de données (aussi appelées jeux de données ou corpus) audiovisuelles, en accès libre, payant ou encore réservées pour des applications de recherche sont disponibles en grande quantité. Dans ce paragraphe, on dresse le contour des éléments à prendre en compte lors du choix d'un corpus.

En ce qui concerne les états affectifs, on peut facilement catégoriser les bases de données dans trois groupes caractérisant, d'après le scénario qui a été mis en œuvre pour la déclencher (on parle de *data-elicitation scenario*), l'émotion exprimée par le sujet. Cette dernière peut être simulée, induite ou spontanée.

— Les émotions sont dites simulées (*posed*) lorsque les sujets adoptent, sur commande, des comportements émotionnels précis, tant au niveau de leurs expressions faciales qu'au niveau de leur voix.
— La deuxième catégorie implique un scénario qui induit, chez le sujet, l'expression d'émotions en particulier. On parle alors d'émotions induites. Traditionnellement, ce type de scénario implique une conversation guidée par un support (documents, série d'images, vidéos, etc.) que le sujet consulte dans un ordre défini pendant la capture.
— Enfin, les émotions sont spontanées lorsqu'aucune mesure n'est prise, lors de la capture, pour conduire le sujet à exprimer une émotion en particulier. Ces données sont notamment capturées lors d'interactions dyadiques (entre deux humains) ou entre un humain et une machine (HCI, *Human-Computer Interaction*).

En plus du scénario de *data-elicitation*, il faut également considérer les conditions dans lesquelles les données ont été capturées, qui sont soit de laboratoire, soit naturelles. Ce paramètre s'applique à la fois aux corpus destinés à l'évaluation des états émotionnels et dépressifs :

— Les données capturées en condition de laboratoire ont toutes des paramètres communs ou très proches, tels que la luminosité, la distance à la caméra ou l'angle de capture.
— Dans l'autre cas, on parle de données *in-the-wild*. Les paramètres cités précédemment peuvent varier d'un échantillon à l'autre, et les sujets peuvent être capturés à leur insu.

En théorie, chacune des deux conditions de capture peut être associée à n'importe quel scénario, bien qu'en pratique, il existe des combinaisons plus courantes que d'autres. Par exemple, il est approprié d'utiliser une base de données induite pour le développement d'un outil visant à analyser les états émotionnels lors de la réalisation de tâches spécifiques.

### 3.3.2 Descripteurs

Un descripteur est une transformation, appliqué à une source de données (par exemple, à une image) et qui produit une description. Les descriptions sont généralement des vecteurs, dont la dimension varie en fonction de la formulation du descripteur associé. Ce sont ces vecteurs qui sont utilisés en entrée de méthodes d'intelligence artificielle.

#### 3.3.2.1 Descripteurs visuels

Les descripteurs visuels sont discutés suivant qu'ils sont spatiaux ou spatio-temporels. Dans chaque catégorie, ils peuvent être géométriques ou de texture.

Le point de départ d'un bon nombre de descripteurs, qu'ils soient spatiaux ou spatio-temporels, est la détection du visage. Le détecteur de Viola et Jones [121] exploite des descripteurs pseudo-Haar (Fig. 3.5) calculés sur des intégrales d'images de régions rectangulaires.

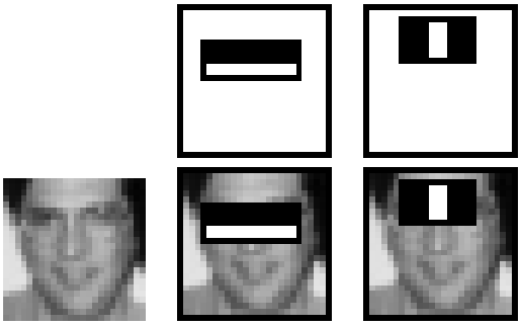


FIGURE 3.5 – Exemples de filtres de Haar appliqués sur un visage pour le calcul de features, d'après [121]

**Descripteurs spatiaux** Les descripteurs spatiaux décrivent des éléments visuels statiques, i.e. image par image s'ils sont extraits d'une vidéo. En ce sens, il n'est pas possible de prendre en compte l'aspect temporel avec cette famille de descripteurs.

Les points d'intérêt faciaux (*facial points of interest* ou *facial landmarks*) sont simples, compréhensibles et rapides à calculer (Fig. 3.7). Ce descripteur géométrique est particulièrement adapté pour le calcul rapide de descripteurs dérivés

(distances, angles, etc.). Le prédicteur de Kazemi et Sullivan [63] utilise une cascade de régresseurs pour affiner la position d'une *mean shape* de départ sur un visage en entrée, jusqu'à ce qu'elle corresponde le mieux possible à ses contours réels (Fig. 3.6).

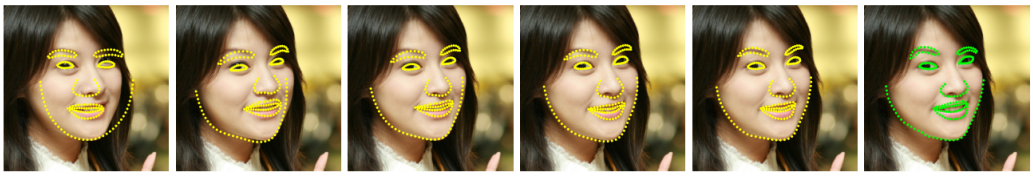


FIGURE 3.6 – Positions successives de points d'intérêts faciaux à plusieurs étapes de la prédiction, d'après [63]

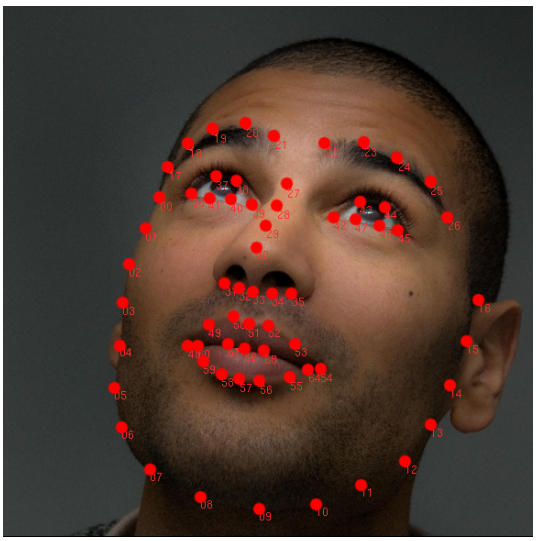


FIGURE 3.7 – Points d'intéret faciaux (illustration de [64]).

L'opérateur LBP (*Local Binary Patterns*) a été popularisé par Ojala *et al.* [86] pour l'analyse de textures et de contours, et est robuste aux variations de luminosité. Il encode la texture présente dans un voisinage circulaire de pixels en un motif binaire sur huit positions, en seuillant la valeur $f_i$ (de l'intensité) de chacun des huit pixels entourant un pixel central $p$ par la valeur $f_p$ du pixel central (Fig. 3.8). L'encodage d'un pixel central $\mathrm{LBP}(p)$ est défini comme suit (Eq. 3.1) :

$$\mathrm{LBP}(p) = \sum_{i=1}^{8} \mathrm{H}(f_i - f_p)2^i \tag{3.1}$$

où H est la fonction de Heaviside tel que $H(x) = 1$ si $x \geq 0$ et $H(x) = 0$ si $x < 0$.

La description LBP d'une image est traditionnellement représentée sous la forme d'un histogramme des fréquences de chacun des $2^8 = 256$ encodages possibles pour un pixel. Les LBP uniformes définissent 59 encodages possibles par pixel, ne retenant que les 59 représentations binaires uniformes (i.e. contenant au plus deux transitions de 0 à 1 ou vice-versa) possibles entre $2^0$ et $2^8$.

Ahonen *et al.* [1] ont montré que ces 59 valeurs suffisaient pour encoder efficacement les micro-mouvements faciaux, en plus de diminuer considérablement la dimension du descripteur.

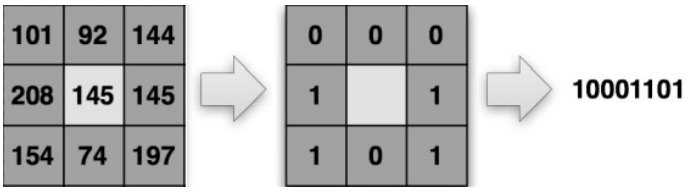



FIGURE 3.8 – Processus d'encodage binaire d'un pixel avec LBP. Les pixels autour du pixel central sont seuillés, puis les valeurs sont concaténées.

Les descriptions exploitant des images de Gabor sont également courantes. Elles consistent essentiellement à appliquer un descripteur quelconque (par exemple, le LBP) sur une ou plusieurs images de Gabor plutôt que sur une image source. Au sens strict, une image de Gabor est issue de la convolution d'une image source avec un filtre de Gabor. L'image de Gabor rend essentiellement compte de la présence de contours dont l'orientation et la fréquence sont celles du filtre.

Ainsi, les LGBP (*Local Gabor Binary Patterns*) sont fondés sur le calcul de LBP sur des images de Gabor. La description est également un histogramme, dont la

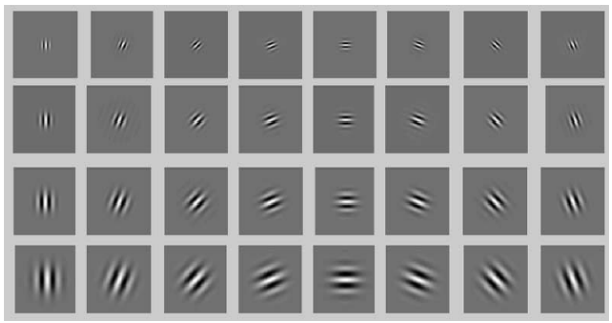


FIGURE 3.9 – Filtres de Gabor de huit orientations et quatre fréquences différentes

dimension est nécessairement plus grande que celle des LBP. Les LGBP héritent néanmoins de la robustesse des LBP.

Les *Histogram of Oriented Gradients* (HOG) sont fondés sur la possibilité de modéliser correctement une forme par la distribution des gradients ou des directions des contours présents à l'image. Dalal et Triggs [21] proposent de diviser l'image en cellules et de calculer, pour chacune, l'histogramme des gradients et des orientations de contours, puis de les combiner.

Le *Scale-invariant Feature Transform* (SIFT, [73]) est un autre opérateur qui permet d'extraire une description à des endroits spécifiques dans une image. Le SIFT en lui-même n'est pas un descripteur mais plutôt une méthode permettant de trouver des points pertinents de l'image où calculer un descripteur (par exemple un HOG). La popularité de cet opérateur est liée à sa robustesse face à l'occultation, ainsi qu'aux changements d'échelle et de luminosité.

**Descripteurs spatio-temporels** Les descripteurs spatio-temporels sont appliqués à des séquences d'images (i.e. à des vidéos) qui se suivent dans le temps. Ils permettent d'incorporer l'aspect temporel et ainsi de rendre compte du mouvement présent dans la séquence.

Dérivés des points d'intérêts faciaux, le suivi de points d'intérêts faciaux est l'encodage de base pour la description de séquences d'images. La position des points est suivie sur chaque image, dans une fenêtre temporelle fixée. On peut ainsi calculer la vitesse des mouvements ou quantifier l'activité musculaire faciale [118].

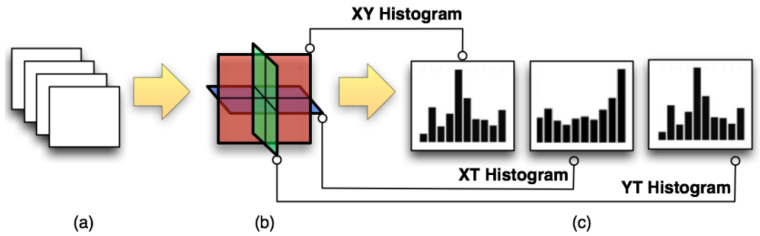



FIGURE 3.10 – Plans considérés pour l'extraction générique de descripteurs avec l'extension TOP d'après [2]. (a) Images successives dans une vidéo, (b) plans $x-y$, $y-t$ et $x-t$, (c) concaténation des trois histogrammes

L'extension TOP (*Three Orthogonal Planes*) permet à un descripteur initialement spatial de prendre en compte l'aspect temporel. Elle a été proposée par Zhao

et Pietikainen [131] pour les LBP, mais peut s'appliquer à n'importe quel opérateur bas-niveau. La démarche consiste à calculer un opérateur non plus seulement sur une image (i.e. dans le plan $x-y$ à un instant $t$), mais aussi sur les plans $y-t$ et $x-t$. Ces plans représentent respectivement les axes $x$ et $y$ en fonction du temps. La figure 3.10 synthétise la procédure de calcul.

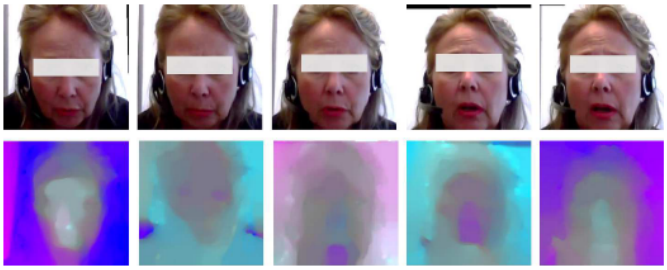


FIGURE 3.11 – Flux optique pour une séquence d'images. L'échelle de couleurs estime la quantité de mouvement : indigo (pas de mouvement), violet (fabile quantité), turquoise, gris et blanc (grande quantité de mouvement).

Parmi les applications du flux optique (*optical flow*), on trouve la représentation sous forme d'image de l'intensité ou de la quantité de mouvement détectée dans une séquence d'images. Il s'agit essentiellement de suivre le déplacement des pixels (ou d'un ensemble de pixels, voire de points d'intérets). On représente ensuite l'intensité des déplacements de chaque pixel dans le plan. Sur la figure 3.11, les intensités sont représentées suivant une échelle de couleurs.

Il existe plusieurs méthodes pour le calcul du flux optique (Lucas et Kanade [74], Farnebäck [35]), basées notamment sur l'hypothèse où le mouvement dans un voisinage d'un point donné (i.e. d'un pixel) sera similaire au mouvement de ce point.

#### 3.3.2.2 Descripteurs auditifs

En Informatique Affective, les descripteurs auditifs dits psychoaccoustiques sont très utilisés. Ils mesurent le signal auditif suivant des fondements en psychologie, c'est-à-dire le signal tel qu'il est perçu par un humain, plutôt que le signal émis par une source.

Des exemples et plus de précisions sont fournis dans le paragraphe 3.3.5.3, dans le cadre d'une base de données précise.

### 3.3.3 Annotation

L'annotation des états dépressifs se fait par séquence vidéo : toutes les images d'un même enregistrement seront associées à un même état dépressif (suivant l'une des échelles décrites en section 3.2).

L'annotation des états émotionnels requiert un dispositif plus avancé, puisqu'on associe à chaque image d'une vidéo donnée un score émotionnel. Si la modélisation émotionnelle retenue est discrète, alors les annotations seront discrètes (par exemple : joie, tristesse, colère, etc.). Sinon, les annotations seront sous la forme d'une trace continue dans l'intervalle $[-1;1]$ pour chaque dimension affective considérée.

Que ce soit pour les états émotionnels ou dépressifs, la présence d'annotations permet d'utiliser des méthodes de classification supervisée pour apprendre les données.

#### 3.3.3.1 Outils pour l'annotation émotionnelle des vidéos

Le processus classique d'annotation des états émotionnels dans le domaine continu implique un annotateur qui visionne un enregistrement vidéo, dans lequel un sujet est en jeu. L'annotateur a pour objectif d'évaluer l'état émotionnel du sujet dans une ou plusieurs dimensions affectives à l'aide d'un outil spécifique.

**FeelTrace** L'une des premières solutions pour l'annotation émotionnelle des vidéos est FeelTrace [19]. Initialement livrée sous forme de code source, FeelTrace permet d'annoter de manière continue une vidéo suivant les dimensions affectives *valence* et *arousal* (Fig. 3.12). L'outil peut être modifié afin d'annoter d'autres dimensions affectives.

Un scénario d'annotation est le suivant. L'annotateur est muni d'un joystick analogique (de type manette de jeu) dont la position au repos correspond à l'origine du repère de la figure 3.12. Pendant le visionnage d'une vidéo où un sujet exprime des émotions, l'annotateur utilise le joystick pour déplacer le disque coloré dans le repère. Par exemple, si le sujet à l'écran est pris d'un fou rire, l'annotateur devrait déplacer le joystick vers la droite.

**ANNEMO** ANNotating EMOtion [101] est un *framework* d'annotation des états affectifs destiné à être utilisé en ligne. L'annotation se fait en visionnant une vidéo

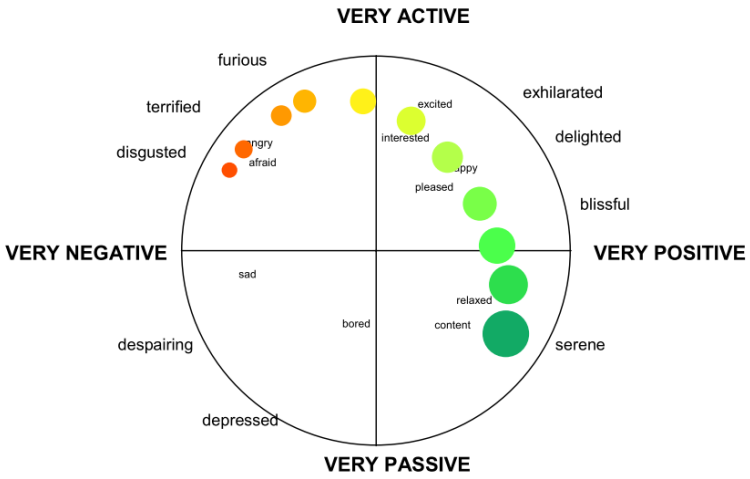



FIGURE 3.12 – Illustration du retour d'affichage avec FeelTrace [19]. Lors de l'annotation, l'annotateur voit évoluer la position du curseur qu'il déplace.

où le signal émotionnel est émis. Contrairement à FeelTrace, l'outil s'apparente à une solution clef en mains, avec une interface de configuration (pour le choix des dimensions à annoter, par exemple), et une interface d'annotation où les annotateurs se connectent pour visualiser et annoter les vidéos. L'annotation est faite au moyen de la souris, via un curseur horizontal à l'écran, pour chaque dimension affective séparément.

#### 3.3.3.2 Un nouveau paradigme d'annotation émotionnelle

Les outils de l'état de l'art permettent l'annotation émotionnelle de vidéos dans le domaine continu. Ces méthodes produisent des valeurs à une fréquence plus élevée que la fréquence de capture des vidéos. Par exemple, pour FeelTrace, la position du joystick est enregistrée mille fois par seconde, alors que les vidéos sont traditionnellement enregistrées à une vitesse de trente images par seconde. Une étape supplémentaire est alors nécessaire pour synchroniser les deux signaux.

Récemment, Yannakakis *et al.* [130] (incluant R. Cowie, l'un des créateurs de FeelTrace) ont critiqué l'ancien paradigme et proposent désormais d'associer des valeurs ordinales aux changements d'intensité émotionnelle présents dans la vidéo, plutôt qu'une valeur réelle dans un intervalle. En ce sens, l'outil AffectRank [130] (Fig. 3.13) invite l'annotateur à estimer la direction vers laquelle s'est déplacée l'émotion du sujet à l'écran parmi un nombre de directions proposées (cercles blancs) en fonction d'un certain nombre d'émotions de référence (cercles bleus).

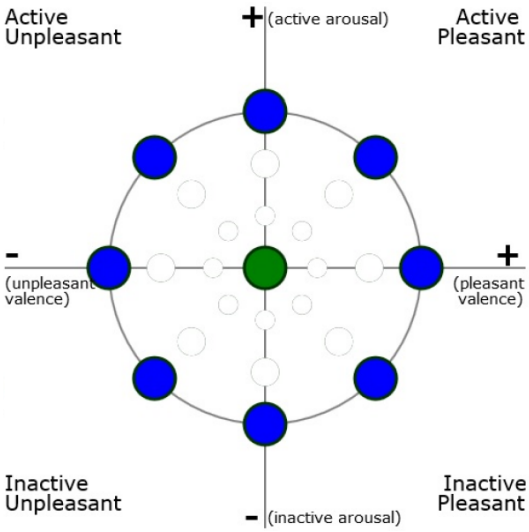



FIGURE 3.13 – Illustration du retour d'affichage avec AffectRank d'après [130].

### 3.3.4 Bases de données courantes

Ne nombreuses bases de données sont décrites dans la littérature, chacune ayant ses propres critères en matière de *data-elicitation*, de description des données ou encore d'annotation. On propose ci-dessous une revue de quelques bases de données parmi les plus répandues pour l'étude des mouvements faciaux, des états émotionnels ou des états dépressifs. Des études approfondies sur les bases de données en Informatique Affective sont proposées par El Ayadi *et al.* [31] ou encore Wu *et al.* [125].

**Cohn-Kanade** Les bases de données Cohn-Kanade [59] et Extended Cohn-Kanade [75] (CK et CK+) proposent 486 séquences de 97 sujets humains différents. La base de données CK contient des séquences où un sujet simule une expression parmi les *basic-six*. Les annotations sont l'émotion demandée (plutôt que l'émotion jouée) ainsi que le code FACS correspondant à l'expression à son *apex*. La base CK+ contient à la fois des expressions simulées et spontanées, toujours annotées en émotions de base et en code FACS.

**AVDLC** L'*AudioVisual Depresive Language Corpus*) [120] comporte 340 vidéos de 292 sujets âgés entre 18 et 63 ans. Les émotions sont modélisées dans le domaine continu et les vidéos sont annotés dans les dimensions affectives *valence*, *arousal* et *dominance* avec l'outil FeelTrace. Les états dépressifs sont modélisés et annotés

en termes de scores BDI-II. Les vidéos sont capturées dans des conditions de laboratoire, et les émotions sont spontanées : les sujets répondent oralement à des questions ou bien ils lisent un texte affiché à l'écran.

**SEWA** La base *Sentiment Analysis in the Wild* [67] est une base de données *in-the-wild* comportant des vidéos de volontaires enregistrées par eux-mêmes. Les émotions sont induites par des publicités que les sujets regardent puis commentent ensemble à l'occassion d'une visioconférence. La base contient ainsi plus de 25h d'enregistrements de 398 individus différents. Elle est notamment annotée en termes d'*action units* pour le domaine discret, et en *valence* et *arousal* pour le domaine continu.

**AFEW** La base *Acted Facial Expressions in the Wild* [25] propose des extraits vidéos issus de films connus de l'industrie du cinéma. Les extraits ont été sélectionnés en estimant leur contenu émotionnel à partir des sous-titres. Plusieurs annotations sont proposées, comme l'expression générale de la scène (parmi les *basic-six*), le nom de l'acteur ou encore la durée de la vidéo. Plus de 1 400 extraits pour 330 sujets sont proposés.

**GEMEP-FERA** La base GEMEP-FERA est une portion de la base de données *GEneva Multimodal Emotion Portrayal* [7] utilisée à l'occasion du challenge *Facial Expression Recognition and Analysis* dont la première édition a eu lieu en 2011. Elle est annotée en FACS et en émotions dans le domaine continu. Les sujets de la base sont dix acteurs, entrainés par des professionnels pour simuler 18 émotions différentes, totalisant plus de 7 000 images.

**RECOLA** La *REmote COLlaborative and Affective interactions* [102] est une base de données mise au point par des checheurs en informatique et en psychologie. Elle comporte quelques 9.5 heures d'enregistrements audio, visuels et physiologiques (activité électrodermale et cardiaque) d'interactions entre 46 participants (français) en train de résoudre une tâche en collaboration.

**DAIC-WOZ** La *Distress Analysis Interview Corpus - Wizard-Of-Oz* est une partie de la base DAIC [40] dont les données ont été acquises lors de sessions où la méthode de *data elicitation* était celle dite du magicien d'Oz. Cette technique

consiste à faire croire aux sujets qu'ils s'entretiennent avec un ordinateur alors qu'un opérateur humain, généralement caché ailleurs, pilote l'interaction. La base DAIC-WOZ inclut 189 sessions d'une durée moyenne de 16 minutes, et comprend la transcription textuelle et les bandes sons des dialogues, ainsi qu'une description faciale des sujets. Les sessions sont annotées avec le score dépressif PHQ-8.

**iBUG-300W** La *intelligent Beahviour Understanding Group* 300 *in-the-Wild* est une base de données d'images capturées dans des conditions naturelles [105]. Elle est notamment destinée à entraîner des modèles de détection de points d'intérêts faciaux. Les paramètres de la capture (luminosité, occlusion, qualité, etc.) varient considérablement d'une image à l'autre, tout comme l'expression faciale des individus à l'image.

### 3.3.5 Les données de la base AVDLC

La série de challenges AVEC invite une fois par an, en marge de la conférence internationale ACM-MultiMedia, les chercheurs à confronter leurs méthodes et leurs résultats sur un jeu de données unique et dans un contexte défini. L'édition 2014 proposait deux challenges [119] : la classification de l'état dépressif et la prédiction de l'état émotionnel.

Le corpus utilisé est issu de la base de données AVDLC et présente trois partitions : apprentissage (*training*), développement (*development*) et test (*testing*). Lors du challenge, les deux premières paritions étaient fournies annotées, et il était d'usage d'utiliser la partition de développement pour évaluer le système en généralisation. La partition de test servait alors pour comparer les performances des participants, ses annontations étant retenues secrètes par les organisateurs.

#### 3.3.5.1 Acquisition

Les vidéos ont été acquises dans des conditions non-invasives et reproductibles. Les sujets font face à un ordinateur portable muni d'une webcam et sont équipés d'un micro-casque. Ils suivent des instructions présentées à l'écran afin d'accomplir l'une des deux tâches suivantes :

— La tâche *Freeform*, qui consiste à répondre à une question du type "quel était votre repas préféré ?" ou "quel a été votre cadeau favori, et pourquoi ?".

— La tâche *Northwind*, qui consiste à lire à voix haute un extrait de la fable "Die Sonne und der Wind".

Pour l'acquisition, seul du matériel grand public a été utilisé (ordinateur portable, webcam intégrée, micro-casque). Toutes les vidéos sont contenues dans un conteneur MPEG-4 (codec H.264), propice à la représentation des visages et des corps humains [32]. Les sujets étaient âgés de 18 à 63 ans (moyenne 31.5 ans et écart-type 12.3 ans).

#### 3.3.5.2 Modalité visuelle

Les images sont capturées à raison de 30 images (*frames*) par seconde, via la webcam de l'ordinateur. Chacune est recentrée sur une région du visage, détectée grâce à la méthode de Viola et Jones [121]. Les images sont redimensionnées de manière à mesurer 200x200 pixels.

Sur les vidéos, le sujet peut éventuellement avoir le visage occulté par des lunettes ou des cheveux. N'étant pas immobile, il peut arriver que son visage sorte du champ de la caméra, ou qu'il ne soit plus visible à l'écran (tête baissée ou tournée vers le côté). L'image 3.14 est issue de l'une des vidéos. Le contexte est similaire dans toutes les vidéos.

Chaque image est décrite au moyen d'un LGBP-TOP. Cependant, dans le souci de réduire la taille des données, seule la dimension $x - y$ a été fournie : les descripteurs ne prennent donc plus en compte l'aspect temporel des vidéos. Les histogrammes de chaque image sont concaténés par vidéo, constituant ainsi les *features baseline* vidéo, i.e. les descripteurs de base.

FIGURE 3.14 – Image exploitable extraite d'une vidéo de AVEC14

#### 3.3.5.3 Modalité auditive

Le signal audio est capturé via un micro-casque branché à la carte son de l'ordinateur, à un taux d'échantillonnage de 41 kHz et 16 bits. Tous les individus s'expriment exclusivement en allemand. L'écoute des vidéos a montré que les sujets ne cherchent pas à parler lentement ou rapidement, ni à modifier leur voix. Certains passages sont silencieux, et on entend éventuellement les sujets rire ou tousser.

Les données auditives ont été extraites via l'outil openEAR [34], et sont composés de 2 268 descripteurs. Ces descripteurs sont issus, dans un premier temps, de l'extraction de descripteurs bas niveaux (*Low Level Descriptors*, LLD) sur lesquels ont été appliquées des fonctions spécifiques afin d'en extraire des informations de plus haut niveau. Le jeu distingue deux catégories de descripteurs bas niveau : ceux relatifs à l'énergie et au spectre, au nombre de trente-deux ; et ceux relatifs à la voix, au nombre de six.

Les descripteurs sont calculés sur des séquences segmentées suivant trois schémas distincts. Les schémas 1 et 2 sont adaptés à la reconnaissance des états émotionnels, et le schéma 3 est adapté à la reconaissance des états dépressifs :

- **Schéma de segmentation 1 :** Les descripteurs sont calculés exclusivement sur des passages où une activité vocale a été détectée. Les passages "silencieux" de plus de 200ms sont considérés comme séparateurs des segments.
- **Schéma de segmentation 2 :** Les descripteurs sont calculés sur de courts segments de 3 secondes se chevauchant d'une seconde.
- **Schéma de segmentation 3 :** Les descripteurs sont calculés sur des segments longs de 20 secondes se chevauchant d'une seconde.

Les tableaux 3.5 et 3.6 présentent les descripteurs bas-niveau et les fonctions appliquées sur ces derniers pour l'obtention des *features* auditives constituant le jeu de données. Ces descripteurs correspondent soit à des descripteurs courants du signal audio, soit à des descripteurs dits psychoacoustiques. D'autres permettent de fournir une information sur la durée, la fréquence ou la tendance des signaux.

#### 3.3.5.4 Annotation

Chaque vidéo est annotée à la fois pour les états emotionnels et dépressifs. Pour les états émotionnels, des valeurs de *valence*, d'*arousal* et de *dominance* sont associées à chaque image, alors que pour les états dépressifs, un niveau de dépression BDI-II est associé à une vidéo entière.

TABLE 3.5 – Tableau des descripteurs bas niveau, d'après [119]

| **Energy & spectral (32)** |
|---|
| loudness (auditory model based),<br>zero crossing rate,<br>energy in bands from 250-650 Hz, 1 kHz-4 kHz,<br>25%, 50%, 75%, and 90% spectral roll-off points,<br>spectral flux, entropy, variance, skewness, kurtosis,<br>psychoacoustic sharpness, harmonicity, flatness,<br>MFCC 1-16 |
| **Voicing related (6)** |
| $F_0$ (sub-harmonic summation, followed by Viterbi smoothing), probability of voicing,<br>jitter, shimmer (local), jitter (delta : "jitter of jitter"),<br>logarithmic Harmonics-to-Noise Ratio (logHNR) |

**Annotation avec des états émotionnels** L'annotation des vidéos avec des états émotionnels est continue et a été synchronisée *a posteriori* sur la fréquence des images dans les vidéos. En ce sens, il y a une valeur comprise entre -1 et 1 par dimension affective et par image, i.e. trente par seconde. L'annotation a été réalisée au moyen d'un joystick et de l'outil FeelTrace [19]. L'utilisation des dimensions *valence*, *arousal* et *dominance* est propice à une représentation dans le circumplex de Russell.

**Annotation avec des états dépressifs** Le sujet de chaque vidéo a passé un test BDI-II, et c'est le score obtenu qui tient lieu d'annotation pour les états dépressifs. Les fournisseurs du corpus n'ont pas précisé la durée écoulée entre le passage du test et la capture vidéo, et les travaux exploitant les données font l'hypothèse que les deux événements ont été réalisés dans un bref laps de temps. En effet, le test BDI-II a une validité de deux semaines. Les scores BDI-II vont de 0 à 46 dans les vidéos, avec une moyenne à 15.0 et 15.6 et un écart-type de 12.3 et 12.0, pour les partitions d'apprentissage et de développement, respectivement. Certains scores entre 0 et 46 ne sont pas représentés.

### 3.3.6 Problèmes liés aux jeux de données

Tous ces jeux de données ont été constitués notamment dans le but d'améliorer la compréhension des phénomènes émotionnels chez l'être humain. Pour ce faire,

TABLE 3.6 – Tableau des fonctions appliquées aux descripteurs bas niveau, d'après [119]

| **Statistical functionals (23)** |
|---|
| positive arithmetic mean, root quadratic mean, standard deviation, flatness, skewness, kurtosis, quartiles, inter-quartile ranges, 1%, 99% percentile, percentile range 1%-99%, percentage of frames contour is above : minimum + 25%, 50%, and 90% of the range, percentage of frames contour is rising, maximum, mean, minimum segment length, standard deviation of segment length |
| **Regression functionals (4)** |
| linear regression slope, and corresponding approximation error (linear), quadratic regression coefficient $a$, and approximation error (linear) |
| **Local minima/maxima related functionals (9)** |
| mean and standard deviation of rising and falling slopes (minimum to maximum), mean and standard deviation of inter maxima distances, amplitude mean of maxima, amplitude range of minima, amplitude range of maxima |
| **Other (6)** |
| LP gain, LPC 1-5 |

l'Intelligence Artificielle propose des méthodes intéressantes. Toutefois, en amont de l'emploi de ces méthodes, il est nécessaire d'identifier un certain nombre de subtilités liées à l'utilisation des données et créées par les conditions d'acquisition, le choix des descripteurs et les méthodes d'annotation retenues pour la constitution d'un jeu de données. La maîtrise de ces subtilités permet non seulement de choisir une base de données de manière judicieuse, mais aussi d'appliquer aux données des prétraitements adéquats.

#### 3.3.6.1 Problèmes liés à l'acquisition

Les conditions d'acquisition des sources de données, i.e. des vidéos, déterminent plusieurs paramètres informatiques, tels que la luminosité, la distance à la caméra

ou l'angle de capture. D'autres paramètres, plus subjectifs, sont également influencés par les conditions de capture : un sujet impressionné par son environnement (Fig. 3.15) aura des comportements émotionnels différents de ceux qu'il aurait dans un environnement familier. D'un autre côté, on estime qu'en situation "normale" (i.e. *in-the-wild*), les individus sont drastiquement moins expressifs, ce qui rend la détection d'*action units* ou l'annotation d'émotions plus délicate.

FIGURE 3.15 – Conditions de laboratoire "extrêmes" pour la capture d'*action units* [88].

Le scénario de *data-elicitation*, qui vise à orienter l'expression émotionnelle du sujet (émotions simulées, induites ou spontanées), doit être étudié précisément. Par exemple, l'utilisation de données simulées par des professionnels (comme c'est le cas de la base GEMEP-FERA) est pertinente pour l'étude des muscles mis en jeu dans des configurations spécifiques. En revanche, les modèles appris sur ces données pourraient ne pas reconnaître des émotions issues de données *in-the-wild*.

#### 3.3.6.2 Problèmes liés à la description

Le rôle primaire d'un descripteur est d'encoder les données sous une forme d'information qu'il sera possible ensuite d'exploiter de manière automatisée. Suivant cette définition, les descripteurs doivent être choisis pour leur capacité à encoder de l'information pertinente au regard du problème à résoudre. L'état de l'art permet de dire (au delà des fondements théoriques des descripteurs) si un descripteur en particulier est adapté ou non à la reconnaissance des émotions. De plus, les

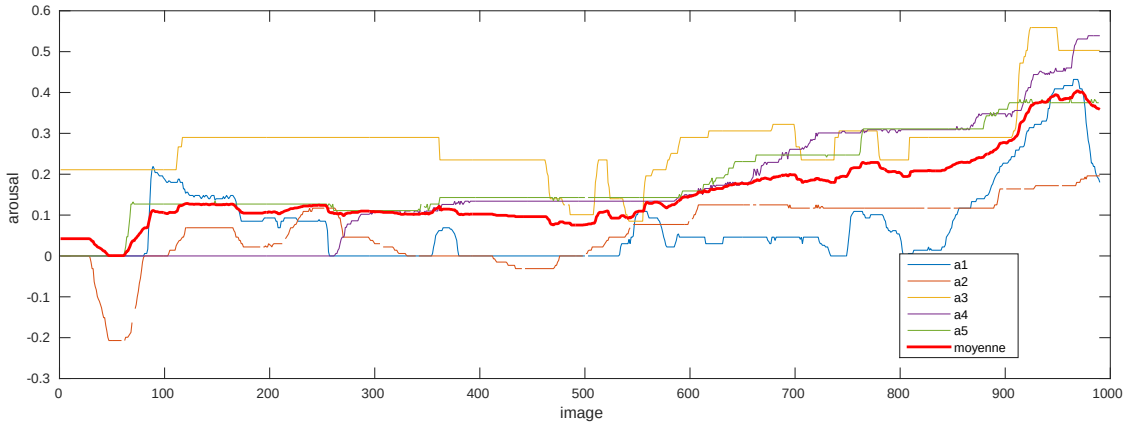



FIGURE 3.16 – Traces de l'*arousal* de cinq annotateurs (de a1 à a5) pour une même vidéo de la base AVLDC.

méthodes de *deep-learning* permettent de s'affranchir du choix des descripteurs, grâce à leur phase d'apprentissage de représentation.

Le paragraphe 3.1.1.1 précise les éléments relatifs à la temporalité des expressions faciales. En particulier, les expressions faciales sont segmentées en quatre phases dont l'*apex*, une phase dite "en plateau" où l'expression faciale est à son apogée. Les phases d'*onset* et d'*offset* sont respectivement celles où l'expression apparaît et disparaît. La phase de l'*apex* représenterait ainsi une fenêtre temporelle pendant laquelle une expression donnée (et, par extension, un état émotionnel) serait la plus visible. Cet élément est intéressant à prendre en compte comme méthode de sous-échantillonage d'un jeu de données, par exemple.

#### 3.3.6.3 Problèmes liés à l'annotation

**Emotion exprimée versus émotion perçue** Les méthodes d'annotation existantes permettent d'annoter des vidéos avec les émotions perçues par l'annotateur, ou avec les émotions que l'on a demandé au sujet de jouer. En aucun cas les annotations ne sont les émotions ressenties par le sujet. Ainsi, on considère que les émotions ressenties et perçues sont extrêmement proches entre elles. Malgré tout, au sens strict du terme, l'usage de ces annotations ne permet pas d'analyser l'expression émotionnelle ressentie par le sujet, mais bien l'expression émotionnelle perçue par l'annotateur.

**Fiabilisation des annotations** Afin de fiabiliser la trace d'annotation, et faire en sorte qu'elle soit la plus proche possible de la réalité, plusieurs annotateurs opèrent sur une même vidéo. La base AVLDC propose la moyenne des traces comme méthode de fiabilisation [120]. La figure 3.16 montre la trace de cinq annotateurs (a1 à a5) ayant annoté une même vidéo dans la dimension affective *arousal*. Pour l'exemple pris, on observe que les traces sont très différentes les unes des autres : dans cet exemple, la moyenne (en rouge) n'est pas une suffisante, statistiquement parlant.

D'autres approches se fondent sur l'hypothèse que les traces sont d'autant plus fiables que les annotateurs sont d'accord. Cette hypothèse introduit le concept de corrélation ou de concordance inter-annotateur (*inter-rater reliability*). Pour la base RECOLA, cette corrélation est calculée par le coefficient de Cronbach $\alpha$ (Eq. 3.2) :

$$\alpha = \frac{A}{A-1}\left(1 - \frac{\sum_{a=1}^{A} \sigma_{\mathbf{y}_a}^2}{\sigma_{\mathbf{Y}}^2}\right) \tag{3.2}$$

avec $A$ le nombre d'annotateurs, $\sigma^2$ la variance, $\mathbf{y}_a$ la trace de l'annotateur $a$ et $\mathbf{Y}$ l'ensemble des traces.

Le coefficient est défini dans $[-1;1]$, et pour $\alpha < 0.8$, on considère que la corrélation est insuffisante.

Ringeval *et al.* [99] ont proposé une approche plus fine, en calculant une trace pondérée $\mathbf{y}_{EWE}$ (EWE pour *Evaluator Weigthed Estimator*) à partir des traces $\mathbf{y}_a$ de chaque annotateur $a$, pour une vidéo $v$ et une dimension affective $d$ fixées. La fiabilité $R_{v,a,d}$ s'exprime comme (Eq. 3.3) :

$$\begin{aligned} \mathbf{y}_{EWE_{v,d}} &= \sum_{a=1}^{A} R_{v,d,a}\ \mathbf{y}_{v,d,a} \\ R_{v,d,a} &= \frac{1}{\sum_{a_i=1}^{A} \rho_{v,d,a_i}}\ \rho_{v,d,a} \\ \rho_{v,d,a} &= \frac{1}{A} \sum_{a_i=1}^{A} \rho_{v,d,(a,a_i)} \end{aligned} \tag{3.3}$$

avec $\rho_{n,d,(a,a_i)}$ le coefficient de corrélation linéaire (de Pearson) entre les traces des annotateurs $a$ et $a_i$.

Lorsque les annotateurs ne sont pas entrainés à reconnaître les émotions, il est indispensable d'utiliser une méthode de fiabilisation. La base de données AVDLC est un exemple de base annotée par des annotateurs naïfs.

**Latence de la trace d'annotation** Un défi d'ampleur est la modélisation du temps de latence (ou *lag*) entre l'instant $t$ où l'émotion est exprimée et l'instant $t+\tau$ où l'annotateur va l'étiqueter. Dans la mesure où les données et les annotations sont synchronisées, un biais temporel est nécessairement créé dans toute base annotée dans le domaine continu (Fig. 3.17). Plusieurs auteurs se sont intéressés à ce

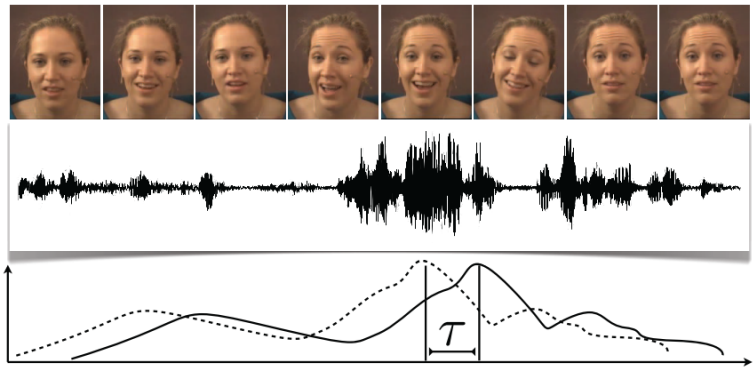



FIGURE 3.17 – Illustration du lag entre l'apparition de l'émotion (pointillés) et son annotation (ligne) [78].

problème. En 2012, Nicolle *et al.* [84] font l'hypothèse que le délai est constant pour un jeu de données et une dimension affective fixés, et le modélisent en calculant la probabilité $P$ que le délai $\tau$ soit optimal grâce au coefficient de corrélation de Pearson $\rho$ (Eq. 3.4) :

$$P(\tau) = \frac{1}{\beta}\sum_{i=1}^{n} \rho\left(x_i(t), y(t-\tau)\right) \tag{3.4}$$

avec $\beta$ un coefficient de normalisation, $x_i, i \in [1,m]$ la $i^e$ dimension d'un vecteur $\mathbf{x}$ et $y(t-\tau)$ l'annotation de $\mathbf{x}$ à l'instant $t$, décalé de $\tau$ secondes.

En 2015, Marioryad et Busso [78] proposent une estimation à tout instant du délai optimal $\bar{\tau}$ entre le signal émotionnel EMO (qui est estimé en quantifiant l'intensité des *action units*) et la trace d'annotation $\mathsf{ANN}^{\bar{\tau}}$ (obtenu en décalant les annotations dans le temps) en maximisant le critère d'information mutuelle entre les deux variables (Eq. 3.5) :

$$\begin{aligned} \mathrm{I}\left[\mathbf{X};\mathbf{Y}\right] &= \sum_{\mathbf{y}\in\mathbf{Y}}\sum_{\mathbf{x}\in\mathbf{X}} p(\mathbf{x},\mathbf{y}) log\left(\frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}\right) \\ \bar{\tau} &= \arg\max_{\tau}\left(\mathrm{I}\left[\mathsf{EMO};\mathsf{ANN}^{\tau}\right]\right) \end{aligned} \tag{3.5}$$

avec $p(.)$ la fonction de masse.

Cette approche permet de corriger le lag à tout instant. Les expérimentations menées par Marioryad et Busso indiquent que leurs délais sont cohérents avec ceux trouvés par Nicolle *et al.*, dont les performances sur un même jeu de données [81] sont proches. En particulier, le retard pour les dimensions *valence* et *arousal* est entre 3 et 4 secondes.

## 3.4 Conclusion

Les définitions multiples de l'émotion, ainsi que le grand nombre de représentations associées à ce terme sont autant de difficultés auxquelles doit faire face la recherche en *Affective Computing*. Aujourd'hui, un conscensus semble se dégager pour l'usage de modélisations continues, mais de nouveaux paradigmes fondés sur la compréhension humaine de la perception des émotions émergent. La dépression connaît un environnement plus cadré et pose moins de problèmes de définitions. Cependant, son étude reste délicate, car elle impose le respect d'un certain nombre de règles, comme le délai de validité d'un diagnostic, par exemple. Que ce soit pour l'analyse automatique des états émotionnels ou des états dépressifs, les données issues de vidéos font désormais l'unanimité. D'un côté, les données vidéos sont simples à acquérir, compte tenu des méthodes et moyens informatiques récents. D'un autre, leur transformation (via un descripteur) et leur annotation demeure délicate et source de biais.

Dans l'état de l'art, plusieurs travaux sur la reconnaissance des états affectifs ou dépressifs exploitent des données issues de vidéos. Le chapitre suivant est consacré à leur présentation, et souligne la difficulté à comparer les méthodes et les performances de systèmes compte tenu de la diversité des données et des représentations.

CHAPITRE 4

# TRAVAUX SUR L'ANALYSE DES ETATS EMOTIONNELS ET DÉPRESSIFS

## 4.1 Reconnaissance des états émotionnels

Pour l'état de l'art sur la reconnaissance des états émotionnels, on regroupe les travaux suivant qu'ils mettent en œuvre des SVM ou d'autres réseaux de neurones. Ces méthodes, qui sont parmi les plus utilisées et les plus efficaces, sont aussi celles que nous avons utilisées pour nos travaux.

### 4.1.1 Travaux à base de SVM

Dans leur effort pour la détection de l'intensité de deux AU spécifiques aux relations mère-enfant, Mahoor *et al.* [77] utilisent des SVM pour classifier l'intensité de deux *action units*. Les données sont décrites par les paramètres de forme et de texture d'un AAM (Active Appearance Model) sur dix-huit minutes de vidéos d'interactions entre des bébés et leur mère et ont subi une réduction de dimensionnalité via une méthode spectrale (RLPI, *Regularized Locality Preserving Indexing*).

Pour la classification des émotions discrètes, Gu *et al.* [42] emploient des SVM sur des représentations graphiques des descripteurs audio, notamment des spectrogrammes (base de données du challenge AVEC2015 [100]). Pour les cinq émotions considérées (colère, joie, neutre, panique et tristesse), le taux de succès moyen est

supérieur à 90%. On note aussi l'emploi d'une méthode de sélection manuelle des descripteurs, ainsi que d'une base de données où les sujets s'expriment en chinois (MAS [126], *Mandarin Affective Speech*).

Dans les deux travaux cités précédemment, l'usage d'une méthode de réduction de la dimension a été nécessaire, et confirme l'intérêt d'un faible nombre de descripteurs pour la classification par des SVM.

Dans un autre cas, Kanluan *et al.* [60] font usage de SVR pour la fusion de prédicteurs neuronaux. Les descripteurs utilisés sont des descripteurs géométriques, calculés sur les données de la base *in-the-wild* SEWA [67] , pour prédire des dimensions affectives dans le domaine continu. On retrouve aussi des usages de SVR limités à la prédiction de certaines modalités, comme l'ont proposé Sánchez-Lozano *et al.* [106] pour des données audio, obtenant une corrélation moyenne de 0.13 en classification sur l'ensemble de test du corpus AVDLC.

Sun *et al.* [113] ont employé des SVR sur les données de la base RECOLA [101] pour la prédiction de la *valence* et de l'*arousal*, et obtiennent des CCC de 0.721 et 0.762, respectivement.

### 4.1.2 Travaux à base de Réseaux de Neurones (RN)

Les méthodes de la seconde grande famille sont capables d'apprendre l'influence du temps sur les variations de l'état affectif. Dans ce contexte, les méthodes de la famille des réseaux de neurones font l'unanimité.

Chao *et al.* [13] ont appris des LSTM-RNN (voir 2.2.1.3) avec les données multimodales temporelles (auditives, visuelles et physiologiques) de la base RECOLA en utilisant une fonction d'optimisation insensible à l'erreur désirée ($\epsilon$-*insensitive loss*). En particulier, la modalité image était composée de descripteurs dynamiques de texture LGBP-TOP et de descripteurs appris par le biais de réseaux de neurones convolutionnels. Cette étude montre qu'il peut être pertinent de combiner ensemble des descripteurs de bas et de haut niveau, puisqu'elle obtient des erreurs RMSE de 0.103 et 0.137 pour la *valence* et l'*arousal*, respectivement.

Une autre dimension affective, nommée *likability*, a aussi été prédite par Chen *et al.* [14] en utilisant les LSTM-RNN. Toutefois, une modalité supplémentaire (transcription des dialogues) et des descripteurs de très haut niveau ont été employés, appris à partir de réseaux convolutionnels tels que VGGFace [89], Re-

sNet [46] et Soundnet [5]. Ils obtiennent des erreurs RMSE de 0.081 et 0.086 pour la *valence* et l'*arousal*, respectivement, et de 0.146 pour la *likability*.

Huang *et al.* [52] obtiennent des erreurs RMSE de 0.093 et 0.085 avec des LSTM-RNN en combinant des descripteurs (LGBP-TOP) à d'autres descripteurs de haut niveau, appris via le réseau convolutionnel AlexNet [68].

Singh *et al.* [111] ont décrit les visages au moyen d'histogrammes de gradients (HOG), de vecteurs SIFT et de descripteurs appris avec des réseaux convolutionnels, pour des scores CCC de 0.466 et 0.306 pour la *valence* et l'*arousal*, respectivement. A titre de comparaison (Singh *et al.* n'ayant pas founi de mesure d'erreur), Chao *et al.* réalisent 0.718 et 0.716, et Chen *et al.* obtiennent 0.756 et 0.672, en termes de CCC pour *valence* et l'*arousal*, respectivement.

On retrouve aussi des ELM (*Extreme Learning Machine*) pour la prédiction de l'état affectif. Kaya et Salah [62] les emploie sur des descripteurs de l'état de l'art, et obtiennent des taux de succès dépassant difficilement les 45% sur les données *in-the-wild* de base de données AFEW [25] pour la reconnaissance d'émotions discrètes.

Les performances entre Chen *et al.* et Huang *et al.* sont proches mais en défaveur de Huang *et al.*, et ce malgré une prise en compte du temps de latence entre l'annotation et l'expression émotionnelle. Cette prise en compte a montré son importance dans la littérature, notamment avec des méthodes non neuronales et des descripteurs de plus bas niveau [78] [84]. Chao *et al.* et Singh *et al.* ont tous deux de moins bonnes performances, et leur point commun notable est l'utilisation de descripteurs bas niveau avec des méthodes de très haut niveau. Le déploiement de ces outils complexes, dont le fonctionnement interne n'est pas interprétable humainement, révèle que l'efficacité de certaines améliorations au niveau des données (comme la prise en compte du temps de latence) dépend de la nature des méthodes employées par la suite.

## 4.2 Reconnaissance des états dépressifs

Les travaux sur la reconnaissance automatique de la dépression sont moins nombreux que ceux concernant la reconnaissance de l'émotion. La dépression est une maladie, reconnue par l'organisation mondiale de la santé et définie par des paramètres stricts dans le DSM-IV [3]. L'étude de la dépression en Informatique

Affective demeure délicate, puisque faisant appel à des concepts médicaux, mais est néanmoins essentielle pour aider à progresser dans la prévention des risques psychosociaux.

Trois éditions du challenge AVEC se sont intéressées à la prédiction des états dépressifs, en 2013, 2014 et 2017. Les données utilisées lors des challenges 2013 et 2014 provenant de la même base, nous étudions ici des méthodes proposées pour les éditions 2014 et 2017 (les plus récentes).

### 4.2.1 Challenge AVEC2014

Pour le challenge AVEC2014 [119], la mesure d'erreur RMSE à battre et fournie come *baseline* était de 9.26 sur l'ensemble de développement (10.86 sur l'ensemble de test) pour la prédiction du score BDI-II associé à chaque vidéo.

Senoussaoui *et al.* [109] ont utilisé des LGBP-TOP (uniquement le plan $x-y$, tel que fourni par les organisateurs) comme descripteur visuel et des *i-vector* [20] comme descripteur audio (les *i-vector* sont une formulation où les composantes des vecteurs sont des modèles de mixtures gaussiennes). En apprenant les états dépressifs séparément des états non-dépressifs à l'aide de modèles linéaires (GLM, *Generalized Linear Models*), de RVM (*Relevance Vector Machines*) et de SVM, ils ont obtenu une erreur de 7.91 sur l'ensemble de développement (10.43 sur l'ensemble de test), faisant mieux que la baseline qui employait des SVR classiques.

Une approche basée sur des MHH (*Motion History Histogram*, Fig. 4.1) concaténant des LBP (*Local Binary Pattern*), des LPQ (*Local Phase Quantization*) et des EOH (*Edge Orientation Histogram*) est proposée par Jan *et al.* [56]. Ces descripteurs sont calculés par frames et permettent une représentation à la fois temporelle et spatiale. Les modalités ont été traitées séparémment avec des régresseurs linéaires (PLS, *Partial Least Squares*) puis les sorties sont fusionnées, pour obtenir une erreur de 9.09 (10.26 sur l'ensemble de test).

Alors que l'approche de Senoussaoui *et al.* vise à obtenir les meilleures performances en testant plusieurs prédicteurs, celle de Jan *et al.* combine des descripteurs et retient la meilleure combinaison. Toutefois, les deux travaux utilisent des éléments et des approches classiques, tant du côté description que du côté classification.

Kächele *et al.* [58] proposent une piste originale en définissant un ensemble de descripteurs complètement nouveaux et surprenants compte tenu du problème.

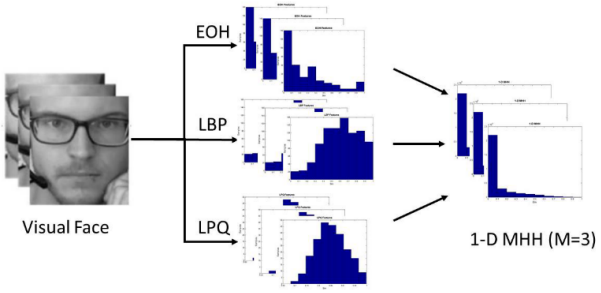



FIGURE 4.1 – Calcul des Motion History Histograms par concaténation de descripteurs classiques proposé par [56]

Parmi eux, on retrouve la durée des vidéos, le poids des sujets, le ratio de compression de la vidéo ou encore du *text mining* via une reconnaissance automatique du discours. En combinant dix-neuf descripteurs de ce style, ils obtiennent une erreur de 9.19 sur l'ensemble de test, ce qui est meilleur que les performances obtenues avec des descripteurs liés à l'expression (faciale ou verbale) du sujet ; qui plus est, avec une méthode de classification par forêts aléatoires classique. Toutefois, cette méthode prend en compte des paramètres liés à la morphologie des sujets ou des caractéristiques qui n'existent que dans les données du challenge (ratio de compression des vidéos, par exemple), et pourrait ne pas donner d'aussi bons résultats dans un contexte différent.

Dans le même courant d'idées, Espinosa *et al.* [33] ont aussi utilisé des caractéristiques liées au clip vidéo (durée, moments parlés, etc.), en plus de descripteurs apportant des informations sur la vitesse moyenne des expressions, calculés grâce au suivi de points d'intérêt faciaux. Le mouvement présent à l'image dans un certain laps de temps est également pris en compte (*Motion History Image*, MHI, et *Motion Average Image*, MAI). Avec des SVR, ils combinent plusieurs approches : apprentissage de l'audio, des vidéos aux moments parlés et non parlés et une fusion des deux. Pour la tâche Freeform, la modalité auditive permet de reconnaître les états dépressifs avec une erreur de 8.97 (ensemble de développement), meilleure que toute autre combinaison, même multimodale. Ils tentent aussi de prédire les scores dépressifs à partir des annotations émotionnelles (*valence*, *arousal*, *dominance*), qui deviennent alors des données. Cette approche est moins performante, avec une erreur de 10.91 (ensemble de développement).

De nouveaux descripteurs sont proposés par Williamson *et al.* [122], tels que le *Facial Activation Rate*, défini comme étant une moyenne des AU sur une durée fixée,

ainsi qu'une mesure de la coordination faciale. Ainsi, ils ont utilisé un détecteur d'AU et choisi avec soin plusieurs descripteurs audio (durée des phonèmes, *pitch slope*). Avec des ELM, une erreur de 8.12 (ensemble de test) est obtenue.

L'échec de la prédiction des états dépressifs à partir des annotations émotionnelles n'est pas surprennant (les émotions étant considérées comme promptes et négativement correlées avec les mesures de dépression, nécessairement mesurées sur la durée), le succès avec les passages audio pour la tâche Freeform peut s'expliquer par le principe même de la tâche (voir partie 3.3.5.1) : elle consiste à répondre à un certain nombre de questions d'ordre général, impliquant à la fois des temps de pause, dont la durée peut être révélatrice de l'état dépressif, et une plus grande variabilité émotionnelle qu'en lisant un passage écrit (qui est le principe de la tâche Northwind). La table 4.1 propose un comparatif des performances des travaux étudiés plus haut.

TABLE 4.1 – Résultats comparés des travaux cités pour le challenge AVEC2014 (Dépression). Les performances sont calculées sur le jeu de test.

| Auteurs | RMSE | Remarque |
|---|---|---|
| Valstar *et al.* [119] | 10.86 | LGBP-TOP, SVR |
| Senoussaoui *et al.* [109] | 10.43 | LGBP-TOP, GLM, RVM |
| Jan *et al.* [55] | 10.26 | LBP, LPQ, PLS |
| Kächele *et al.* [58] | 9.19 | Descripteurs morph., fôrets aléatoires |
| Espinosa *et al.* [33] | 10.82 | MHI, MAI, SVR |
| Williamson *et al.* [122] | **8.12** | FAR, descr. auditifs, ELM |

### 4.2.2 Challenge AVEC2017

Le challenge AVEC2017 [99] est différent en plusieurs points de celui de 2014. Le label à prédire est désormais un score PHQ-8 associé aux sujets, et la base de données est la base DAIC-WOZ. Pour ce challenge, des données visuelles, auditives, textuelles et physiologiques sont fournies. L'erreur RMSE à battre et fournie comme *baseline* est de 6.62 (7.05 sur l'ensemble de test).

Dans une proposition hybride, Yang *et al.* [128] emploient à la fois des réseaux de neurones profonds (Fig. 4.2) et une régression multivariée afin d'apprendre séparément les états dépressifs des états non-dépressifs. Les descripteurs employés sont, en plus de ceux fournis, des descripteurs appris via des réseaux de neurones

convolutifs (pour les données visuelles) et un comptage du nombre de mots, de soupirs et de rires. Une erreur RMSE de 3.09 est obtenu (5.4 pour l'ensemble de test).

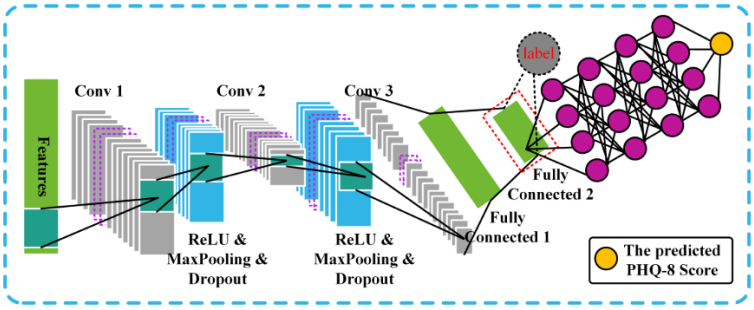



FIGURE 4.2 – Des réseaux de neurones profonds similaires sont mis en œuvre pour chacune des modalités disponibles (vidéo, audio et textuelle) pour la prédiction du score dépressif, tel que proposé par [128]

Deux autres approches utilisent des forêts aléatoires, mais sur des descripteurs différents. Sun et Wang [114] ont proposé d'utiliser les transcriptions pour récupérer un certain nombre d'informations, comme la qualité du sommeil, si le sujet prend un traitement médical ou encore s'il a été diagnostiqué comme étant dépressif. Ces éléments traduisent souvent une probabilité élevée pour la dépression, tout comme l'information apportée par le mouvement du sujet, également exploitée (accélération des points du visage, notamment). Cette approche obtient une erreur RMSE de 4.7 (4.98 en test).

Outre les forêts aléatoires, Gong et Poellabauer [38] ont utilisé des SVR et des descentes de gradient pour retenir le meilleur régresseur sur les descripteurs de la *baseline*, auxquelles ils ont adjoint des éléments sur le thème de la conversation. Les transcriptions ont ainsi été étudiées afin de modéliser le thème (*topic modeling*) suivant plusieurs paramètres (vocabulaire, nombre de mots, etc.). En fonction du topic, un certain nombre de descripteurs était retenu pour l'apprentissage et la généralisation. Ils ont testé leur méthode à la fois en validation croisée (en concaténant les ensembles d'apprentissage et de développement) et de manière classique. Les meilleures performances sont obtenues par la descente de gradient, qui est de 2.94 en validation croisée, 3.54 sur l'ensemble de développement et 4.99 sur l'ensemble de test.

## 4.3 Discussion

Dans ce chapitre, nous avons présenté plusieurs travaux sur la reconnaissance des états émotionels et des états dépressifs. La grande diversité des méthodes et approches montre l'absence de consensus, tant sur les modèles de classifieurs que sur les types de descripteurs à privilégier. L'évaluation de la performance des systèmes est également différente d'une tentative à l'autre, ce qui entrave leur comparaison.

L'existence de challenges, fournissant des bases de données partagées et des procédures communes pour l'évaluation, est une piste intéressante pour l'avancement du domaine.

Dans les prochaines années, les tentatives feront vraisemblablement toutes appel à des méthodes de très haut niveau, mettant en œuvre des réseaux de neurones convolutionnels, car leurs performances sur les données visuelles ou auditives dépassent celles des méthodes désormais "historiques" (réseaux de neurones traditionnels, SVM, etc.). Malgré tout, ces dernières ont quelques avantages, dont celui de modéliser les phénomènes de manière humainement compréhensible et de nécessiter moins de ressources informatiques que les méthodes de haut niveau. En plus d'automatiser l'analyse des comportements, elles fournissent des pistes intéressantes pour leur interprétation.

CHAPITRE 5

# CLASSIFICATION DES INTENSITÉS EMOTIONNELLES

Dans ce chapitre, on cherche à classifier l'état émotionnel de sujets présents dans des vidéos. Les classes représentent des niveaux d'intensités émotionnelles, qui se veulent plus interprétables que les traditionnelles valeurs réelles entre -1 et 1 utilisées pour représenter l'émotion.

## 5.1 Spécification des données et des classes

### 5.1.1 Données utilisées

Le jeu de données utilisé est extrait de la base AVDLC, décrit dans la section 3.3.5). Seules les données issues de la tâche Freeform sont utilisées, dans la mesure où l'état de l'art montre qu'elles sont plus adaptées que celles de la tâche Northwind pour la reconnaissance des états émotionnels. Le jeu entier a été nettoyé des données aberrantes et redondantes avant utilisation. En particulier, les données nulles (correspondant à des visages non détectés), les vidéos trop courtes (de moins de dix secondes) et les données non synchronisées avec leurs annotations ont été supprimées de la base. Afin de réduire la redondance dans les données (les vidéos ayant été enregistrées à 30 images par secondes), on retient une image sur deux de chaque vidéo. Au total, on dispose de 60 000 exemples concernant 51 sujets distincts, annotés avec l'information émotionnelle (trois attributs : *valence, arousal,*

*dominance* ; valeurs continues) et l'identifiant du sujet (un attribut).

La base de données AVDLC a été annotée par des volontaires naïfs, ce qui peut remettre en question la fiabilité des traces affectives produites (voir paragraphe 3.3.6.3). Nous avons implémenté la méthode de fiabilisation proposée par Ringeval *et al.* [99] afin d'évaluer la fiabilité des annotations. La table 5.1 reprend les coefficients $R_{v,d,a}$ de corrélation inter-annotateurs (Eq. 3.3, toutes vidéos confondues par annotateur et par dimension affective).

TABLE 5.1 – Coefficients $R$ de corrélation inter-annotateurs pour les données support. Chaque valeur est la corrélation d'un annotateur avec tous les autres annotateurs, pour l'ensemble des vidéos dans la dimension affective considérée.

| Annotateur | Valence | Arousal | Dominance |
|---|---|---|---|
| a1 | 0.69 | 0.62 | -0.05 |
| a2 | 0.84 | 0.72 | 0.31 |
| a3 | 0.83 | 0.70 | 0 |
| a4 | 0.75 | 0.74 | 0.28 |
| a5 | 0.81 | 0.79 | - |

Malgré leur naïveté, les annotateurs sont plutôt d'accord concernant la *valence* et l'*arousal* des sujets notés, ce qui induit une forme de cohérence dans la moyenne des traces. La corrélation pour la *dominance*, elle, est en revanche faible. La méthode prévoit de calculer la trace moyenne sans les traces produites par les annotateurs pour lesquels $R \leq 0$. Cette procédure a uniquement été appliquée pour la dimension affective *dominance*. De manière globale, ces résultats peuvent s'expliquer par la difficulté pour un individu lambda à interpréter la notion de *dominance*, contrairement aux deux autres dimensions affectives dont la définition est plus intuitive.

Initialement fournis dans des fichiers plats (au format CSV), les fichiers ont été convertis au format libre H5 [1], adapté aux données de grande taille. L'usage de ce format diminue considérablement les temps d'accès et d'écriture des données, tout en réduisant l'espace disque nécessaire à leur stockage.

### 5.1.2 Spécification des classes

Le choix d'utiliser des classes d'intensités émotionnelles plutôt que des valeurs continues est notamment motivé par l'argument de Yannakakis *et al.* [130], que

1. https ://www.hdfgroup.org/

nous partageons, et qui propose qu'une information de classe est plus fiable qu'une valeur continue.

De manière intuitive, une émotion intense est annotée, dans la dimension affective considérée, par une valeur proche des bornes de son domaine de définition : 1 ou $-1$. *A contrario*, une émotion neutre est annotée avec une valeur proche de 0. La distribution des annotations (Fig. 5.1) nuance cette approche, dans la mesure où les annotations sont en grande majorité (plus de 90%) dans l'intervalle $[-0.2;0.2]$. De plus, les émotions dont la valeur est proche de 0 sont bien plus nombreuses que celles dont la valeur est proche de $-0.2$ et 0.2.

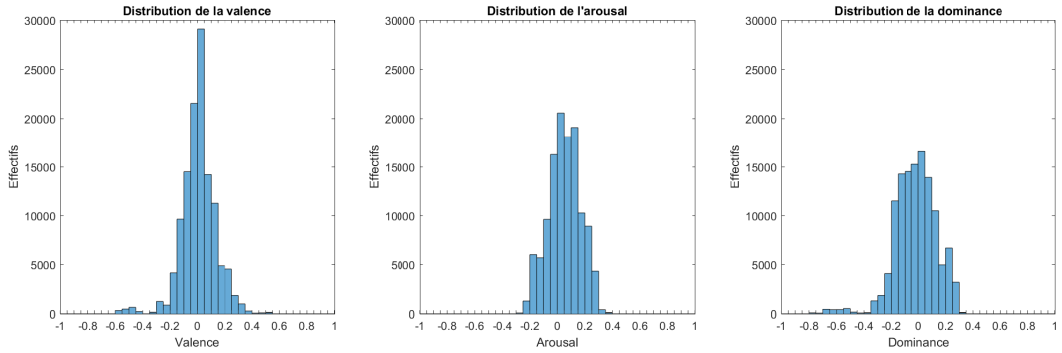



FIGURE 5.1 – Distribution des intensités émotionnelles entre -1 et 1.

En ce sens, la définition des classes d'intensités émotionnelles doit respecter les deux crièteres suivants : ne pas introduire de déséquilibre de population entre classes (ce qui pourrait nuire à la qualité de la classification) et correspondre à l'interprétation humaine des intensités émotionnelles (plutôt qu'à leur distribution uniforme théorique dans $[-1;1]$).

On considère six classes d'intensités émotionnelles, définies sur $[-1;1]$ telles que toute émotion dont la valeur est comprise dans [min;max[ appartienne à la classe correspondante (Tab. 5.2 et Fig. 5.2) :

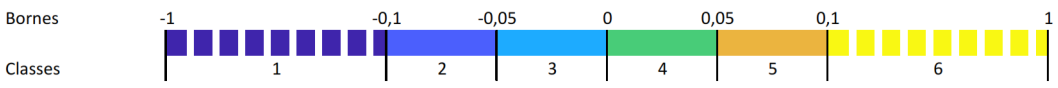



FIGURE 5.2 – Illustration des bornes de définition des classes.

La figure 5.3 présente la distribution des intensités émotionnelles dans les classes. Les classes d'intensité émotionnelle permettent également de rééquilibrer la distribution des données.

TABLE 5.2 – Définition des classes d'intensités émotionnelles pour les dimensions *valence*, *arousal* et *dominance*.

| Classe | Minimum | Maximum | Remarque |
|---|---|---|---|
| 1 | -1 | -0.1 | Intensité fortement négative |
| 2 | -0.1 | -0.05 | Intensité négative à neutre |
| 3 | -0.05 | 0 | Intensité neutre |
| 4 | 0 | 0.05 | Intensité neutre |
| 5 | 0.05 | 0.1 | Intensité neutre à positive |
| 6 | 0.1 | 1 | Intensité fortement positive |

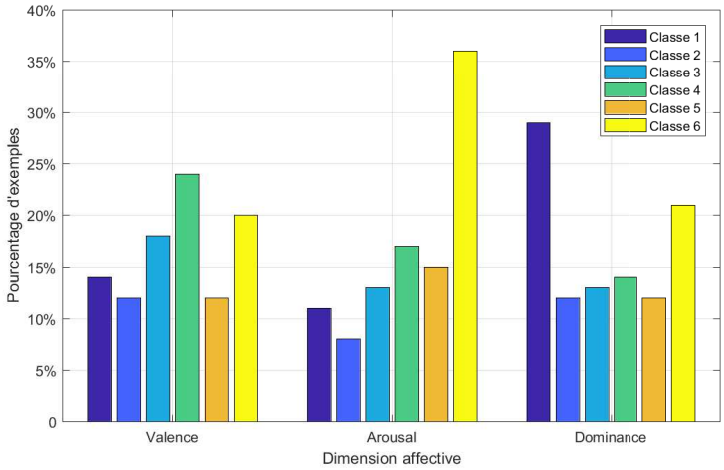



FIGURE 5.3 – Distribution des intensités émotionnelles dans les classes.

## 5.2 Première approche

### 5.2.1 Conditions expérimentales

Compte tenu de la dimensionnalité élevée des données (16 992 dimensions), on réalise une *kernel*-PCA [108] afin d'en réduire la dimension. L'analyse en composantes principales [90] (ACP, ou *princial component analysis*, PCA) est une méthode de réduction de dimension fondée sur le changement de l'espace de représentation des données. Sur le critère de la variance, on détermine quels attributs sont moins corrélés avec les autres. Ces attributs, qui maximisent la variance expliquée, sont appelés composantes principales et sont utilisés comme axes de représentation des données.

Le noyau retenu pour la *kernel*-PCA est le noyau intersection d'histogramme, afin de tenir compte de la nature des données.

Les 3 639 composantes dont la variance permet d'expliquer 95% de la variance totale ont été retenues, soit une réduction d'un facteur 5. On a également réalisé une analyse en composantes principales classique, afin de comparer l'apport du noyau.

Les expérimentations ont pour objectif de classifier l'intensité émotionnelle des sujets présents dans les données, suivant les intensités définies au paragraphe 5.1.2. On utilise une SVM à noyau linéaire afin de classifier les données issues de la *kernel*-PCA. Afin de comparer leurs performances, on classifie les données issues de la PCA classique avec une SVM à noyau linéaire et une SVM à noyau RBF. La SVM à noyau RBF n'est pas utilisée sur les données issues de la *kernel*-PCA, ces dernières ayant déjà fait l'objet d'un calcul de noyau. Dans cette sous-section, la dimension affective *valence* est prise comme "pilote" afin de présenter les résultats intermédiaires, par souci de clarté.

### 5.2.2 Optimisation des classfieurs

Une recherche en grille est menée sur 20% du jeu de données réduit (12 000 exemples, dimension 3 639) pour chaque classifieur afin de déterminer les meilleurs hyperparamètres. A chaque essai, le modèle appris est généralisé sur des sujets inconnus. Pour les SVM linéaires, seul le paramètre de porosité de la marge $C$ est optimisé ; pour la SVM à noyau RBF, la recherche en grille est menée pour les paramètres $C$ et $\gamma$.

#### 5.2.2.1 Optimisation du paramètre $C$ pour les SVM linéaires

On fait varier $C$ entre $1.10^{-2}$ et $1.10^{4}$ de manière logarithmique pour les données issues de la *kernel*-PCA et celles issues de la PCA classique. La figure 5.4 montre les taux de succès calculés sur l'ensemble d'apprentissage (ligne bleue) et sur l'ensemble de généralisation (ligne orange).

On note que pour une grande valeur de $C$, l'apprentissage se fait correctement mais produit une grande quantité de vecteurs de support (au dessus de 70%) : il y a donc surapprentissage. Dans tous les cas, les taux de succès en généralisation demeurent mauvais, avec un maximum de 0.32 pour $C = 1$.

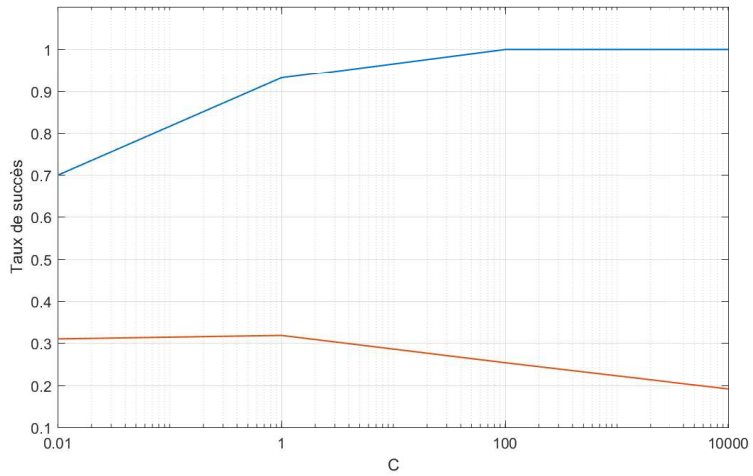



FIGURE 5.4 – Résultats de l'optimisation de la SVM linéaire

#### 5.2.2.2 Optimisation des paramètres $C$ et $\gamma$ pour la SVM RBF

On fait varier $C$ entre $1.10^{-2}$ et $1.10^{4}$ et $\gamma$ entre $1.10^{-4}$ et $1.10^{3}$ de manière logarithmique pour les données issues de la PCA classique. La figure 5.5a montre les taux de succès calculés sur l'ensemble d'apprentissage et la figure 5.5b les taux de succès calculés sur l'ensemble de généralisation, pour la recherche en grille.

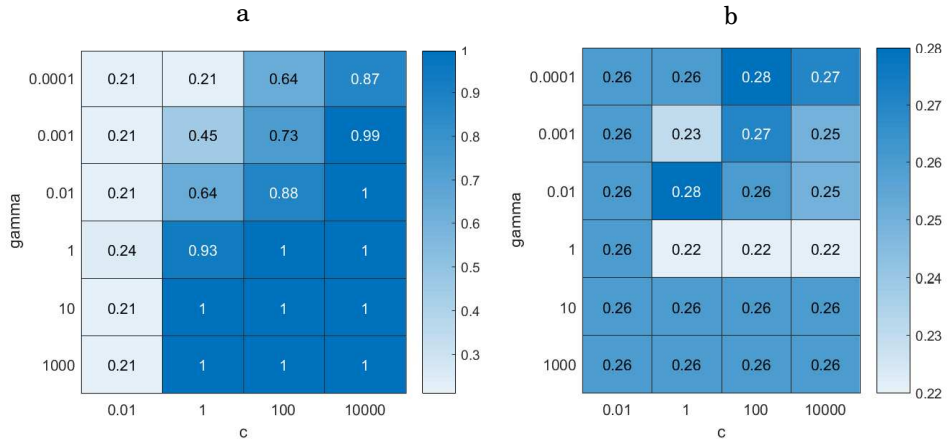



FIGURE 5.5 – (a) Taux de succès en apprentissage. (b) Taux de succès en généralisation.

On note que les taux de succès en apprentissage atteingnent 100% pour certaines combinaisons de paramètres, alors que le taux maximal en généralisation est de 28%. Ces taux sont à contraster avec le nombre de vecteurs de support, qui reste

élevé (toujours supérieur à 50%), et en particulier pour les taux d'apprentissage à 100% (entre 70 et 100% de vecteurs de support).

### 5.2.3 Résultats et éléments d'interprétation

Chacune des trois expériences a été réalisée sur la totalité du jeu de données, en validation croisée *leave-one-subject-out*. Les hyperparamètres retenus sont $C = 1$ pour la SVM linéaire et $C = 1, \gamma = 0.01$ pour la SVM RBF.

#### 5.2.3.1 Classification avec une SVM linéaire

La figure 5.6 montre la répartition des taux de succès (par sujet) en généralisation réalisés avec une SVM linéaire sur les données issues de la PCA classique (Fig. 5.6a) et sur celles issues de la *kernel*-PCA (Fig. 5.6b). Avec un taux de succès moyen ne dépassant pas 0.25 dans les deux cas, cette tentative n'a pas réussi à classifier les intensités émotionnelles.

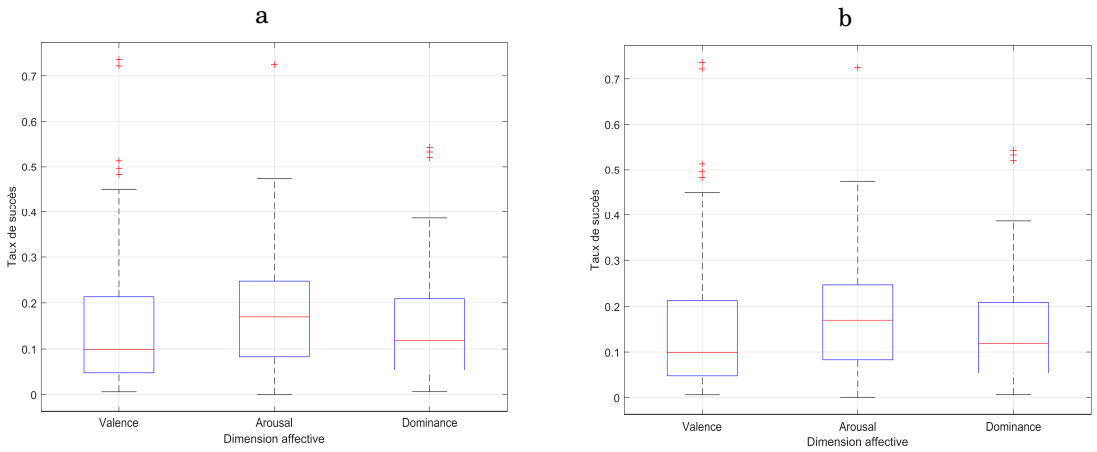



FIGURE 5.6 – Synthèse des performances avec une SVM linéaire. (a) PCA classique. (b) *kernel*-PCA.

D'autres indicateurs de performance (voir Eq. 2.24, 2.25, et 2.26) permettent de mieux interpréter ces premiers résultats (Tab.5.3). Par exemple, pour l'*arousal* (dimension ayant permis d'atteindre les meilleurs performances), il apparaît que les classifieurs réussissent à prédire la non-appartenance des exemples à une classe (spécificités élevées), mais ont du mal à prédire leur appartenance à la bonne classe (rappel faible).

TABLE 5.3 – Indicateurs de performance - Arousal

| | Classe 1 | Classe 2 | Classe 3 | Classe 4 | Classe 5 | Classe 6 |
|---|---|---|---|---|---|---|
| | PCA classique + SVM linéaire | | | | | |
| Précision | 0,04 | 0,06 | 0,16 | 0,17 | 0,23 | 0,32 |
| Rappel | 0,07 | 0,12 | 0,12 | 0,16 | 0,24 | 0,20 |
| Spécificité | 0,79 | 0,85 | 0,90 | 0,83 | 0,85 | 0,77 |
| | *kernel*-PCA + SVM linéaire | | | | | |
| Précision | 0,08 | 0,09 | 0,15 | 0,2 | 0,27 | 0,29 |
| Rappel | 0,11 | 0,11 | 0,13 | 0,14 | 0,21 | 0,22 |
| Spécificité | 0,67 | 0,87 | 0,88 | 0,89 | 0,9 | 0,81 |

Les coefficients de Pearson et de Spearman sont également médiocres pour chacune des trois dimensions affectives.

#### 5.2.3.2 Classification avec une SVM à noyau RBF

La SVM à noyau RBF ($C = 1, \gamma = 0.01$) n'a pas non plus réussi à discriminer les classes émotionnelles des sujets du jeu de données (réduit à l'aide d'une PCA classique). La synthèse des taux de succès par sujet (Fig. 5.7) fait état de résultats équivalents à ceux obtenus avec une SVM linéaire.

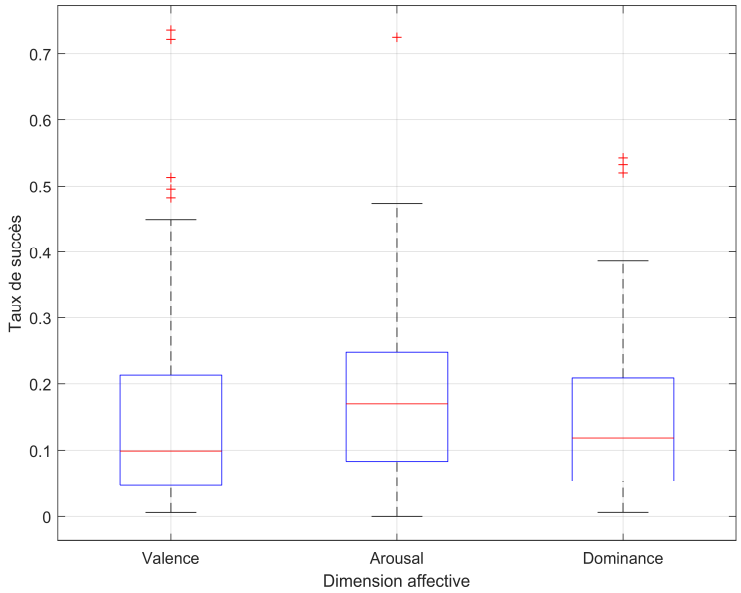



FIGURE 5.7 – Synthèse des performances avec une SVM RBF

La dimension affective *arousal* demeure celle qui offre les meilleures perfor-

mances, néanmoins médiocres et avec le même profil que pour la SVM linéaire : les exemples appartenant à une classe donnée ont du mal à être reconnus comme tels.

TABLE 5.4 – Indicateurs de performance - Arousal (PCA classique + SVM RBF)

| | Classe 1 | Classe 2 | Classe 3 | Classe 4 | Classe 5 | Classe 6 |
|---|---|---|---|---|---|---|
| Précision | 0,01 | 0,07 | 0,14 | 0,16 | 0,2 | 0,29 |
| Rappel | 0,12 | 0,15 | 0,15 | 0,18 | 0,21 | 0,18 |
| Spécificité | 0,78 | 0,82 | 0,89 | 0,86 | 0,86 | 0,79 |

#### 5.2.3.3 Eléments d'interprétation

Aucune des trois expériences de cette première approche n'a su offrir de bonnes performances. Toutefois, elles apportent des indications intéressantes quant à la possibilité de classifier ces données. En effet, chacune des trois expériences a tenté une projection différente dans l'espace des données (grâce aux réductions dimensionnelles) et avec différents types de frontières (grâce aux noyaux des SVM), mais aucune combinaison ne s'est avérée fructueuse. Une piste à étudier plus en profondeur est la séparabilité du jeu de données, et en particulier sa séparabilité vis-à-vis des annotations émotionnelles.

## 5.3 Etude de la séparabilité des classes

La variabilité émotionnelle d'un sujet est forte lorsque son signal affectif passe par plusieurs classes, et est faible dans le cas contraire. En moyenne, dans les données, un sujet passe par quatre intensités émotionnelles sur six. Cette forte variabilité, couplée au grand nombre de sujets et aux mauvais résultats obtenus en classification, impose d'étudier la séparabilité des données.

Les descripteurs de bas niveau sont connus dans la littérature pour être sujets au biais d'identité [1], i.e. à l'encodage d'une information essentiellement morphologique (relative au sujet) au détriment d'une information sémantique (relative, dans notre cas, aux intensités émotionnelles). Une solution rapide serait, dans ce cas, de changer de description. Toutefois, cela ne permettrait pas de documenter les véritables causes des mauvais taux de succès, à savoir s'ils sont dûs :

— Aux données, et plus précisément à leur description, ou

— Au processus de classification (réduction dimensionnelle, calcul de noyau et algorithme de classification utilisé).

Dans cette section, on applique plusieurs outils formels et visuels afin d'estimer la séparabilité des données.

### 5.3.1 Estimation formelle

La mesure de divergence de Kullback-Leibler [70] estime la différence entre deux distributions de probabilités. Cette mesure peut être utilisée en classification supervisée afin d'évaluer la divergence dans les données de deux classes distinctes. Si on note $P$ et $Q$ des distributions de probabilité (définies sur un même espace de probabilité) d'une variable aléatoire discrète, la divergence (symétrique) de Kullbach-Leibler est définie par (Eq. 5.1) :

$$\begin{aligned} D_{KL}(P,Q) &= D_{KL}(P||Q) + D_{KL}(Q||P) \\ &= -\sum_{\mathbf{x}} P(\mathbf{x})\ log\left(\frac{Q(\mathbf{x})}{P(\mathbf{x})}\right) + \sum_{\mathbf{x}} Q(\mathbf{x})\ log\left(\frac{P(\mathbf{x})}{Q(\mathbf{x})}\right) \end{aligned} \tag{5.1}$$

A partir de cette définition, on calcule la divergence entre les données des six classes d'intensités émotionnelles, pour chaque dimension affective (Fig. 5.8). On note d'emblée la très faible divergence, de l'ordre de $10^{-2}$, des données d'une classe à l'autre.

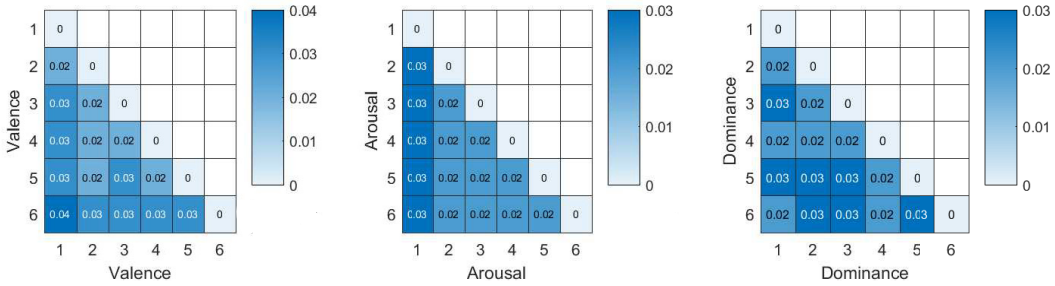



FIGURE 5.8 – Mesures de divergence entre les classes.

A titre de comparaison, la divergence entre les sujets est $10^{-1}$, soit un ordre de grandeur plus élevé que la divergence entre les classes. Cette différence est significative, puique le calcul est effectué à l'échelle logarithmique. La divergence entre sujets est plus élevée que la divergence entre intensités émotionnelles, ce qui est peu propice à la classification des intensité émotionnelles.

### 5.3.2 Estimation visuelle

#### 5.3.2.1 Projection des composantes principales

L'analyse en composantes principales permet de "concentrer" une grande quantité d'information dans les premières composantes du jeu de données réduit. Toutefois, dans la mesure où la PCA réalise une réduction non-supervisée, aucune information concernant les classes n'est introduite lors du traitement. En conséquence, la grande quantité d'information présente dans les premières composantes n'est pas nécessairement liée aux classes.

Afin de visualiser les éventuelles séparations existant dans les données réduites, on projette la première composante en fonction de la seconde. On colore ensuite les points d'après leur classe de valence (Fig. 5.9a) et d'après le sujet (Fig. 5.9b).

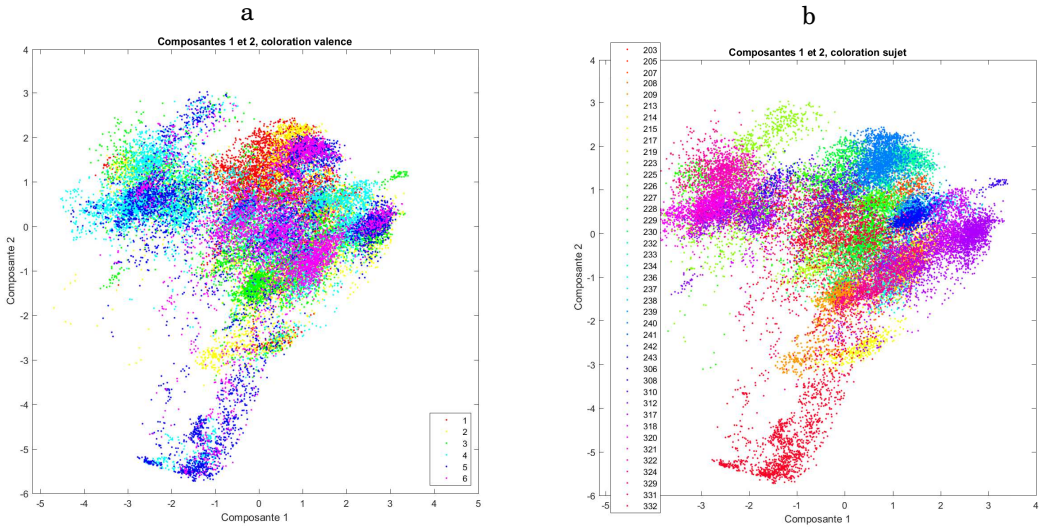



FIGURE 5.9 – (a) Coloration par classe valence. (b) Coloration par sujet.

On note que les sujets sont plutôt bien séparés, malgré une représentation dans un espace de seulement deux composantes. En revanche, les intensités émotionnelles ne définissent pas de zones, et sont réparties sur l'ensemble du graphique.

#### 5.3.2.2 Visualisation par *t-distributed stochastic neighbor embedding*

Une autre méthode d'estimation est la *t-distributed stochastic neighbor embedding* (t-SNE) [76], qui permet de visualiser un jeu de données **X** de grande

dimension $m$ par une carte de deux ou trois dimensions, notée $\mathbf{z}$. On note $n$ le nombre d'exemples.

L'algorithme repose sur la "conversion" de la distance entre deux points en une probabilité conditionnelle, représentant la similarité entre ces deux points : deux points seront similaires s'ils sont proches. On note $p_{ij}$ la probabilité que deux points $x_j$ et $x_i$ de l'espace de grande dimension soient proches et $q_{ij}$ la probabilité que leurs représentations $z_j$ et $z_i$ dans l'espace de faible dimension soient également proches. Dans l'espace de grande dimension, la conversion est réalisée en modélisant la distribution des points par une loi normale centrée en $x_i$ (Eq. 5.2), alors que dans l'espace de faible dimension, elle est modélisée par une distribution de Student (Eq. 5.3).

$$p_{ij} = \frac{exp\left(-||x_i - x_j||^2/2\sigma_i^2\right)}{\sum_{k \neq i} exp\left(-||x_k - x_j||^2/2\sigma_i^2\right)} \tag{5.2}$$

$$q_{ij} = \frac{\left(1 + ||y_i - y_j||^2\right)^{-1}}{\sum_{k \neq i} \left(1 + ||y_k - y_i||^2\right)^{-1}} \tag{5.3}$$

La représentation de faible dimension est correcte si et seulement si, pour deux points $x_j$ et $x_i$ donnés, $p_{ij} = q_{ij}$. Cette égalité est résolue en minimisant la somme $S_{D_{KL}}$ des divergences de Kullback-Leibler sur l'ensemble du jeu de données (Eq. 5.4).

$$S_{D_{KL}} = \sum_i \sum_j p_{ij} log\left(\frac{p_{ij}}{q_{ij}}\right) \tag{5.4}$$

On projette les composantes produites par t-SNE puis, comme précédemment, on colore les points suivant leur classe de valence (Fig. 5.10a) et suivant le sujet (Fig. 5.10b).

On observe que les sujets définissent des zones tranchées, confirmant les résultats obtenus via les estimations précédentes. Les intensités émotionnelles, en revanche, n'ont pas de caractère discriminant vis-à-vis des zones. Compte tenu de ces éléments, il est peu probable de pouvoir reconnaître autre chose que les sujets avec ces données.

La séparation *naturelle* existant dans les données, i.e. celle encodée par la description utilisée, permet essentiellement de discriminer les sujets. Afin de

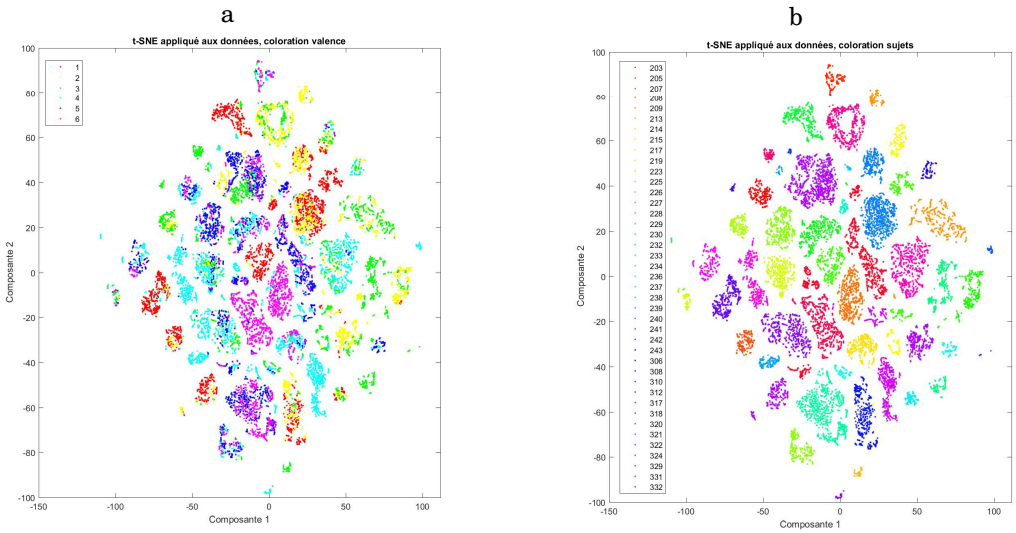



FIGURE 5.10 – (a) Coloration par classe de valence. (b) Coloration par sujet.

discriminer les intensités émotionnelles, le jeu de données doit être séparable en ce sens. Dans la section suivante, on propose d'utiliser une méthode de réduction dimensionnelle supervisée.

## 5.4 Seconde approche

Les résultats et conclusions des sections précédentes montrent que les données véhiculent une grande part d'information morphologique, empêchant de discriminer les intensités émotionnelles en classification. Dans cette section, on tire profit de ces éléments afin de représenter les données dans un espace où elles sont séparables d'après leur classe. La séparabilité des données est à nouveau vérifiée, en amont de la classification.

### 5.4.1 Analyse Discriminante Linéaire

L'analyse discriminante linéaire (*linear discriminant analysis*, LDA), est une technique de réduction dimensionnelle supervisée [115].

Soit un jeu de données annoté $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ avec $\mathbf{x}_i \in \mathrm{R}^m$ un vecteur de données et $y_i \in \{k_1, \ldots, k_c\}$ la classe de $\mathbf{x}_i$. L'action de la LDA consiste à trouver un espace de dimension $l \leq (c-1)$ dans lequel les données dont la classe est commune sont proches.

On note $\boldsymbol{S}_W$ la matrice des variances intra-classes (*scatter-within matrix*) et $\boldsymbol{S}_B$ la matrice des variances inter-classes (*scatter-between matrix*) définies par (Eq. 5.5) :

$$\begin{aligned} \boldsymbol{S}_W &= \sum_{q=1}^{c} \boldsymbol{S}_q \,; \quad \boldsymbol{S}_B = \sum_{q=1}^{c} \left( n_y \left(\boldsymbol{\mu}_y - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}_y - \boldsymbol{\mu}\right)^T \right) \\ \text{avec } \boldsymbol{S}_q &= \sum_{\mathbf{x} \setminus y = k_q} \left(\mathbf{x} - \boldsymbol{\mu}_y\right)\left(\mathbf{x} - \boldsymbol{\mu}_y\right)^T \text{ et } \boldsymbol{\mu}_y = \frac{1}{n_y} \sum_{\mathbf{x} \setminus y = k_q} \mathbf{x} \\ \text{et } \boldsymbol{\mu} &= \frac{1}{n} \sum_{\forall \mathbf{x}} \mathbf{x} \end{aligned} \tag{5.5}$$

où $n_y$ est le nombre d'exemples de la classe $y$.

La matrice de transformation $\mathbf{W}$ vers un espace de dimension réduite peut être obtenue en calculant le critère de Fischer [36] (Eq. 5.6), qui peut être reformulé comme (Eq. 5.7) :

$$\mathbf{W} = \arg\max \frac{|\mathbf{W}^T \boldsymbol{S}_B \mathbf{W}|}{|\mathbf{W}^T \boldsymbol{S}_W \mathbf{W}|} \tag{5.6}$$

$$\boldsymbol{S}_W \mathbf{W} = \lambda \boldsymbol{S}_B \mathbf{W} \tag{5.7}$$

où $\lambda$ représente les valeurs propres de la matrice $\mathbf{W}$. Les axes de projection du nouvel espace de dimension réduit sont les vecteurs propres de $\mathbf{W} = \boldsymbol{S}_W^{\top} \boldsymbol{S}_B$, idéalement ordonnés par valeurs propres décroissantes.

On réduit le nombre de dimensions du jeu de données par analyse discriminante linéaire, afin de maximiser la séparation entre les classes d'intensités émotionnelles. Le jeu de données réduit, de dimension 5, est utilisé pour les expérimentations. La figure 5.11 donne une estimation visuelle de la séparation des données pour la dimension affective valence : on note que les données sont effectivement séparées d'après leur intensité émotionnelle.

### 5.4.2 Expérimentations et résultats

On utilise les données réduites à l'aide de la LDA afin de classifier l'intensité émotionnelle des sujets. Pour cette tentative, une SVM linéaire avec $C = 1$ est utilisée. La LDA a été calculée séparément pour chacune des dimensions affectives, à partir des données originales (de dimension 16 992).

La plupart des exemples d'une classe donnée sont classifiés dans la bonne classe (le taux de succès moyen pour la classification de la *valence* est de 0.99). Les erreurs

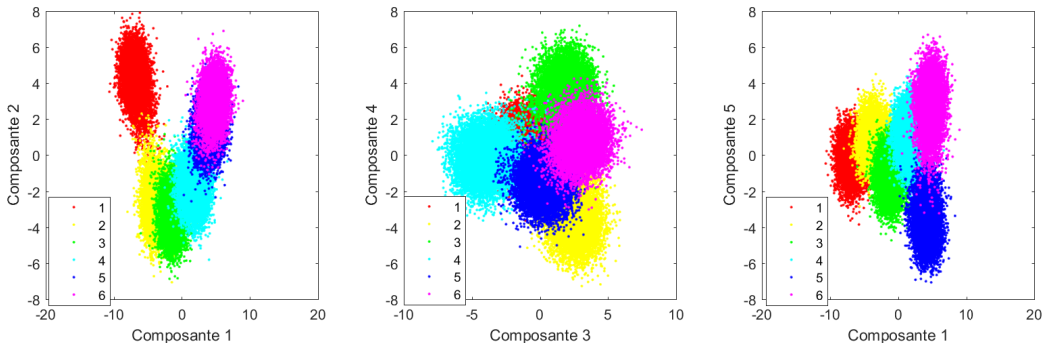



FIGURE 5.11 – Quelques exemples d'espaces de visualisation du jeu de données dans l'espace de dimension cinq obtenu par LDA.

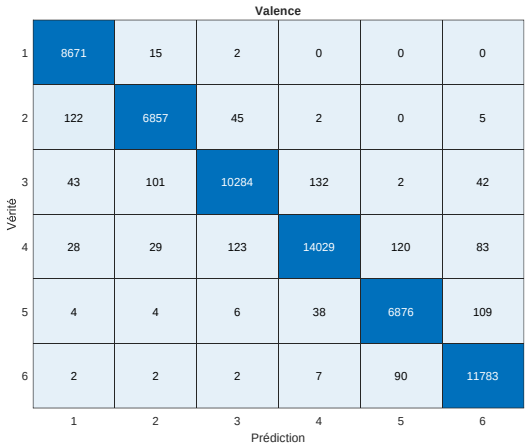



FIGURE 5.12 – Matrice de confusion - Classification de la valence (LDA + SVM)

de classification, lorsqu'elles se produisent, placent les exemples dans une classe adjacente à leur vraie classe (Fig. 5.12 et Tab. 5.5) .

TABLE 5.5 – Indicateurs de performance - Valence (LDA + SVM)

| | Classe 1 | Classe 2 | Classe 3 | Classe 4 | Classe 5 | Classe 6 |
|---|---|---|---|---|---|---|
| Précision | 0,98 | 0,98 | 0,98 | 0,99 | 0,97 | 0,98 |
| Rappel | 1,00 | 0,98 | 0,97 | 0,97 | 0,98 | 0,99 |
| Spécificité | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 0,99 |

Ce phénomène est mesurable avec le coefficient de corrélation de Spearman, qui est de 0.97. Ces mesures indiquent que les sorties du classifieur sont globalement bonnes, mais qu'en cas d'erreur, la distance entre prédiction et vérité est faible.

Les bons résultats pour le rappel et la spécificité indiquent que les classifieurs

arrivent à reconnaître à la fois l'appartenance d'un exemple à une classe et sa non-appartenance aux autres classes.

Les résultats pour les dimensions affectives *arousal* (Fig. 5.13 et Tab. 5.6) et *dominance* (Fig. 5.14 et Tab. 5.7) suivent le même profil.

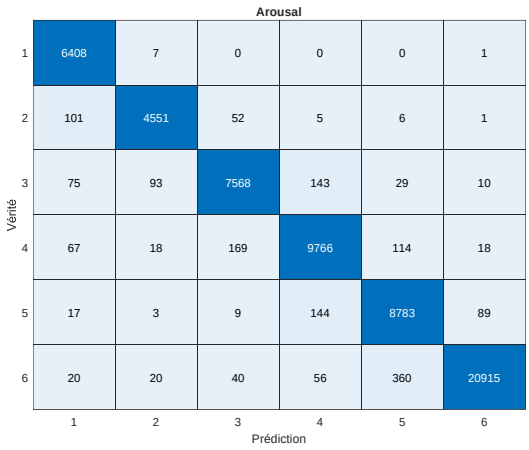



FIGURE 5.13 – Matrice de confusion - Classification de l'arousal (LDA + SVM)

TABLE 5.6 – Indicateurs de performance - Arousal (LDA + SVM)

| | Classe 1 | Classe 2 | Classe 3 | Classe 4 | Classe 5 | Classe 6 |
|---|---|---|---|---|---|---|
| Précision | 0,96 | 0,97 | 0,97 | 0,97 | 0,95 | 0,99 |
| Rappel | 1,00 | 0,97 | 0,96 | 0,96 | 0,97 | 0,98 |
| Spécificité | 0,99 | 0,99 | 0,99 | 0,99 | 0,99 | 1,00 |

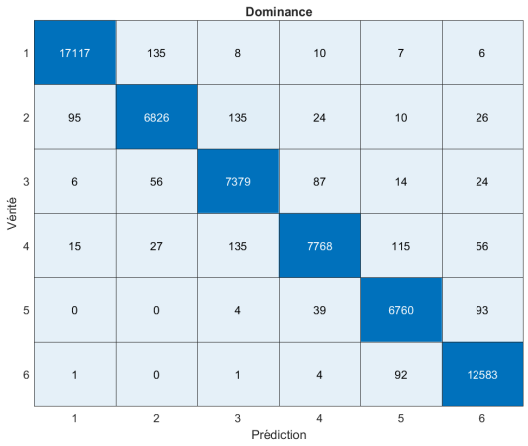



FIGURE 5.14 – Matrice de confusion - Classification de la dominance (LDA + SVM)

TABLE 5.7 – Indicateurs de performance - Dominance (LDA + SVM)

| | Classe 1 | Classe 2 | Classe 3 | Classe 4 | Classe 5 | Classe 6 |
|---|---|---|---|---|---|---|
| Précision | 0,99 | 0,97 | 0,96 | 0,98 | 0,97 | 0,98 |
| Rappel | 0,99 | 0,96 | 0,98 | 0,96 | 0,98 | 0,99 |
| Spécificité | 1,00 | 1,00 | 0,99 | 1,00 | 1,00 | 1,00 |

L'analyse discriminante linéaire a permis de projeter les données dans un espace où il est possible de discriminer les exemples d'après leur intensité émotionnelle. Ces résultats montrent qu'en dépit de la grande quantité d'information morphologique encodée dans les données LGBP, il demeure possible d'en tirer des informations émotionnelles indépendantes des sujets.

## 5.5 Discussion

Dans ce chapitre, nous avons utilisé un échantillon de la base de données annotée AVDLC afin de classifier les intensités émotionnelles de cinquante et un sujets différents. Les annotations ont été fiabilisées grâce à l'implémentation d'une méthode de calcul de la corrélation inter-annotateur de l'état de l'art. Les données ont également subi un traitement de préparation afin d'en réduire la redondance et de maximiser la variabilité émotionnelle.

Les classes d'intensités ont été établies sur les dimensions affectives *valence*, *arousal* et *dominance* en respectant la perception humaine d'une émotion intense d'une part, et neutre d'autre part. La classification multiclasse, réalisée à l'aide de SVM un-contre-un, a fourni des éléments d'interprétation intéressants sur les données. Elle a démontré l'intérêt de classifier les intensités plutôt que de prédire le signal affectif continu. On a retenu une méthode de validation croisée *leave-one-subject-out*, afin de construire un modèle capable à la fois d'utiliser toutes les données à disposition et de reconnaître les intensités émotionnelles de sujets inconnus.

Nous avons observé que le biais morphologique présent dans les données (de type histogrammes bas-niveau) empêchait la discrimination des intensités émotionnelles dans les données. Les changements de représentation linéaires (PCA) ou non-linéaires (*kernel*-PCA) n'ont pas permis de résoudre le problème, que ce soit avec une SVM linéaire ou à noyau gaussien. Ces résultats sont en accord avec les

observations de l'état de l'art, qui préconisent d'autres types de descripteurs pour la reconnaissance émotionnelle.

Toutefois, on a montré qu'il est possible de changer l'espace de représentation de ces données, *a priori* non adaptées, afin de les rendre séparables d'après un critère différent, celui des intensités émotionnelles, par exemple. Le changement de représentation des données a été réalisé avec une analyse discriminante linéaire, qui est une méthode de réduction dimensionnelle supervisée. En généralisation sur des sujets inconnus, les taux de succès sont supérieurs à 99%.

CHAPITRE 6

# CLASSIFICATION DES ETATS DÉPRESSIFS

Dans ce chapitre, on cherche à classifier l'état dépressif de sujets présents dans des vidéos. Les classes sont définies pour correspondre aux critères du Beck Depression Inventory II.

## 6.1 Spécification des données et des classes

On considère les vidéos mises à disposition par la base AVDLC, et non les descripteurs LGBP-TOP utilisés dans le chapitre 5. On conserve le partitionnement de la base détaillé dans la section 3.3.5, où une partie des données est consacrée à l'apprentissage, et l'autre à la généralisation. On rappelle également que les sujets présents dans une partition ne le sont pas dans l'autre, ce qui rend l'approche *subject independent*. Enfin, les vidéos des deux tâches (Freeform et Northwind) sont considérées.

### 6.1.1 Construction d'un nouveau descripteur visuel

Les points d'intérêt faciaux (voir paragraphe 3.3.2.1) sont des descripteurs géométriques simples et rapides à calculer. Ils permettent notamment le calcul de descripteurs dérivés, qu'ils soient spatiaux ou spatio-temporels. Malgré leur connotation rudimentaire, ils sont très utilisés dans l'état de l'art (voir sections 4.1 et 4.2).

Dans une image, un visage est détecté par la méthode de Viola et Jones [121] (voir 3.3.2.1). Ensuite, on met en œuvre le prédicteur de points faciaux de Kazemi et Sullivan [63], qui a été entrainé sur la base de données iBUG 300W [105]. Le prédicteur donne la position de 68 points d'intérêts faciaux sur chaque visage détecté.

Le calcul des points d'intérêt produit, pour chaque image, un vecteur de taille 136 où les abscisses et les ordonnées de 68 points sont concaténées. La numérotation et la localisation des points est montrée sur la figure 6.1.

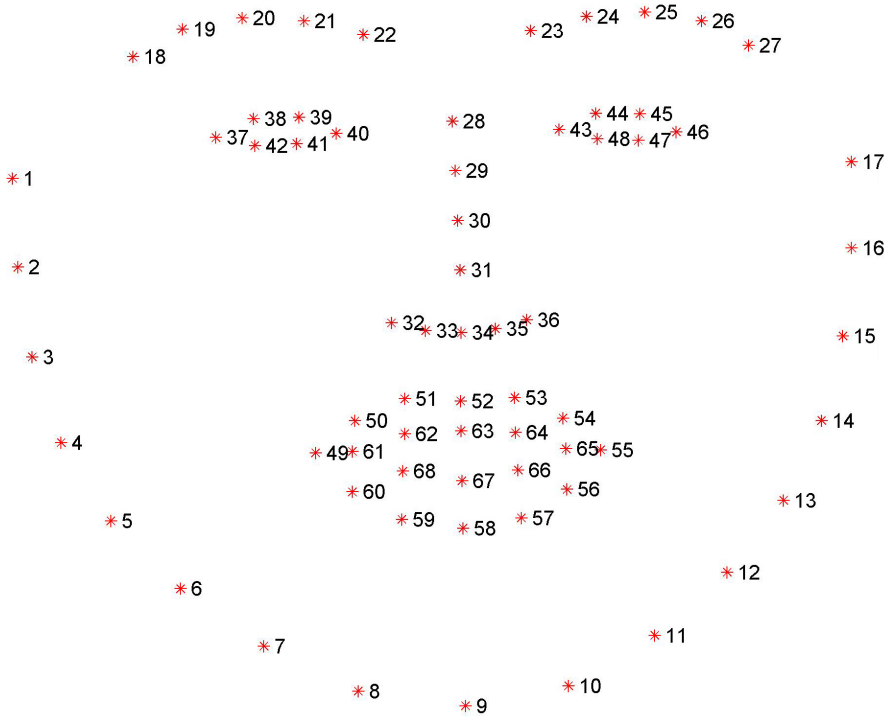



FIGURE 6.1 – Numérotation et localisaton des points d'intérêts faciaux.

**Alignement du visage** Lorsque l'on traite une vidéo image par image, ou lorsque l'on souhaite comparer plusieurs images entre elles, il est intéressant d'aligner les visages. Conceptuellement, l'alignement (en deux dimensions) consiste en trois étapes :

— Placer le visage à la verticale dans l'image. Une méthode classique est d'aligner les yeux sur une ligne horizontale. Pour ce faire, on réalise une rotation d'angle $\theta$ de l'ensemble du visage afin que la droite passant par les yeux soit parallèle à l'axe des abscisses.

— Redimensionner le visage. Cela permet à tous les visages traités d'être à la même échelle. Pour ce faire, on exprime une position recherchée pour les deux yeux en fonction de la hauteur $h$ et de la largeur $w$ de l'image. Le changement d'échelle d'un facteur $s$ est appliqué à l'ensemble des points.
— Centrer les visages. Cette étape consiste à choisir un point $t$ qui varie peu d'une image à l'autre, par exemple le point interocculaire ($P_{28}$ sur la Fig. 6.1). Ce point est translaté au centre de l'image, et l'ensemble des points subit la même translation.

Les trois opérations, bien qu'indépendantes, peuvent être exprimées sous la forme d'une matrice de transformation $M_{align}$ (Eq. 6.1)

$$M_{align} = \begin{bmatrix} s_x cos(\theta) & sin(\theta) & t_x \\ -sin(\theta) & s_y cos(\theta) & ty \end{bmatrix} \tag{6.1}$$

En fonction du problème à traiter, et des informations que l'on souhaite extraire des visages, il peut être intéressant de ne pas réaliser certaines des étapes ci-dessus. Par exemple, si l'on souhaite estimer la quantité de mouvement de la tête, la troisième étape (centrage des visages) ne doit pas être réalisée.

**Estimation de la quantité de mouvement** Le calcul de la quantité de mouvement des points faciaux vise à saisir la dynamique temporelle des expressions faciales, en estimant la variation des coordonnées des points. A partir des vecteurs de coordonnées de points faciaux, on calcule l'écart-type des composantes dans une fenêtre chevauchante, dont les paramètres de taille $T_{fen}$ et de $T_{pas}$ sont fixés. On note ce descripteur R-STD (pour *Rolling-STandard Deviation*). On note $x_{jt}$ la $j^{ème}$ composante d'un exemple à l'instant $t$. La composante correspondante $x'_{jt}$ du descripteur R-STD s'exprime comme suit (Eq. 6.2) :

$$\begin{aligned} x'_{jt} &= \sigma\left(x_{jt}, x_{j(t+1)}, \ldots, x_{j(t+T_{fen})}\right) \\ x'_{j(t+1)} &= \sigma\left(x_{j(t+T_{pas})}, x_{j(t+T_{pas}+1)}, \ldots, x_{j(t+T_{pas}+T_{fen})}\right) \end{aligned} \tag{6.2}$$

où $\sigma$ représente l'écart-type.

Le nombre d'exemples obtenus après la transformation est nécessairement inférieur au nombre d'exemples initial. Si on note $n$ le nombre d'images statiques, le nombre d'exemples $n_{R\text{-}STD}$ est donné par (Eq. 6.3) :

$$n_{\text{R-STD}} = \frac{n - \mathsf{T}_{\text{fen}}}{\mathsf{T}_{\text{pas}}} + 1 \tag{6.3}$$

### 6.1.2 Données utilisées et spécification des classes

La table 6.1 donne des détails sur les quantités de vidéos, de sujets et de scores de la base de données. En particulier, on note que plusieurs sujets ont le même score.

TABLE 6.1 – Quantité de vidéos, de sujets et de scores dans les données utilisées

| Tâche | Apprentissage | Généralisation |
|---|---|---|
| Freeform | 100 vidéos | 100 vidéos |
| + | 50 sujets | 43 sujets |
| Northwind | 31 scores | 29 scores |

#### 6.1.2.1 Données visuelles

Pour les expérimentations, on part des images de visages extraites des vidéos (Tab 6.2) afin de construire les données. Le descripteur utilisé change en fonction de l'étude menée :

— étude de l'impact des mouvements de la tête : descripteur spatial (points d'intérêt faciaux, voir Fig. 6.1) ;
— étude de l'impact de la dimension temporelle : descripteur spatio-temporel R-STD ;
— étude de l'impact du biais morphologique, avec le descripteur qui aura permis d'obtenir les meilleurs résultats (points d'intérêts faciaux ou R-STD).

TABLE 6.2 – Quantité d'images de visages issues des vidéos.

| Tâche | Apprentissage | Généralisation |
|---|---|---|
| Freeform | 91 111 images | 71 711 images |
| Northwind | 65 127 images | 63 206 images |

#### 6.1.2.2 Données auditives

Les données auditives sont celles fournies dans l'extrait de la base AVDLC à disposition (voir paragraphe 3.3.5.3 pour une présentation générale).

Contrairement aux données visuelles, les données auditives comportent plusieurs descripteurs, chacun capturant un aspect précis du signal sonore. Un exemple auditif est de dimension 2 268, et combine des informations sur l'énergie et le spectre du signal sonore, sur les caractéristiques de la voix et sur les durées parlées. Ces éléments constituent un ensemble standard de descripteurs pour le traitement du signal dans le contexte de l'Informatique Affective [119].

Les données auditives ont été calculées sur des fenêtres longues de vingt secondes, se chevauchant par pas de deux secondes. Un tel schéma de segmentation est propice à la capture d'informations sur les changements longs et lents, caractéristiques des sujets dépressifs. Dans les données auditives, certains sujets ou vidéos n'apparaissent plus, soit parce que la durée de la vidéo était inférieure à vingt secondes, soit parce que le sujet n'a pas parlé pendant la capture. Le nombre d'exemples auditifs de la base est proposé dans la table 6.3.

TABLE 6.3 – Quantité d'exemples auditifs issus des vidéos.

| Tâche | Apprentissage | Généralisation |
|---|---|---|
| Freeform | 1 036 ex. | 745 ex. |
| Northwind | 629 ex. | 622 ex. |

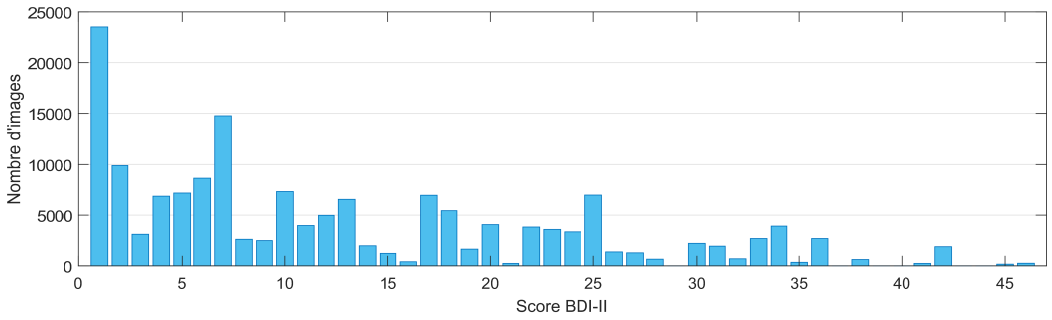
#### 6.1.2.3 Spécification des classes

Chaque vidéo est annotée par un unique score BDI-II, évaluant l'intensité de l'état dépressif du sujet enregistré. L'objectif de la classification sera de reconnaître le score BDI-II associé à un sujet. Dans les données, les scores vont de 0 à 46, alors que théoriquement, ils peuvent aller jusqu'à 63 (état dépressif maximal, voir Tab. 3.2). De plus, certains scores entre 0 et 46 ne sont pas représentés (Fig. 6.2).

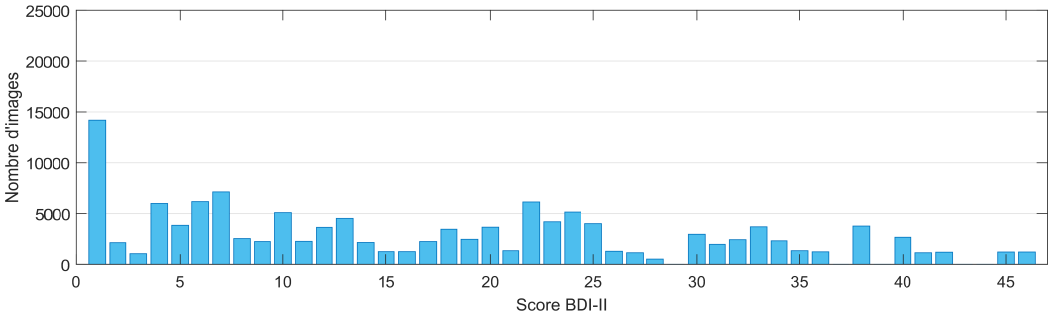
Dans la suite, on retient deux spécifications de classes :

— La première consiste à considérer autant de classes qu'il y a de scores BDI-II dans les données. Cette spécification a l'avantage d'être plus fine. En revanche, si un score en particulier n'est pas présent dans la partition d'apprentissage, il ne pourra pas être reconnu en phase de généralisation.

— La seconde consiste à considérer quatre classes, correspondant aux interprétations des scores BDI-II de la table 3.2, de la dépression minimale (classe 1) à la dépression sévère (classe 4).




(a)




(b)

FIGURE 6.2 – Distribution des scors BDI-II dans les images de la tâche Freeform (a) et dans les images de la tâche Northwind (b) sur l'ensemble du jeu de données.

### 6.1.3 Evaluation de la dépendance aux sujets des descripteurs

Afin d'estimer la séparation "naturellement" présente dans les données, on a calculé la visualisation t-SNE [76] des descripteurs visuels et auditifs. En particulier, chaque point de la visualisation est coloré suivant le score dépressif ou le sujet de l'exemple correspondant. Les couleurs ont été choisies afin de maximiser leur séparation visuelle [49]. Les mesures de divergence de Kullback-Leibler (Tab. 6.4) sont cohérentes avec les visualisations : la séparation "naturelle" des données ne favorise ni les classes dépressives, ni les sujets.

La visualisation de la séparabilité des données visuelles (Fig. 6.3) et des données auditives (Fig. 6.4) montre l'absence de groupes *a priori*. En classification, le modèle ne sera donc pas biaisé par la séparation naturelle des données, qui ne favorise ni la séparation des sujets, ni la séparation des scors dépressifs.

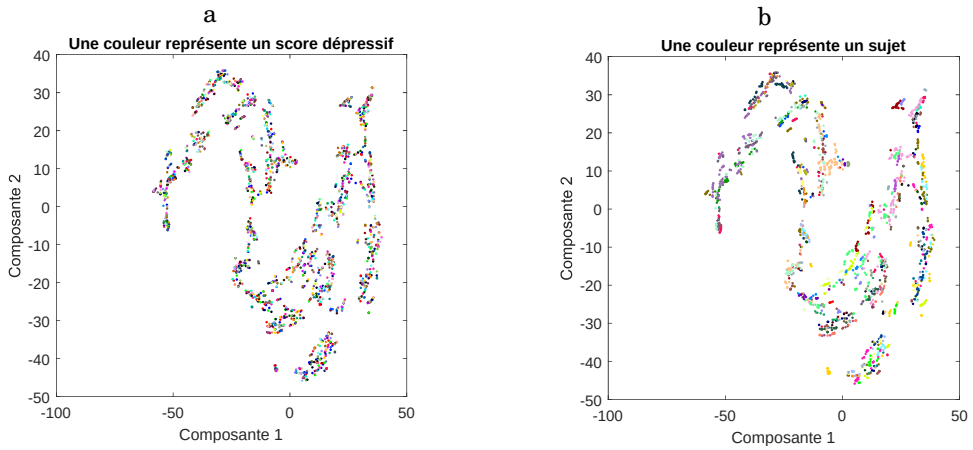



FIGURE 6.3 – Visualisation t-SNE des données visuelles pour la tâche Freeform. (a) Coloration par score et (b) par sujet.

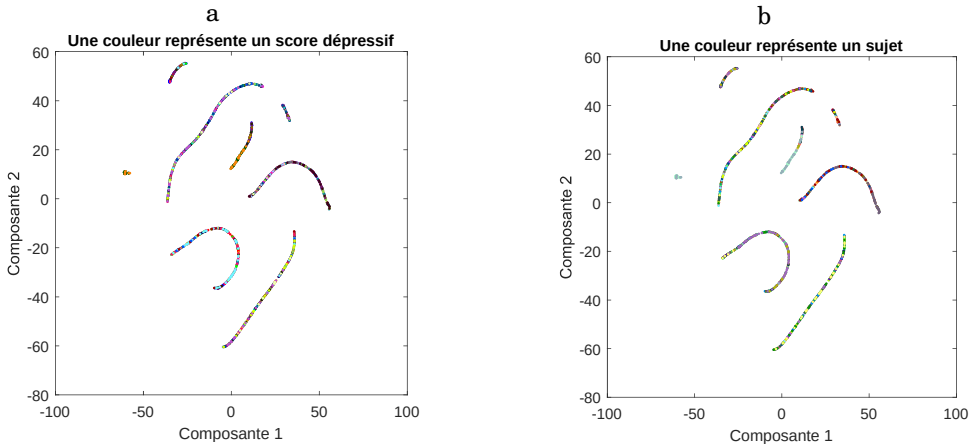



FIGURE 6.4 – Visualisation t-sne des données auditives pour la tâche Freeform. (a) Coloration par score et (b) par sujet.

### 6.1.4 Protocole de prédiction des classes

Le protocole de prédiction des classes décrit ici est valable aussi bien pour les expérimentations sur la modalité visuelle que sur la modalité auditive. Pour rappel, chaque vidéo met en jeu un unique sujet, et est annotée d'un unique score BDI-II. Il est possible que certains sujets soient présents dans plus d'une vidéo, et qu'un même sujet ait des scores différents dans les vidéos le concernant.

Dans la suite des expérimentations unimodales, le mot "exemple" désigne un vecteur de données visuelles ou auditives. Dans la mesure où plusieurs descripteurs

TABLE 6.4 – Divergences de Kullback-Leibler entre groupes de données.

| Modalité | Tâche | Sujets | Scores |
|---|---|---|---|
| Visuelle | Freeform | 0.31 | 0.79 |
| | Northwind | 0.22 | 0.31 |
| Auditive | Freeform | 0.14 | 0.15 |
| | Northwind | 0.09 | 0.11 |

sont utilisés, sa nature précise (points d'intérêts faciaux, R-STD, descripteurs auditifs ou descripteurs auditifs de dimension réduite) sera précisée pour chaque étude.

Une fois le modèle appris sur l'ensemble d'apprentissage, on lui présente les exemples correspondant à une vidéo afin de reconnaître son score BDI-II. Ces exemples ont par conséquent la même sortie désirée. Le classifieur calcule pour chaque exemple un score BDI-II, dit score (ou sortie) intermédiaire. La prédiction finale (ou score final) du classifieur pour la vidéo correspond à la moyenne des scores interménaires. Le processus est itéré pour chaque vidéo. Les erreurs en classification (RMSE et MAE, voir Eq. 2.28 et Eq. 2.29) sont calculées sur les scores finaux, et non sur les scores intermédiaires.

Ce protocole de calcul et les indicateurs d'erreurs sont conformes à ceux du challenge AVEC2014 [119] et permettent de comparer les résultats obtenus à ceux issus du challenge.

## 6.2 Classification de la modalité visuelle

On réalise une série d'expérimentations afin de reconnaître l'intensité dépressive des sujets à partir de la modalité visuelle des vidéos de la base de données. Le modèle de classifieur utilisé est le classifieur neuronal incrémental à base de prototypes (INN), qui a été présenté dans la section 2.4.

### 6.2.1 Premières expérimentations

On réalise une première série d'expériences afin d'étudier l'impact des mouvements de la tête et l'impact de la dimension temporelle sur les performances du modèle.

#### 6.2.1.1 Etude de l'impact des mouvements de la tête

On cherche à déterminer si l'encodage d'informations liées aux mouvements de la tête dans la description des données bénéficie à la reconnaissance des états dépressifs. Pour ces expériences, on considère deux descriptions spatiales des visages :

— Visages alignés de manière fixe : les visages sont centrés dans l'image. En conséquence, la position du point $P_{28}$ est la même dans tous les exemples.

— Visages alignés de manière mobile : les visages ne sont pas centrés dans l'image.

Dans chacune des deux descriptions, un exemple est représenté par un vecteur $\mathbf{x}$ de dimension 136.

**Conditions expérimentales et optimisation du classifieur** L'optimisation du classifieur correspond à la recherche du meilleur de chacun des trois hyperparamètres du modèle : seuil d'influence $s_{inf}$, seuil de confusion $s_{conf}$ et coefficient de rapprochement $\alpha$. On réalise une recherche en grille sur l'ensemble d'apprentissage. La recherche d'une configuration optimale permet à la fois de viser les meilleures performances en généralisation et d'étudier le comportement du classifieur sur un grand nombre d'essais. La recherche est menée séparément pour chaque tâche (Freeform d'une part, Northwind d'autre part) et pour chaque alignement (fixe d'une part, mobile d'autre part), soit quatre recherches en grille au total. Les domaines de recherche pour chaque hyperparamètre sont donnés dans la table 6.5.

TABLE 6.5 – Domaines de recherche des hyperparamètres $s_{inf}$, $s_{conf}$ et $\alpha$ pour la recherche en grille. Pour $s_{conf}$, les valeurs de 0.1 à 1 par pas de 0.1 ont également été testées.

| Hyperparamètre | Départ | Fin | Pas |
|---|---|---|---|
| $s_{inf}$ | 10 | 150 | 10 |
| $s_{conf}$ | 1 | 10 | 2 |
| $\alpha$ | 0.1 | 1 | 0.2 |

Un bon indicateur de la qualité d'apprentissage du modèle est le ratio $\rho_{proto}$ entre le nombre de prototypes et le nombre d'exemples d'apprentissage, un ratio élevé étant synonyme de surapprentissage. Les valeurs retenues pour les hyperparamètres (Tab. 6.6) font un compromis entre une erreur (RMSE et MAE) et un ratio de prototypes ($\rho_{proto}$) faibles. En particuler, les valeurs de $s_{conf} < 1$ produisent des

erreurs légèrement inférieures à celles retenues, mais avec des ratio de prototypes significativement plus élevés.

TABLE 6.6 – Résultats de la recherche en grille pour l'étude de l'impact des mouvements de la tête.

| Modalité | Tâche/Alignement | $s_{inf}$ | $s_{conf}$ [1] | $\alpha$ | RMSE | MAE | $\rho_{proto}$ |
|---|---|---|---|---|---|---|---|
| Image | Freeform Mob. | 80 | 1 | 0.1 | 8.81 | 6.97 | 0.18 |
| | Freeform Fix. | 60 | 1 | 0.1 | 9.45 | 8.14 | 0.14 |
| | Northwind Mob. | 60 | 1 | 0.1 | 9.42 | 8.11 | 0.14 |
| | Northwind Fix. | 140 | 3 | 0.1 | 9.46 | 7.72 | 0.14 |

La proximité à zéro des meilleurs seuils de confusion peut s'expliquer par la proximité des classes dans l'espace des données, qui nécessitent un seuil contraint pour les discriminer. On a analysé la corrélation entre les hyperparamètres ($s_{inf}$, $s_{conf}$, $\alpha$) d'une part et l'erreur et le ratio de prototypes (RMSE, MAE, $\rho_{proto}$) d'autre part. On n'observe pas de corrélation entre les hyperparamètres et l'erreur. En revanche, le ratio de prototypes et le coefficient de rapprochement sont fortement corrélés : plus le ratio est grand, plus le nombre de prototypes est élevé (Tab. 6.7).

TABLE 6.7 – Corrélations entre les hyperparamètres, l'erreur et le ratio de prototypes, calculées sur l'ensemble des essais réalisés pour la recherche en grille sur la modalité image

| | Alignement mobile | | | Alignement fixe | | |
|---|---|---|---|---|---|---|
| | RMSE | MAE | $\rho_{proto}$ | RMSE | MAE | $\rho_{proto}$ |
| $s_{inf}$ | 0.06 | 0.09 | -0.07 | -0.06 | -0.07 | -0.14 |
| $s_{conf}$ | 0.29 | 0.16 | 0.32 | 0.16 | -0.06 | **0.60** |
| $\alpha$ | 0.06 | 0.12 | **0.92** | 0.36 | 0.38 | **0.75** |

**Résultats** Les modèles appris avec les meilleurs hyperparamètres (Tab. 6.6) sont utilisés pour les évaluer sur l'ensemble de généralisation.

Les résultats (Tab. 6.8) montrent que l'alignement mobile favorise la reconnaissance des états dépressifs, puisque les erreurs obtenues pour cette version sont plus petites que celles obtenues pour l'alignement fixe. Aussi, on note que les erreurs obtenues pour la tâche Freeform sont plus petites que celles obtenues pour la tâche

1. Les valeurs de $s_{conf} < 0.1$ ont produit des modèles avec un fort ratio de prototypes. Ces modèles n'ont par conséquent pas été retenus.

TABLE 6.8 – Résultats pour les alignements mobiles et fixes en généralisation.

| | Freeform | | Northwind | |
|---|---|---|---|---|
| Alignement | Mobile | Fixe | Mobile | Fixe |
| RMSE | **9.24** | 10.56 | 9.44 | **9.42** |
| MAE | **6.97** | 9 | **7.61** | 8.08 |
| $\rho_{\text{proto}}$ | 0.12 | 0.11 | 0.11 | 0.10 |

Northwind. Cette dernière observation est cohérente avec les observations faites dans l'état de l'art (voir 4.2.1) : les sujets sont plus expressifs lorsqu'ils exécutent la tâche Freeform. Ces premiers résultats sont encourageants, car ils sont proches des erreurs de la *baseline* du challenge AVEC2014 où l'erreur RMSE/MAE était de 9.31/7.57 (pour la tâche Freeform ; la tâche Northwind ayant été écartée pour le calcul de la *baseline*)

#### 6.2.1.2 Etude de l'impact de l'aspect temporel

Le descripteur utilisé pour les expériences précédentes a permis d'obtenir des résultats intéressants. Toutefois, les coordonnées de points faciaux n'encodent pas d'information sur l'aspect temporel des micro-mouvements. Un consensus se dégage dans l'état de l'art sur le bénéfice de la prise en compte de l'aspect temporel pour l'étude des états dépressifs. Pour cette seconde série d'expériences, on utilise le descripteur spatio-temporel R-STD présenté dans le paragraphe 6.1.1, qui estime la quantité de mouvement dans un intervalle temporel.

**Conditions expérimentales et optimisation du classifieur** Le modèle de classifieur l'INN, et on recherche les meilleurs hyperparamètres $s_{\text{inf}}$, $s_{\text{conf}}$ et $\alpha$ pour les données R-STD, calculées sur les deux tâches Freeform et Northwind. Dans la suite, on considère uniquement l'alignement mobile, qui produit de meilleurs résultats que l'alignement fixe.

Le descripteur R-STD possède deux paramètres pour son calcul : la taille de la fenêtre temporelle $\tau_{fen}$ et le pas de chevauchement $\tau_{pas}$. Ces deux paramètres sont également considérées pour la recherche en grille. Au total, huit paramètres doivent être optimisés pour chacune des deux tâches, soit seize recherches en grille au total. Les domaines de recherche pour les hyperparamètres de l'INN sont inchangés (Tab. 6.5), et on fait varier $\tau_{fen}$ de 5 à 20 secondes par pas de 5 secondes, et $\tau_{pas}$ de 1 à 2 secondes par pas de 1 seconde.

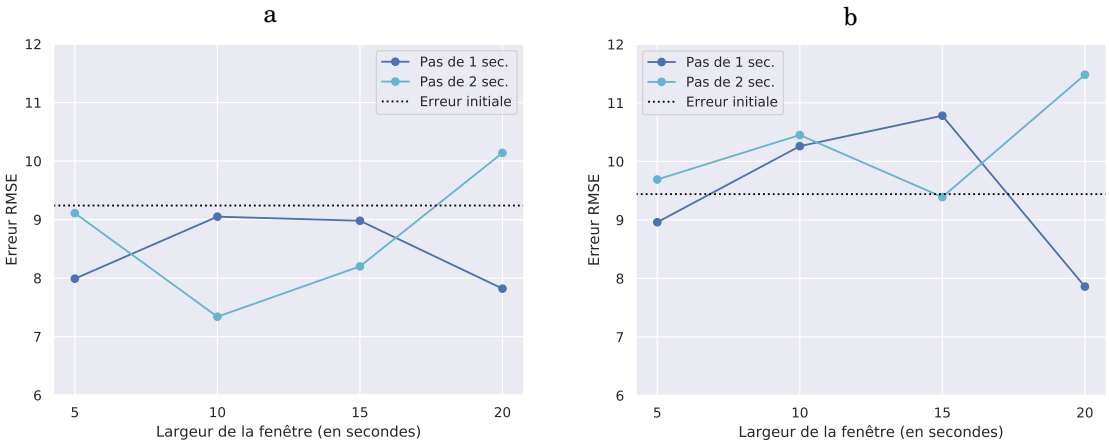



FIGURE 6.5 – Représentation graphique de l'erreur en classification obtenue pour chaque fenêtre et chaque pas lors de la recherche en grille pour la modalité image (descripteur R-STD) pour la tâche Freeform (a) et Northwind (b).

La figure 6.5 présente les erreurs en apprentissage pour les meilleurs paramètres trouvés dans chacune des seize recherches en grille. Pour la tâche Freeform, l'erreur obtenue avec le descripteur R-STD est presque toujours inférieure à l'erreur obtenue avec avec le descripteur spatial. Pour la tâche Northwind, l'erreur est plus rarement améliorée.

La table 6.9 propose une synthèse des valeurs des paramètres pour les meilleures erreurs obtenues.

On note que les ratio de prototypes $\rho_{\text{proto}}$ sont plus élevés avec le descripeur R-STD qu'avec le descripteur spatial. Cela peut s'expliquer par le nombre d'exemples moins nombreux pour le R-STD, et par la nature plus complexe de l'information qu'il encode. Les meilleurs hyperparamètres pour la tâche Freeform sont (80, 1, 0.1) avec un couple ($\top_{pas}$, $\top_{fen}$) de (1, 20). Pour la tâche Northwind, on retient les valeurs (20, 1, 0.30) et (1, 20).

**Résultats** On utilise les meilleurs paramètres dans chaque tâche afin d'évaluer le modèle sur l'ensemble de généralisation. Les performances obtenues avec le descripteur R-STD (Tab. 6.10) sont meilleures que celles obtenues avec le descripteur spatial (Tab. 6.8). Aussi, elles donnent de nouveau l'avantage à la tâche Freeform pour la reconnaissance des états dépressifs.

Dans le corpus, chaque sujet a réalisé les deux tâches. Valstar *et al.* [119] suggèrent de fournir un score par sujet plutôt qu'un score par vidéo. L'état dépressif

TABLE 6.9 – Grid search

| Tâche | $s_{inf}$ | $s_{conf}$ | $\alpha$ | RMSE | MAE | $\rho_{proto}$ | pas | fenetre |
|---|---|---|---|---|---|---|---|---|
| Freefrom | 140 | 1 | 0,10 | 7,99 | 6,03 | 0.71 | 1 | 5 |
| | 20 | 8 | 0,90 | 9,05 | 7,03 | 0.87 | 1 | 10 |
| | 80 | 1 | 0,10 | 8,98 | 7,23 | 0.44 | 1 | 15 |
| | **80** | **1** | **0,10** | **7,82** | **5,72** | **0.42** | **1** | **20** |
| | 30 | 6 | 0,90 | 9,11 | 7,38 | 0.95 | 2 | 5 |
| | 30 | 1 | 0,70 | 7,34 | 5,77 | 0.76 | 2 | 10 |
| | 20 | 2 | 0,90 | 8,20 | 6,19 | 0.76 | 2 | 15 |
| | 20 | 6 | 0,90 | 10,14 | 7,04 | 0.77 | 2 | 20 |
| Northwind | 30 | 1 | 0,10 | 8,96 | 7,50 | 0.72 | 1 | 5 |
| | 90 | 3 | 0,90 | 10,26 | 8,03 | 0.76 | 1 | 10 |
| | 100 | 1 | 0,10 | 10,78 | 8,33 | 0.50 | 1 | 15 |
| | **20** | **1** | **0,30** | **7,86** | **6,54** | **0.49** | **1** | **20** |
| | 130 | 1 | 0,70 | 9,69 | 7,69 | 0.84 | 2 | 5 |
| | 60 | 3 | 0,90 | 10,45 | 8,33 | 0.85 | 2 | 10 |
| | 40 | 1 | 0,30 | 9,39 | 7,66 | 0.65 | 2 | 15 |
| | 40 | 1 | 0,10 | 11,48 | 8,66 | 0.61 | 2 | 20 |

TABLE 6.10 – Erreurs en généralisation avec le descripteur R-STD.

| | Freeform | Northwind | F + N |
|---|---|---|---|
| RMSE | 7.31 | 8.80 | 7.00 |
| MAE | 5.43 | 7.30 | 6.37 |
| $\rho_{proto}$ | 0.24 | 0.31 | - |

de certains sujets est mieux reconnu dans une tâche que dans l'autre. La colonne "F + N" de la table 6.10 calcule l'erreur en considérant les meilleures prédictions finales pour chaque sujet, plutôt que la moyenne des prédictions sur les deux tâches. Ce résultat montre que, malgré le caractère adapté de la tâche Freeform, l'association de la tâche Northwind demeure intéressant puisqu'elle améliore les performances.

#### 6.2.1.3 Discussion

Les deux études menées ont permis de mettre en avant deux aspects qui s'avèrent incontournables pour la classfication des états dépressifs. D'un côté, la prise en compte des mouvements de la tête est nécessaire et cohérente avec la caractérisation même des troubles dépressifs. La prise en compte de ces macro-

mouvements, en complément des seuls micro-mouvements faciaux, a permis d'obtenir des erreurs RMSE/MAE de 9.24/6.97, contre 10.56/9.0 pour les micro-mouvements seuls (tâche Freeform). D'un autre côté, la dimension temporelle est également bénéfique à une meilleure discrimination des états dépressifs. Pour les mêmes extraits vidéos, on obtient des erreurs de 7.31/5.43 (tâche Freeform) en utilisant un descripteur capable d'estimer la quantité de mouvement des points faciaux d'intérêt (R-STD). Pour les deux expériences, la tâche Freeform convient mieux que la tâche Northwind. Cela est cohérent avec la nature des deux tâches, déjà évoquée dans l'état de l'art (voir 4.2.1).

### 6.2.2 Vers un modèle plus robuste

On cherche désormais à approfondir l'étude en contruisant un modèle sur l'ensemble des données à disposition, tout en conservant la propriété *subject independent* de l'approche.

On met en œuvre une validation croisée *leave-one-subject-out* sur l'ensemble des données spatio-temporelles (décrites à l'aide du R-STD), avec alignement mobile du visage. Pour cette expérience, on ne tient plus compte de l'appartenance d'une image à une vidéo, mais plutôt du sujet enregistré. Ainsi, la stratégie de validation consistera à apprendre sur tous les sujets sauf un, qui sera utilisé pour la généralisation.

Les hyperparamètres utilisés sont les mêmes que pour l'expérience précédente (voir Tab. 6.9). Les résultats (Tab. 6.11) sont globalement moins bons que les précédents. C'est explicable notamment par le changement de stratégie de validation, qui implique un plus grand nombre d'exemples d'apprentissage mais aussi une plus grande diversité de sujets.

TABLE 6.11 – Résultats de la validation croisée *Leave One Subject Out* avec les descripteurs R-STD sur l'ensemble des données alignées de manière mobile

| | Freeform | Northwind |
|---|---|---|
| RMSE | 10.31 | 10.75 |
| MAE | 9.81 | 10.36 |
| $\rho_{\text{proto}}$ | 0.26 | 0.34 |

La répartition des erreurs par sujet (Fig. 6.6) montre que l'état dépressif de certains sujets est parfaitement reconnu (erreur proche de zéro), alors que l'état

dépressif d'autres sujets, moins nombreux, n'est pas reconnu et engendre une erreur élevée. Le fait que le classifieur ne soit pas capable de reconnaître des scores BDI-II qu'il n'a pas appris est une piste d'explication. En effet, certains sujets sont les seuls représentants d'un score donné : pour ceux-là, la classification a très peu de chances de réussir.

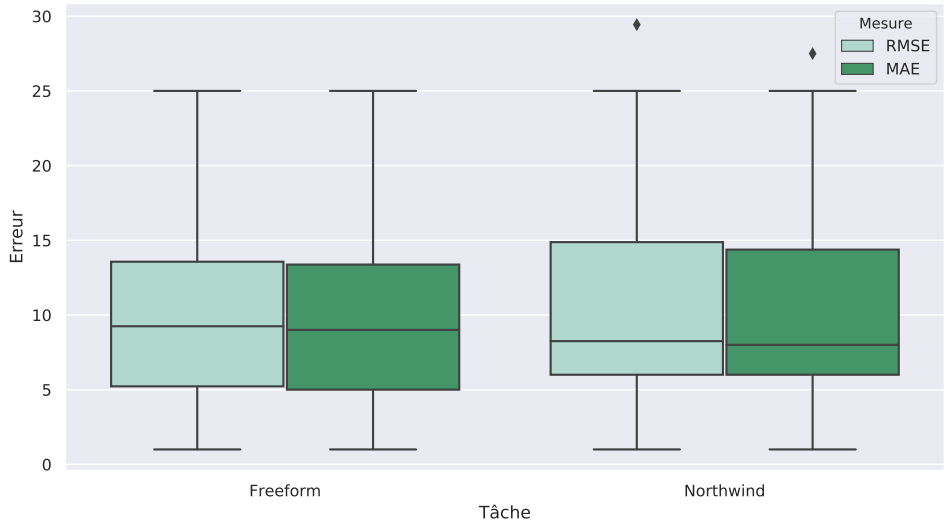



FIGURE 6.6 – Répartition des erreurs pour la validation croisée *Leave One Subject Out* avec le descripteur R-STD sur l'ensemble des données alignées de manière mobile

La figure 6.2 montrait que la plupart des scores représentés par un seul sujet étaient des scores dépressifs élevés. Dans l'état de l'art sur les représentations des états dépressifs (voir 3.2.1), on expliquait l'interprétation des scores dépressifs en quatre groupes (Tab. 3.2) : dépression minimale, pour un score allant de 0 à 13 ; modérée, pour un score allant de 14 à 19 ; moyenne, pour un score allant de 20 à 28 ; et sévère, pour un score allant de 29 à 63.

Afin de déterminer si certains intervalles de scores sont spécifiquement mal reconnus, on calcule l'erreur par interprétation de l'état, plutôt que par score dépressif (Tab. 6.12). Cette analyse montre que les intensités minimales et modérées sont les mieux détectées, avec une erreur entre 8 et 8.31 (RMSE). Les intensités maximales, qui sont également celles les plus souvent représentées par un seul sujet, ne sont pas bien détectées, avec une erreur de 20.8.

TABLE 6.12 – Erreur moyenne par sévérité dépressive pour la tâche Freeform, en validation croisée LOSO

| Intensité dépressive | Erreur moyenne (RMSE) |
|---|---|
| Minimale | 8.31 |
| Modérée | 8.00 |
| Moyenne | 11.37 |
| Sévère | 20.80 |

### 6.2.3 Etude comparative

Afin de comparer les performances du classifieur incrémental, on entraîne quatre autres modèles sur les données R-STD (alignement mobile). Les modèles (SVM à noyau linéaire, fôrets aléatoires - 200 arbres, k plus proches voisins - $k = 1$, et perceptron multicouche) sont appris sur l'ensemble d'apprentissage et généralisés sur l'ensemble de test. Les classifieurs sont comparés en termes d'erreur en généralisation (Fig. 6.7) et de vitesse d'exécution moyenne (Tab. 6.13).

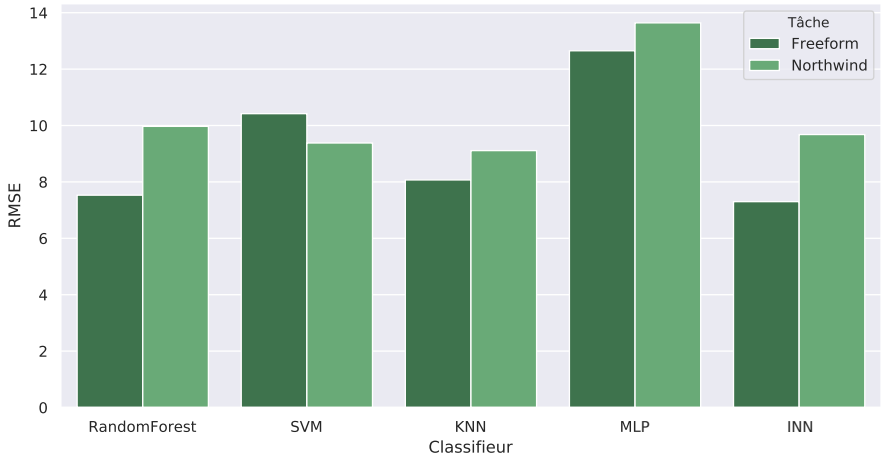



FIGURE 6.7 – Comparaison des performances de plusieurs classifieurs avec notre méthode (INN)

Des cinq classifieurs comparés, l'INN est celui qui obtient l'erreur la plus petite. Les k-NN réalisent une erreur proche (néanmoins supérieure) de celle de l'INN, ce qui met en valeur l'intérêt des seuils d'influence et de confusion pour augmenter le pouvoir discriminant du classifieur.

TABLE 6.13 – Temps moyen d'exécution et erreurs RMSE des différentes méthodes comparées

| Méthode | Temps (secondes) | RMSE (Freeform) | RMSE (Northwind) |
|---|---|---|---|
| Random Forest | 11.22 | 7.51 | 9.98 |
| SVM | 1.03 | 10.3 | 9.37 |
| KNN | **0.07** | 8.12 | 9.26 |
| MLP | 1.17 | 12.67 | 13.64 |
| INN | 0.32 | **7.39** | 9.63 |

Les temps d'exécution moyens soulignent aussi la rapidité de l'INN, qui n'est toutefois pas le plus rapide. Cependant, ce modèle de classifieur demeure celui qui présente le meilleur compromis entre qualité et rapidité.

Les performances de la méthode proposée sont également comparées avec celles des méthodes de l'état de l'art. La table 6.14 synthétise les performances obtenues sur le même jeu de données avec des méthodes différentes. On compare d'une part les travaux exploitant uniquement la tâche Freeform (F) et d'autre part ceux exploitant les deux tâches (F + N).

TABLE 6.14 – Résultats comparés de la *baseline* "Dépression", des travaux de challengers et des nôtres sur les données du challenge AVEC2014. Toutes les performances présentées ont été obtenues en généralisatin sur la partition de développement.

| | RMSE | MAE | Remarque |
|---|---|---|---|
| Valstar *et al.* [119] (F) | 9.31 | 7.57 | SVR et LBGP-TOP (baseline) |
| Senoussaoui *et al.* [109] (F) | 8.52 | 6.95 | SVR et LGBP-TOP |
| Notre méthode (F) | **7.31** | **5.43** | INN et Rolling-STD |
| Jain *et al.* [54] (F + N) | 8.7 | 6.97 | ELM et Moore-Penrose inverse |
| Kaya *et al.* [61] (F + N) | 10.27 | 8.21 | SVM et Fisher Vector Encoding |
| Jan *et al.* [55] (F + N) | 9.49 | 7.36 | Partial Least Squares et MHH |
| Senoussaoui *et al.* [109] (F + N) | 8.88 | 7.24 | SVR et LGBP-TOP |
| Notre méthode (F + N) | **7.00** | **6.37** | INN et Rolling-STD |

Notre méthode, i.e. le descripteur spatio-temporel associé au classifieur incrémental, obtient des erreurs inférieures à celles de l'état de l'art. Il reste important de préciser que ces performances sont calculées sur le jeu de développement (que nous considérons comme ensemble de généralisation), et non sur le jeu de test officiel du challenge, auquel nous n'avons pas accès.

## 6.3 Classification de la modalité auditive

Les données auditives sont construites sur des segments temporels longs, destinés à capturer la dynamique des changements lents et à long terme, caractéristiques des états dépressifs. Les exemples correspondent à des segments audio de vingt secondes se chevauchant par pas de deux secondes. Chacune des 2 268 dimensions correspond à un traitement précis (voir la description de la base de données, paragraphe 3.3.5.3). Ces traitements sont pour la plupart de nature statistique et encodent des éléments comme le maximum, le minimum ou la moyenne de mesures également encodées dans la description. Par conséquent, on peut s'attendre à une forte corrélation des variables dans les données auditives.

On cherche à entraîner un INN afin de reconnaître automatiquement l'état dépressif à partir des données auditives. En particulier, deux problématiques sont traitées :

— La problématique de la dépendance des données aux sujets plutôt qu'aux états dépressifs.
— La problématique de la forte dimensionnalité, coïncidant avec la présence de variables corrélées dans les données.

Des éléments de réponse ont été apportés concernant la première problématique dans la section 6.1.3 où on a mesuré la divergence entre les données correspondant à des classes (i.e. d'états dépressifs) différentes, d'une part, et à des sujets différents, d'autre part. Cette première analyse a montré qu'il n'y avait pas de séparation *a priori* favorable aux classes ou aux sujets dans les données. Dans nos expérimentations, on cherche à construire un modèle sur l'ensemble d'apprentissage (validation *hold-out*), puis sur la totalité des données utilisables (validation *leave-one-subject-out* - LOSO).

La seconde problématique motive l'utilisation d'une méthode de réduction dimensionnelle non supervisée telle que la PCA. En effet, on souhaite conserver la séparation qui existe "naturellement" dans les données : si cette séparation était favorable aux sujets, il aurait été nécessaire d'utiliser une réduction dimensionnelle supervisée comme dans le chapitre sur la classification des états émotionnels (voir 5.4).

### 6.3.1 Premières expérimentations

Dans un premier temps, on cherche à reconnaître l'état dépressif dans les données auditives sans réduire leur dimension. Pour ce faire, on recherche les meilleurs hyperparamètres pour l'INN via une recherche en grille.

On fait varier $s_{inf}$ de 50 à 140 par pas de 10, $s_{conf}$ de 1 à 9 par pas de 1 et $\alpha$ de 0.1 à 1 par pas de 0.1. Les meilleures combinaisons (Tab 6.15) sont très différentes pour les tâches Freeform et Northwind, ce qui souligne de nouveau la différence entre les deux tâches.

TABLE 6.15 – Meilleurs hyperparamètres pour la modalité auditive

| Tâche | $s_{inf}$ | $s_{conf}$ | $\alpha$ |
|---|---|---|---|
| Freeform | 140 | 9 | 0.3 |
| Northwind | 50 | 7 | 0.5 |

En validation *hold-out* (Tab. 6.16), les erreurs (RMSE et MAE) varient beaucoup d'une tâche à l'autre, à l'avantage de la tâche Freeform, plus adaptée à la reconnaîssance des états dépressifs dans les données auditives que la tâche Northwind (RMSE de 7.38 pour Freeform, contre 9.4 pour Northwind, avec des paramètres optimaux). Pour les deux tâches, le nombre de prototypes est toutefois élevé (entre 30% et 38%).

TABLE 6.16 – Erreurs en apprentissage et en généralisation (*hold-out*) pour la reconnaissance des états dépressifs à partir des données auditives

| | Apprentissage | | | Généralisation | |
|---|---|---|---|---|---|
| Tâche | RMSE | MAE | $\rho_{proto}$ | RMSE | MAE |
| Freeform | 0.23 | 0.05 | 0.3 | 7.38 | 5.74 |
| Northwind | 1.52 | 0.56 | 0.30 | 9.4 | 7.6 |

En validation croisée LOSO (Tab. 6.17), les erreurs sont calculées pour chaque sujet, et les erreurs finales sont les moyennes des erreurs de chaque sujet. On note que ce mode de validation réussit un peu moins bien que le précédent, ce qui est sans doute dû à l'évaluation de certains scores dépressifs par des modèles n'ayant pas rencontré ces scores en apprentissage.

Le calcul des erreurs par sévérité dépressive montre qu'elles sont plus faibles pour les sévérités dépressives minimales (Fig. 6.8). La répartition des erreurs

TABLE 6.17 – Erreurs en apprentissage et en généralisation (*leave-one-subject-out*) pour la reconnaissance des états dépressifs à partir des données auditives

| | Apprentissage | | | Généralisation | |
|---|---|---|---|---|---|
| Tâche | RMSE | MAE | $\rho_{\text{proto}}$ | RMSE | MAE |
| Freeform | 0.31 | 0.69 | 0.32 | 9.91 | 9.6 |
| Northwind | 1.99 | 0.91 | 0.39 | 10.97 | 10.46 |

montre de plus grands écarts pour les états dépressifs les plus sévères, ce qui traduit une difficulité à reconnaître ces situations.

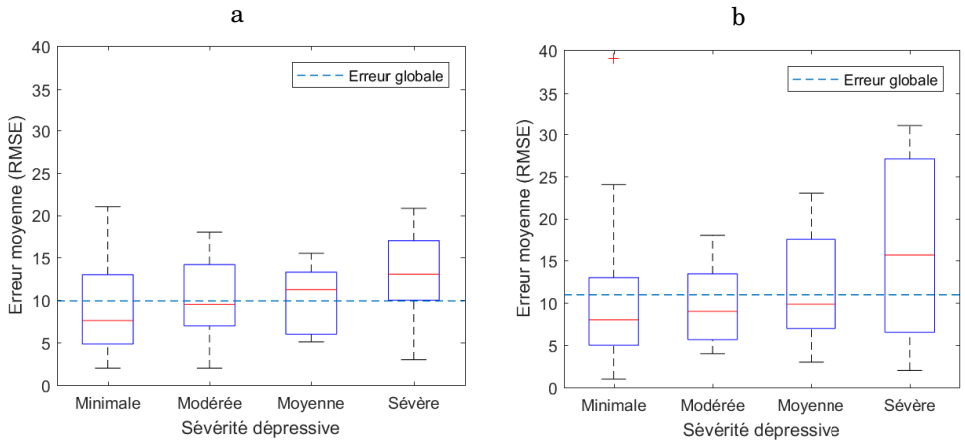



FIGURE 6.8 – Répartition de l'erreur en validation croisée LOSO - Données auditives (a) Freeform et (b) Northwind.

### 6.3.2 Réduction dimensionnelle

Les données auditives ont une dimension 17 fois supérieure à celle des données visuelles. La présence de variables corrélées dans les données auditives motive l'application d'une analyse en composantes principales pour en réduire la dimension, d'une part, et pour décorréler les variables, d'autre part. On souhaite réduire au minimum le rapport entre la dimension des données de chaque modalité. Pour ce faire, on réduit la dimension des données auditives en conservant entre 100 et 200 composantes principales par pas de 25. Pour chaque essai, on recherche les meilleurs hyperparamètres, puis on généralise le meilleur modèle sur les données d'apprentissage et de test. Les domaines de recherche des hyperparamètres sont

l'union des domaines précédents (voir 6.3.1) et des domaines suivants : $s_{inf}$ de 10 à 40 pas pas de 10, $s_{inf}$ de 0.1 à 1 par pas de 0.1 et $\alpha$ de 0.1 à 1 par pas de 0.1.

TABLE 6.18 – Erreurs en apprentissage et en généralisation (*hold-out*) pour la reconnaissance des états dépressifs à partir des données auditives (avec réduction dimensionnelle)

| Tâche | Dim. | Hyperparam. | | | Apprentissage | | | Généralisation | |
|---|---|---|---|---|---|---|---|---|---|
| | | $s_{inf}$ | $s_{conf}$ | $\alpha$ | RMSE | MAE | $\rho_{proto}$ | RMSE | MAE |
| Freeform | 100 | 100 | 2 | 0.8 | 3.57 | 1.76 | 0.36 | 8.78 | 6.53 |
| | 125 | 40 | 0.3 | 0.2 | 2.75 | 0.97 | 0.2 | **9.05** | 6.76 |
| | 150 | 40 | 0.9 | 0.1 | 0.51 | 0.16 | 0.16 | 9.12 | 6.5 |
| | 175 | 40 | 0.5 | 0.1 | 1.09 | 0.34 | 0.18 | 9.88 | 7.05 |
| | 200 | 40 | 0.7 | 0.3 | 1.61 | 0.74 | 0.26 | 9.78 | 7.35 |
| Northwind | 100 | 50 | 7 | 0.1 | 0.53 | 0.12 | 0.19 | 9.63 | 7.7 |
| | 125 | 50 | 4 | 0.1 | 0.57 | 0.16 | 0.19 | 9.6 | 7.62 |
| | 150 | 60 | 5 | 0.1 | 0.63 | 0.2 | 0.19 | **9.16** | 7.34 |
| | 175 | 60 | 1 | 0.1 | 1.5 | 0.58 | 0.19 | 9.73 | 7.88 |
| | 200 | 30 | 0.1 | 0.2 | 0.0 | 0.0 | 0.36 | 9.71 | 7.32 |

La table 6.18 synthétise les meilleures performances et configurations. Les valeurs en gras sont les meilleures erreurs RMSE sur l'ensemble de test. Ces valeurs correspondent à des dimensions différentes. Toutefois, c'est la dimension 150 qui obtient l'erreur et le ratio de prototypes moyens les plus petits. Afin d'harmoniser la démarche, et afin de ne pas avoir une dimensionnalité différente pour chaque tâche, on réalise la suite des expériences avec les données auditives de dimension réduite à 150. La réduction d'un facteur 15 a dégradé les performances entre un et deux points pour chaque tâche, mais a permis de réduire le nombre de prototypes d'environ un tiers.

La figure 6.9 confirme notre hypothèse concernant la forte corrélation des variables dans les données auditives : on observe qu'il suffit de 45 composantes (sur 2 268) pour expliquer 100% de la variance cumulée, quelle que soit la tâche.

La reconnaissance des états dépressifs au moyen de l'INN fonctionne mieux avec les données auditives de la tâche Freeform qu'avec les données visuelles ou avec les données de l'autre tâche. Ce résultat est cohérent avec les observations faites dans l'état de l'art, et s'explique notamment par une meilleure qualité des données auditives. Les erreurs (RMSE/MAE de 7.38/5.74) obtenues avec l'INN sont significativement meilleures que celles de la *baseline* (11.52/8.93 sur l'ensemble de

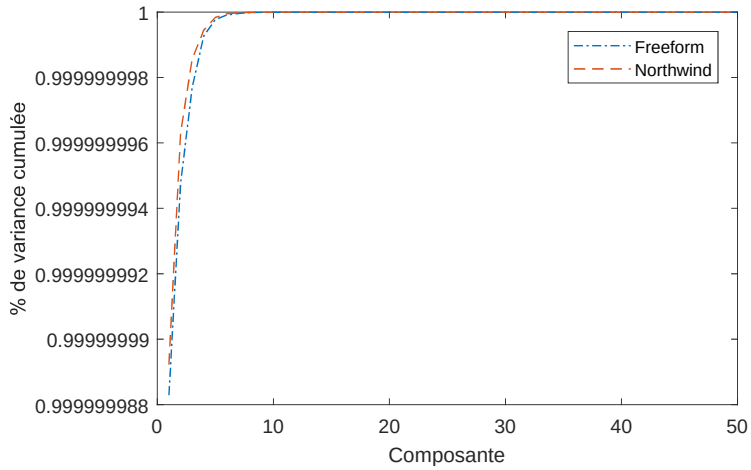



FIGURE 6.9 – Cumul du pourcentage de variance expliquée par chaque composante - Données auditives pour Freeform et Northwind.

développement). La réduction de dimension des données auditives n'a pas permis d'augmenter l'efficacité du classifieur en termes d'erreur, mais a permis en revanche d'augmenter la qualité de l'apprentissage : le nombre de prototypes est réduit d'un tiers environ, par rapport au nombre de prototypes nécessaires à l'apprentissage des données de pleine dimension.

## 6.4 Un outil d'aide à l'évaluation des états dépressifs en temps réel

La méthode proposée permet l'évaluation des états dépressifs en temps réel, dans la mesure où :

— elle est interprétable. Le descripteur R-STD encode la quantité de micro-mouvements (points d'intérêts faciaux) et de macro-mouvements (mouvements de la tête), qui sont toutes les deux connues pour être spécifiques chez les sujets dépressifs ;

— elle est rapide. Le calcul des descripteurs et la classification peuvent s'exécuter très rapidement sur un ordinateur du marché, grâce à leur calcul itératif et incrémental.

— elle est efficace. Les performances calculées sur le jeu de données à disposition sont encourageantes et prouvent la pertinence de l'approche.

### 6.4.1 Cas d'utilisation

Les risques psychosociaux évoqués dans l'introduction sont notamment caractérisés par la présence de troubles dépressifs. La détection et le suivi de ces troubles peuvent contribuer à améliorer l'environnement de travail des individus. On se place, pour ce cas d'utilisation, dans le contexte d'une entreprise où les collaborateurs utilisent régulièrement un ordinateur équipé d'une webcam pour leurs tâches courantes. Le poste de travail peut alors être équipé de l'outil dont il est question ici, afin qu'il soit utilisé à intervalles réguliers pour l'évaluation de l'état dépressif des volontaires.

Sur le plan psychologique, la démarche doit effectivement être volontaire et consentie : les résultats du système peuvent être faussés par un comportement exagéré, gêné ou simulé. Sur le plan éthique, l'accord de l'utilisateur est aussi indispensable. Enfin, sur le plan de la santé, l'usage d'un tel outil peut s'inscrire comme une étape vers l'amélioration de l'état dépressif et du bien-être au travail, s'il est exploité correctement.

Dans le concept proposé, la collaboration d'un professionnel de la santé est requise, car il est le seul à pouvoir faire un suivi des sujets sur le plan médical. L'outil, lui, se pose comme facilitateur et aide de premier recours pour l'usager. A intervalles réguliers, l'utilisateur peut interagir avec le système. A chaque session, le système évalue le score dépressif du sujet et le lui fait connaître. L'évaluation tient non seulement compte de la session en cours, mais aussi des sessions des deux semaines passées, conformément à la définition médicale de la dépression. Idéalement, l'intervention de l'expert est requise à deux niveaux :

— Le premier niveau concerne directement l'aide aux utilisateurs. L'outil est capable de lever des alertes auprès de l'expert dans la situation où le score dépressif d'un usager est inquiétant. En réponse, l'expert peut proposer à l'usager un entretien, par exemple.

— Le second niveau concerne l'outil d'évaluation et son perfectionnement. L'outil, de par la nature incrémentale de l'algorithme utilisé, est capable de se spécialiser sur un sujet en particulier ou d'intégrer de nouveaux éléments d'apprentissage. L'action de l'expert consiste ici à avoir une action "corrective" sur le système en indiquant, par exemple, si une évaluation est bonne ou non. Une auto-expertise est également possible en questionnant l'usager.

La figure 6.10 est une vue d'ensemble du concept décrit ci-dessus.

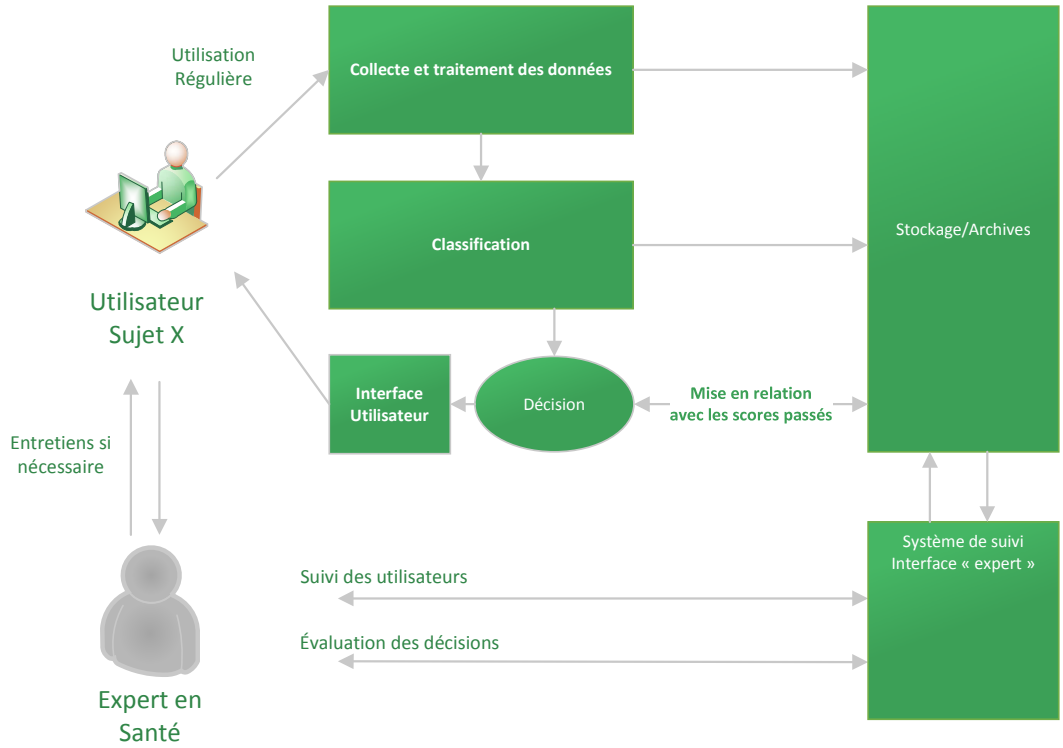



FIGURE 6.10 – Schéma d'ensemble du concept tel que proposé pour une utilisation dans un contexte professionnel

### 6.4.2 Versions développées

Nous avons proposé une premier prototype [15] à la Conférence Nationale en Intelligence Artificielle. Il exploitait les descripteurs spatiaux. Chaque seconde, 10 images (sur 30) étaient décrites et classifiées, et le score proposé à un instant $t$ était calculé par un vote majoritaire sur les 20 dernières images. Un indicateur affichait à l'utilisateur si son visage était détecté, et un autre affichait le score dépressif évalué.

La seconde version exploite les descripteurs R-STD sur des visages alignés de manière mobile. Le descripteur est calculé par fenêtres de $\mathsf{T}_{fen} = 20s$ avec un chevauchement de $\mathsf{T}_{pas} = 1s$. Ces paramètres sont justifiés par les expériences précédentes (voir 6.2.1.2) qui ont mis en avant l'importance de la dimension temporelle et des mouvements de la tête. Le score affiché à un intant $t$ tient compte de la

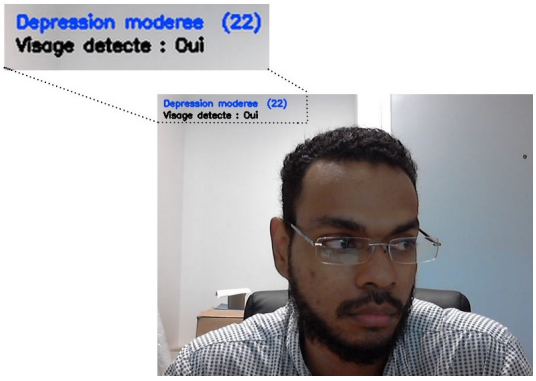



FIGURE 6.11 – Interface utilisateur de la première version [15]

moyenne des scores obtenus dans les fenêtres temporelles précédentes. L'affichage est différent, reposant sur d'autres motivations que dans la première version.

**Workflow** La faisabilité de l'évaluation en temps réel repose sur la rapidité des traitements. La capture est effectuée dans un thread dédié, ce qui permet une meilleure gestion de la fréquence d'acquisition. Le calcul des R-STD peut être effectué à une fréquence plus faible que la fréquence d'acquisition. La classification, elle, est réalisée toutes les secondes (à partir de la vingtième seconde).

L'affichage de l'utilisateur filmé a été supprimé, car nous avons constaté que le "retour image" captivait les sujets, au détriment de l'objet de l'expérimentation. En lieu et place, on reproduit le contexte de la tâche Freeform : une question aléatoire d'ordre général est affichée. Le schéma détaillé du workflow est proposé dans la figure 6.12.

**Démonstration** Le système a été testé par plusieurs utilisateurs volontaires. Dans la configuration retenue, la durée de la session n'est ni prédéfinie, ni bornée : elle dépend entièrement du temps mis par l'utilisateur pour répondre aux questions. L'utilisateur démarre et progresse à l'aide de la souris dans le fil des questions. Le résultat de la session est affiché uniquement à la fin, de manière à ce que l'utilisateur ne cherche pas à altérer son comportement pour faire varier un indicateur intermédiaire. Toutefois, tous les scores intermédiaires sont retenus pour le calcul du score final de la session.

Les données stockées sont un idenfiant pour l'utilisateur et les scores obtenus pour ses sessions. Les vidéos ou images ne sont pas stockées (comme l'impose le

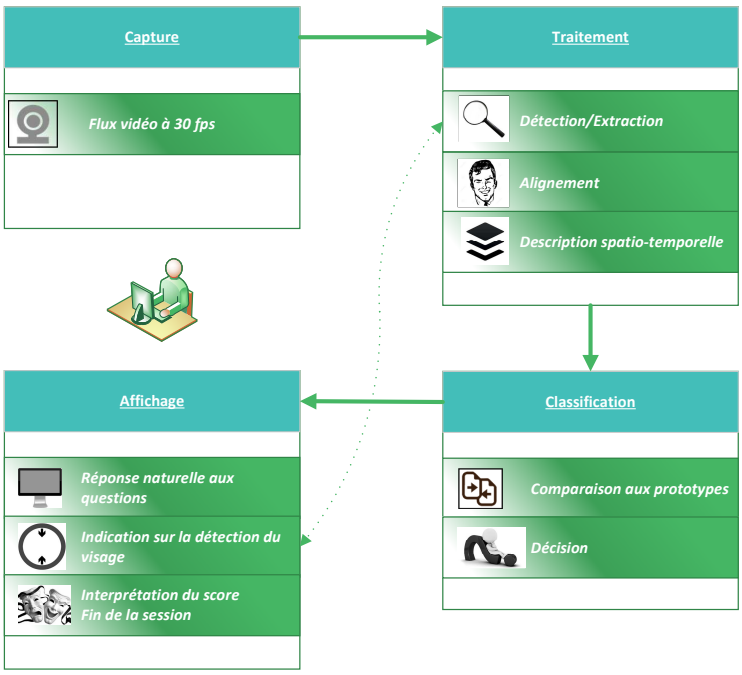



FIGURE 6.12 – Schéma détaillé du fonctionnement de l'outil

RGPD[2]) même pendant la session : elles servent uniquement au calcul à la volée des descripteurs R-STD. La figure 6.13 est une vue schématique de l'interface utilisateur de l'outil. L'interface expert n'a pas encore été réalisée.

## 6.5 Conclusion

Dans ce chapitre consacré à la reconnaissance des états dépressifs à partir de données unimodales, on a étudié l'impact de plusieurs paramètres sémantiques sur l'efficacité du modèle informatique mis en œvre.

Pour la modalité visuelle, l'usage de descripteurs encodant à la fois les micro-mouvements et les macro-mouvements du visage est pertinent, car ils ont un pouvoir discriminant supérieur aux seuls micro-mouvements. La prise en compte de l'aspect temporel bénéficie également aux performances des classifieurs, tout en réduisant le nombre d'exemples de la base de données. L'estimation des performances sur l'ensemble des données, via une validation croisée *leave-one-subject-out*, révèle que les états dépressifs peu intenses (score BDI-II entre 0 et 13) sont plus

2. Règlement Général sur la Protection des Données. Voir les textes officiels : https ://www.cnil.fr/fr/textes-officiels-europeens-protection-donnees

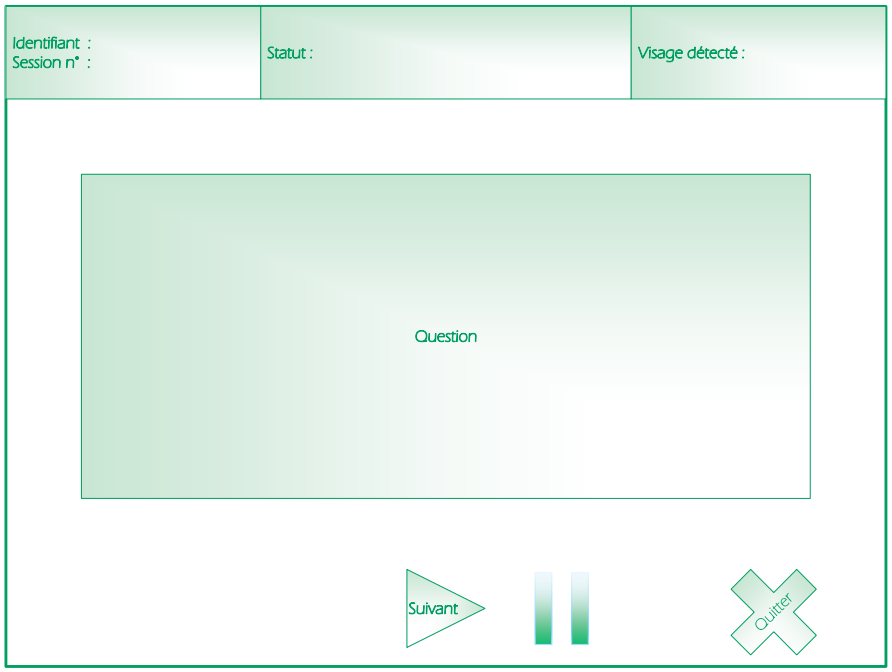



FIGURE 6.13 – Interface utilisateur de l'outil pour les sessions d'évaluation

facilement reconnaissables que les états dépressifs plus sévères (scores entre 14 et 63). Un tel modèle est utilisable pour la prédiction des états dépressifs en temps réel, bien que les scores prédits nécessitent l'intervention d'un professionnel de la santé pour être validés.

Pour la modalité auditive, dont les données encodent l'aspect temporel de manière naturelle, la reconnaissance des états dépressifs est plus performante. Cependant, l'acquisition et le traitement du signal auditif, non expérimenté dans ce travail, est plus complexe.

Dans les deux cas, l'INN a su montrer son efficacité pour la reconnaissance des états dépressifs, tant au niveau des performances (avec des erreurs proches ou en deçà de ceux de l'état de l'art) qu'au niveau de la vitesse de classification des données.

Dans le chapitre suivant, on s'intéresse à la fusion des modalités pour la reconnaissance des états dépressifs, avec comme défi l'amélioration des performances par rapport aux approches unimodales.

CHAPITRE 7

# VERS UNE FUSION DES MODALITÉS POUR LA CLASSIFICATION DES ETATS DÉPRESSIFS

Les recherches sur l'analyse des états dépressifs en Informatique Affective montrent que l'usage de plusieurs modalités augmente l'efficacité des modèles de prédiction. En particulier, les modalités visuelles et auditives sont appréciées pour leur facilité d'acquisition et leur caractère non invasif. De manière générique, la fusion consiste à réunir ou aggréger des informations provenant de différentes sources, en vue de les exploiter pour prendre une décision [96].

Dans ce chapitre, on expérimente plusieurs approches de fusion des modalités pour la classification des états dépressifs.

## 7.1 Approches traditionnelles pour la fusion

Traditionnellement, on observe deux stratégies de fusion :

— La stratégie dite *score-level fusion*, où chaque modalité est apprise séparément. Les modèles construits prédisent, pour chaque exemple, un score qui mesure la probabilité pour l'exemple d'appartenir à une classe. Ces prédictions doivent être traitées afin d'obtenir la décision finale du système. On parle aussi de fusion "aval".

— La stratégie dite *data-level fusion*, où les différentes modalités sont apprises par un modèle unique. Les sources de données sont concaténées en amont de

l'apprentissage, et les modèles construits prédisent, pour chaque exemple, la décision finale du système. On parle aussi de fusion "amont".

### 7.1.1 Fusion des scores par l'opérateur moyenne

Dans les sections 6.2 et 6.3, on a construit des classifieurs unimodaux capables de reconnaître l'état dépressif de sujets, à partir de données visuelles et auditives, respectivement. Chacun des modèles calcule un score BDI-II par exemple classifié (dit score intermédiaire), puis calcule un score final pour la vidéo correspondante en moyennant les scores intermédiaires.

Un opérateur intuitif pour la combinaison des scores intermédiaires est la moyenne. On note $M$ une modalité, avec $M \in \{V, A\}$ représentant les modalités visuelles et auditives. Pour une vidéo $k$, $\mathbf{y}'_{M,k}$ désigne le vecteur des $N_{M,k}$ scores intermédiaires calculés par un classifieur unimodal. La fusion *score-level* $FS_k$ des modalités de la vidéo $k$ par l'opérateur moyenne peut s'écrire comme (Eq. 7.1) :

$$FS_k = 0.5 * \left( \frac{1}{N_{V,k}} \sum \mathbf{y}'_{V,k} + \frac{1}{N_{A,k}} \sum \mathbf{y}'_{A,k} \right) \tag{7.1}$$

On calcule la fusion des scores pour les données visuelles et auditives utilisées pour les expérimentations *hold-out* des sections 6.2 et 6.3. En particulier, on utilise les données ayant permis d'obtenir les meilleures performances. Pour les données visuelles, il s'agit de celles décrites au moyen du descripteur R-STD calculé avec par fenêtres de 20s par pas de 1s. Pour les données auditives, les données sans réduction dimensionnelle et calculées sur des fenêtres de 20 par pas de 2s sont retenues. Dans les deux cas, c'est la tâche Freeform qui a obtenu les erreurs les plus petites. Les résultats comparés (tâche Freeform) sont présentés dans la table 7.1.

TABLE 7.1 – Comparaison des erreurs unimodales et multimodales, avec une fusion par moyenne des scores intermédiaires.

| Modalité | Tâche | Généralisation | |
|---|---|---|---|
| | | RMSE | MAE |
| Visuelle | Freeform | 7.31 | 5.43 |
| | Northwind | 8.8 | 7.0 |
| Auditive | Freeform | 7.38 | 5.74 |
| | Northwind | 9.4 | 7.6 |
| Fusion *score-level* (FS) | Freeform | **7.28** | **5.74** |
| | Northwind | 9.19 | 8.76 |

Cette approche, pourtant naïve *a priori*, réussit à améliorer la meilleure modalité (visuelle dans ce cas). Le gain, bien que faible, s'explique par l'utilisation de l'opérateur moyenne sur les scores intermédiaires plutôt que sur les scores finaux de chaque vidéo. Les scores intermédiaires évaluent l'état dépressif dans chaque vidéo de manière plus fine (exemple par exemple) que les scores finaux, puisqu'il existe plusieurs exemples visuels et auditifs pour une même vidéo.

Aussi, cette approche ne nécessite pas la création de "paires sémantiques" d'exemples, où chaque exemple d'une modalité correspond à un exemple de l'autre modalité. Sur ces données en particulier, c'est une méthode adaptée, puisque :

— Certaines vidéos sont muettes ou n'ont pas produit de descripteur visuel (en raison de l'absence de détection de visage, par exemple).

— Les configurations optimales pour le calcul des descripteurs visuels et auditifs sont différentes et aboutissent à un nombre d'exemples différent, pour une même vidéo.

### 7.1.2 Fusion des données par concaténation des exemples

La concaténation des exemples est une autre approche de fusion traditionnelle, qui consiste à unir deux exemples hétérogènes pour ne former qu'un seul exemple. Une fois tous les exemples concaténés, un modèle unique est entraîné sur les données. Pour cette approche, et contrairement à la fusion *score-level*, les exemples concaténés ensemble doivent former des paires sémantiques.

On forme des paires d'exemples à partir de données décrites selon une même configuration temporelle, avec une fenêtre de 20s et un pas de 2s. Cette configuration désavantage la modalité visuelle, puisque ce n'est pas celle qui a permis d'obtenir les meilleures performances sur cette modalité. Pour la tâche Freeform, 1 036 et 735 paires d'exemples de dimension 2 404 (soit 136 + 2 268) sont utilisables pour l'apprentissage et la généralisation. Pour la tâche Northwind, les paires sont moins nombreuses, avec 629 et 622 éléments pour l'apprentissage et la généralisation, respectivement. On recherche dans un premier temps les meilleurs hyperparamètres pour entraîner un INN à reconnaître les états dépressifs dans les exemples multimodaux obtenus par concaténation des modalités. La recherche en grille aboutit à la sélection des hyperparamètres ($s_{inf}$, $s_{conf}$, $\alpha$) = (90, 5, 0.5) pour Freeform et (60, 7, 0.3) pour Northwind.

Les résultats de la fusion par concaténation des exemples (Tab. 7.2) montre que

cette approche réussit moins bien que les approches unimodales et multimodale testées jusqu'ici, hormis pour la tâche Northwind.

TABLE 7.2 – Résultats de la fusion par concaténation des exemples.

| Modalité | Tâche | Apprentissage | | | Généralisation | |
|---|---|---|---|---|---|---|
| | | RMSE | MAE | $\rho_{\text{proto}}$ | RMSE | MAE |
| Fusion *data-level* (FD) | Freeform | 3.07 | 1.55 | 0.33 | 10.22 | 8.39 |
| | Northwind | 2.12 | 0.86 | 0.29 | 9.44 | 7.72 |

Toutefois, ces résultats ne sont pas directement comparables avec les précédents, puisqu'ils sont été calculés sur un jeu de données différent.

### 7.1.3 Comparaison des approches traditionnelles pour la fusion

La comparaison des différentes approches réalisées jusqu'ici demande leur réexpérimentation sur un jeu de données uniques. On retient le jeu de données le plus restrictif, i.e. celui utilisé pour la fusion des modalités par concaténation des exemples.

TABLE 7.3

| Modalité | Tâche | Apprentissage | | | Généralisation | |
|---|---|---|---|---|---|---|
| | | RMSE | MAE | $\rho_{\text{proto}}$ | RMSE | MAE |
| Visuelle | Freeform | 1.14 | 0.38 | 0.24 | 9.05 | 6.89 |
| | Northwind | 1.34 | 0.47 | 0.24 | 11.52 | 9.21 |
| Auditive | Freeform | 0.23 | 0.05 | 0.3 | 7.38 | 5.74 |
| | Northwind | 1.52 | 0.56 | 0.30 | 9.4 | 7.6 |
| Auditive PCA, dimension 150 | Freeform | 0.51 | 0.16 | 0.16 | 9.12 | 6.5 |
| | Northwind | 0.63 | 0.2 | 0.19 | 9.16 | 7.34 |
| Fusion *score-level* | Freeform | - | - | - | 8.15 | 6.23 |
| | Northwind | - | - | - | 11.12 | 8.53 |
| Fusion *data-level* | Freeform | 3.07 | 1.55 | 0.33 | 10.22 | 8.39 |
| | Northwind | 2.12 | 0.86 | 0.29 | 9.44 | 7.22 |

Sur le jeu de données constitué de paires d'exemples, on note qu'aucune des approches multimodales ne réussit à améliorer la reconnaissance des états dépressifs de la meilleure modalité, i.e. la modalité auditive sans réduction dimensionnelle

(Tab. 7.3). Le grand nombre de configurations différentes testées indique que le modèle de classifieur utilisé a atteint sa limite sur ces données en particulier.

## 7.2 Un framework modulaire pour la fusion abstraite multimodale

Les approches de fusion traditionnelles, i.e. basées sur la fusion des scores ou des données, ne prennent pas en compte les corrélations cachées qui peuvent exister entre les modalités. Nous proposons ici une notion de "fusion abstraite", en nous appuyant sur module de mémoire associative multimodale.

La m-BAM, en tant que mémoire associative multimodale, est capable d'apprendre des paires de vecteurs hétérogènes. Pour chaque paire, une représentation abstraite est construite à partir des couches sensorielles (où l'on présente les modalités) vers une couche d'intégration. Ce fonctionnement est assimilable à une fusion, puisqu'il tire profit des liens existant entre des modalités distinctes pour encoder, en une seule représentation multimodale, une grande quantité d'information. L'INN, en tant que classifieur, produit également des représentations intermédiaires : les prototypes. Ces derniers encodent également une grande quantité d'information sur la classe à laquelle ils sont rattachés.

La stratégie de fusion retenue est issue des travaux de Reynaud [97]. Elle opère au niveau des représentations intermédiaires contruites par les réseaux de neurones employés, comme le ferait le cerveau lors de l'intégration d'informations issues de canaux sensoriels multiples.

### 7.2.1 Présentation des modules et fonctionnement

#### 7.2.1.1 Vue d'ensemble

Le framework est conçu pour la classification d'exemples présentés selon plusieurs modalités. La première étape consiste en la construction de prototypes grâce à des INN. Chaque INN construit des prototypes pour la modalité qu'il reçoit en entrée. Par exemple, pour la classification du score dépressif dans des données visuelles et auditives, on met en œuvre deux INN (*INN1* et *INN2* sur la Fig. 7.1), l'un pour la modalité visuelle $V$ et l'autre pour la modalité auditive $A$. On note $V^{(p)}$ et $A^{(p)}$ les membres unimodaux d'une même paire sémantique d'exemples.

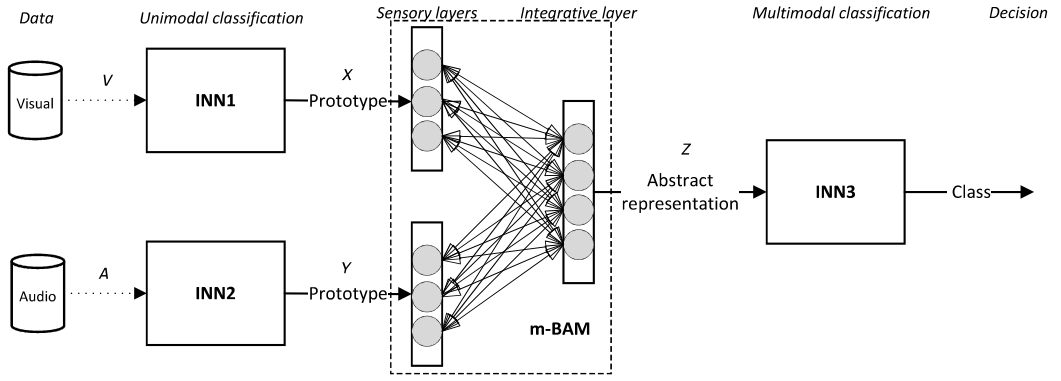



FIGURE 7.1 – Framework modulaire pour la fusion des modalités [16]

La seconde étape implique l'usage de la m-BAM comme module de fusion des représentations unimodales (i.e. les prototypes). Chaque paire de prototypes $X^{(p)} = \mathsf{P}_{\text{meilleur}}\left(V^{(p)}\right)$ et $Y^{(p)} = \mathsf{P}_{\text{meilleur}}\left(A^{(p)}\right)$ activée par les exemples en entrée de *INN1* et *INN2* est apprise par la m-BAM. Lorsque toutes les paires sont apprises, on peut récupérer les états stables $Z^{(p)}$ de la couche d'intégration, qui représente la fusion des modalités.

A la troisième étape, une représentation abstraite multimodale $Z^{(p)}$ a été produite pour chaque paire d'exemples unimodaux fournie en entrée de *INN1* et *INN2*. Le framework fait alors usage d'un troisième INN ($INN3$) afin d'associer chaque $Z^{(p)}$ à l'étiquette de classe de la paire correspondante.

Lors du processus, les paires de vecteurs doivent être congruentes, i.e. elles doivent être sémantiquement liées. Cette contrainte a été évoquée pour la fusion *data-level* des modalités (voir 7.1.2). Pour la m-BAM, elle se traduit ainsi :

— Pour les paires de vecteurs de données (en entrée de *INN1* et *INN2*), les membres d'une même paire doivent correspondre aux données visuelles et auditives d'un même extrait vidéo.
— Pour les paires de prototypes (en entrée de la m-BAM), les membres d'une même paire doivent correspondre aux prototypes activés par une paire congruente de vecteurs de données.

Dans la description ci-dessus, on utilise deux modalités. Toutefois, il n'y a pas de limitation théorique quant au nombre de modalités que la m-BAM peut intégrer dans une représentation multimodale [95].

#### 7.2.1.2 Apprentissage et généralisation

L'algorithme 1 est proposé pour l'apprentissage des données multimodales au moyen du framework modulaire de fusion. On note m-BAMX, m-BAMY et m-BAMZ les couches sensorielles (X et Y) et la couche d'intégration (Z) de la BAM.

**Algorithme 1** Apprentissage du framework de fusion

**pour** $V^{(p)}$ with $p = 1,\ldots,n$ **faire**
Proposer $V^{(p)}$ en entrée de *INN1*
Mettre à jour ou créer des prototypes
**fin pour**
**pour** $A^{(p)}$ with $p = 1,\ldots,n$ **faire**
Proposer $A^{(p)}$ en entrée de *INN2*
Mettre à jour ou créer des prototypes
**fin pour**
**Conséquence :** L'apprentissage des modèles unimodaux achevé

---

**Prérequis :** Les poids et seuils sont initialisés par compression (voir 7.2.2.1)
**répéter**
**pour** chaque paire $(V^{(p)},A^{(p)})$ avec $p = 1,\ldots,n$ **faire**
Proposer $X^{(p)} = \mathsf{P}_{\text{meilleur}}\left(V^{(p)}\right)$ en entrée de m-BAMX
Proposer $Y^{(p)} = \mathsf{P}_{\text{meilleur}}\left(A^{(p)}\right)$ en entrée de m-BAMY
Mettre à jour les poids et les seuils
**fin pour**
**jusqu'á ce que** Les poids et les seuils soient stables
**Conséquence :** La fusion des modalités est achevée

---

**pour** chaque paire $(V^{(p)},A^{(p)})$ avec $p = 1,\ldots,n$ **faire**
Proposer $V^{(p)}$ en entrée de *INN1* et $A^{(p)}$ en entrée de *INN2*
Injecter $X^{(p)}$ dans m-BAMX et $Y^{(p)}$ dans m-BAMY
Récupérer les motifs de fusion $Z^{(p)}$
Proposer $Z^{(p)}$ en entrée de *INN3*
Mettre à jour ou créer des prototypes
**fin pour**
**Conséquence :** L'apprentissage du modèle multimodal est achevé

La généralisation du modèle se fait simplement en présentant le motif de fusion produit pour une donnée inconnue à *INN3*.

### 7.2.2 Initialisation des couches du réseau

On parle de convergence en apprentissage lorsque le nombre d'époques nécessaires à l'apprentissage des paires est un nombre fini. De même, on parle de convergence en rappel (vers un état stable) lorsque le nombre de réverbérations nécessaires au rappel d'une paire est un nombre fini. Le réseau ne converge pas nécessairement, et il est commun d'utiliser un garde-fou lors de ces phases.

La m-BAM converge vers un état stable si et seulement si ses hyperparamètres et son état initial sont favorables. Le faible nombre d'hyperparamètres (au nombre de deux, voir 2.5.1.2) autorise la réalisation d'une recherche en grille pour l'optimisation du réseau. En revanche, les états initiaux des neurones sur les couches X, Y et Z du réseau ont un impact sur la vitesse à laquelle il converge. Traditionnellement, les états des couches X et Y sont initialisés avec les exemples à apprendre. Pour la couche Z, on distingue trois méthodes d'initialisation :

— L'initialisation aléatoire : la couche Z est initialisée avec un motif aléatoire.
— L'initialisation par compression des données : la couche Z est initialisée avec un code de liage (Eq. 7.2) qui encode les états de X et Y
— L'initialisation par zones d'affectation : les états de la couche Z sont initialisés en réservant une "zone" contiguë de neurones dont les états seront mis à 1 pour une paire d'exemples à apprendre, et à -1 pour les autres paires.

Trois indicateurs peuvent être observés pour comparer les méthodes d'initialisation. En premier lieu, le nombre d'époques, i.e. le nombre de fois que l'on présente la base d'apprentissage au réseau, est témoin de la vitesse d'apprentissage. Un nombre d'époques trop élevé n'est pas intéressant pour un classifieur. En second lieu, le nombre de réverbérations, i.e. le nombre de fois que l'activation d'une paire d'exemples est propagée dans le réseau avant de converger vers un état stable. Plus le nombre de réverbérations est élevé, moins le réseau est efficace. Le troisième indicateur est l'erreur de rappel, c'est-à-dire la distance entre une paire de motifs apprise et la paire de motifs rappellée. Une distance adaptée à cette évaluation est la distance de Hamming. Des études théoriques de la m-BAM [98] ont montré que ces indicateurs sont corrélés : plus le nombre d'époques est élevé, plus le nombre de réverbérations est important ; et plus le nombre de réverbérations est important, plus l'erreur de rappel est grande.

#### 7.2.2.1 Méthodes d'initialisation

L'initialisation par compression des données consiste à initilialiser la couche Z avec un code de liage des états des couches X et Y. Le code de liage rend compte des états des neurones des couches sensorielles, et dépend par conséquent des données proposées en entrée du réseau et des poids et seuils aléatoires à cette phase. Une proposition est l'utilisation du seuillage de l'activation de la couche Z correspondant aux états initiaux de X et Y comme code de liage 7.2 [97] :

$$S_{Z_k}^{(p)} = \sum_{dimX}^{i=1} w_{ki}^{Z \to X} X_i^{(p)} + \sum_{dimY}^{j=1} w_{kj}^{Z \to Y} Y_j^{(p)} - \theta_{Z_k} \tag{7.2}$$

Cette approche est identifée par Reghis *et al.* [95] comme fournisant les meilleurs résultats mais ne garantissant pas la convergence de la m-BAM. Ils proposent une initialisation par zones d'affectaction, qui repose sur les éléments suivants :

— Chaque paire en entrée possède une zone de neurones contigus sur la couche Z, initialisée à 1. Le reste des neurones de Z, correspondant aux zones des autres paires, est initialisé à -1.

— La distance de Hamming entre deux motifs sur la couche Z doit être égale à au moins deux fois la taille de la zone d'affectation.

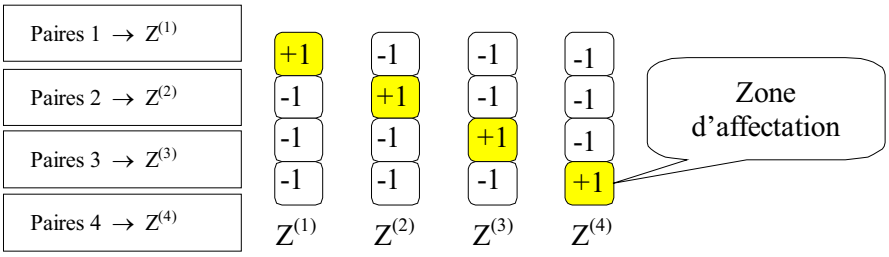



FIGURE 7.2 – Exemple de zones d'affectation d'après [95].

En conséquence de ces éléments, la taille de la couche Z ne peut pas être fixée librement et dépend du nombre de paires à apprendre. Par exemple, pour une couche Z de 4 neurones, on peut affecter au maximum 4 paires à des zones d'une taille de 1 neurone (Fig. 7.2). Pour cette proposition, les états de Z ne dépendent plus des données.

#### 7.2.2.2 Expérimentations et résultats

Afin de comparer les méthodes d'initialisation par compression et par affectation, on utilise un jeu de données neutre pour apprendre et rappeler des paires apprises au moyen d'une m-BAM initialisée de manière aléatoire, par compression des données ou par zones d'affectations. Les données sont issues de la base *Optical Recognition of Handwritten Digits Data Set* [1], composée de bitmaps d'images de chiffres de 0 à 9 de 8x8 pixels. On entraîne le réseau à associer les paires représentant des 0 et des 5, des 1 et des 6, des 2 et des 7, des 3 et des 8, et des 4 et des 9 (le choix de ces paires est arbitraire).

Les couches X et Y de la m-BAM sont de taille 64 (i.e. adaptées à la dimension des données), et la taille de la couche Z est fixée à 64 cellules. On fait varier les hyperparamètres $\xi$ de 15 à 200 par pas de 5 et $\lambda$ de 0.5 à 2 par pas de 2. Pour chaque essai, on relève le nombre d'époques, le nombre de réverbérations et le taux d'erreur de rappel. Les paires présentées au réseau lors de la phase de rappel correspondent exactement aux paires apprises, sans ajout de bruit. On fixe comme garde-fou 1000 époques et 1000 réverbérations.

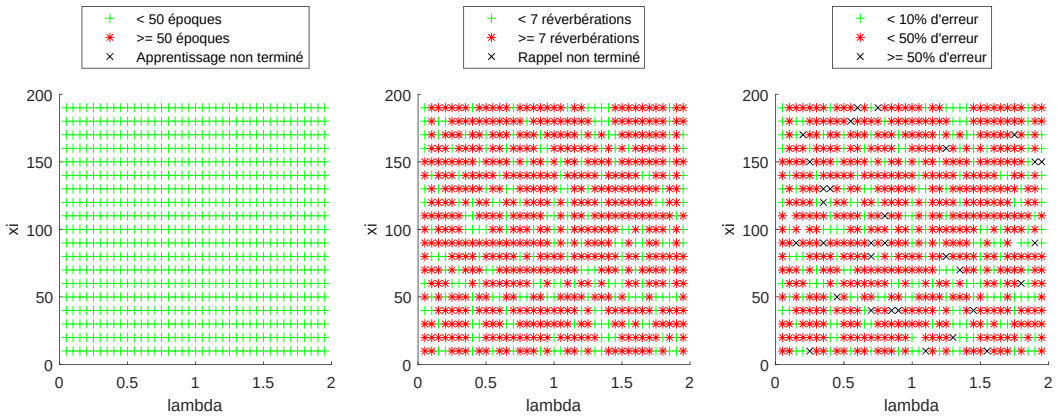



FIGURE 7.3 – Résultats pour l'initialisation par compression. A gauche, l'apprentissage ; au centre, le rappel ; et à droite, le taux d'erreur en rappel.

Pour l'initialisation par compression (Fig. 7.3), la phase d'apprentissage est très rapide et est réalisée en moins de 10 époques. Cependant, les valeurs de $\xi$ et $\lambda$ testées n'ont pas permis au réseau de rappeler les paires en moins de 1000 itérations. Aussi, la phase de rappel n'aboutit pas et les états reconstitués sont éloignés des paires apprises.

1. https ://archive.ics.uci.edu/ml/datasets/optical+recognition+of+handwritten+digits

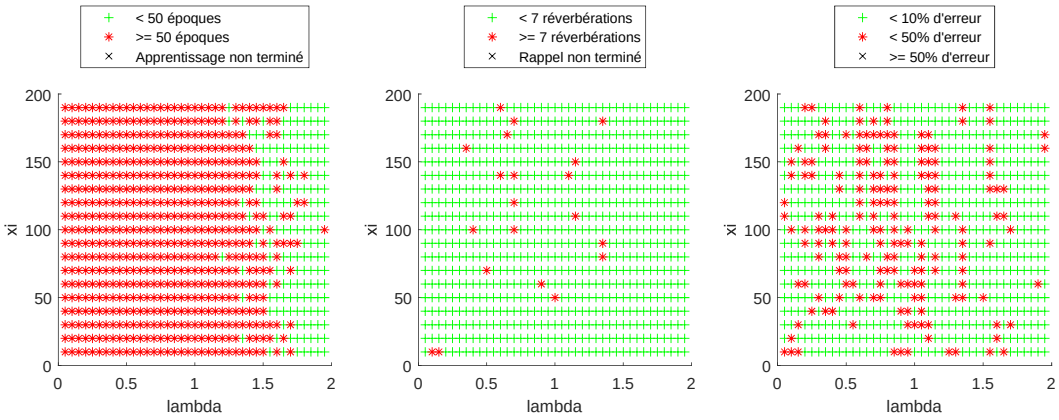



FIGURE 7.4 – Résultats pour l'initialisation par zones d'affectations. A gauche, l'apprentissage ; au centre, le rappel ; et à droite, le taux d'erreur en rappel.

L'initialisation par zones d'affectation présente un autre profil (Fig. 7.4). La phase d'apprentissage est parfois plus longue pour cette méthode d'initialisation, mais les paires sont presque toujours rappellées dans les limites du garde-fou. La phase de rappel ayant abouti, l'erreur en rappel est faible.

Ces résultats sont en accord avec ceux de Reghis *et al.* [95] sur la non-garantie de convergence pour l'initialisation par compression. Bien que la méthode par zone d'affectations converge plus facilement, elle ne permet pas de fixer librement le nombre de neurones sur la couche Z. Cependant, cet élément n'est pas handicapant, compte tenu de la (faible) capacité de stockage de la BAM.

### 7.2.3 Classification multimodale de l'état dépressif

L'un des points forts du framework est de pouvoir opérer sur des données hétérogènes. C'est le cas des données visuelles et auditives utilisées pour les expérimentations unimodales des sections 6.2 et 6.3, qui ont des dimensions et des domaines de définition différents. En particulier, les données visuelles sont de dimension 136, alors que les données auditives sont de dimension 2 268.

Dans cette section, on utilise le framework de fusion afin de réaliser la classification multimodale de l'état dépressif [16]. L'usage du framework requiert quelques modifications au niveau des données, à savoir :

— Le seuillage des données, afin que chaque neurone reçoive la valeur -1 ou 1. Pour ce faire, le seuillage retenu opère par composante ("par colonne") dans

les données : la moyenne de chaque composante est calculée, puis les valeurs de la composante qui sont supérieures à la moyenne sont mises à -1, et les autres sont mises à 1.

— L'utilisation de paires congruentes, comme c'était le cas pour la fusion *data-level* (voir 7.1.2).

**Expérimentations pilotes sur la tâche Freeform** Afin d'étudier le comportement du framework sur des données moins abstraites, on l'utilise afin de fusionner les modalités visuelles et auditives de la tâche Freeform et reconnaître l'état dépressif des sujets. Pour rappel, la tâche Freeform dispose de 1036 et 735 paires d'exemples pour une validation *hold-out*. Les modèles pour la classification unimodale sont ceux appris pour la comparaison des approches traditionnelles de fusion (voir 7.1.3), car ils ont été appris sur des paires d'exemples sémantiques.

On lance l'apprentissage de la totalité des 1 036 paires d'apprentissage avec une m-BAM dont les paramètres étaient fixés à $\xi = 100$ et $\lambda = 1.7$. Les motifs de la couche d'intégration sont initialisés par compression (voir 7.2.2.1). Malgré un apprentissage des poids assez rapide (16 époques), la m-BAM ne parvient pas à rappeler les paires. Ce constat est cohérent avec la capacité théorique de la m-BAM, estimé à environ 1.68 fois la taille de la couche la plus petite [98], soit un peu plus de 228 paires pour les données utilisées.

On réduit progressivement le nombre de paires à apprendre, jusqu'à ce que la m-BAM soit capable de les rappeler correctement. Pour le cas spécifique des données utilisées, la capacité observée de la m-BAM est autour de 50. On apprend la base d'apprentissage par paquet de 50 paires, puis on récupère pour chacune le motif de fusion correspondant et on passe au paquet suivant. Conformément au fonctionnement du framework, les motifs sont classifiés au moyen d'un INN afin de les associer à une classe en sortie.

Afin de déterminer la taille optimale de la couche d'intégration, on réalise l'expérience pour plusieurs tailles allant de 150 à 1050 par pas de 100 (Fig. 7.5). Les meilleures performances sont obtenues pour une couche Z de 450 neurones (Tab. 7.4).

L'utilisation du framework comme méthode de fusion a permis d'améliorer légèrement l'erreur en généralisation de la melleure modalité, auditive dans ce cas. Cependant, les données de la modalité auditive ont une dimension 17 fois

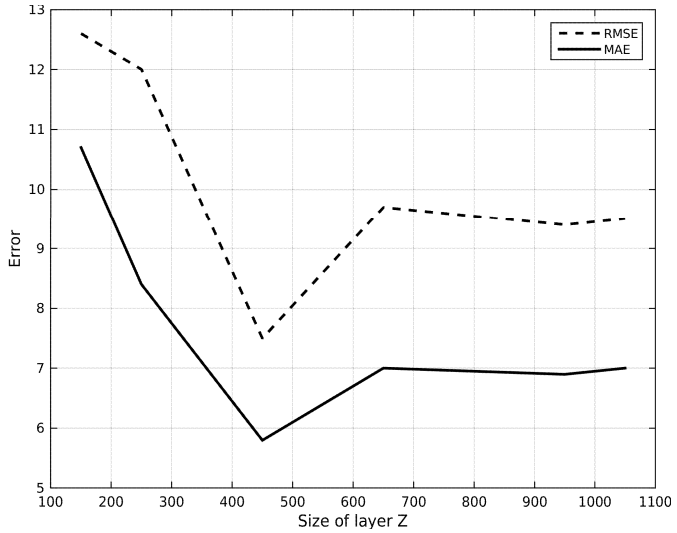



FIGURE 7.5 – Erreurs en généralisation pour différentes tailles de la couche d'intégration [16].

TABLE 7.4 – Résultats unimodaux et multimodaux pour la tâche Freeform [16].

| Modalité | RMSE | MAE |
|---|---|---|
| Visuelle | 10.56 | 8.59 |
| Auditive | 7.53 | 5.85 |
| Fusion abstraite | **7.47** | **5.78** |

supérieure à celle des données de la modalité visuelle. En conséquence, le réseau a sans doute été influencé plus largement par la modalité auditive.

A l'issue de ces essais, on peut formuler les remarques suivantes :

— La m-BAM est capable d'apprendre un petit nombre de paires d'exemples dans un temps "raisonnable". Pour l'apprentissage d'un grand nombre d'exemples (par rapport à la taille de la plus petite couche), il faut les apprendre par petits paquets.

— La recherche des hyperparamètres de la m-BAM est fastidieuse : des petites variations de $\xi$, $\lambda$ et de $dimZ$ influent beaucoup sur la rapidité du réseau et sur sa capacité à apprendre et à rappeler des paires.

— Les fondements théoriques du modèle ne permettent pas de l'utiliser avec des données réelles (elles doivent être binaires ou bipolaires). Une étape de seuillage est par conséquent nécessaire, ce qui peut être perçu comme une "incohérence", en fonction de la nature des données.

**Fusion abstraite avec des couches sensorielles de tailles équivalentes** Lors de la classification des données auditives, leur forte dimensionnalité avait déjà fait l'objet d'une étude (voir 6.3.2) qui montrait que 45 composantes sur 2 268 suffisaient pour expliquer 100% de la variance totale des données. Les résultats montraient également que le nombre de prototypes était réduit d'environ un tiers lors de l'apprentissage des données réduites, par rapport aux données de pleine dimension. Avec une dimension de 150 (réduction d'un facteur 15), les performances sur les données auditives du jeu de données composé de paires sémantiques (Tab. 7.3) sont réduites pour la tâche Freeform (de 7.38/5.74 à 9.12/6.5, en termes de RMSE/MAE) mais améliorées pour la tâche Northwind (de 9.4/7.6 à 9.16/7.34).

On utilise les mêmes données visuelles que précédemment avec les données auditives réduites pour la classification multimodale de l'état dépressif au moyen du framework. Le protocole expérimental est le suivant : on fait varier $dimZ$ et pour chaque taille, les paires sont apprises par paquet de 50 (couche Z initialisée par compression). Lorsque tous les motifs de fusion sont produits, on réalise une recherche en grille pour trouver les meilleurs hyperparamètres pour leur apprentissage au moyen de l'INN en sortie du framework. La figure 7.6 montre les erreurs en généralisation et le ratio de prototypes du meilleur essai pour chaque taille de la couche Z.

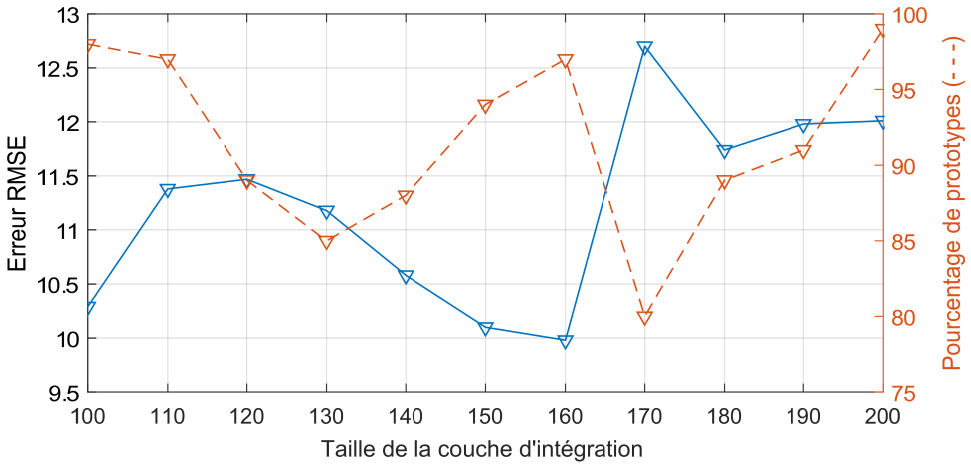



FIGURE 7.6 – Erreur en généralisation et nombre de prototypes (pointillés) en fonction de la taille de la couche d'intégration (couches de tailles équivalentes)

Les résultats pour cet essai révèlent que la réduction dimensionnelle des données auditives n'a pas bénéficié au framework : les erreurs sont plus grandes, et les

INN sont en proie au surapprentissage ($\rho_{\text{proto}} > 0.8$).

## 7.3 Synthèse et discussion

Le tableau 7.5 compare les performances otenues pour chaque stratégie de fusion réalisée.

TABLE 7.5 – Comparaison des méthodes de fusion

| Modalité | Tâche | Apprentissage | | | Généralisation | |
|---|---|---|---|---|---|---|
| | | RMSE | MAE | $\rho_{\text{proto}}$ | RMSE | MAE |
| Fusion abstraite | Freeform | 1.14 | 0.38 | 0.34 | **7.47** | **5.78** |
| | Northwind | 1.34 | 0.47 | 0.37 | 10.39 | 9.25 |
| Fusion *score-level* | Freeform | - | - | - | 8.15 | 6.23 |
| | Northwind | - | - | - | 11.12 | 8.53 |
| Fusion *data-level* | Freeform | 3.07 | 1.55 | 0.33 | 10.22 | 8.39 |
| | Northwind | 2.12 | 0.86 | 0.29 | **9.44** | **7.22** |

On note que c'est la fusion abstraite qui réalise les plus petites erreurs, suivie de près par la fusion *score-level*.

L'utilisation du framework pour la fusion des modalités est une approche intéressante : elle utilise des représentations intermédiaires pour un usage qui ne leur est initialement pas destiné. Le succès de l'approche confirme que la m-BAM est un modèle pertinent pour la fusion des modalités, et que les INN sont des modèles pertinents pour l'émulation de ces modalités. Toutefois, le framework nécessite beaucoup de prétraitements au niveau des données, notamment de seuillage et de création de paires d'exemples. Les méthodes de fusion traditionnelles, et notamment la fusion *score-level* offre une plus grande souplesse mais des résultats moins intéressants. La fusion abstraite et *score-level* se positionne mieux que la *baseline* sur ces données, où l'erreur (RMSE/MAE) à battre était de 8.34/6.68. Cependant, les résultats ne sont pas directement comparables, car nous avons dû supprimer un certain nombre d'exemples de la base de données.

CHAPITRE 8

# CONCLUSION

Entre le début et la fin de cette thèse, l'*Affective Computing* a connu une progression fulgurante. Nous avons exploré ce domaine à travers deux applications pour lesquelles l'intérêt du public et des chercheurs n'a fait que croître : l'analyse des état émotionnels et l'analyse des états dépressifs.

Côté *computing*, l'évaluation automatique des états émotionnels et dépressifs requiert l'usage de méthodes d'Intelligence Artificielle. Nous avons expliqué le fonctionnement des principaux modèles de classifieurs utilisés dans l'état de l'art. En particulier, deux modèles issus des sciences cognitives ont retenu notre attention : le classifieur neuronal incrémental à base de prototypes et la mémoire associative multimodale. L'implémentation de ces méthodes a été l'un des moments notables de la thèse, puisqu'elle les fait sortir du cadre de la traditionnelle "boîte noire". Dans un contexte résolument pluridisciplinaire, cet avantage se manifeste par la possibilité d'expliquer et d'interpréter facilement le fonctionnement et les résultats des traitements mis en œuvre.

La représentation des états émotionnels et dépressifs dispose d'un très large éventail de possibilités. Pour la représentation des états émotionnels, plusieurs courants de psychologie et de psychiatrie se confrontent ou se contredisent, aboutissant à des représentations différentes. Nous avons étudié cet aspect, purement *affective*, et montré qu'il ne bénéficie pas toujours à l'avancée de la recherche : en l'absence de fondements qu'il n'est pas possible de remettre en cause, il est

délicat de construire des approches qui ne seront elles-mêmes pas remises en cause. En particulier, la question de l'annotation des états affectifs de manière continue ou discrète pose un problème intéressant qui illustre parfaitement la croisée de l'*affective* et du *computing*. Pour nos travaux, nous avons choisi de représenter les états émotionnels par classes d'intensités, car l'information de classe, qui comprend un segment de plusieurs valeurs, est plus fiable et plus interprétable qu'une valeur continue. Dans le cadre de la prévention des risques psychosociaux, l'état émotionnel n'est pas un indicateur suffisant, à cause du caractère spontané de l'émotion.

La représentation des états dépressifs est plus cadrée et il existe des passerelles entre les différentes représentations. Nous avons mis au point un descripteur visuel spatio-temporel qui encode la quantité de micro et de macro-mouvements du visage. Sa rapidité de calcul lui permet d'être construit à la volée, pour la classification en temps réel des états dépressifs. Le prototype proposé doit encore faire l'objet de développements, mais les performances obtenues dans un contexte *subject-idenpendent* sont prometteuses. Nous avons également étudié plusieurs stratégies pour la fusion des modalités visuelles et auditives, et l'usage des *features* intermédiaires produites par les réseaux de neurones a su fournir de meilleurs résultats que les stratégies de fusion traditionnelles. La dépression étant, par définition, observée sur un période longue, le suivi de la sévérité dépressive est par conséquent pertinent pour la prévention des risques psychosociaux.

Une suite intéressante à ces travaux consisterait en une phase plus large de test de l'outil d'évaluation en temps réel : son utilisation, en plus de valider les méthodes de classification employées, ouvrirait la porte à de nouvelles clefs pour la compréhension des risques psychosociaux et de leur survenue.

# BIBLIOGRAPHIE


[1] Timo AHONEN, Abdenour HADID et Matti PIETIKAINEN : Face Description with Local Binary Patterns : Application to Face Recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 28(12):2037–2041, 2006.

[2] Timur R. ALMAEV et Michel F. VALSTAR : Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 356–361, 2013.

[3] AMERICAN PSYCHIATRIC ASSOCIATION : *Diagnostic and statistical manual of mental disorders*. 4th edition édition, 1994.

[4] American Psychiatric ASSOCIATION : *Diagnostic and statistical manual of mental disorders (4th ed.)*. Washington, DC, 1994.

[5] Yusuf AYTAR, Carl VONDRICK et Antonio TORRALBA : Soundnet : Learning sound representations from unlabeled video. *In Advances in Neural Information Processing Systems*, pages 892–900, 2016.

[6] A. AZCARRAGA et A. GIACOMETTI : A prototype-based incremental network model for classification tasks. *In Proceedings of Neuro-Nîmes*, pages 121–134, 1991.

[7] Tanja BÄNZIGER et Klaus R. SCHERER : Introducing the geneva multimodal emotion portrayal (gemep) corpus. *Blueprint for affective computing : A sourcebook*, pages 271–294, 2010.

[8] Per BECH, Marianne KASTRUP et Ole J. RAFAELSEN : Mini-compendium of rating scales for states of anxiety, depression, mania, schizophrenia with corresponding dsm—iii syndromes. *Acta Psychiatrica Scandinavica*, 1986.

[9] Aaron T. BECK : *Depression : Causes and Treatment*. University of Pennsylvania Press, 1972.

[10] Aaron T. BECK, Robert A. STEER et Gregory K. BROWN : Beck depression inventory-II. *San Antonio*, 78(2):490–498, 1996.

[11] Michael BIEHL, Barbara HAMMER et Thomas VILLMANN : Prototype-based models in machine learning. *Wiley Interdisciplinary Reviews : Cognitive Science*, 7(2):92–111, 2016.

[12] Gary. BRADSKI : The OpenCV Library. *Dr Dobbs Journal of Software and Tools*, 25:120–125, 2000.

[13] Linlin CHAO, Jianhua TAO, Minghao YANG, Ya LI et Zhengqi WEN : Long Short Term Memory Recurrent Neural Network based Multimodal Dimensional Emotion Recognition. *In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge - AVEC'15*, pages 65–72, Brisbane, 2015. ACM.

[14] Shizhe CHEN, Qin JIN, Jinming ZHAO et Shuai WANG : Multimodal Multi-task Learning for Dimensional and Continuous Emotion Recognition. *In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge - AVEC'17*, pages 19–26, Mountain View, 2017. ACM.

[15] Stephane CHOLET et Helene PAUGAM-MOISY : Démonstration du diagnostic automatique de l'état dépressif. *In Conférence Nationale d'Intelligence Artificielle*, volume 2133, pages 116–117, Nancy, 2018. CEUR-WS.

[16] Stephane CHOLET, Helene PAUGAM-MOISY et Sebastien REGIS : Bidirectional associative memory for multimodal fusion : a depression evaluation case study. *In Proceedings of the 2019 International Joint Conference on Neural Networks*. To be published, 2019.

[17] Gilles COHEN : Méthodes à noyau, 2007. URL https://www.lri.fr/~antoine/Courses/Master-ISI/chap-14-svm.pdf.

[18] Corinna CORTES et Vladimir VAPNIK : Support-Vector Networks. *Machine Learning*, 297(20):273–297, 1995.

[19] Roddy COWIE, Ellen DOUGLAS-COWIE, Susie SAVVIDOU, Edelle MCMAHON, Martin SAWEY et Marc SCHRÖDER : "Feeltrace" : An instrument for recording perceived emotion in real time. *In ISCA tutorial and research workshop (ITRW) on speech and emotion*, pages 19–24, 2000.

[20] Nicholas CUMMINS, Julien EPPS, Vidhyasaharan SETHU et Jarek KRAJEWSKI : Variability compensation in small data : Oversampled extraction of i-vectors for the classification of depressed speech. *In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 970–974, 05 2014.

[21] Navneet DALAL et Bill TRIGGS : Histograms of oriented gradients for human detection. *In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, June 2005.

[22] Charles DARWIN : *L'expression des émotions chez l'homme et les animaux*. 1877. URL http://gallica.bnf.fr/ark:/12148/bpt6k772010.

[23] Naiyang DENG, Yingjie TIAN et Chunhua ZHANG : *Support vector machines : optimization based theory, algorithms, and extensions*. Chapman and Hall/CRC, 2012.

[24] Robert J. DERUBEIS, Greg J. SIEGLE et Steven D. HOLLON : Cognitive therapy vs. medications for depression : Treatment outcomes and neural mechanisms. *Natural Review of Neuroscience*, 9(10):788–796, oct 2008.

[25] Abhinav DHALL, Roland GOECKE, Simon LUCEY et Tom GEDEON : Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19 (3):34–41, 2012.

[26] Jeff DONAHUE, Lisa Anne HENDRICKS, Sergio GUADARRAMA, Marcus ROHRBACH, Subhashini VENUGOPALAN, Kate SAENKO, Trevor DARRELL, U. T. AUSTIN, Umass LOWELL et U. C. BERKELEY : Long-term Recurrent Convolutional Networks for Visual Recognition and Description. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634, 2015.

[27] Harris DRUCKER, Christopher JC BURGES, Linda KAUFMAN, Alex J SMOLA et Vladimir VAPNIK : Support vector regression machines. *In Advances in neural information processing systems*, pages 155–161, 1997.

[28] Dirk W. EILERT : Action units on face. URL https://www.mimikresonanz24.com/en/. [Consulté le 02-04-2018].

[29] Paul EKMAN : Differential Communication Of Affect By Head And Body Cues. *Journal of Personality and Social Psychology*, 2(5):726–735, 1965.

[30] Paul EKMAN et Wallace V. FRIESEN : *Facial action coding system*. Consulting Psychologists Press, Stanford University, Palo Alto, 1977.

[31] Moataz EL AYADI, Mohamed S. KAMEL et Fakhri KARRAY : Survey on speech emotion recognition : Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[32] Marc ESCHER, Igor PANDZIC et Magnenat N. THALMANN : Facial deformations for MPEG-4. *Proceedings Computer Animation'98*, (February):56–62, 1998.

[33] Humberto Pérez ESPINOSA, Hugo Jair ESCALANTE, Luis VILLASEÑOR-PINEDA, Manuel MONTES-Y-GÓMEZ, David PINTO-AVEDAÑO et Veronica REYES-MEZA : Fusing Affective Dimensions and Audio-Visual Features from Segmented Video for Depression. *In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC'14*, pages 49–55, Orlando, 2014. ACM.

[34] Florian EYBEN, Martin WÖLLMER et Björn SCHULLER : Openear—introducing the munich open-source emotion and affect recognition toolkit. *In 3rd international conference on Affective computing and intelligent interaction and workshops - ACII 2009*, pages 1–6. IEEE, 2009.

[35] Gunnar FARNEBÄCK : Two-frame motion estimation based on polynomial expansion. *In* Josef BIGUN et Tomas GUSTAVSSON, éditeurs : *Image Analysis*, pages 363–370, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[36] Ronald A. FISHER : The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[37] Johnny R. J. FONTAINE, Klaus R. SCHERER, Etienne B. ROESCH et Phoebe C. ELLSWORTH : The World of Emotions Is Not Two-Dimentional. *Psychological Science*, 18(12):1050–1057, 2007.

[38] Yuan GONG et Christian POELLABAUER : Topic Modeling Based Multi-modal Depression Detection. *In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge - AVEC'17*, pages 69–76. ACM, 2017.

[39] Ian GOODFELLOW, Yoshua BENGIO et Aaron COURVILLE : *Deep Learning*. MIT Press, 2016.

[40] Jonathan GRATCH, Ron ARTSTEIN, Gale LUCAS, Giota STRATOU, Stefan SCHERER, Angela NAZARIAN, Rachel WOOD, Jill BOBERG, David DEVAULT, Stacy MARSELLA, David TRAUM, Skip RIZZO et Louis-philippe MORENCY : The Distress Analysis Interview Corpus of human and computer interviews. *In Language Resources and Evaluation Conference*, pages 3123–3128. Citeseer, 2014.

[41] Stephen GROSSBERG : Competitive learning : From interactive activation to adaptive resonance. *Cognitive Science*, 11:23–63, 1987.

[42] Yu GU, Eric POSTMA et Hai-Xiang LIN : Vocal Emotion Recognition with Log-Gabor Filters. *In Proceedings of the 5th International Workshop on Audio / Visual Emotion Challenge - AVEC'15*, numéro January, pages 25–31. ACM, 2015.

[43] Max HAMILTON : Scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, 23(1):56–62, 1960.

[44] Max HAMILTON : Development of a rating scale for primary depressive illness. *British journal of social and clinical psychology*, 6(4):278–296, 1967.

[45] Kaiming HE, Xiangyu ZHANG, Shaoqing REN et Jian SUN : Deep residual learning for image recognition. *In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[46] Kaiming HE, Xiangyu ZHANG, Shaoqing REN et Jian SUN : Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[47] Carl Herman HJORTSJÖ : *Man's Face and Mimic Language*. Studen litteratur, 1969.

[48] Sepp HOCHREITER et Jürgen SCHMIDHUBER : Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[49] Tim HOLY : Generate maximally perceptually-distinct colors. URL https://se.mathworks.com/matlabcentral/fileexchange/29702-generate-maximally-perceptually-distinct-colors. [Consulté le 10-01-2019].

[50] Chih-Wei HSU et Chih-Jen LIN : A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, 13(2):415–425, 2002.

[51] Frederick Y. HUANG, Henry CHUNG, Kurt KROENKE, Kevin L. DELUCCHI et Robert L. SPITZER : Using the patient health questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. *Journal of General Internal Medicine*, 21(6):547–552, 2006.

[52] Jian HUANG, Ya LI, Jianhua TAO, Zheng LIAN, Zhengqi WEN, Minghao YANG et Jianyan YI : Continuous Multimodal Emotion Prediction Based on Long Short Term Memory Recurrent Neural Network. *In Proceedings of the 7th Annual Workshop*

*on Audio/Visual Emotion Challenge - AVEC'17*, pages 11–18, Mountain View, 2017. ACM.

[53] John D. HUNTHER : Matplotlib : A 2d graphics environment. *Computing in Science Engineering*, 9(3):90–95, May 2007.

[54] Varun JAIN, James L. CROWLEY, Anind DEY et Augustin LUX : Depression Estimation Using Audiovisual Features and Fisher Vector Encoding. *In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC'14*, pages 87–91. Academic Press, 2014.

[55] Asim JAN, Hongying MENG, Yona FALINIE, Binti ABD, Fan ZHANG et Saeed TURABZADEH : Automatic Depression Scale Prediction using Facial Expression Dynamics and Regression. *In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC'14*, numéro August, pages 73–80, Orlando, 2014. ACM.

[56] Asim JAN, Hongying MENG, Yona FALINIE, Binti A. GAUS et Fan ZHANG : Automatic Intelligent System for Automatic Depression Level Analysis Through Visual and Vocal Expressions. *IEEE Transactions on Cognitive and Developmental Systems*, 10 (3):668–680, 2018.

[57] Eric JONES, Travis OLIPHANT, Pearu PETERSON *et al.* : SciPy : Open source scientific tools for Python, 2001–. URL http://www.scipy.org/. [Consulté le 01-01-2019].

[58] Markus KÄCHELE, Martin SCHELS et Friedhelm SCHWENKER : Inferring Depression and Affect from Application Dependent Meta Knowledge. *In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC'14*, pages 41–48, New York, New York, USA, 2014. ACM.

[59] Takeo KANADE, Yingli TIAN et Jeffrey F. COHN : Comprehensive database for facial expression analysis. *In IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53. IEEE, 2000.

[60] Ittipan KANLUAN, Michael GRIMM et Kristian KROSCHEL : Audio-visual emotion recognition using an emotion space concept. *In 2008 16th European Signal Processing Conference*, pages 1–5. IEEE, 2008.

[61] Heysem KAYA, Fazilet CILLLI et Albert Ali SALAH : Ensemble CCA for Continuous Emotion Prediction. *In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC'14*, numéro February 2016, pages 19–26, Orlando, 2014. ACM.

[62] Heysem KAYA et Albert Ali SALAH : Combining Modality-Specific Extreme Learning Machines for Emotion Recognition in the Wild. *Journal on Multimodal User Interfaces*, 10(2):139–149, 2016.

[63] Vahid KAZEMI et Josephine SULLIVAN : One millisecond face alignment with an ensemble of regression trees. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

[64] Davis. E. KING : Dlib-ml : A Machine Learning Toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[65] Paul R. KLEINGINNA et Anne M. KLEINGINNA : A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and emotion*, 5 (4):345–379, 1981.

[66] Bart KOSKO : Bidirectional associative memories. *IEEE Transactions on Systems, man, and Cybernetics*, 18(1):49–60, 1988.

[67] Jean KOSSAIFI, Robert WALECKI, Yannis PANAGAKIS, Jie SHEN, Maximilian SCHMITT, Fabien RINGEVAL et Jing HAN : SEWA DB : A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild.

[68] Alex KRIZHEVSKY, Ilya SUTSKEVER et Geoffrey E. HINTON : Imagenet classification with deep convolutional neural networks. *In Advances in neural information processing systems*, pages 1097–1105, 2012.

[69] Kurt KROENKE, Tara W. STRINE, Robert L. SPITZER, Janet B. W. WILLIAMS, Joyce T. BERRY et Ali H. MOKDAD : The phq-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1-3):163–173, 2009.

[70] S. KULLBACK et R. A. LEIBLER : On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 03 1951.

[71] Yann LECUN : Generalization and network design strategies. Technical Report CRG-TR-89-4, Department of Computer Science, University of Toronto, 1989.

[72] Pamela W. LEE, Herbert C. SCHULBERG, Patrick J. RAUE et Kurt KROENKE : Concordance between the PHQ-9 and the HSCL-20 in depressed primary care patients. *Journal of Affective Disorders*, 99(1-3):139–145, 2007.

[73] D. G. LOWE : Object recognition from local scale-invariant features. *In Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2, Sep. 1999.

[74] Bruce D. LUCAS et Takeo KANADE : An iterative image registration technique with an application to stereo vision. *In Proceedings of the 7th International Joint Conference on Artificial Intelligence*, volume 2 de *IJCAI'81*, pages 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc.

[75] Patrick LUCEY, Jeffrey F. COHN, Takeo KANADE, Jason SARAGIH, Zara AMBADAR et Iain MATTHEWS : The extended cohn-kanade dataset (ck+) : A complete dataset for action unit and emotion-specified expression. *In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 94–101. IEEE, 2010.

[76] Laurens van der MAATEN et Geoffrey HINTON : Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[77] Mohammad H. MAHOOR, Steven CADAVID, Daniel S. MESSINGER et Jeffrey F. COHN : A framework for automated measurement of the intensity of non-posed facial action units. *In Computer Vision and Pattern Recognition*, pages 74–80, 2009.

[78] Soroosh MARIOORYAD et Carlos BUSSO : Correcting Time-Continuous Emotional Labels by Modeling the Reaction Lag of Evaluators. *IEEE Transactions on Affective Computing*, 6(2):97 – 108, 2015.

[79] G. B. MCBRIDE : A proposal for strength-of-agreement criteria for lins concordance correlation coefficient. *NIWA Client Report*, may 2005.

[80] Warren S. MCCULLOCH et Walter PITTS : A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[81] Gary MCKEOWN, Michel VALSTAR, Roddy COWIE, Maja PANTIC et Marc SCHRODER : The semaine database : Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012.

[82] Wes MCKINNEY : Data structures for statistical computing in python. *In Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.

[83] NATIONAL INSTITUTE OF MENTAL HEALTH : Depression, 2018. URL https://www.nimh.nih.gov/health/topics/depression/. [Consulté le 02-04-2018].

[84] Jérémie NICOLLE, Vincent RAPP, Kévin BAILLY, Lionel PREVOST et Mohamed CHETOUANI : Robust continuous prediction of human emotions using multiscale dynamic cues. *In Proceedings of the 14th ACM International Conference on Multimodal Interaction*, pages 501–508, Santa Mocina, 2012. ACM.

[85] Heekuck OH et Suresh C. KOTHARI : Adaptation of the relaxation method for learning in bidirectional associative memory. *IEEE Transactions on Neural Networks*, 5(4):576–583, 1994.

[86] Timo OJALA, Matti PIETIKÄINEN et David HARWOOD : A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, jan 1996.

[87] Travis E. OLIPHANT : *A Guide to Numpy*. Trelgol Publishing, 2006.

[88] Maja PANTIC et Leon J. M. ROTHKRANTZ : Facial action recognition for facial expression analysis from static face images. *IEEE Transactions on Systems, Man, and Cybernetics, Part B : Cybernetics*, 34(3):1449–1461, 2004.

[89] Omkar. M. PARKHI, Andrea. VEDALDI et Andrew ZISSERMAN : Deep face recognition. *In British Machine Vision Conference*, volume 1, page 6, 2015.

[90] Karl PEARSON : On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2 (11):559–572, 1901.

[91] Fabian PEDREGOSA, Gaël VAROQUAUX, Alaxandre GRAMFORT, Vincent MICHEL, Bertran THIRION, Olivier GRISEL, Mathieu BLONDEL, Peter PRETTENHOFER, Ron WEISS, Vincent DUBOURG, Jake VANDERPLAS, Alexandre PASSOS, David COURNAPEAU, Matthieu BRUCHER, Matthieu. PERROT et Édouard DUCHESNAY : Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[92] Rosalind W. PICARD : Affective Computing. *M.I.T Media Laboratory Perceptual Computing Section Technical Report*, (321):1–16, 1995.

[93] Robert PLUTCHIK : The Nature of Emotions plutchiknatureofemotions 2001. *American Science*, 89(4):344–350, 2001.

[94] Didier PUZENAT : *Parallélisme et modularité des modèles connexionnistes*. Thèse de doctorat, École Normale Supérieure de Lyon, 1997.

[95] Abdelmalek REGHIS, Frédéric ALEXANDRE, Yann BONIFACE, Abdelmalek REGHIS et A : Un réseau de neurones pour des associations multimodales. *In 14ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle-RFIA'2004*, pages 10–, Toulouse, 2004.

[96] Sebastien REGIS, Richard EMILION et Andrei DONCESCU : Opérateurs d'agrégation pour les données multiples. *Fouille de Données Complexes*, E(27):1–22, 2014.

[97] Emanuelle REYNAUD : *Modélisation connexionniste d'une mémoire associative multimodale*. Thèse de doctorat, Institut National Polytechnique de Grenoble, 2002.

[98] Emanuelle REYNAUD et Helene PAUGAM-MOISY : A multiple BAM for hetero-association and multisensory integration modelling. *In Proceedings of the International Joint Conference on Neural Networks*, volume 4, pages 2117–2122, Montreal, Canada, 2005. IEEE.

[99] Fabien RINGEVAL, Maja PANTIC, Björn SCHULLER, Michel VALSTAR, Jonathan GRATCH, Roddy COWIE, Stefan SCHERER, Sharon MOZGAI, Nicholas CUMMINS et Maximilian SCHMITT : AVEC 2017 - Real-life Depression, and Affect Recognition Workshop and Challenge. *In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge - AVEC'17*, pages 3–9, 2017.

[100] Fabien RINGEVAL, Björn SCHULLER, Michel VALSTAR, Shashank JAISWAL, Erik MARCHI, Denis LALANNE, Roddy COWIE et Maja PANTIC : AV+EC 2015 : The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data. *In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge - AVEC'15*, 2015.

[101] Fabien RINGEVAL, A. SONDEREGGER, J. SAUER et D. LALANNE : Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. *In 10th IEEE international conference and workshops on automatic face and gesture recognition*, pages 1–8. In Automatic Face and Gesture Recognition, 2013.

[102] Fabien RINGEVAL, Andreas SONDEREGGER, Juergen SAUER et Denis LALANNE : Introducing the recola multimodal corpus of remote collaborative and affective interactions. *In Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013.

[103] James RUSSELL : A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[104] James A. RUSSELL, Jo-Anne BACHOROWSKI et José-Miguel FERNÁNDEZ-DOLS : Facial and Vocal Expressions of Emotion. *Annual Review of Psychology*, 54(1):329–349, 2003.

[105] Christos SAGONAS, Epameinondas ANTONAKOS, Georgios TZIMIROPOULOS, Stefanos ZAFEIRIOU et Maja PANTIC : 300 Faces In-The-Wild Challenge : database and results. *Image Vision Computing*, 47:3–18, 2015.

[106] Enrique SÁNCHEZ-LOZANO, Paula LOPEZ-OTERO, Laura DOCIO-FERNANDEZ, Enrique ARGONES-RÚA et José Luis ALBA-CASTRO : Audiovisual three-level fusion for continuous estimation of Russell's emotion circumplex. *In Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge - AVEC'13*, pages 31–40. ACM Press, 2013.

[107] Harold SCHLOSBERG : Three Dimensions of Emotion. *The Psychological Review*, 61 (2):81–88, 1954.

[108] Bernhard SCHÖLKOPF, Alexander SMOLA et Klaus-Robert MÜLLER : Kernel principal component analysis. *In International conference on artificial neural networks*, pages 583–588. Springer, 1997.

[109] Mohammed SENOUSSAOUI, Milton SARRIA-PAJA, João F. SANTOS et Tiago H. FALK : Model Fusion for Multimodal Depression Classification and Level Detection. *In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC'14*, pages 57–63, Orlando, 2014. ACM.

[110] Hongchi SHI, Yunxin ZHAO et Xinhua ZHUANG : A general model for bidirectional associative memories. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 28(4):511–519, 1998.

[111] Narotam SINGH, Nittin SINGH et Abhinav DHALL : Continuous Multimodal Emotion Recognition Approach for AVEC 2017. *In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge - AVEC'17*, pages 11–18, Mountain View, 2017. ACM.

[112] Robert L. SPITZER, Kurt KROENKE et Janet B.W. WILLIAMS : Validation and utility of a self-report version of prime-md : the phq primary care study. *Jama*, 282 (18):1737–1744, 1999.

[113] Bo SUN, Siming CAO, Liandong LI, Jun HE et Lejun YU : Exploring Multimodal Visual Features for Continuous Affect Recognition. *In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC'16*, pages 83–88. ACM, 2016.

[114] Bo SUN et Zhaoying WANG : A Random Forest Regression Method With Selected-Text Feature For Depression Assessment. *In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge - AVEC'17*, pages 61–68. ACM, 2017.

[115] Alaa THARWAT, Tarek GABER, Abdelhameed IBRAHIM et Aboul Ella HASSANIEN : Linear discriminant analysis : A detailed tutorial. *Ai Communications*, 30:169–190„ 05 2017.

[116] Ying-li TIAN, Takeo KANADE et Jeffrey F. COHN : Facial Expression Analysis. *In Handbook of face recognition*, chapitre 11, pages 247–275. Springer, 2005.

[117] Michael E. TIPPING : Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

[118] Michel VALSTAR et Maja PANTIC : Fully Automatic Facial Action Unit Detection and Temporal Analysis. *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06).*, 2006.

[119] Michel VALSTAR, Björn SCHULLER, Kirsty SMITH, Timur ALMAEV, Florian EYBEN, Jarek KRAJEWSKI, Roddy COWIE et Maja PANTIC : Avec 2014 : 3d dimensional affect and depression recognition challenge. *In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC '14*, pages 3–10. ACM, 2014.

[120] Michel VALSTAR, Björn SCHULLER, Kirsty SMITH, Florian EYBEN, Bihan JIANG, Sebastian SCHNIEDER, Roddy COWIE et Maja PANTIC : AVEC 2013 – The Continuous Audio/Visual Emotion and Depression Recognition Challenge *. *ACM-Multimedia*, 2013.

[121] Paul VIOLA et Michael JONES : Fast Multi-view Face Detection. *In IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[122] James R. WILLIAMSON, Thomas F. QUATIERI, Brian S. HELFER, Gregory CICCARELLI et Daryush D. MEHTA : Vocal and Facial Biomarkers of Depression based on Motor Incoordination and Timing. *In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC'14*, pages 65–72, Orlando, 2014. ACM.

[123] Svante WOLD, Michael SJÖSTRÖM et Lennart ERIKSSON : Pls-regression : a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130, 2001.

[124] WORLD HEALTH ORGANIZATION : Depression, 2018. URL http://www.who.int/mediacentre/factsheets/fs369/fr/. [Consulté le 02-04-2018].

[125] Chung-Hsien WU, Jen-Chun LIN et Wen-Li WEI : Survey on audiovisual emotion recognition : databases, features, and data fusion strategies. *APSIPA Transactions on Signal and Information Processing*, 3:12, 2014.

[126] Tian WU, Yingchun YANG, Z. WU et Dongdong LI : Masc : A speech corpus in mandarin for emotion analysis and affective speaker recognition. *In 2006 IEEE Odyssey-the speaker and language recognition workshop*, pages 1 – 5, 07 2006.

[127] Zong-Ben XU, Yee LEUNG et Xiang-Wei HE : Asymmetric bidirectional associative memories. *IEEE transactions on systems, man, and cybernetics*, 24(10):1558–1564, 1994.

[128] Le YANG, Hichem SAHLI, Xiaohan XIA, Ercheng PEI, Meshia Cédric OVENEKE et Dongmei JIANG : Hybrid Depression Classification and Estimation from Audio Video and Text Information. *In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC'14*, pages 45–51. ACM, 2017.

[129] Guermeur YANN : *SVM Multiclasses , Théorie et Applications*. Thèse de doctorat, 2007.

[130] Georgios. N. YANNAKAKIS, Roddy COWIE et Carlos BUSSO : The ordinal nature of emotions. *In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, volume 00, pages 248–255, Oct. 2017.

[131] G. ZHAO et M. PIETIKAINEN : Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, June 2007. ISSN 0162-8828.