UNIVERSITÉ DE LORRAINE | BIBLIOTHÈQUES UNIVERSITAIRES

AVERTISSEMENT

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite encourt une poursuite pénale.

Contact bibliothèque : ddoc-theses-contact@univ-lorraine.fr
(Cette adresse ne permet pas de contacter les auteurs)

LIENS

Code de la Propriété Intellectuelle. articles L 122. 4
Code de la Propriété Intellectuelle. articles L 335.2- L 335.10
http://www.cfcopies.com/V2/leg/leg_droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Reconnaissance automatique sans-contact de l'état affectif de la personne par fusion physio-visuelle à partir de vidéo du visage

THÈSE

présentée et soutenue publiquement le 15 juin 2023

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention Automatique, Traitement du signal et des images, Génie informatique)

par

Yassine Ouzar

Composition du jury

Président :	Bruno Emile (Pr., Laboratoire Prisme, Université d'Orléans)
Rapporteurs :	Yannick Benezeth (HDR., Laboratoire ImViA, Université Bourgogne Frédéric Vanderhaegen (Pr., Laboratoire LAMIH, Université Polytechnique Hauts-de-France)
Examinateurs :	Samia Bouchafa-Bruneau (Pr., Laboratoire IBISC, Université d'Évry (Paris-Saclay)) Bruno Emile (Pr., Laboratoire Prisme, Université d'Orléans)
Encadrants :	Frédéric Bousefsaf (MCF., Laboratoire LCOMS, Université de Lorraine) Choubeila Maaoui (Pr., Laboratoire LCOMS, Université de Lorraine)

Laboratoire de Conception, Optimisation et Modélisation des Systèmes, EA 7306

Mis en page avec la classe thesul.

Remerciements

Je tiens à exprimer ma profonde gratitude pour l'honneur et l'opportunité qui m'ont été accordés de soumettre mon manuscrit de thèse intitulé "Reconnaissance automatique sans-contact de l'état affectif de la personne par fusion physio-visuelle à partir de vidéo du visage". C'est avec un mélange de fierté, d'humilité et de reconnaissance que je partage ce travail qui représente des années de recherche, d'efforts et de dévouement.

Je tiens à remercier chaleureusement Mme Choubeila MAAOUI et Mr Frédéric BOUSEFSAF pour leur soutien indéfectible, leurs conseils éclairés et leur expertise précieuse tout au long de ce parcours. Votre encadrement attentif et vos encouragements constants ont été essentiels pour l'aboutissement de ce projet. Votre engagement envers l'excellence académique et votre passion pour la recherche ont été une source d'inspiration inestimable.

Mes remerciements vont également à mes collègues de laboratoire, mes amis et ma famille, qui m'ont soutenu tout au long de ce voyage. Votre encouragement, votre soutien émotionnel et votre croyance en mes capacités ont été des moteurs indispensables pour surmonter les défis et les moments de doute.

Je remercie vivement les personnes avec qui j'ai partagé le bureau pendant ces années, Djameleddine, Lynda, Imene, Rachid, Youssefe, Hafida, Yanis, Anis et Gaetan. Merci particulièrement à Djameleddine, Narimane et Anis pour votre soutien et tous les moments passés ensemble et qui m'ont permis d'oublier le stress du travail.

Enfin, je voudrais remercier tous les membres du laboratoire LCOMS. Je remercie particulièrement les camarades doctorants, post-doctorants et stagiaires que j'ai eu à fréquenter pendant ces dernières années. Je les remercie pour leurs gentillesses, leurs conseils et surtout pour leurs amitiés. Je suis heureux d'avoir pu partager avec eux de nombreux moments de convivialité, de débats enrichissants et d'activités diverses.

Encore une fois, merci du fond du cœur à tous ceux qui ont contribué de près ou de loin à l'achèvement de ce travail. C'est avec une immense gratitude que je présente ce manuscrit de thèse, en espérant qu'il contribuera de manière significative au domaine de l'informatique affective et ouvrira la voie à de nouvelles avancées passionnantes.

To all those who make me feel good

The question is not whether
intelligent machines can have emotions,
but whether machines can be
intelligent without any emotions.
Marvin Minsky

Résumé

La reconnaissance automatique de l'état affectif reste un sujet difficile en raison de la complexité des émotions / stress, qui impliquent des éléments expérientiels, comportementaux et physiologiques. Comme il est difficile de décrire l'état affectif de la personne de manière exhaustive en termes de modalités uniques, des études récentes se sont concentrées sur des stratégies de fusion afin d'exploiter la complémentarité des signaux multimodaux.

L'objectif principal de cette thèse consiste à étudier la faisabilité d'une fusion physio-visuelle pour la reconnaissance automatique de l'état affectif de la personne (émotions / stress) à partir des vidéos du visage. La fusion des expressions faciales et des signaux physiologiques permet de tirer les avantages de chaque modalité. Les expressions faciales sont simples à acquérir et permettent d'avoir une vision externe de l'état affectif, tandis que les signaux physiologiques permettent d'améliorer la fiabilité et relever le problème des expressions faciales contrefaites.

Les recherches développées dans cette thèse se situent à l'intersection de l'intelligence artificielle, l'informatique affective ainsi que l'ingénierie biomédicale. Notre contribution s'axe sur deux aspects. Nous proposons en premier lieu une nouvelle approche bout-en-bout permettant d'estimer la fréquence cardiaque à partir d'enregistrements vidéo du visage à l'aide du principe de photopléthysmographie par imagerie (iPPG). La méthode repose sur un réseau spatio-temporel profond (X-iPPGNet) qui apprend le concept d'iPPG à partir de zéro, sans incorporer de connaissances préalables ni passer par l'extraction manuelle des signaux iPPG.

Le second aspect porte sur une chaine de traitement physio-visuelle pour la reconnaissance automatique des émotions spontanées et du stress à partir des vidéos du visage. Le modèle proposé comprend deux étages permettant d'extraire les caractéristiques de chaque modalité. Le pipeline physiologique est commun au système de reconnaissance d'émotion et celui du stress. Il est basé sur MTTS-CAN, une méthode récente d'estimation du signal iPPG. Deux modèles neuronaux distincts ont été utilisés pour prédire les émotions et le stress de la personne à partir des informations visuelles contenues dans la vidéo (e.g. expressions faciales) : un réseau spatio-temporel combinant le module Squeeze-Excitation et l'architecture Xception pour estimer l'état émotionnel et une approche d'apprentissage par transfert pour l'estimation du niveau de stress. Cette approche a été privilégiée afin de réduire les efforts de développement et surmonter le problème du manque de données. Une fusion des caractéristiques physiologiques et des expressions faciales est ensuite effectuée pour prédire les états émotionnels ou de stress.

Mots-clés: émotion, stress, fusion physio-visuelle, vidéo, photoplethysmographie par imagerie, fréquence cardiaque, expressions faciales, apprentissage profond.

Abstract

Human affective state recognition remains a challenging topic due to the complexity of emotions, which involves experiential, behavioral, and physiological elements. Since it is difficult to comprehensively describe emotion in terms of single modalities, recent studies have focused on artificial intelligence approaches and fusion strategy to exploit the complementarity of multimodal signals using artificial intelligence approaches.

The main objective is to study the feasibility of a physio-visual fusion for the recognition of the affective state of the person (emotions/stress) from facial videos. The fusion of facial expressions and physiological signals allows to take advantage of each modality. Facial expressions are easy to acquire and provide an external view of the affective state, while physiological signals improve reliability and address the problem of falsified facial expressions.

The research developed in this thesis lies at the intersection of artificial intelligence, affective computing, and biomedical engineering. Our contribution focuses on two points. First, we propose a new end-to-end approach for instantaneous pulse rate estimation directly from facial video recordings using the principle of imaging photoplethysmography (iPPG). This method is based on a deep spatio-temporal network (X-iPPGNet) that learns the iPPG concept from scratch, without incorporating prior knowledge or going through manual iPPG signal extraction.

The second contribution focuses on a physio-visual fusion for spontaneous emotions and stress recognition from facial videos. The proposed model includes two pipelines to extract the features of each modality. The physiological pipeline is common to both the emotion and stress recognition systems. It is based on MTTS-CAN, a recent method for estimating the iPPG signal, while two distinct neural models were used to predict the person's emotions and stress from the visual information contained in the video (e.g. facial expressions) : a spatio-temporal network combining the Squeeze-Excitation module and the Xception architecture for estimating the emotional state and a transfer learning approach for estimating the stress level. This approach reduces development effort and overcomes the lack of data. A fusion of physiological and facial features is then performed to predict the emotional or stress states.

Keywords: physio-visual fusion, iPPG, video, deep learning, emotion, stress, pulse rate.

الملخص

يكمن البحث الذي تم تطويره في هذه الأطروحة في تقاطع الذكاء الاصطناعي والحوسبة العاطفية والهندسة الطبية الحيوية. الهدف الرئيسي هو دراسة جدوى الاندماج الفيزيو- بصري للتعرف على الحالة العاطفية للإنسان من مقاطع فيديو الوجه. مساهمتنا تركز على نقطتين أولاً ، اقترحنا نهجًا جديدًا شاملًا لتقدير معدل النبض اللحظي مباشرةً من تسجيلات فيديو الوجه باستخدام مبدأ تصوير التصوير الضوئي (iPPG). تعتمد الطريقة على شبكة مكانية وزمنية عميقة (X-iPPGNet) تتعلم مفهوم iPPG من الصفر ، دون دمج معارف مسبقة أو المرور عبر استخراج إشارة iPPG .

تركز المساهمة الثانية على الاندماج الفيزيو- بصري للتعرف على العواطف العفوية و الإجهاد من مقاطع فيديو الوجه. يتضمن النموذج المقترح خطي أنابيب لاستخراج ميزات كل طريقة. خط الأنابيب الفسيولوجي مشترك لكل من أنظمة التعرف على العاطفة والتوتر و يعتمد على MTTS-CAN ، وهي طريقة حديثة لتقدير إشارة iPPG ، بينما تم استخدام نموذجين عصبيين خاصين بالتنبؤ بمشاعر الشخص والتوتر من المعلومات المرئية الواردة في الفيديو (مثل تعابير الوجه) : شبكة مكانيةزمانية تجمع بين وحدة Squeeze-Excitation وهيكل Xception لتقدير الحالة العاطفية ونهج نقل التعلم لتقدير مستوى الإجهاد. تم تفضيل هذا النهج من أجل تقليل جهود التطوير والتغلب على مشكلة نقص البيانات. في الأخير يتم دمج الخصائص الفسيولوجية وتعابير الوجه للتنبؤ بالحالات العاطفية أو التوتر.

Liste des publications

Revues internationales

- **OUZAR, Yassine**, DJELDJLI, Djamaleddine, BOUSEFSAF, Frédéric, and MAAOUI, Choubeila. X-iPPGNet : A novel one stage deep learning architecture based on depthwise separable convolutions for video-based pulse rate estimation. *Computers in Biology and Medicine, 2023.* **(IF = 6.698)**
- BOUSEFSAF, Frédéric, DESQUINS, Théo, DJELDJLI, Djamaleddine, **OUZAR, Yassine**, MAAOUI, Choubeila, and PRUSKI, Alain. Estimation of Blood Pressure Waveform from Facial Video Using a Deep U-Shaped Network and the Wavelet Representation of Imaging Photoplethysmographic Signals, *Biomedical Signal Processing and Control, 2021.* **(IF = 5.076)**
- BOUSEFSAF, Frédéric, DJELDJLI, Djamaleddine, **OUZAR, Yassine**, MAAOUI, Choubeila, and PRUSKI, Alain. iPPG 2 cPPG : Reconstructing contact from imaging photoplethysmographic signals using U-Net architectures. *Computers in Biology and Medicine, 2021, vol. 138, p. 104860.* **(IF = 6.698)**

Conférences internationales

- **OUZAR, Yassine**, Lagha, Lynda, BOUSEFSAF, Frédéric, and MAAOUI, Choubeila. Multimodal stress state detection from facial videos using physiological signals and facial features. *In : Proceedings of the IEEE/CVF International Conference on Pattern Recognition, 2022.*
- **OUZAR, Yassine**, BOUSEFSAF, Frédéric, DJELDJLI, Djamaleddine, and MAAOUI, Choubeila. Video-based multimodal spontaneous emotion recognition using facial expressions and physiological signals, *In : Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition, 2022.*

- **OUZAR, Yassine**, DJELDJLI, Djamaleddine, BOUSEFSAF, Frédéric, and MAAOUI, Choubeila. LCOMS Lab's approach to the Vision for Vitals (V4V) Challenge. *In : Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021. p. 2750-2754.*

Conférences nationales et colloques

- BOUSEFSAF, Frédéric, DJELDJLI, Djamaleddine, **OUZAR, Yassine**, MAAOUI, Choubeila, and PRUSKI, Alain.Transformée en ondelettes et IA pour la reconstruction d'un signal PPG en contact à partir de sa version sans contact. *In GRETSI'22, XXVIIIème Colloque Francophone de Traitement du Signal et des Images, Nancy, Nov. 2022.*
- BOUSEFSAF, Frédéric, DESQUINS, Théo, DJELDJLI, Djamaleddine, **OUZAR, Yassine**, MAAOUI, Choubeila, and PRUSKI, Alain. Estimation of blood pressure waveform from facial video using a deep U-shaped network and the wavelet representation of imaging photoplethysmographic signals, *Conférence Handicap 2022, 129–134, Paris, Juin 2022.*
- **OUZAR, Yassine**, BOUSEFSAF, Frédéric, and MAAOUI, Choubeila. Mesure sans contact de la fréquence par caméra basée sur l'apprentissage profond. *Colloque Jeunes Chercheurs IFRATH, 2021.*
- **OUZAR, Yassine**, BOUSEFSAF, Frédéric, and MAAOUI, Choubeila. Reconnaissance multimodale des émotions spontanées par caméra basée sur les expressions faciales et les signaux physiologiques. *In Journées de printemps de la SAGIP 2022, Bidart, May 2022.*
- **OUZAR, Yassine**, BOUSEFSAF, Frédéric, and MAAOUI, Choubeila, and CHELGHOUM, Kamel. Système bimodal pour la reconnaissance des émotions basé sur l'apprentissage profond. *In 27e Journées STP du GdR MACS, Nantes, Feb. 2020.*

Table des matières

Introduction générale

Le visage humain est une riche source d'informations. Il se caractérise entre autres par sa grande expressivité dans la transmission des émotions. Bien que les émotions soient multi componentielles et se manifestent sur différents canaux (verbal, non verbal et vocal), les expressions faciales sont la modalité la plus étudiée et elle est considérée comme le canal majeur de communication émotionnelle car elle est visible et facilement observable. Les travaux de la littérature scientifique exhibent des résultats impressionnants sur des bases de données actées et acquises dans des conditions contrôlées. Néanmoins, les recherches se sont rarement confrontées à des situations réelles ou à des environements naturels, les performances et la fiabilité pouvant se dégrader considérablement. Outre les conditions environnementales (angles de caméra, conditions d'éclairage et occultation du visage) et la capacité des personnes à contrôler et à simuler leurs émotions, les expressions faciales sont également plus affectées par les différences sociales et culturelles. L'expressivité humaine peut varier d'un individu à l'autre et peut être exprimée différemment selon la situation et l'état psychologique de la personne. Dans certaines cultures, les expressions telles que la colère ou le chagrin sont considérées comme déshonorantes et sont découragées, ce qui conduit la personne à remplacer son sentiment par un faux sourire [1]. De plus, les expressions faciales peuvent être un mélange de différents états émotionnels qui se produisent simultanément ou peuvent ne pas être exprimées du tout. Par conséquent, l'utilisation d'expressions faciales pour identifier l'état affectif d'une personne peut conduire à des fausses inférences.

Les émotions et le stress entraînent également des changements physiologiques en réponse à des stimuli externes. Ces réactions peuvent être capturées par des dispositifs appropriés et peuvent à leur tour être utilisées pour déduire l'état affectif du sujet. Nombreuses études ont exploité les signaux physiologiques pour la reconnaissance des émotions et du stress afin de surmonter les limitations des expressions faciales. L'avantage d'utiliser des paramètres physiologiques pour évaluer l'état affectif plutôt que les expressions faciales réside principalement dans leur grande fiabilité. Ils s'activent et s'inhibent en réponse au système nerveux autonome, qui

est activé involontairement et ne peut donc pas être contrôlé. Cependant, ils sont contraignants dans leur acquisition et ne peuvent être utilisés en dehors du laboratoire.

Afin d'améliorer les performances des systèmes affectifs, plus d'attention a été accordée à la fusion de deux ou plusieurs modalités pour tirer les avantages de chacune. La combinaison des expressions faciales et des signaux physiologiques (physio-visuelle) peut améliorer la précision et offrir une plus grande fiabilité en exploitant les caractéristiques de chaque modalité. Elle permet également de recueillir continuellement des informations sur l'état affectif de la personne malgré d'éventuelles acquisitions manquantes ou des données corrompues. Celles-ci peuvent survenir lors de l'utilisation d'une seule modalité dans un environnement bruyant ou dans le cas d'une expression faciale contrefaite. En outre, la fusion physio-visuelle peut aider à compenser les erreurs et à résoudre les ambiguïtés en apprenant des représentations utiles de données de nature différente. Cependant, la principale limitation est liée à l'intrusion des dispositifs d'acquisition de données physiologiques qui sont psychologiquement stressants et peuvent perturber l'utilisateur. Cela peut modifier l'état émotionnel et donc affecter la précision des estimations.

Les signaux physiologiques sont habituellement mesurés à l'aide de dispositifs en contact qui doivent être attachés à des parties spécifiques du corps et nécessitent certaines conditions pour obtenir des mesures précises. Cependant, de nombreuses recherches au cours de la dernière décennie ont montré que des technologies alternatives sans contact, telles que les caméras standards, peuvent être utilisées pour mesurer l'activité cardiaque d'une personne. En plus d'être abordables, les caméras sont déjà intégrées dans de nombreux périphériques informatiques courants tels que les ordinateurs portables et les smartphones. Bien que les premiers travaux montrant la possibilité de mesurer l'activité cardiaque à l'aide d'une simple caméra remontent à une quinzaine d'années, les recherches sur cette thématique sont en plein essor depuis les cinq dernières années, notamment grâce aux avancées réalisées dans le domaine de l'intelligence artificielle pour les tâches de vision par ordinateur. L'intérêt pour cette méthode de mesure ne cesse de croître et de plus en plus de laboratoires et d'équipes de recherche s'y intéressent.

Le travail de thèse s'inscrit dans le cadre de l'informatique affective qui est une thématique reconnue dans le domaine de l'intelligence artificielle. L'objectif principal de la thèse consiste à étudier la faisabilité de la fusion physio-visuelle pour la reconnaissance de l'état affectif à partir des vidéos du visage. Le but essentiel est de surmonter le problème de falsification des émotions et améliorer la précision à travers la fusion de signaux externes (les expressions faciales) et de signaux internes (les signaux physiologiques). A la différence des travaux existants, les paramètres physiologiques seront mesurés à distance en utilisant le principe de la photopléthysmographie par imagerie (IPPG). Il s'agit d'une technique de mesure non invasive permettant d'estimer un

ensemble de fonctions vitales par analyse video. Elle repose sur l'observation des fines fluctuations de la couleur de la peau d'une personne associées à la variation du volume sanguin.

Notre contribution s'axe sur deux points. Nous proposons en premier lieu une nouvelle approche bout-en-bout permettant d'estimer la fréquence cardiaque à partir de vidéos du visage. Il s'agit d'un réseau spatio-temporel profond (X-iPPGNet) qui apprend le concept d'iPPG à partir de zéro, sans incorporer de connaissances préalables ni passer par l'extraction manuelle des signaux iPPG. X-iPPGNet fusionne l'extraction du signal iPPG et la prédiction de la fréquence cardiaque en une seule étape. Nous nous appuyons sur la capacité des modèles d'apprentissage profond à apprendre implicitement des informations utiles directement à partir des données brutes.

Le second point de notre contribution est axé sur une chaine de traitement physio-visuelle pour la reconnaissance automatique des émotions spontanées et du stress à partir des vidéos du visage. Le modèle proposé comprend deux pipelines permettant d'extraire les caractéristiques de chaque modalité. Le pipeline physiologique est commun au système de reconnaissance d'émotion et celui du stress. Il est basé sur une méthode de pointe appelée MTTS-CAN. Deux pipelines visuels différents sont proposés : (i) nous proposons un réseau de neurones spatio-temporel qui combine le module Squeeze-Excitation et l'architecture Xception pour la reconnaissance des émotions et (ii) nous adoptons une stratégie d'apprentissage par transfert pour la reconnaissance du stress. Finalement, nous fusionnons les caractéristiques physiologiques et les expressions faciales pour prédire les états émotionnels ou de stress.

Le manuscrit est organisé en quatre chapitres regroupés en deux parties de deux chapitres chacune.

Partie I

La premère partie présente un état de l'art sur les émotions et le stress et sur la mesure sans contact de la fréquence cardiaque basée sur la vidéo.

- Le premier chapitre introduit en premier lieu des notions de base sur les émotions et le stress telles que leurs définitions, leurs représentations et les techniques d'élicitation. Ensuite, un résumé des travaux de la littérature scientifique est présenté, des méthodes unimodales aux approches multimodales.
- Le deuxième chapitre présente un état de l'art sur les techniques de mesure de l'activité cardiaque en mettant en avant la photopléthysmographie par imagerie, ses applications et ses défis. Puis, nous décrivons en détails la chaine de mesure de la fréquence cardiaque par

iPPG. Nous clôturons le chapitre par une revue des travaux de la littérature scientifique.

Partie II

La deuxième partie présente la contribution de notre travail. Elle s'articule autour de la mesure de l'activité cardiaque et de la reconnaissance multimodale de l'état affectif basée sur les expressions faciales et les signaux physiologiques.

- Le troisième chapitre est consacré à la mesure visuelle de la fréquence cardiaque à l'aide de l'iPPG. Nous présentons d'abord les bases de données utilisées pour l'apprentissage et la validation du modèle proposé. Nous décrivons ensuite l'architecture générique de notre modèle (X-iPPGNet) et les différents détails d'implémentation, puis nous exposons et comparons noss résultats à ceux des méthodes concurrentes. Les expériences approfondies menées pour explorer les critères impactant les performances de notre approche. Enfin, nous clôturons le chapitre par une discussion et une conclusion des résultats obtenus.
- Le quatrième et le dernier chapitre porte sur la fusion physio-visuelle pour la reconnaissance de l'état émotionnel et de stress en utilisant une approche sans contact et mono-capteur. Ce chapitre se compose de deux parties. La première section est consacrée à la reconnaissance des émotions, tandis que la seconde section est dédiée à l'identification de l'état de stress. Chacune des sections décrit le pipeline visuel et physiologique constituant le système affectif et présente les résultats expérimentaux en utilisant les modalités visuelles ou physiologiques seules et ensembles.

Première partie

Etat de l'art

1

Etat de l'art sur l'émotion et le stress

Sommaire

1.1 Introduction

Ce chapitre présente un état de l'art sur l'émotion et le stress :

La première section présente d'abord quelques notions concernant les émotions telles que leurs définitions, leurs théories, et leurs composantes. Ensuite, nous présentons une description des méthodes de la littérature scientifique sur la reconnaissance automatique des émotions.

Dans la deuxième section, nous définissons en premier lieu le stress et nous présentons ses différentes théories et ses techniques d'élicitation. Par la suite, nous présentons un état de l'art

sur les systèmes de reconnaissance de stress.

1.2 Emotion

1.2.1 Définition de l'émotion

Il existe plusieurs définitions pour le terme « émotion » et les théoriciens et les psychologues des émotions ne s'accordent pas sur la définition du concept émotionnel. En effet, la complexité de la nature de l'émotion humaine, qui fait intervenir des éléments cognitifs, comportementaux et physiologiques, ainsi que la diversité des formes avec lesquelles elle s'exprime ont longtemps empêché l'accès à une définition univoque du concept. Au début des années 1980, Kleinginna [2] avait déjà recensé près d'une centaine de définitions différentes présentées dans la littérature scientifique. La plupart des définitions décrivent l'émotion comme un processus psychophysiologique qui se produit automatiquement en tant que réaction de l'organisme à un stimulus ou à une situation particulière. Les psychologues pensent que l'émotion est une attitude subjective générée par l'expérience d'une personne à des évènements extérieurs, ainsi qu'une réponse instinctive coordonnée faite par le corps, qui peut inclure les effets conjoints du langage, du comportement et de l'esprit [3]. Le psychologue Ekman, qui a passé la majeure partie de sa carrière à comprendre les nuances physiques et sociales de l'émotion, attribue la fonction de l'émotion à la mobilisation d'un organisme pour faire face rapidement à des rencontres interpersonnelles importantes [4].

Les théories des émotions existantes peuvent fondamentalement être divisées en deux catégories, selon lesquelles les émotions sont soit cognitives (c'est-à-dire liées au cerveau), soit physiques (liées au corps) [5]. D'après Larousse [6], l'émotion est définie comme une réaction affective transitoire d'assez grande intensité, habituellement provoquée par une stimulation venue de l'environnement. Le Dictionary of Cognitive Psychology [7] définit l'émotion comme un état mental tandis que Cannon la considère comme étant essentiellement une réponse somatique [8]. William James [9] a associé les émotions aux changements du système nerveux autonome (SNA) qui se produisent à un niveau viscéral après une stimulation. Schachter et Singer [10] ont souligné que le cerveau est strictement lié aux états affectifs. Cabanac [11] a interprété les émotions comme un processus continu de sentiments subjectifs.

À l'heure actuelle, la définition des émotions ne fait pas consensus. Dans de nombreux cas, les émotions sont généralement associées à la personnalité, à l'humeur et au désir d'une personne [12, 13]. Dans le cadre de cette thèse, on considère que l'état émotionnel des humains peut être obtenu à partir d'un large éventail d'indices comportementaux et de signaux qui sont disponibles par le biais d'une expression ou d'une perception visuelle, auditive et physiologique.

1.2.2 Théories des émotions

L'étude des phénomènes émotionnels remonte à plusieurs siècles et se poursuit jusqu'à aujourd'hui, gagnant en complexité et en ampleur. Les notions fondamentales dans la recherche sur les émotions se trouvent dans des théories relativement récentes. De nombreuses théories sont apparues entre le XIIIe et le XVIIe siècle, mais les plus populaires sont celles de René Descartes et de Baruch Spinoza, considérés comme des figures importantes de l'histoire de la philosophie en général et des émotions en particulier. Descartes a proposé une théorie fondée sur l'idée que les émotions sont une combinaison de primitifs émotionnels [14]. Spinoza, quant à lui, définit un espace émotionnel dans lequel toutes les expériences émotionnelles peuvent être décrites selon 3 dimensions : joie, désir et tristesse [15]. Malgré la nette distinction entre leurs définitions fondamentales, le désaccord majeur entre les deux théories réside dans le fait que Descartes établit une séparation nette entre l'esprit (cognition, cerveau rationnel) et le corps (émotion, instincts), tandis que Spinoza les unifie et voit les émotions comme les fondements de l'esprit. Plus tard, au XIXème siècle, Darwin introduit la théorie de l'évolution [16]. Selon lui, les émotions sont innées, universelles et font partie d'un patrimoine génétique. Il lie également les émotions au système nerveux.

William James et Carl Lange associent l'émotion à une approche physiologique [9]. Pour eux la prise de conscience des modifications physiologiques suite à la perception d'un stimulus constitue l'émotion. Les émotions sont donc secondaires et résultent de phénomènes physiologiques. Un stimulus tel que la présence d'un serpent, provoque une augmentation du rythme cardiaque qui à son tour évoque l'émotion de la peur. Cannon [8] et Bard [17] ont exprimé un désaccord avec la théorie de James-Lange, disant que les émotions ne sont pas simplement des phénomènes physiologiques. Pour eux les réponses physiologiques ne peuvent pas expliquer l'expérience émotionnelle. L'argument est que la réponse physiologique comme le changement du rythme cardiaque est trop lente pour induire une expression émotionnelle aussi rapide et intense. Lors de la présence d'un stimuli externe une activation de certains processus dans le cerveau implique une réponse physiologique et en même temps une réponse émotionnelle. Plus tard, Schachter et Singer [10] ont proposé une théorie sur les émotions basée sur deux facteurs. L'évaluation de l'expérience physiologique définie et détermine l'expérience émotionnelle. Après qu'une réaction physiologique ait lieu vis-à-vis d'un stimuli externe, l'interprétation de cette réponse physiologique est le facteur déterminant quant à l'émotion exprimée. La figure 1.1 montre une comparaison schématique des théories physiologiques de James-Lange, Cannon-Bard et Schachter-Singer. Certains des concepts fournis par Spinoza, Descartes et Darwin sont les fondements de théories et de modèles plus récents. Les théories de James-Lange, Cannon-Bard et Schachter-Singer, toutes trois fondées sur

une approche physiologique, ont eu une influence considérable sur la recherche d'une définition de l'émotion. Ce domaine de recherche regroupe aujourd'hui des chercheurs de domaines aussi variés que la philosophie, la psychologie, la sociologie, la neurophysiologie mais aussi l'intelligence artificielle et la robotique.

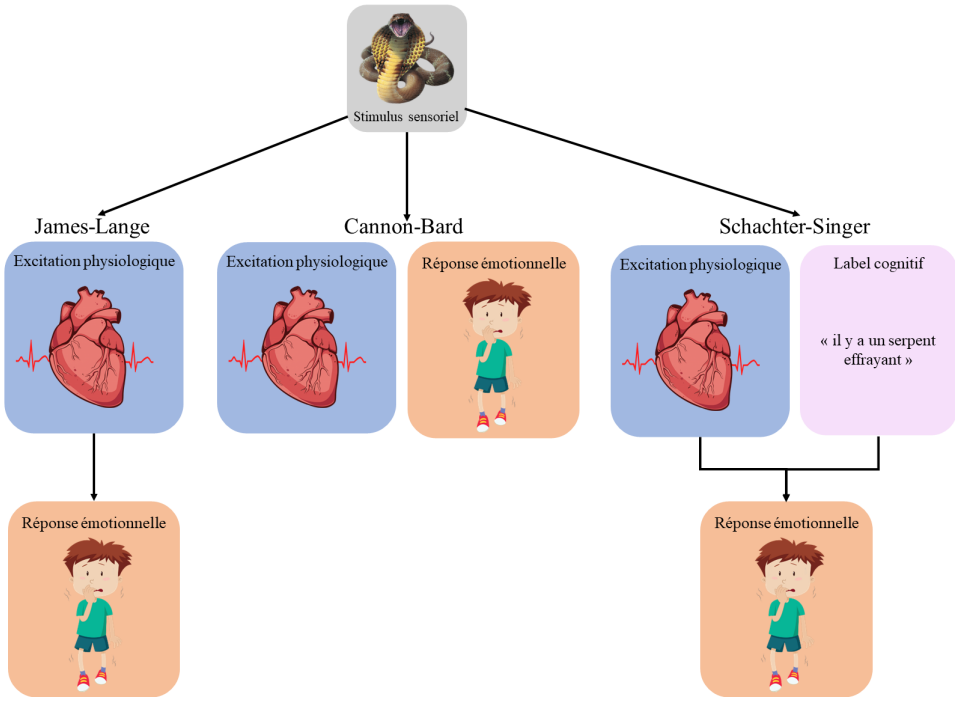


FIGURE 1.1 – Les théories physiologiques de l'émotion.

1.2.3 Elicitation de l'émotion

L'élicitation de l'émotion est la première étape pour mener des expériences sur la reconnaissance de l'état affectif de la personne. Différentes stratégies d'induction de l'émotion ont été mises en œuvre. Elles peuvent être classées en deux catégories : les émotions induites et celles naturellement suscitées. Les émotions induites font référence aux types d'émotions provoquées par des stimuli délibérément choisis pour induire différentes émotions chez les sujets, tandis que les expressions naturalistes font référence à des situations naturelles et à des stimuli incontrôlables. D'autre part, au lieu d'évoquer des émotions chez les sujets, certaines études ont utilisé des expressions actées, ce qui signifie que des émotions spécifiques sont exprimées intentionnellement par des sujets sélectionnés dans un environnement de laboratoire contrôlé.

Les émotions induites sont très populaires dans la recherche sur la reconnaissance des émotions. Picard et al. [18] ont déclaré que la principale préoccupation de l'élicitation des émotions

consiste à choisir un stimulus approprié pour induire une émotion spécifique. Les méthodes d'élicitation les plus courantes consistent à montrer aux utilisateurs diverses ressources telles que des sons [19], des vidéos [20], des images [21] et à invoquer des émotions par le biais de souvenirs passés grâce au rappel autobiographique [22]. Il est également possible d'utiliser d'autres moyens pour induire des émotions, par exemple les odeurs [23], les conversations ou les débats [24], la conduite automobile [25] et la réalité virtuelle [26]. Quelle que soit la capacité d'un stimulus à induire l'émotion, la conscience qu'a le sujet du but de l'expérience pourrait avoir un impact sur la fiabilité des données obtenues.

1.2.4 Représentation des émotions

Les recherches sur les émotions sont florissantes et ne cessent de progresser, mais le débat se poursuit sur la nature et les causes des émotions, leurs mécanismes biologiques, leurs catégories et leur rôle dans nos activités quotidiennes [27]. La représentation de l'émotion est l'un des sujets fondamentaux de la recherche sur les émotions qui permet de les définir et de les modéliser de manière quantitative en attribuant des étiquettes à chaque état émotionnel. Dans la littérature scientifique, il y a généralement deux types de représentations des émotions qui sont couramment utilisées. La première représentation comprend le modèle dimensionnel tandis que la seconde représentation est catégorielle. Ces deux représentations ont des caractéristiques spécifiques que nous détaillerons dans ce qui suit.

1.2.4.1 Approche dimensionnelle

L'approche dimensionnelle a été introduite par Wilhelm Wundt [28]. Elle décrit les émotions comme une combinaison de plusieurs dimensions psychologiques où chaque dimension représente une propriété fondamentale commune à toutes les émotions. Les dimensions peuvent être un axe de plaisir et de déplaisir, d'éveil ou d'ennui, de nervosité, de puissance et bien d'autres selon les modèles. Le principe de cette représentation est illustré sur la Figure 1.2.

Le modèle dimensionnel le plus connu et le plus utilisé est le modèle dimensionnel de la valence et de l'activation. La valence représente la polarité de l'émotion, de désagréable à agréable. En revanche, la dimension de l'activation représente le niveau d'excitation ou d'inhibition des émotions, allant de la somnolence ou de l'ennui à une excitation extrême. La valence permet de distinguer les émotions positives et agréables, comme la joie, des émotions négatives et désagréables, comme la colère. L'activation représente le niveau d'excitation corporelle, qui transparaît par nombre de réactions physiologiques, comme l'accélération du cœur ou la transpiration. Certains travaux ajoutent toutefois une troisième dimension, trouvant les deux premières insuffisantes. Cette troi-

sième dimension est nommée contrôle ou dominance [29]. Elle correspond à l'effort de la personne pour contrôler son émotion et permet de distinguer les émotions provoquées par le sujet lui-même ou par l'environnement.

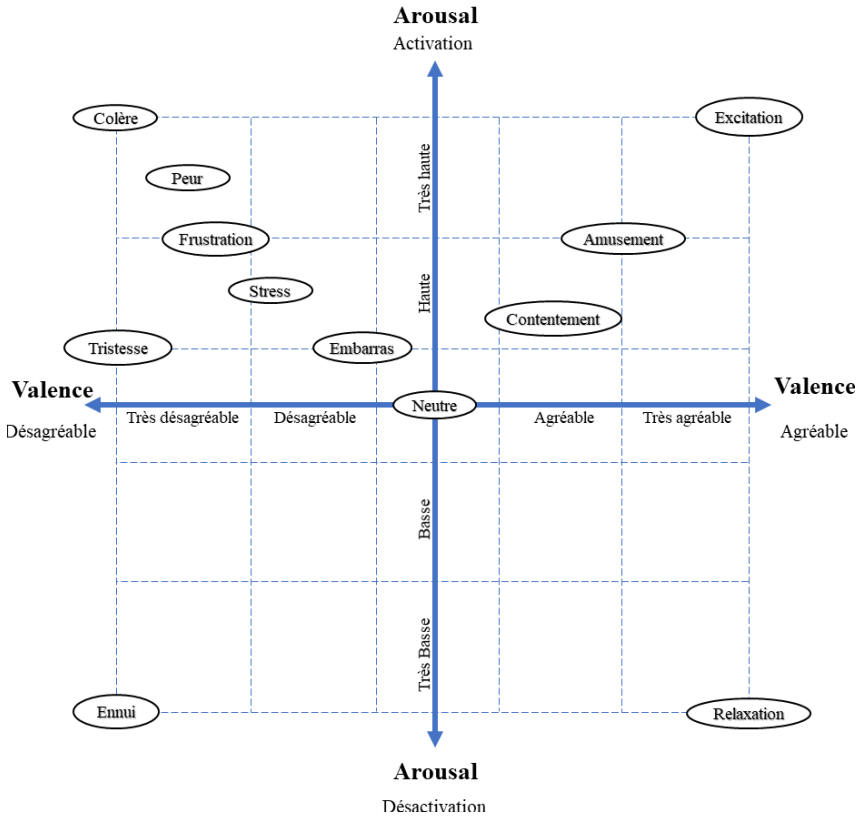


FIGURE 1.2 – La représentation de quelques émotions sur deux axes en utilisant l'approche dimensionnelle [30].

1.2.4.2 Approche catégorielle

La représentation catégorielle, appelée aussi modèle émotionnel discret, décrit la présence de certaines émotions de base universellement répandues dans toutes les cultures et sont considérées par la plupart des théoriciens comme universelles et innées [4]. Plusieurs modèles émotionnels ont été élaborés pour permettre à la communauté scientifique de différencier les émotions en utilisant les caractéristiques explicites à chacune d'entre elles. En 1962, Tomkins a suggéré qu'il existait huit émotions de base : la peur, la colère, l'angoisse, la joie, le dégoût, la surprise, la confiance et l'anticipation [31]. Plus tard, Plutchik a proposé un modèle multidimensionnel de roue qui comprend huit émotions de base différentes de celles de Tomkins : la peur, la colère, la tristesse, la joie, le dégoût, la surprise, la confiance et l'anticipation [32]. Ce modèle de roue

repose sur trois notions : l'aspect dimensionnel dans lequel on dénombre l'intensité, la similitude et la polarité ; la notion de persistance et la notion de pureté (Figure 1.3).

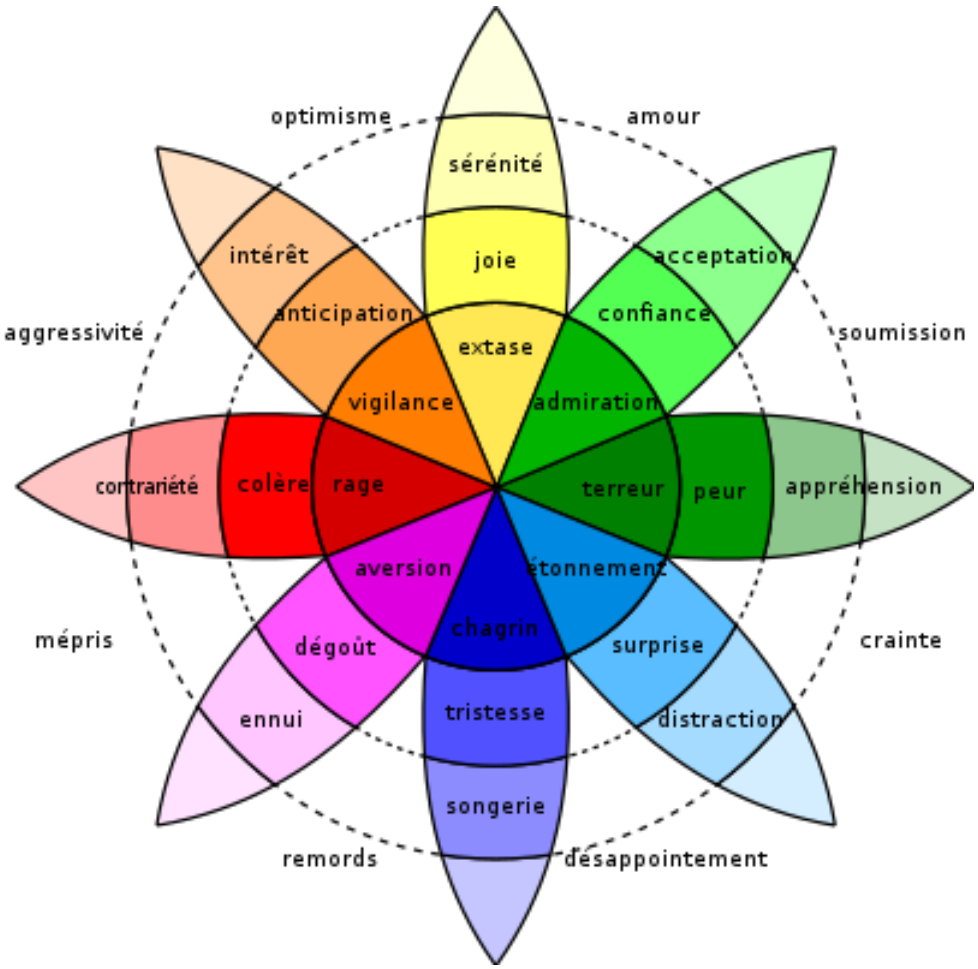


FIGURE 1.3 – La roue des émotions de Plutchik [33].

Les émotions étant rarement pures mais plutôt issues de combinaisons, toutes les émotions humaines peuvent être formées par une combinaison d'une ou plusieurs de ces émotions de base. Ekman a identifié seulement six émotions primaires : la peur, la colère, la tristesse, la joie, le dégoût et la surprise [4]. Ces émotions sont universelles, biologiquement vécues par tous les humains et largement acceptées dans ce domaine de recherche.

1.2.5 Composantes des émotions

Les émotions sont des processus psychophysiologiques complexes associés à de nombreuses activités externes et internes. Le processus émotionnel comporte trois composantes principales impliquant des éléments cognitifs, comportementaux et physiologiques.

1.2.5.1 Composantes comportementales

La composante comportementale fait référence à la façon dont nous exprimons et montrons nos émotions au moyen de traits comportementaux externes. Ces signaux traduisent ce que nous ressentons et nous apprenons à connaître les émotions des autres en les observant. Les émotions se manifestent sur différents canaux et sont exprimées à la fois verbalement par des mots et non verbalement par des expressions faciales, des gestes, des postures corporelles et des mouvements. D'après les travaux de Mehrabian [34], la communication des émotions est à 7% verbale, à 38% vocale (tonalité de la voix) et 55% non verbale (expressions faciales, expressions corporelles). Les expressions faciales sont la modalité la plus étudiée et elle est considérée comme le canal majeur de communication émotionnelle car elle est visible et facilement observable.

1.2.5.2 Composantes physiologiques

La composante physiologique est la façon dont le corps réagit à une émotion suite à une excitation externe. Ces réactions sont utilisées comme témoins d'une manifestation émotionnelle et sont le résultat de la réponse du système nerveux autonome à l'état émotionnel spécifique que vit la personne. Contrairement aux modalités comportementales, les signaux physiologiques sont très difficiles voire impossibles à contrefaire car le système nerveux autonome est activé de manière involontaire et ne peut donc pas être contrôlé. En d'autres termes, il est difficile pour les sujets de simuler leurs réactions physiologiques pour détourner leur état émotionnel. Parmi les réponses physiologiques couramment observées, l'activité cardiaque et l'activité cérébrale sont des indicateurs souvent utilisés pour reconnaître les émotions.

1.2.5.3 Composantes subjectives

La composante subjective est décrite comme la façon dont nous interprétons les émotions et réfléchissons aux situations ou stimulations. Cet élément se base sur le ressenti de la personne et sa perception de la situation ou l'événement qui a déclenché la réponse émotionnelle. La mesure du ressenti émotionnel peut être effectuée grâce à des échelles comme le SAM (Self Assesment Manikin) [35] et le VAS (Visual Analog Scale) [36] qui permettent d'évaluer les états émotionnels phasiques et toniques respectivement. Des questionnaires peuvent aussi être utilisés pour reconnaître l'état émotionnel notamment pour la mesure de l'anxiété ou du stress [37].

1.2.6 Reconnaissance automatique des émotions

1.2.6.1 Expression et perception de l'émotion humaine

Les émotions sont fondamentales dans la vie quotidienne car elles influencent notre communication, notre apprentissage, notre perception et notre prise de décision. Elles s'expriment à la fois volontairement et involontairement. D'une part, les gens peuvent communiquer verbalement leur état émotionnel de manière explicite par des mots. D'autre part, les émotions peuvent être exprimées de manière involontaire à travers les changements physiologiques et comportementaux liés à l'état affectif de la personne. Dans les sous-sections suivantes, nous détaillons les modalités couramment utilisées pour la reconnaissance des émotions.

Les expressions faciales

Les expressions faciales sont un aspect important du comportement et de la communication non verbale et constituent le moyen le plus évident de l'expression émotionnelle. Elles jouent un rôle majeur dans la recherche sur les émotions et dans les interactions sociales pour diverses raisons : elles sont visibles et contiennent de nombreuses fonctionnalités utiles pour la reconnaissance des émotions et il est plus facile de collecter un grand ensemble de données de visage que d'autres moyens de communication humaine [38, 39]. Pour cette raison, la plupart des chercheurs acceptent que les expressions faciales sont le moyen le plus efficace permettant de reconnaître les émotions. Lors de l'expérience d'une émotion par un individu, des muscles spécifiques du visage sont activés permettant ainsi de communiquer l'émotion ressentie. Ekman [40] a développé le Facial Action Coding System (FACS), qui associe les mouvements mesurables des muscles faciaux appelés des unités d'actions (Action Units, AU) à un espace émotionnel discret. Chaque configuration de mouvement isolé d'un ou plusieurs muscles du visage permet d'identifier les six émotions de base et une multitude d'émotions secondaires.

Ekman [41] a étudié la relation entre les émotions et l'expression faciale. Il a identifié ensuite un sous-ensemble d'émotions corrélé à certaines expressions faciales spécifiques. Dans sa théorie, Ekman distingue les émotions de base des émotions plus complexes, morales et prosociales. Les premières, acquises au cours de l'évolution, sont innées et universelles, en ce sens qu'elles ont des propriétés communes à toutes les espèces. Néanmoins, l'être humain peut produire une multitude d'expressions faciales autres que les six émotions primaires d'Ekman. Dans la vraie vie, il y a une complexité et un mélange d'émotions. Nous sommes rarement dans une colère furieuse, dans une tristesse extrême ou dans une joie délirante, mais souvent dans un mélange de peur et de soulagement, d'amusement et de colère. L'humain exprime donc non seulement les émotions

primaires mais il peut également exprimer des expressions composées résultantes d'un mélange des expressions de base avec des niveaux d'intensités expressives variés. En effet, une simple expression peut exprimer plusieurs états selon son niveau d'intensité.

De manière générale, il existe deux classes d'expressions faciales : les macro- et micro-expressions. La différence majeure entre ces deux classes réside à la fois dans leur durée et leur intensité. Les macro-expressions sont volontaires, durent généralement entre 0,5 et 4 secondes [42], et sont réalisées à l'aide de mouvements faciaux sous-jacents qui couvrent une grande surface du visage [43]. Elles peuvent donc être clairement distinguées des autres informations sur le visage tel que le clignement des yeux . En revanche, les micro-expressions sont des expressions involontaires et rapides [44, 45], dont la durée typique est comprise entre 0,065 et 0,5 seconde [46]. Elles apparaissent généralement lors de situations où les enjeux qui en découlent sont élevés. Bien que les gens puissent intentionnellement dissimuler ou retenir leurs véritables émotions en déguisant leur macro-expressions, les micro-expressions sont considérées comme étant fiables pour estimer l'émotion car il est quasiment impossible de les feindre ou les contrefaire [44, 45, 47, 48]. Néanmoins, les micro-expressions sont très difficiles à identifier à l'œil nu en raison de leurs propriétés inhérentes (courte durée, caractère involontaire et faible intensité) ; même les experts ayant reçu une formation intensive ne peuvent pas toujours les distinguer [49]. Aussi, l'analyse humaine des micro-expressions est longue, coûteuse et sujette aux erreurs. Il est donc hautement souhaitable de développer des systèmes automatiques d'analyse des micro-expressions basés sur des techniques de vision par ordinateur et d'intelligence artificielle [50].

Les signaux physiologiques

Les émotions influencent l'activité du système nerveux autonome qui, à son tour, régule divers paramètres corporels [1]. Chaque émotion est caractérisée par une variation physiologique particulière associée à des modifications du SNA, comme par exemple la modification du rythme cardiaque, de la pression artérielle, de la température corporelle [51]. Les signaux physiologiques présentent certains avantages par rapport aux modalités comportementales en ce sens qu'ils sont difficiles ou impossibles à contrefaire car ils s'activent ou s'inhibent en réponse au système nerveux autonome. Ce dernier est activé de manière involontaire et ne peut pas être contrôlé [52]. Les signaux physiologiques peuvent intégrer des informations liées à l'état émotionnel interne qui n'est pas forcément traduit par des manifestations extérieures observables. Ils sont également moins affectés par les différences sociales et culturelles [4]. Le principal inconvénient de l'utilisation des signaux physiologiques pour la reconnaissance de l'émotion réside dans le fait que les mesures sont prises de manière intrusive, au moyen de capteurs situés sur le corps de l'utilisateur qui

peuvent interférer avec les sujets et modifier leur état émotionnel, ce qui n'est pas approprié dans la plupart des contextes d'interaction naturelle. De plus, la complexité de la mesure et la sensibilité des capteurs limitent fortement leur champ d'application, puisqu'ils sont contraignants dans leur mise en œuvre et ne peuvent être utilisés en dehors du laboratoire.

Nicolas Simonazzi [53] a listé les signaux physiologiques les plus couramment utilisés dans la reconnaissance des émotions selon trois activités :

- Activité électrodermale : l'activité électrodermale (Electrodermal activity, EDA) est un signal physiologique qui peut facilement être mesuré à partir de la surface du corps. L'EDA représente l'activité du système nerveux autonome. Elle caractérise les changements des propriétés électriques de la peau dus à l'activité des glandes sudoripares et est physiquement interprétée comme une conductance. Les glandes sudoripares réparties sur la peau ne reçoivent des informations que du système nerveux sympathique, ce qui en fait un bon indicateur du niveau d'activation dû à des stimuli sensoriels et cognitifs externes [54].
- Activité cardiaque : les changements cardiovasculaires sont nécessaires pour se préparer à l'action et reflètent les expériences émotionnelles. La fréquence cardiaque (FC) et sa variabilité (VFC) sont de bons indicateurs de la valence émotionnelle [55]. Elles peuvent être mesurées à l'aide de l'électrocardiographie (ECG) ou de la photopléthysmographie (PPG). L'ECG permet d'identifier les effets relatifs des composantes parasympathique et sympathique au niveau des nœuds en mesurant les impulsions électriques à la surface de la peau provoquées par la dépolarisation des cellules du myocarde à chaque battement cardiaque. La PPG est une mesure optique de l'activité cardiaque permettant d'observer les variations de volume sanguin dans un tissu biologique de manière non-invasive grâce à des photorécepteurs. La forme d'onde du signal PPG peut fournir des informations sur les changements de l'activation sympathique [56]. Cette activité, agissant sur le diamètre des vaisseaux sanguins, entraîne des modifications du volume et du débit sanguin. Pour cette raison, l'activité cardiaque est prise en compte dans la reconnaissance de l'état émotionnel.
- Activité cérébrale : les signaux des électroencéphalogrammes (EEG) font référence à l'activité du système nerveux central et sont l'un des paramètres les plus fiables pour détecter les émotions avec une grande précision. Des marqueurs émotionnels sont présents dans les signaux EEG. Ils ne peuvent pas être facilement trompés par les actions volontaires d'un utilisateur. Récemment, l'intérêt pour l'évaluation de l'activité cérébrale à l'aide de l'EEG a augmenté en raison de la disponibilité de casques portables peu coûteux et de leur facilité d'utilisation. Cependant, la mesure est contraignante car ces casques nécessitent la mise en place de plusieurs paires d'électrodes sur le cuir chevelu (8, 16 ou 32) à l'aide d'un adhésif

conducteur d'électricité. Par conséquent, ils ne sont pas adaptés aux applications pratiques en dehors du laboratoire en raison de leur grande sensibilité aux artefacts de mouvements et aux conditions environnementales.

Il existe également plusieurs autres signaux physiologiques qui sont utiles pour la détection et la classification des émotions :

- Fréquence respiratoire : elle référe au volume d'air associé aux différentes phases du cycle respiratoire. Le repos et la relaxation entraînent des respirations plus lentes et plus superficielles. L'excitation émotionnelle et les activités physiques génèrent des respirations plus profondes. Les émotions à valence négative provoquent une respiration irrégulière.
- Température de la peau : elle réfère aux fluctuations de la température du corps humain liées à la vasodilatation des vaisseaux sanguins périphériques induite par une activité accrue du système nerveux sympathique.
- Electromyographie (EMG) : les deux prinicpaux muscles faciaux fortement liés à l'expérience de la valence affective sont le grand zygomatique (le muscle qui tire les coins des lèvres vers le haut en produisant un sourire) et le corrugator supercilii (le muscle qui abaisse les sourcils en produisant un froncement). Les activations de ces muscles peuvent être captées à l'aide de l'EMG qui mesure l'activité électrique des muscles en utilisant généralement des électrodes de surface.

La parole

La parole est le moyen le plus naturel d'exprimer et transmettre les émotions [57]. Plusieurs types d'informations peuvent être déduits de la parole tels que l'identité du locuteur, ce que le locuteur a dit et comment il l'a dit [58].

Les indices affectifs sont transmis dans la parole à la fois par des messages explicites (linguistiques) et implicites (paralinguistiques). L'information linguistique reflète la sémantique du message. Certaines informations sur l'état affectif du locuteur peuvent être déduites directement des caractéristiques de surface des mots, qui sont résumées dans les dictionnaires de mots affectifs et l'affinité lexicale [59, 60]. Le reste des informations affectives se trouve sous la surface du texte et ne peut être détecté que lorsque le contexte sémantique (par exemple, les informations sur le discours) est pris en compte [61]. Par ailleurs, la voie linguistique de la parole n'est pas universelle, d'où la nécessité de développer un processeur vocal en langage naturel différent pour chaque langue ou dialecte. Ensuite, elle est sensible à la dissimulation, car les gens ne sont pas toujours sincères dans leurs émotions. Si l'on ne considère que le message linguistique, sans prêter

attention à l'aspect non linguistique du discours et sans tenir compte de la manière dont il a été prononcé, on risque de passer à côté d'aspects importants de l'énoncé pertinent et même de mal comprendre le message oral.

En ce qui concerne le contenu paralinguistique, l'information affective est transmise par la prosodie et reflète la façon dont les mots sont prononcés. Les mesures paralinguistiques de la parole telles que la hauteur, le volume, l'intonation, la fréquence, etc. sont largement explorées dans le domaine de l'informatique affective et offrent une grande précision de reconnaissance [62, 63]. Ce résultat semble raisonnable étant donné que les états affectifs impliquent des réactions physiologiques (par exemple, des changements dans les systèmes nerveux autonomes et somatiques), qui modifient à leur tour différents aspects du processus de production de la voix. Par exemple, l'excitation sympathique associée à un état de colère produit souvent des changements dans la respiration et une augmentation de la tension musculaire, qui influencent la vibration des plis vocaux et la forme du conduit vocal, affectant ainsi les caractéristiques acoustiques de la parole [64]. Le principal problème auquel sont confrontés les systèmes de reconnaissance des émotions basés sur la parole est lié aux conditions de bruit ambiant, qui se produisent fréquemment dans des environnements naturels.

La reconnaissance des émotions par la parole est un outil puissant lorsque les commandes entre la machine et l'homme sont faites verbalement. Néanmoins, il existe de nombreux problèmes liés aux changements de voix en fonction du sexe et de l'âge, et des variations d'expression dans différentes langues. De plus, dans la plupart des applications informatiques, les utilisateurs utilisent rarement des commandes verbales et/ou des commandes dactylographiées. Par ailleurs, la majorité des systèmes existants sont entraînés et testés sur des bases de données actées et donc plus éloignées des contextes réels. Les corpus des scènes naturelles existants sont limités et le nombre d'exemples et de sujets est généralement faible.

1.2.6.2 Reconnaissance unimodale des émotions

Cette section présente les études bibliographiques sur la reconnaissance des émotions à l'aide de modalités individuelles, à savoir les signaux visuels, audio et physiologiques, qui sont les canaux les plus informatifs en termes d'expressions et de prédictions de l'émotion [65, 66, 67].

1.2.6.2.1 Reconnaissance des expressions faciales

Les recherches initiales sur les systèmes unimodaux de reconnaissance des émotions se sont concentrées sur les expressions faciales en raison de l'expressivité du visage et de leur importance dans nos communications quotidiennes. En effet, le visage est la partie la plus visible du

corps humain et il est plus facile de collecter un grand ensemble de données faciales que d'autres modalités comportementales ou physiologiques. Les méthodes couramment adoptées pour la reconnaissance des expressions faciales sont soit des approches conventionnelles soit basées sur l'apprentissage profond [68, 69]. Nous explorons en détail dans cette sous section les deux approches de reconnaissances des expressions faciales.

Méthodes conventionnelles

Avant la résurgence des approches basées sur l'apprentissage profond, l'extraction manuelle des caractéristiques dominait le domaine de la vision par ordinateur. Par conséquent, des approches conventionnelles visant à représenter les caractéristiques faciales par leur forme ou leur apparence ont été proposées [70, 68]. Le pipeline de base de ses approches se compose principalement de trois modules clés : détection du visage suivi de l'extraction de caractéristiques et enfin la classification et la reconnaissance de l'émotion. La structure du système est illustrée en Figure 1.4.

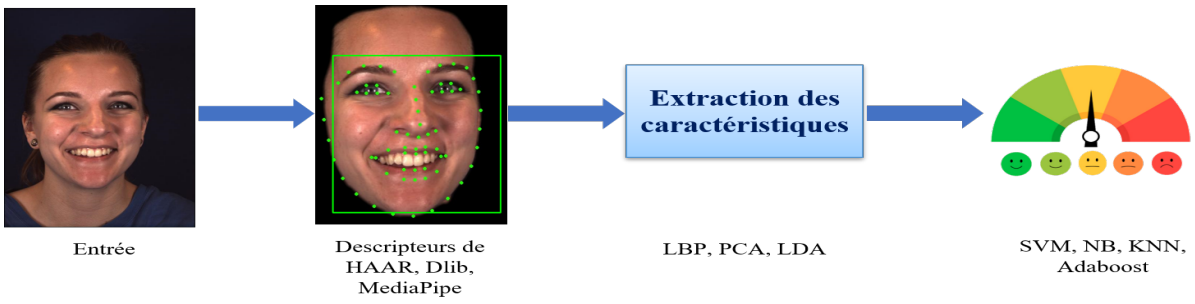


FIGURE 1.4 – La structure de base d'un système conventionnel de reconnaissance des émotions basé sur les expressions faciales.

Dans la majorité des systèmes conventionnels de reconnaissance des expressions faciales, l'accent est mis sur le choix des meilleures méthodes d'extraction de caractéristiques et également des algorithmes d'apprentissage et de classification adaptés. Leurs modèles s'inspirent des algorithmes de vision par ordinateur classiques divisés en étapes. Tout d'abord, le visage est détecté dans l'image, puis il est recadré. Parmi les détecteurs de visage les plus utilisés, nous trouvons le classifieur en cascade de Haar [71], Dlib [72], MTCNN [73]. Quant à l'extraction de caractéristiques, de nombreuses approches traditionnelles ont été utilisées, telles que les motifs binaires locaux (Local Binary Pattern, LBP) [74], l'histogramme de gradient orienté (Histogram of Oriented Gradients, HOG) [75], les ondelettes de Gabor [76] et la transformation de caractéristiques visuelles invariantes à l'échelle (Scale Invariant Feature Transform, SIFT) [77]. Ces descripteurs

TABLE 1.1 – Tableau comparatif des différents systèmes conventionnels de reconnaissance des expressions faciales de la littérature.

Méthode	**Base de données**	**Extraction de caractéristiques**	**Classifieur**	**Précision (%)**
Shan et al. [74]	CK+	LBP	SVM	88.4
Greche et al. [83]	CK+	HOG	LDA	87.78
Verma et al. [76]	JAFFE	Gabor	ANN	85.7
Berretti et al. [77]	BU-3DFE	SIFT	SVM	78.43
Ionescuet al. [78]	FER	BoW	SVM	67.48
Wang et al. [79]	CK+	LBP-TOP	SVM	87.74
Matamoros et al. [84]	KDEF	Gabor	PCA & logique floue	98.8
Kamarol et al. [85]	CK+	AAM	HMM	82.4

ont été étendus pour capturer les caractéristiques spatio-temporelles, comme le sac de mots visuels (Bag of Words, BoW) sur les caractéristiques SIFT dans [78] et LBP-Three Orthogonal Planes (TOP) dans [79]. Pour la classification, différents algorithmes classiques d'apprentissage automatique ont été appliqués tels que le séparateur à vaste marge (SVM) [80], les forêts aléatoires [81] et le modèle de markov caché [82]. Le tableau 1.1 présente un résumé des systèmes conventionnels de la littérature scientifique pour la reconnaissance des expressions faciales.

De telles méthodes basées sur l'extraction manuelle des caractéristiques dépendent moins de la quantité des données et du matériel, ce qui présente des avantages lorsqu'on utilise des bases de données et des ressources matérielles limitées. Cependant, leurs capacités de généralisation et de discrimination pour capturer la physionomie du visage ne sont pas suffisantes, surtout dans les situations du monde réel, caractérisées par un grand degré de variabilité et de difficultés.

Méthodes basées sur l'apprentissage profond

Au cours de la dernière décennie, grâce à la disponibilité des bases de données volumineuses et à l'augmentation spectaculaire de la puissance de calcul, les réseaux de neurones profonds (Deep Neural Networks, DNN) ont révolutionné le domaine de la vision par ordinateur. Une transition considérable a eu lieu dans les études de l'état de l'art où les méthodes basées sur l'apprentissage en profondeur ont remplacé avec succès les approches conventionnelles. Cette transition est motivée par les résultats impressionnants obtenus par les modèles basés DNN dans un large éventail d'applications telles que l'imagerie médicale, la reconnaissance du visage et

bien d'autres encore. Récemment, les méthodes basées sur l'apprentissage profond ont dominé le domaine de la reconnaissance des expressions faciales, en raison de la capacité des réseaux de neurones profonds à apprendre une hiérarchie de caractéristiques qui va des représentations de bas niveau aux représentations de haut niveau. Les DNN ont permis des avancées sans précédent et ont largement dépassé les résultats des algorithmes conventionnels. La puissance des techniques de l'apprentissage profond réside dans leur grande capacité de généralisation pour de nouvelles données et de leur aptitude à extraire automatiquement des caractéristiques robustes et à apprendre des représentations non linéaires complexes. Aujourd'hui, les méthodes basées sur l'apprentissage profond permettent de réaliser une catégorisation des expressions faciales avec une précision de l'ordre de 98% dans des situations contrôlées [86]. Néanmoins, plusieurs problèmes liés aux conditions environnementales et aux artefacts de mouvement peuvent dégrader la précision de la reconnaissance, comme les variations d'éclairage ou de mouvements significatifs de la tête [87]. De plus, les algorithmes d'apprentissage profond échouent souvent dans le cas de visages inexpressifs ou d'expressions contrefaites et leurs performances dépendent de la taille et la diversité des bases de données.

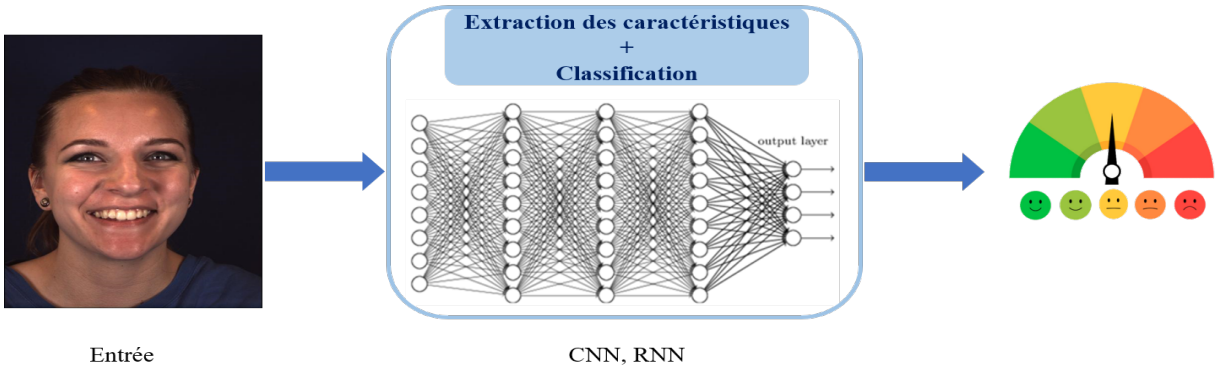


FIGURE 1.5 – La structure de base d'un système de reconnaissance des expressions faciales à partir des images statiques basé sur l'apprentissage profond.

Les méthodes existantes de reconnaissance des expressions faciales basées sur l'apprentissage profond peuvent être divisées en deux catégories : la reconnaissance statique et la reconnaissance dynamique des expressions faciales. La première catégorie apprend les caractéristiques spatiales seulement à partir d'images individuelles, tandis que la deuxième catégorie extrait des caractéristiques spatio-temporelles par un encodage de séquences d'images. Le tableau 1.2 résume les approches de classification des expressions faciales basées sur l'apprentissage profond.

Reconnaissance des expressions faciales basée sur les images statiques

Les premiers travaux sur la reconnaissance des expressions faciales se sont concentrés sur les images statiques en raison des bases de données disponibles ainsi que la simplicité du traitement par rapport aux images dynamiques. Dans cette catégorie, les réseaux de neurones convolutifs (Convolutional Neural Network, CNN) 2D sont l'architecture principale pour la description des caractéristiques spatiales, raison pour laquelle ils dominent la première catégorie de représentation des expressions faciales à partir d'images statiques.

Les CNN sont des architectures puissantes pour extraire des caractéristiques spatiales des images faciales. Les premières couches apprennent des représentations de bas niveau comme les bords, les textures et les objets en plus du visage, tandis que les couches suivantes captent les différentes parties du visage en repérant les points faciaux importants (par exemple les yeux, les lèvres, et les sourcils etc.). Cette partie du réseau apprend des caractéristiques plus spécifiques qui peuvent être interprétées sémantiquement pour décrire les expressions faciales sous forme d'un vecteur de caractéristiques expressives. Ce dernier est ensuite utilisé pour la classification. La structure du système est illustrée en Figure 1.5.

Contrairement aux approches traditionnelles dans lesquelles les caractéristiques sont extraites manuellement, les CNN apprennent à extraire les caractéristiques directement de données à l'aide d'algorithmes itératifs comme la descente de gradient. Une architecture CNN typique comprend généralement un empilement des couches alternées de convolution et de mise en commun (pooling en anglais), suivies d'une ou plusieurs couches entièrement connectées (fully connected en anglais) à la fin. Dans certains cas, une couche entièrement connectée est remplacée par une couche de mise en commun par moyenne globale (Global Average Pooling). En plus des fonctions d'activation, différentes unités de régularisation telles que la normalisation par lots (batch-normalization en anglais) et l'abandon (dropout en anglais) sont également incorporées pour optimiser les performances des CNN. L'image d'entrée est d'abord introduite dans la couche de convolution, qui est une opération linéaire permettant d'extraire des cartes de caractéristiques. Généralement, une fonction d'activation suit la convolution, habituellement, une ReLU permettant d'effectuer une mise en correspondance non linéaire. Les sorties générées par chaque couche de convolution passent à une couche de sous-échantillonnage appelée mise en commun ou pooling pour réduire la dimensionnalité des cartes de caractéristiques. La couche entièrement connectée est principalement utilisée à la fin du réseau pour la classification. Contrairement à la convolution et à la mise en commun, il s'agit d'une opération globale. Elle prend en entrée le vecteur de caractéristiques et analyse globalement la sortie de toutes les couches précédentes. Par conséquent, elle effectue une combinaison non linéaire des caractéristiques sélectionnées, et utilise ce résultat

pour la classification des données.

Depuis le succès retentissant d'AlexNet [88] dans le défi de classification ImageNet en 2012, diverses architectures CNN profondes (e.g. VGG [89], ResNet [90], Inception [91]) ont été proposées et atteignent une grande précision dans de nombreux défis de reconnaissance d'objets dans les images. Inspirés par le grand succès de l'apprentissage profond, les systèmes de reconnaissance des expressions faciales ont incorporé les différentes architectures CNN dans le pipeline conventionnel pour extraire automatiquement les caractéristiques faciales. Bénéficiant des puissances de ces architectures, les étapes d'extraction de caractéristiques et de classification des émotions peuvent être réalisées de bout en bout sans connaissances préalables ni étapes supplémentaires. Nguyen et al. [92] ont proposé un modèle CNN à 18 couches similaire à l'architecture VGG. Mollahosseini [93] a développé un réseau de neurones basé sur l'architecture Inception tandis que Hardjadinata et al. ont utilisé Xception et DenseNet [94]. Chang et al. [95] ont construit un modèle CNN basé sur l'architecture ResNet qui comprend des connexions résiduelles permettant de réduire l'impact de la disparition du gradient. Arriaga et al. [96] se sont inspirés des modèles VGG [89] et GoogleNet [97] et ont utilisé une combinaison des deux modèles pour la dérivation d'un modèle final moins profond. Shao et al. [98] ont construit un réseau peu profond (light-CNN) basé sur des convolutions séparables en profondeur avec des connexions résiduelles pour réduire efficacement les coûts de calcul et les besoins en mémoire. En outre, les modèles d'attention ont été appliqués avec succès afin d'explorer et se concentrer sur les régions significatives et les caractéristiques les plus discriminantes du visage [99].

Reconnaissance des expressions faciales basée sur une séquence d'images dynamiques

La première catégorie présentée dans la sous-section précédente 1.2.6.2 considère l'utilisation des caractéristiques spatiales et négligent les informations temporelles caractérisant les mouvements des muscles du visage. En revanche, la deuxième catégorie extrait à la fois des caractéristiques spatiales et temporelles à partir d'une séquence d'images. En effet, les expressions faciales dynamiques contiennent plus d'informations que les expressions statiques [100]. La reconnaissance dynamique des expressions faciales consiste à suivre les changements de position des points de repères dans les images de la séquence vidéo. Cette analyse, cependant, est sensible au mouvement de la tête et la variation de l'éclairage pendant l'expression.

Alors que la première catégorie repose sur les CNN 2D qui capturent seulement les informations spatiales dans l'image, la prise en compte des caractéristiques spatio-temporelles peut être effectuée par l'extension des modèles CNN standard pour représenter des volumes 3D. Cela est réalisé soit en utilisant des CNN 3D, des réseaux récurrents en conjonction avec des CNN 2D

TABLE 1.2 – Tableau comparatif des différents systèmes de reconnaissance des expressions faciales basés sur l'apprentissage profond.

Méthode	**Bases de données**	**Type du réseau**	**Caractéristique**	**Précision (%)**
Ouellet et al. [105]	CK+	AlexNet	Statique	94.4
Li et al. [106]		RBM	Statique	96.4
Nasri et al. [107]		Xception	Statique	97.5
Hasani et al. [108]		3D-Inception + ResNet	Dynamique	95.53
Liu et al. [109]	JAFFE	DBN	Statique	91.8
Hardjadinata et al. [94]		DenseNet	Statique	71.02
Hamester et al. [110]		Auto-Encodeur	Statique	95.8
Mollahosseini al. [93]	MMI	Inception	Statique	77.9
Yang et al. [111]		GAN	statique	73.23
Minaee et al. [112]	FER2013	Mécanisme d'attention	Statique	70.02
Panagiotis et al. [113]		GoogleNet	Statique	65.2
Jung et al. [114]	Oulu-CASIA	DTAGN-Joint	Dynamique	81.46
Liu et al. [115]		CNN + RNN	Dynamique	88.33
Yan et al. [116]	AFEW	VGG16 + LSTM	Dynamique	44.46
Vielzeuf et al. [117]		3D-CNN + LSTM	Dynamique	43.2

standards, ou en utilisant un réseau à deux étages en parallèle où le premier extrait les caractéristiques spatiales et le deuxième les caractéristiques temporelles. Les CNN 3D sont capables de capturer les mouvements des muscles faciaux dans une séquence d'images [101]. Cependant, ces modèles présentent des inconvénients en termes de temps de calcul et de besoins en mémoire. Une approche hybride combinant un CNN 2D pour les caractéristiques spatiales des images individuelles et un réseau de neurones récurrents (Recurrent Neural Network, RNN) pour les caractéristiques temporelles des images consécutives, a été proposée [102, 103, 104]. Les modèles récurrents tels que RNN, LSTM (Long Short-Term Memory) et GRU (Gated Recurrent Unit) sont capables d'apprendre des dépendances à long terme pour encoder le mouvement des muscles du visage à travers les séquences d'images.

La combinaison d'un CNN 2D avec un réseau récurrent peut être réalisée de manière séquentielle ou en parallèle. Kahou et al. [118] ont présenté une architecture hybride RNN-CNN pour le défi "Emotion Recognition in the Wild (EmotiW) 2015". Kim et al. [119] tirent parti du CNN 2D pour apprendre les caractéristiques spatiales, puis une LSTM est utilisée pour apprendre les caractéristiques temporelles de la représentation des caractéristiques spatiales. Kang et Ma [120] ont utilisé un CNN de type VGG16 combiné avec une GRU pour apprendre les caractéristiques spatio-temporelles. Zhao [121] a proposé un modèle à deux étages composé de deux CNN et un

mécanisme d'attention partagé. Cela permet au réseau de repérer de manière autonome et accorder plus d'attention aux zones plus pertinentes du visage qui sont utiles pour la catégorisation des expressions faciales. Jung et al. [122] ont proposé une architecture de réseaux profonds à deux flux qui collaborent l'un avec l'autre. Le premier réseau est nommé DTAN. Il se base sur les apparences de plusieurs images, tandis que le second réseau est nommé DTGN. Il extrait les caractéristiques géométriques temporelles utiles des points de repère faciaux. Ces deux modèles sont combinés à l'aide d'une nouvelle méthode d'intégration basée sur le réglage fin (fine-tuning en anglais) afin d'améliorer les performances de la reconnaissance des expressions faciales.

1.2.6.2.2 Reconnaissance des émotions basée sur les signaux physiologiques

Les systèmes de reconnaissance des émotions utilisant des signaux physiologiques comme modalités d'entrée peuvent être divisés en deux grandes catégories : les méthodes conventionnelles et les méthodes basées sur l'apprentissage profond [123].

Méthodes conventionnelles

Les méthodes conventionnelles requièrent une extraction manuelle des caractéristiques pour classer les émotions à partir des signaux physiologiques. En général, les caractéristiques statistiques du domaine temporel (par exemple la moyenne, l'écart type, la première dérivée du signal [124, 125]) et du domaine fréquentiel (par exemple la transformée de Fourier (Fast Fourier Transform, FFT) du signal et son énergie [126]) sont couramment utilisées pour la reconnaissance. Le tableau 1.3 présente un résumé des systèmes conventionnels de la littérature scientifique pour la reconnaissance des émotions à partir des signaux physiologiques.

Pour la classification des émotions basée sur l'activité cérébrale, les caractéristiques statistiques [127], les caractéristiques en ondelettes [128] et les caractéristiques basées sur la décomposition en modes empiriques (Empirical Mode Decomposition, EMD) [129] ont été extraites à partir des signaux EEG. Ishino et Hagiwara [130] ont utilisé la transformée en ondelettes, la transformée de Fourier et des statistiques telles que la moyenne et la variance comme caractéristiques et ont employé un simple réseau de neurones pour classer quatre émotions. Schaaff [131] et Chanel et al. [132] ont utilisé la transformée de Fourier fenêtrée (Short Time Fourier Transform, STFT) pour extraire les caractéristiques, et les SVM comme classifieur. Zhuang et al. [133] ont présenté une méthode d'extraction de caractéristiques basée sur l'EMD. Les signaux EEG sont décomposés automatiquement à l'aide de l'EMD en fonctions de mode intrinsèques (FMI). Les informations multidimensionnelles des FMI sont considérées comme des caractéristiques pour la reconnaissance des émotions. Trois types de caractéristiques ont été extraites à savoir la pre-

mière différence des séries temporelles, la première différence de phase et l'énergie normalisée. La première différence de la série temporelle décrit l'intensité du changement du signal dans le domaine temporel. La première différence de phase mesure l'intensité du changement de phase tandis que l'énergie normalisée décrit le poids de la composante d'oscillation actuelle. Les trois caractéristiques constituent un vecteur de caractéristiques qui est introduit dans un classifieur SVM pour la détection de l'état émotionnel.

TABLE 1.3 – Tableau comparatif des différents systèmes conventionnels de reconnaissance des émotions à partir des signaux physiologiques.

Méthode	**Modalités**	**Base de données**	**Caractéristiques**	**Extraction de caractéristiques**	**Classifieur**	**Précision (%)**
Murugappan et al. [127]	Activité cérébrale	Privée	Temporelle fréquentielle	Transformée en ondelettes	kNN	83.04
Samara et al. [134]		DEAP	Temporelle fréquentielle	Statistiques, PSD, HOC	SVM	82
Bhatti et al. [135]		Privée	Temporelle fréquentielle	Statistiques, PSD, Transformée en ondelettes	MLP	78.11
Agrafioti et al. [136]	Activité cardiaque	Privée	Fréquentielle	EMD Transformée de Hilbert	LDA	Valence : 76.19 Arousal : 89
Guo et al. [137]		Privée	Temporelle fréquentielle poincaré statiques	PCA, Statistiques, FFT, Poincaré	SVM	71.40
Ismail et al. [138]		A2ES	Temporelle fréquentielle	Statistiques, PSD	SVM	Valence : 64.94 Arousal : 64.61
Cheng et al. [139]	EMG	AuBT	Temporelle fréquentielle	Transformée en ondelettes	ANN	75
Wu et al. [140]	RESP	Privée	Temporelle	MIDTW	kNN	88
Ping et al. [141]	ECG, EMG RESP, EDA	Privée	Temporelle fréquentielle	Statistiques, EMD, Transformée en ondelettes	DT	92
Guendil et al. [142]	ECG, EMG RESP, EDA	AuBT	Temporelle fréquentielle	Transformée en ondelettes	SVM	95
Lisetti et Nasoz [143]	ECG, EDA TEMP	AuBT	Temporelle	Statistiques	kNN LDA ANN	72 75 84

Des méthodes conventionnelles similaires pour la reconnaissance des émotions basées sur l'activité cardiaque ont également été appliquées dans la littérature scientifique. L'ECG et la PPG sont les principaux moyens pour mesurer l'activité cardiaque. L'ECG détecte l'activité électrique du cœur, tandis que la PPG mesure les changements de volume sanguin pendant l'activité cardiaque. Différents types de caractéristiques peuvent être extraits à partir du signaux ECG et PPG. Ils sont utilisés pour discriminer l'état émotionnel de la personne, à savoir les caractéristiques statistiques [144, 145], les caractéristiques de la variabilité cardiaque [146, 147],

les caractéristiques en ondelettes [148], les caractéristiques basées sur la décomposition en modes empiriques [149], et bien d'autres encore. Quant aux classifieurs, les algorithmes classiques de l'apprentissage automatique ont été appliqués avec succès. Par exemple, Agrafioti et al. [136] ont utilisé l'algorithme d'analyse discriminante linéaire (Linear Discriminant Analysis, LDA) et les caractéristiques statistiques dans le domaine temporel et fréquentiel pour reconnaître la valence et l'activation. Jerritta et al. [149] discutaient les avantages de l'utilisation des caractéristiques du domaine fréquentiel par rapport aux caractéristiques basées sur l'EMD. Deux algorithmes de classification, à savoir les k plus proches voisins (k-Nearest Neighbours, KNN) et LDA ont été utilisés pour la reconnaissance des six émotions de base. Dans [147], les caractéristiques de la variabilité cardiaque dans le domaine temporel et fréquentiel extraites du signal PPG ont été introduites dans un SVM pour classifier trois émotions, à savoir joie, tristesse et neutre. Ismail et al. [138] ont évalué les performances de leur système en utilisant des données ECG et PPG unimodales provenant de dispositifs portables. La toolbox AUBT [150] a été utilisée pour extraire 80 caractéristiques des données ECG. Simultanément, 17 caractéristiques du signal PPG ont été extraites à l'aide de la toolbox TEAP [151]. Des modèles d'apprentissage automatique tels que SVM, KNN, le modèle de Bayes naïf (Naive Bayes, NB) et les arbres de décision (Decision Tree, DT) ont été entraînés pour reconnaître l'activation et la valence à partir de ces caractéristiques extraites.

Les signaux cérébraux et cardiaques ont été beaucoup plus utilisés que les autres signaux physiologiques tels que la respiration, l'activité électrodermale et l'electromyographie. Néanmoins, certains travaux ont examiné l'utilisation de ces signaux ensemble ou séparément. Lisetti et Nasoz [143] ont utilisé les signaux de l'activité électrodermale, de fréquence cardiaque et de température pour reconnaître les émotions humaines suscitées par des séquences de films et des questions mathématiques difficiles. Zhao et al. [152] ont extrait 223 caractéristiques de 4 signaux physiologiques (ECG, EEG, EDA et EMG) pour reconnaître la valence et l'activation des utilisateurs en utilisant plusieurs algorithmes classiques de classification tels que SVM et NB. Jimenez et al. [153] ont sélectionné 13 caractéristiques de PPG (4 dans le domaine temporel, 9 dans le domaine fréquentiel) et 14 caractéristiques de EDA (toutes dans le domaine temporel) pour reconnaître six émotions de base. Dans l'étude [154], une fusion de caractéristiques a été appliquée pour combiner quatre signaux physiologiques, à savoir l'ECG, la respiration (RESP), l'EDA et la température de la peau. Le résultat de la fusion est introduit dans un SVM pour reconnaître neuf émotions , à savoir joie, amusement, neutre, peur, anxiété, crainte, dégoût, surprise et tristesse.

Méthodes basées sur l'apprentissage profond

L'apprentissage profond permet non seulement de réduire la charge de travail liée à l'extraction manuelle des caractéristiques, mais aussi à améliorer la précision de la reconnaissance. Les approches basées sur l'apprentissage profond connaissent un grand succès dans la tâche de reconnaissance des émotions, en particulier, les approches d'apprentissage profond utilisant les signaux EEG. Lin et al. [155] et Liu et al. [156] ont converti les signaux EEG en format d'image 2D et ont utilisé les modèles d'apprentissage profonds pré-entraînés AlexNet et ResNet pour l'extraction des caractéristiques et la classification. Kwon et al. [157] ont utilisé un simple CNN pour l'extraction de caractéristiques et la classification des émotions en utilisant des signaux EEG bruts. Ces méthodes susmentionnées ignorent les caractéristiques temporelles des signaux EEG, c'est pourquoi des méthodes d'extraction de caractéristiques spatio-temporelles ont été proposées. Salama [158] a mis en œuvre un CNN 3D pour l'extraction et la classification des caractéristiques spatio-temporelles des signaux EEG. Wang [159] a converti les canaux EEG en une plaque topologique d'électrodes en 2D qui pourrait inclure des informations sur la position topologique et a utilisé un CNN 3D pour l'extraction et la classification des caractéristiques spatio-temporelles. Yang [160] a utilisé un CNN 2D combiné avec un module LSTM pour extraire les caractéristiques spatiales et temporelles respectivement et a fusionné les caractéristiques pour la classification. Un modèle hybride composé d'un CNN et d'un RNN a été défini dans [161]. Une architecture récente basée sur les Transformers [162] a été mise en œuvre dans [163]. Les signaux EEG sont d'abord convertis en images à l'aide d'une transformée en ondelettes continue. Ensuite, les images sont introduites dans le Transformer pour prendre en compte les informations contextuelles en vue de la prédiction des émotions.

Des méthodes d'apprentissage profond basées sur l'activité cardiaque ont été également proposées. Une approche adoptée par Santamaria et al. [164], consiste à utiliser un CNN 1D dédié à l'extraction des caractéristiques, suivi d'un réseau entièrement connecté (FCN) utilisé comme classifieur pour prédire les émotions. Harper et Southern [165] emploient également un CNN 1D avec une LSTM. Siddharth et al. [166] convertissent d'abord les signaux en une image en utilisant des spectrogrammes, puis utilisent un CNN 2D pour l'extraction de caractéristiques, suivi d'une machine à apprentissage extrême [167] pour la classification. Dans une autre étude [168], le spectrogramme du PPG a été utilisé comme entrée du modèle ResNet+BiLSTM. Une nouvelle technique basée sur l'apprentissage auto-supervisé a été présentée dans [169] pour effectuer la reconnaissance des émotions à partir de l'ECG. Des réseaux spatio-temporels profonds CNN 3D ont également été proposés pour mesurer sans contact les caractéristiques de la variabilité cardiaque à partir de vidéos faciales [170]. Les caractéristiques de la variabilité cardiaque dans

le domaine temporel et fréquentiel ont été utilisées pour reconnaître la valence et l'activation.

En plus de l'activité cardiaque et cérébrale, d'autres signaux physiologiques ont également été utilisés de manière unimodale ou multimodale. Zhang et al. [171] ont étudié la capacité à détecter l'état émotionnel en utilisant les signaux de respiration seuls. Un système basé sur l'apprentissage profond a été proposé pour extraire et reconnaître les informations émotionnelles de la respiration. Le réseau proposé comprend un auto-encodeur combiné avec une régression logistique pour la reconnaissance de la valence et l'activation. Une tentative a été faite par Ganapathy et al. [172] pour classifier les états émotionnels en utilisant des signaux d'activité électrodermale et des réseaux de neurones convolutifs multi-échelles. Un réseau de neurones convolutifs a été adopté par Lee et al. [173] pour l'extraction des caractéristiques et la classification du niveau de la valence et l'activation à partir de l'électromyogramme et le photopléthysmogramme. Nakisa et al. [174] ont proposé une approche de fusion multimodale temporelle avec un modèle d'apprentissage profond basé sur CNN et LSTM pour capturer la corrélation émotionnelle non linéaire dans et entre les signaux EEG et BVP. Vijayakumar et al. [175] ont examiné les performances de classification de six combinaisons différentes de signaux périphériques (EOG, EMG, BVP et respiration) en utilisant un modèle CNN unidimensionnel (CNN 1D). Le tableau 1.4 résume les méthodes de reconnaissance des émotions à partir des signaux physiologiques basées sur l'apprentissage profond.

1.2.6.3 Reconnaissance multimodale des émotions

La recherche sur la reconnaissance unimodale des émotions a quasiment atteint son point de saturation en termes de précision notamment avec l'émergence de la reconnaissance multimodale des émotions qui est considérée comme la prochaine tendance dans ce domaine de recherche. Bien que la reconnaissance unimodale des émotions présente de grands succès, elle a également révélé certaines limites au fil du temps. Par exemple, elle ne peut pas décrire entièrement une émotion alors que l'utilisation de plusieurs modalités pour décrire une émotion particulière sera plus complète et représentative [182]. Les systèmes multimodaux sont plus adaptés pour l'analyse et la reconnaissance des émotions car ils sont plus représentatifs, efficaces, non ambigus et ont un domaine d'application plus large par rapport aux systèmes unimodaux. La reconnaissance multimodale s'inspire de la capacité des humains à reconnaître et déchiffrer les émotions en combinant les signaux provenant de différentes sources. Il a été démontré que la fusion de plusieurs modalités renforce la robustesse et améliore la précision de la reconnaissance [183].

Certaines émotions sont plus faciles à reconnaître à partir d'expressions faciales et d'autres à partir de données physiologiques ou à partir de la parole. La joie, par exemple, est difficile

TABLE 1.4 – Tableau comparatif des différents systèmes de reconnaissance des émotions à partir des signaux physiologiques basés sur l'apprentissage profond.

Méthode	Modalités	Base de données	Caractéristique	Type du réseau	Précision (%)
Li et al. [161]	Activité cérébrale	DEAP	Ondelettes	CNN-RNN	Valence : 72.06 Arousal : 74.12
Tao et al. [176]		DEAP DREAMER	Signaux bruts	Mécanisme d'attention	DEAP : 93 DREAMER : 97
Tripathi et al. [177]		DEAP	Signaux bruts	CNN	Valence : 81.41 Arousal : 73.36
Liu et al. [178]		DEAP	PSD Entropie différentielle	Auto-encodeur	83.25
Sarkar et al. [169]	Activité cardiaque	AMIGOS DREAMER WESAD SWELL	Signaux ECG	CNN auto-supervisé	AMIGOS : 78.95 DREAMER : 76 WESAD : 95 SWELL : 93.2
Siddharth al. [166]		DEAP AMIGOS MAHNOB DREAMER	Caractéristiques HRV	CNN-LSTM	DEAP : 45.55 AMIGOS : 58.08 MAHNOB : 57.23 DREAMER :57.73
Wang et al. [168]		Privée	Spectrogramme PPG	ResNet-BiLSTM	84.70
Ganapathy et al. [113]	EDA	DEAP	Temporelles, fréquentielles, morphologiques, STFT	CNN	Valence : 72 Arousal : 75
Jia et al. [179]	ECG, EEG, EOG, EDA	DEAP MAHNOB	Temporelles, fréquentielles	GNN-GRU	DEAP : 97.48 MAHNOB : 93.63
Liu et al. [115]	ECG, EDA	AMIGOS	Signaux bruts	CNN	Valence : 75 Arousal : 76
Kawde et al. [180]	EEG, EOG, EMG	DEAP	Signaux bruts	DBN-Auto_encodeur	Valence : 78 Arousal : 73
Lin et al. [155]	EEG, ECG, EDA, EMG, EOG, RESP, TEMP	DEAP	Temporelles, fréquentielles	CNN	Valence : 85.50 Arousal : 87.30
Ma et al. [181]	EEG, EOG, EMG	DEAP	Signaux bruts	ResNet-LSTM	Valence : 92.30 Arousal : 92.87

à interpréter à partir d'une expression faciale car de nombreuses personnes sourient en cas de frustration naturelle [199] tandis que la même émotion peut être déduite facilement en utilisant d'autres modalités tels que les signaux physiologiques ou la parole. C'est pourquoi les travaux récents se focalisent sur des données multimodales pour la reconnaissance des émotions [200, 201, 194]. Plusieurs modalités, telles que les expressions faciales, la parole, les signaux physiologiques, la gestuelle et la posture, sont désormais combinées pour développer des systèmes plus précis et plus proches du monde physique [202]. L'une des combinaisons les plus naturelles et les plus

TABLE 1.5 – Tableau comparatif des systèmes multimodaux de reconnaissance des émotions basés sur les expressions faciales et les signaux physiologiques.

Méthode	**Modalités**	**Base de données**	**Classifieur**	**Précision (%)**
Koelstra et Patras [184]	EEG + EF	MAHNOB-HCI	NB	Valence : 70.90 Arousal : 73
Huang et al. [185]		MAHNOB-HCI	SVM	Valence : 66.28 Arousal : 63.22
Huang et al. [186]		DEAP	SVM-CNN	Valence : 80 Arousal : 74
Li et al. [187]		SEED	LSTM	CCC = 0.63
Yang et al. [188]		Privée	DBN	99.92
Tivatansakul et al. [189]	CARD + EF	AuBT	kNN	87.92
Yan et al. [190]		PSFE	SVM	84.44
Yu al. [191]		DEAP	3D-CNN	Valence : (CCC = 0.25) Arousal : (CCC = 0.10)
Du et al. [192]		Privée	CNN-Bi-LSTM	87.3
Abdat et al. [193]	CARD + EDA + EMG + TEMP + RESP + EF	Privée	SVM	52.91
Zhong et al. [194]	CARD + RESP+ EDA + TEMP + EF	MAHNOB-HCI	SVM-AdaBoost	Valence : 70.53 Arousal : 73.53
Cimtay et al. [195]	EEG + EDA + EF	DEAP LUMED-2	InceptionResNetV2-DT	53.87 74.2
Li et al. [196]	CARD + EDA + EF	RECOLA	DBN_3D-CNN	Valence : 59.7 Arousal : 61.9
Zhu et al. [197]	CARD + EDA + EEG + EMG + EOG + TEMP + RESP + EF	DEAP	CNN	Valence : 78.47 Arousal : 72.2
Saffaryazdi et al. [198]	CARD + EDA + EEG + EF	DEAP	3D-CNN_LSTM	Valence :57 Arousal : 60.9

- CARD : Signaux cardiaques (ECG, PPG, HR, HRV)
- DBN : Réseau de croyances profond (Deep Belief Network)
- EF : Expressions faciales
- CCC : Coefficient de Corrélation de Concordance

utilisées est celle des expressions faciales et de la voix [200]. Néanmoins, presque toutes les combinaisons ont été examinées dans la littérature scientifique pour étudier l'impact de la fusion multimodale sur les performances. En s'appuyant sur l'expressivité du visage, les expressions faciales ont largement été utilisées en combinaison avec la parole [200], les signaux physiologiques [194] et les mouvements du corps [203]. Par ailleurs, Yang et al. [204] ont proposé un système bimodal basé sur l'audio et le texte tandis que Bakhshi et al. [205] ont fusionné les signaux physiologiques et la parole. Des travaux utilisant plus de deux modalités ont également été proposés. Deng et al. [206] ont combiné trois modalités à savoir les expressions faciales, l'audio et le texte tandis que Ranganathan [207] ont fusionné quatres modalités à savoir les expressions faciales, les gestes, la voix et les signaux physiologiques. Le tableau 1.5 résume les systèmes multimodaux de reconnaissance des émotions basés sur les expressions faciales et les signaux physiologiques.

1.3 Stress

1.3.1 Définition du stress

Il existe un grand nombre d'expériences humaines qui peuvent être regroupées sous le terme de "stress". Ce terme est fréquemment utilisé, tant dans les milieux scientifiques que dans le langage courant, pour désigner un certain nombre de processus différents, qui sont liés mais distincts. Par exemple, le terme "stress" est parfois utilisé pour désigner des événements ou des situations de la vie réelle qui arrivent à une personne, comme la perte d'un emploi ou le divorce.

Le stress est une réaction naturelle du corps humain à un facteur de perturbation extérieur. Il peut être causé par des situations émotionnelles, mentales ou physiques. Bien que les faibles niveaux de stress peuvent avoir des effets bénéfiques sur l'organisme, le stress a souvent un impact négatif sur l'attention, la mémoire et la prise de décision [208]. Des niveaux élevés de stress à long terme sont également corrélés à divers effets négatifs sur la santé, notamment l'anxiété, la dépression et le vieillissement prématuré [209]. Par conséquent, les personnes présentant un risque élevé de stress doivent faire l'objet d'une surveillance continue afin de détecter tout signe de stress avant qu'il n'entraîne des problèmes de santé.

Comme pour les émotions, il n'existe pas de définition commune du stress car il s'agit d'un phénomène subjectif qui est difficile à définir ou de s'accorder sur sa définition [210]. La définition du stress dépend en partie des personnes interrogées et de leur discipline principale. Elle peut inclure l'antécédent, le stimulus ou la réponse. Par conséquent, toute mesure qui enregistre une perception ou une réponse biologique ou neuronale peut être qualifiée de mesure du stress.

Hans Selye a défini le stress comme la réponse neuroendocrinienne non spécifique de l'organisme à toute demande de changement [211]. Depuis, d'autres définitions prenant en compte les capacités d'adaptation de chaque individu ont été exposées [212], dont celle de McEwen [213] qui définit le stress comme un événement menaçant pour un individu, qui provoque des réponses physiologiques et comportementales. Le dictionnaire Merriam Webster définit le stress comme un facteur physique, chimique ou émotionnel qui provoque une tension corporelle ou mentale et peut être un facteur de causalité d'une maladie. D'après les dictionnaires Oxford [214], le stress est défini comme "un état de tension mentale ou émotionnelle résultant de circonstances défavorables ou exigeantes. Lazarus Folkman [212] ont considéré le stress comme "un sentiment éprouvé lorsqu'une personne perçoit que les exigences dépassent les ressources personnelles et sociales que l'individu est capable de mobiliser, ce qui concerne principalement l'émotion humaine et le sentiment de stress. D'autre part, les psychologues cognitifs identifient le stress de manière analytique à partir des composants fondamentaux de la vie mentale, tels que l'attention et son

allocation, les systèmes de mémoire, la résolution de problèmes et la prise de décision [215].

Le concept de stress est donc difficile à cerner en raison des aspects psychologiques et physiologiques qu'il comporte. L'aspect psychologique a été décrit par de multiples modèles tels que le modèle de la demande et du contrôle [216] et le modèle du déséquilibre effort-récompense [217]. Sur le plan physiologique, le stress peut notamment être décrit par l'activité du SNA.

1.3.2 Elicitation du stress

Afin d'étudier le stress dans des environnements de laboratoire, il est demandé aux sujets d'effectuer certaines tâches afin d'induire le stress, tout comme les émotions mais avec des méthodes différentes. De nombreux tests d'élicitation du stress ont été validées. La méthode la plus fréquemment utilisée est probablement le test d'interférence couleur-mot de Stroop [218], suivi par des tâches d'arithmétique mentale [219]. Le test des mots colorés de Stroop est une méthode fiable et valide pour induire des niveaux moyens de stress [220, 221]. Dans ce test, les noms des couleurs sont présentés dans des couleurs différentes de celles correspondant à leurs noms, et le sujet doit choisir les couleurs dans lesquelles les noms apparaissent (voir Figure 1.6 (a)). Le test de calcul mental est un stresseur standard d'intensité modérée utilisé en physiologie pour détecter les changements dans la fonction du SNA [222]. Des nombres à trois ou quatre chiffres sont affichés à l'écran et les chiffres sont additionnés de manière répétée jusqu'à obtenir un nombre à un chiffre. Enfin, le sujet utilise un clavier pour indiquer si le résultat final est un nombre pair ou impair (voir Figure 1.6 (b)). Afin de générer un stress plus important chez les sujets, 5 s sont comptées pour chaque stimulus avant de passer au stimulus suivant, et l'écran affiche si la réponse du sujet est bonne ou mauvaise pour chaque stimulus.

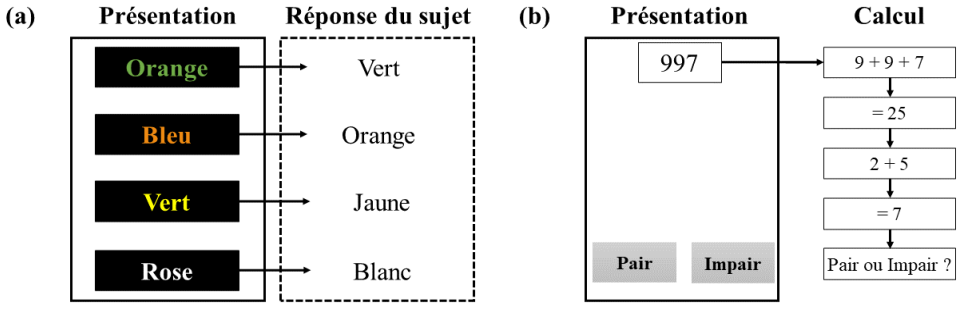


FIGURE 1.6 – Les méthodes d'induction du stress : (a) le test des mots colorés de Stroop, et (b) une illustration du test de calcul mental

Le test au froid est un autre inducteur de stress couramment utilisé en raison de sa facilité d'utilisation [223]. Il implique l'immersion de la main ou d'un membre dans l'eau froide pendant

une durée de 2 à 3 minutes. Au cours de cette expérience, le sujet se sent mal à l'aise et il est douloureux de s'adapter à une température particulière pendant un certain temps. Cette experience déclenche l'activation du système nerveux sympathique qui augmente la pression artérielle, la fréquence cardiaque et la conductance de la peau du corps humain [224]. La conduite automobile est également considérée comme une tâche génératrice de stress [25], tout comme le fait de regarder des films au contenu stressant [225, 226] ou de jouer à des jeux vidéo [227]. Pour le stress social, des stimilus traitant de l'exposition sociale sont utilisés, tels que des entretiens ou l'implication dans des situations où les participants seraient jugés ou évalués pour leurs performances [228].

1.3.3 Reconnaissance unimodale du stress

La majorité des travaux de détection du stress existants sont basés sur l'utilisation d'une seule modalité couplée à des étiquettes binaires stress/non-stress ou plusieurs niveaux de stress (faible, moyen ou élevé). Cependant, le stress n'est pas un état affectif isolé, mais est étroitement lié à l'expression et à la régulation des émotions humaines [229]. Cette section présente les études de l'état de l'art sur la reconnaissance unimodale du stress en utilisant les réponses comportementales et physiologiques.

1.3.3.1 Réponses physiologiques

Activité cardiaque

La fréquence cardiaque et sa variabilité sont des signaux pertinents pour la détection du stress. Elles peuvent être mesurées par un capteur en contact ou à distance à l'aide d'une caméra. Des caractéristiques dans le domaine temporel (RR moyen, écart-type RR), dans le domaine fréquentiel (LF/HF, différence LF HF normalisée) et non linéaires ont été utilisées dans la littérature scientifique. Zubair et al. [230] ont développé un système de détection du stress à cinq niveaux basé sur les signaux PPG collectés à l'aide d'un capteur de pouls. Costin et al. [231] ont ajouté des caractéristiques de variabilité morphologique aux caractéristiques susmentionnées. La variabilité morphologique est une technique complémentaire qui permet d'améliorer les résultats des caractéristiques de la VFC en utilisant les changements dans les données morphologiques des battements [231]. Bousefsaf et al [232] ont montré que le stress mental peut être estimé à partir de la variabilité de la fréquence cardiaque mesurée à distance par une caméra à faible coût. Mcduff et al [233] ont étudié la relation entre le stress cognitif et l'hémodynamique périphérique et la vasomotion mesurées par caméra.

Activité électrodermale

L'activité électrodermale est l'un des signaux les plus utilisés pour la détection du stress et elle permet d'obtenir un excellent aperçu de l'activité du système nerveux [234]. L'EDA correspond à une mesure des variations des caractéristiques électriques de la peau, résultant de l'activité des glandes sudoripares, enregistrées sous forme de variations du potentiel ou de variations de la résistance cutanée [235]. En cas d'excitation émotionnelle, de charge cognitive accrue ou d'activité physique, le niveau de transpiration augmente, ce qui modifie les propriétés de la peau, c'est-à-dire qu'il augmente la conductance et diminue la résistance [234]. Même une petite quantité de sueur, qui n'est pas perceptible à la surface de la peau humaine, entraîne un changement de sa conductivité électrique. L'une des recherches les plus pionnière sur la détection du stress a été menée par Healey et al. [25]. L'EDA ainsi que trois autres mesures physiologiques, à savoir l'ECG, l'EMG du muscle trapèze et la respiration ont été utilisées pour la classification de trois niveaux de stress à l'aide du classifieur LDA. L'EDA et la VFC se sont avérées être les meilleurs corrélats du stress par rapport aux autres signaux physiologiques. Le même classifieur a été utilisé avec la projection de Fisher pour discriminer entre trois niveaux de stress : faible, moyen et élevé en se basant sur le signal d'activité électrodermale [236]. Un système de mesure du stress chronique utilisant l'EDA est présenté dans [237]. Les états stressés et détendus sont discriminés en utilisant le modèle de Bayes naïf, les arbres de décision et les forêts aléatoires.

Activité cérébrale

L'activité cérébrale a une forte relation avec le stress [238]. Elle peut être enregistrée à l'aide de l'imagerie par résonance magnétique fonctionnelle, la tomographie par émission de positrons ou l'EEG. Parmi toutes ces méthodes, l'EEG est la plus utilisée en raison de son faible coût et de sa nature non invasive. Le signal EEG comprend cinq bandes de fréquences différentes où le comportement de chaque bande dépend de la situation à laquelle la personne est confrontée [239]. La gamme de fréquences du signal EEG est comprise entre 0.1 Hz et 50 Hz. En fonction des plages de fréquences l'ordre croissant des ondes cérébrales est delta, thêta, alpha, bêta et gamma. Lors de situations stressantes, l'activité alpha diminue car elle reflète un état de relaxation, tandis que l'activité bêta augmente avec la charge mentale [239]. Le stress a également été associé à des modifications de l'activité frontale droite provoquant une asymétrie frontale. L'analyse de l'asymétrie du signal EEG permet de discriminer les différents états psychologiques [240].

Al-Shargie et al. [241] ont examiné la faisabilité des signaux EEG dans la détection des

niveaux de stress mental. La transformée en ondelettes a été appliquée pour décomposer les signaux EEG en quatre bandes de sous-fréquence. Un SVM a été utilisé pour classifier deux types de caractéristiques extraites dans chaque sous-bande des signaux EEG, à savoir la densité spectrale de puissance et l'énergie moyenne. Une analyse du stress à trois niveaux a été étudiée par Jun et al. [242] en se basant sur le test de Stroop et le test de calcul mental respectivement. La différence relative de la puissance bêta et alpha a été utilisée comme caractéristique et les SVM comme classifieur. Calibo et al. [243] ont calculé la tension quadratique moyenne dans les bandes bêta, alpha et thêta. Ces caractéristiques ont ensuite été utilisées pour classifier les état de stress et non-stress à l'aide de la régression logistique et le kNN. Une méthode de mesure du stress basée sur l'EEG et la transformée de Hilbert a été proposée dans [244]. Trois algorithmes de classification différents, à savoir les SVM, les réseaux de neurones et les forêts aléatoires, ont été utilisés pour la mesure du stress du conducteur. Le tableau 1.6 présente un résumé des systèmes de reconnaissance du stress à partir des signaux physiologiques.

TABLE 1.6 – Tableau comparatif des systèmes de reconnaissance du stress à partir des signaux physiologiques.

Méthode	**Modalité**	**Base de données**	**Caractéristique**	**Classifieur**	**Précision (%)**
Karthikeyan et al. [245]	CARD	Privée	Ondelettes	kNN	96.3
Mcduff et al. [246]		Privée	Fréquentielles	SVM	85
Tanev et al. [247]		Privée	Temporelles Fréquentielles	NB	90
Keshan al. [248]		PhysioNet	Temporelles	SVM	88
Giannakakis et al. [249]		Privée	Temporelles	AdaBoost	91.68
Calibo et al. [243]		Privée	Fréquentielles	LR	73.96
Saeed et al. [250]		Privée	Fréquentielles	RF	75.12
Al-shargie et al. [251]		Privée	Ondelettes	SVM	94
Setz et al. [252]	EDA	Privée	Temporelles	LDA	82.80
Ren et al. [253]		Privé	Temporelles	NB	85.5
Panigrahy et al. [254]		Privée	Temporelles	DT	76.50
Zhai et Baretto [255]	TEMP	Privée	Temporelles Fréquentielles	SVM	90.10
Fernandez et Anishechenko [256]	RESP	Privée	Temporelles Fréquentielles	MLP	94.44
Karthkeyan et al. [257]	EMG	Privée	Ondelettes	kNN	90.70

1.3.3.2 Réponses comportementales

Les expressions faciales

Le stress a une corrélation importante avec les expressions faciales. Il a été constaté que le stress provoque des réactions involontaires sur les muscles faciaux [258]. L'intensité moyenne du sourire, l'activité des sourcils et l'activité de la bouche sont les principales caractéristiques faciales permettant de détecter le stress [239]. Les mouvements des sourcils et les asymétries de la bouche ont été extraits par Dinges et al. [259]. Ces caractéristiques ont été utilisées comme entrées d'un modèle de Markov caché pour quantifier l'état de stress élevé et faible. Zhang et al. [260] ont proposé un système temps réel pour détecter le stress en reconnaissant les expressions faciales associées à la colère, la peur et la tristesse. Ils ont utilisé quatre couches de réseaux convolutifs connectés, qui combinent des caractéristiques de bas niveau avec des caractéristiques de haut niveau pour entraîner un réseau profond visant à identifier les expressions faciales. Giannakakis et al. [249] ont développé un système capable de détecter les états de stress/anxiété grâce à des indices faciaux non volontaires et semi-volontaires. Une multitude de techniques, telles que le modèle actif d'apparence et le flux optique ont été appliquées pour l'extraction des caractéristiques tandis que la meilleure précision de classification a été obtenue avec le classifieur KNN. Un système de détection du stress a été proposé par Og et al. [261] pour les représentants du service client basé sur les expressions faciales et l'apprentissage profond. Il se compose principalement d'un module de détection des visages en temps réel, d'un module de classification des émotions et d'un module de surveillance qui visualise uniquement les niveaux de stress. Des modèles CNN tels que VGG16, VGG19 et ResNet V2 ont été utilisés dans [262] pour mener des expériences sur la reconnaissance des expressions faciales pour la détection du stress. La colère, le dégoût et la peur ont été considérés comme des émotions de stress tandis que neutre et tristesse ont été associées à un état de non-stress. Un modèle de classification a été créé en utilisant une méthode d'apprentissage par transfert et une technique de fine-tuning. Sur la base d'une approche d'apprentissage profond, une méthode légère et fiable pour la détection du stress a été proposée [263]. Un apprentissage par transfert basé sur MobileNet V2 a été utilisé pour la quantification binaire de l'état du stress. Viegas et al. [264] ont extrait 17 unités d'action liées au niveau supérieur du visage. Ensuite, quatre méthodes classiques d'apprentissage automatique, à savoir les forêts aléatoires, LDA, Gaussian Naive Bayes et les arbres de décision ont été utilisées pour détecter le stress mental. La distance euclidienne des points de repère faciaux a été utilisée dans [265] pour l'estimation de l'ennui et le stress d'un joueur de jeu vidéo.

TABLE 1.7 – Tableau comparatif des systèmes de reconnaissance du stress à partir des signaux comportementaux.

Méthode	**Modalité**	**Base de données**	**Classifieur**	**Précision (%)**
Gao et al. [266]	Expressions faciales	Radboud	SVM	90.50
Viegas et al. [264]		Privée	RF	75
Zhang et al. [260]		CK+	CNN	91.2
		Oulu-CASIA		80.4
		KMU-FED		99.3
Og et al. [261]		KETI, KDEF	Xception	98.67
Almeida et al. [262]		KDEF, CK+	VGG16	92
Soury et al. [267]	Parole	Privée	SVM	72
Lu et al. [268]		Privée	GMM	81
Simantiraki et al. [269]		SUSAS	RF	90.06
Ren et al. [253]	Mouvement des yeux	Privé	NB	84.21
Pedrotti et al. [270]		Privée	ANN	79.20
Baltaci et Gokcay [271]		Privée	AdaBoost	72.10
Aigrain et al. [272]	Mouvement du corps	Privée	SVM	77
Giannakakis et al. [273]		Privée	kNN	98.6

La parole

Le stress provoque des changements dans le mécanisme de génération de la voix humaine [274]. L'analyse de la parole a suscité un grand intérêt principalement parce qu'elle peut être mesurée facilement et de manière totalement discrète et non-invasive [275, 276]. Néanmoins, les systèmes de reconnaissance du stress basés sur la parole peuvent être inefficaces dans des situations de silence ou dans des environements bruyants [275].Les caractéristiques vocales peuvent être classées en trois composantes, à savoir l'excitation de la parole, le tractus vocal et le signal vocal. Lors des situations de stress, la première composante est affectée par la tension accrue du muscle des plis vocaux, la deuxième composante est affectée par le changement de position des articulateurs du tractus vocal et la troisième composante est due à son lien avec les deux autres composantes [277]. La hauteur, le débit de parole, l'énergie et les caractéristiques spectrales sont affectés par les événements stressants. La hauteur (moyenne, écart-type), le rapport des bandes de fréquences supérieures, la vitesse d'élocution, l'intensité de la voix, l'énergie lissée, le rapport voix-non-voix, les coefficients cepstraux de fréquence Mel (MFCC) sont les caractéristiques les plus affectées par les événements stressants et sont largement utilisées pour détecter les niveaux de stress [239]. Fernandez et al. [278] ont utilisé cinq modèles de Markov cachés différents pour reconnaître quatre niveaux de stress des conducteurs à partir des formes d'onde de leur voix.

Les caractéristiques physiques des cordes vocales d'une personne ont été examinées par Yao et al. [279] pour identifier le stress dans la parole. Dans [280], l'auteur suggère que le changement de vitesse de la fréquence fondamentale de la voix est la caractéristique la plus importante pour la mesure du stress. La méthode proposée utilise les coefficients de puissance log-fréquence à court terme (LFPC) pour représenter les signaux vocaux et un modèle de Markov caché discret comme classifieur. Un autre système de classification du stress basé sur la parole et utilisant des caractéristiques du domaine spatial et spectral est présenté dans [281]. Dans [267], les auteurs ont extrait des caractéristiques prosodiques, de qualité de la voix et spectrales sur des fenêtres de taille variable, afin de prédire les états de stress pendant une intervention en public à l'aide d'un SVM. Des méthodes basées sur l'apprentissage profond ont été également proposées. Han et al. [282] ont développé un algorithme pour déterminer l'état de stress via les caractéristiques MFCC de la voix en utilisant un critère de décision binaire impliquant deux couches de réseaux neuronaux récurrents à mémoire à long terme (LSTM-RNN) et un classifieur SVM. Le tableau 1.7 résume les systèmes de reconnaissance du stress à partir des signaux comportementaux.

1.3.4 Reconnaissance multimodale du stress

La reconnaissance multimodale du stress humain a fait l'objet d'un large éventail d'études dans la littérature scientifique. Carneiro et al. [283] affirment que pour une mesure suffisamment précise et exacte du stress, une approche multimodale doit être envisagée. Les modalités couramment utilisées pour la détection du stress comprennent la vidéo, l'audio, le texte et les signaux physiologiques [239]. Cependant, les signaux physiologiques s'avèrent plus représentatifs que les autres modalités. Une pléthore de systèmes multimodaux se sont basés sur la fusion des données physiologiques uniquement. L'une des combinaisons de signaux physiologiques les plus couramment utilisées pour détecter le stress est la fusion de l'activité cardiaque et de l'activité électrodermale. Sierra et al. [284] ont proposé un système adapté aux applications en temps réel basé sur les signaux de l'activité électrodermale et la fréquence cardiaque. Une tâche de discours en public est utilisée comme stimulus et le classifieur KNN est appliqué pour détecter l'état de stress et de relaxation. Les caractéristiques temporelles et fréquentielles des signaux PPG et EDA sont utilisées pour identifier le stress chez les sujets dans une étude menée dans [285]. Un classifieur de type SVM a été utilisé pour classer le stress en deux niveaux.

D'autres schémas de fusion ont été également proposés en combinant deux signaux physiologiques ou plus. Un système de reconnaissance du stress des conducteurs en utilisant l'activité cardiaque, l'activité électrodermale et la respiration a été proposé dans [298]. Les caractéristiques sont extraites dans le domaine temporel et fréquentiel et des ondelettes, tandis que les

TABLE 1.8 – Tableau comparatif des systèmes multimodaux de reconnaissance du stress.

Méthode	**Modalité**	**Base de données**	**Caractéristique**	**Classifieur**	**Précision (%)**
Xia et al. [286]	EEG + ECG	Privée	Temporelles Fréquentielles	SVM	79.54
Kurniawan et al. [287]	Audio + EDA	Privée	Temporelles Fréquentielles	SVM	92.47
Anusha et al. [288]	EDA + ECG + TEMP	Privée	Temporelles Fréquentielles	kNN	95.86
Zhang et al. [289]	EF + Audio + ECG	Privée	Temporelles Fréquentielles	Att-ResNet50	85.7
Giakoumis et al. [290]	ECG + EDA + Acc + Video	Privée	Temporelles	LDA	100
Zhai et al. [291]	DP + BVP + EDA + TEMP	Privée	Temporelles	SVM	90.10
Mazos et al. [292]	EDA + Acc + BVP + Audio	Privée	Temporelles Fréquentielles	AdaBoost	94
Bhatti al. [293]	ECG + EDA + RESP + TEMP	WESAD	Temporelles	VGG	89.57
Walambe et al. [294]	EF + Posture + VFC + Int	SWELL-KW	Temporelles Fréquentielles	ANN	96.67
Bobade et Vani [295]	Acc + ECG + BVP + TEMP + RESP + EMG + EDA	WESAD	Temporelles	ANN	95.21
Koldijk et al. [296]	EF + Posture + FC + EDA + Int	SWELL-KW	Temporelles	SVM	90
Bara et al. [297]	Audio + Video + FC + TEMP + EDA + RESP	MuSE	Temporelles Fréquentielles	AE-GRU	88.5

- Acc : signaux d'accélération - EF : Expressions faciales - AE : Auto-encodeur - DP : Diamètre de pupille
- Int : Intéraction avec l'ordinateur - Att : Mécanisme d'attention - TEMP : Température de la peau

algorithmes Analyse en Composantes Principales et Apprentissage Bayésien Clairsemé ont été appliqués pour la sélection des caractéristiques et SVM pour la classification. Healey et al. [25] ont combiné plusieurs caractéristiques physiologiques provenant de l'activité cardiaque, l'activité électrodermale, l'activité électrique des muscles et la respiration. Gjoreski et al. [299] ont fusionné cinq signaux physiologiques à savoir le signal PPG, la fréquence cardiaque, la température de la peau, l'activité électrodermale et la respiration. Une tâche d'arithmétique mentale est utilisée pour induire du stress et les forêts aléatoires ont été utilisées pour la classification des 63 caractéristiques extraites dans le domaine temporel et fréquentiel. Deux expériences ont été réalisées pour la validation de leur système. La première expérience se déroule en laboratoire tandis que la seconde est effectuée hors laboratoire. La fusion des signaux EEG, ECG, EMG et EOG a été réalisée par Akhonda et al. [300]. Un réseau de neurones a été utilisé comme classifieur pour la reconnaissance du stress en trois niveaux.

La fusion des signaux physiologiques avec des données contextuelles a été également proposée.

Maier et al. [301] ont combiné les caractéristiques de la variabilité caradiaque issues de l'ECG et les données du GPS et de l'accéléromètre. Un autre schéma de mesure multimodale du stress a été proposée dans [302]. Des données d'accéléromètre et de l'activité électrodermale obtenues à partir des capteurs de poignet, les données d'appel et de SMS, la localisation et les caractéristiques d'activation et de désactivation de l'écran obtenues à partir des téléphones mobiles, et le score de stress subjectif à l'aide de questionnaires ont été combinés. Dans un autre travail, le signal PPG et les données des capteurs inertiels de l'accéléromètre, du gyroscope et du magnétomètre ont été utilisées pour la classification de l'état de stress [303]. Les caractéristiques du domaine temporel et fréquentiel sont extraites des données acquises et la classification du stress est effectuée à l'aide d'un classifieur SVM avec une fonction de base radiale (RBF).

Contrairement aux émotions où quasiment toutes les combinaisons ont été traitées, on trouve beaucoup moins de schémas de fusion pour la reconnaissance du stress. Vizer et al. [304] ont combiné les frappes du clavier ainsi que des caractéristiques linguistiques du texte écrit tandis que la fusion des caractéristiques de l'apparence physique, des signaux physiologiques et des données comportementales a été proposée dans une étude menée par Liao et al. [305]. Les caractéristiques du diamètre des pupilles et des signaux physiologiques ont été fusionné dans [306] tandis que les données de l'accéléromètre et les vidéos sont analysées dans [290] pour surveiller la réponse au stress. Le tableau 1.8 présente un résumé des schémas de classification multimodale du stress humain disponibles dans la littérature scientifique.

1.4 Conclusion

Ce chapitre avait pour objectif de donner un aperçu général sur l'émotion et le stress. Nous avons d'abord évoqué les notions de bases telles que les définitions de l'émotion et du stress, leurs théories et leurs techniques d'élicitations, puis présenté les méthodes de la littérature scientifique sur les systèmes de reconnaissance automatique des émotions et du stress. Nous nous sommes focalisés sur la reconnaissance de l'état affectif à partir des signaux physiologiques et les expressions faciales ainsi que sur la fusion des deux modalités. Les premiers travaux sur la reconnaissance de l'état affectif ont été principalement des approches unimodales. Bien qu'il existe des avancées dans la reconnaissance unimodale des émotions, en raison de la nature multimodale de l'expression des émotions et de stress, de telles approches restent incapables dans certaines circonstances notamment dans les conditions non controlées. Les études récentes ont montré que l'exploitation de la complémentarité des différentes modalités conduit à de meilleurs résultats et surpasse les approches unimodales.

2

Mesure sans contact de la fréquence cardiaque par caméra

Sommaire

2.1 Introduction

La fréquence cardiaque (FC) est l'un des indicateurs importants pour évaluer l'état émotionnel et l'état de santé d'une personne. La FC est régulièrement surveillée pour identifier différents problèmes de santé. L'électrocardiographie (ECG) et la photopléthysmographie (PPG) sont les principaux moyens permettant de mesurer l'activité cardiaque. Ces deux techniques utilisent des capteurs en contact qui doivent être attachés aux parties du corps et nécessitent certaines conditions pour obtenir de bonnes mesures. Malgré la grande précision et la robustesse fournies par ces dispositifs invasifs, le contact avec la peau peut être gênant voire infaisable en raison de brûlures, d'ulcères cutanés ou de maladies contagieuses. Ces contraintes limitent leur champ d'utilisation.

Ce chapitre présente en premier lieu le fonctionnement du coeur et les principales techniques pour l'estimation de la fréquence cardiaque. Nous nous concentrons ensuite sur la photoplethysmographie par imagerie que nous allons développer tout au long de cette thèse. Nous présenterons en second lieu un état de l'art sur les systèmes de mesure sans contact de la fréquence cardiaque basés sur la photoplethysmographie par imagerie.

2.2 Fonctionnement du cœur

Le cœur est un organe complexe possédant un automatisme permettant la contraction et le relâchement pseudo-périodique. Les contractions du cœur sont déclenchées par des impulsions électriques qui proviennent d'une région située dans l'oreillette droite, appelée le nœud sino-atrial (SAN) et qui sont fortement régulées par le système nerveux autonome (SNA). Le SAN agit comme un stimulateur cardiaque naturel, tandis que son taux d'oscillation s'adapte aux changements de la demande énergétique de l'organisme et aux facteurs environnementaux. Le courant électrique généré est ensuite transmis par des voies spécifiques dans tout le cœur, ce qui permet une contraction et une relaxation régulières. Ce courant électrique peut être détecté à la surface du corps grâce à des électrodes adhésives, notamment sur la paroi thoracique, en utilisant

l'électrocardiographie. A chaque battement cardiaque, le cœur pompe en permanence du sang vers le reste du corps. La circulation du sang dans les vaisseaux entraîne une modification de l'absorption de certaines longueurs d'onde lors de son passage dans un tissu et par conséquent modifie le niveau de l'absorption de la peau. Cette variation est proportionnelle à la modification temporelle du volume sanguin dans le tissu cutané observé et permet de mesurer l'activité cardiaque grâce à la photopléthysmographie.

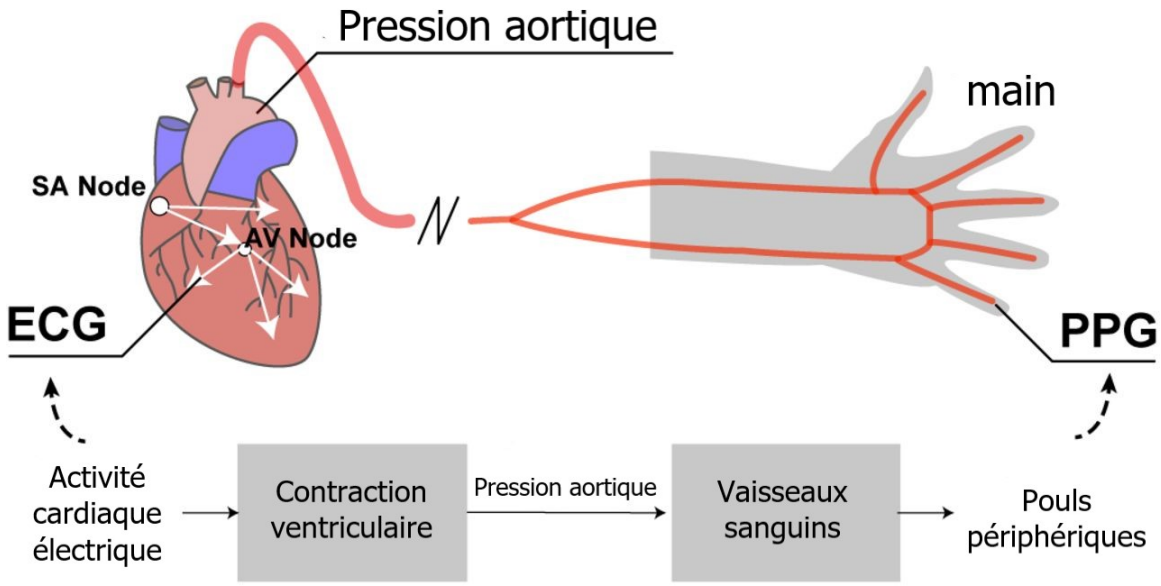


FIGURE 2.1 – Une visualisation de la relation entre l'ECG et la PPG.

2.3 Fréquence cardiaque

La fréquence cardiaque désigne le nombre de contractions du cœur par unité de temps, généralement en battements par minute (Bpm). Un adulte en bonne santé a généralement une fréquence cardiaque au repos comprise entre 60 et 80 bpm [307]. La fréquence cardiaque maximale est fortement corrélée à l'âge [308]. L'équation pour l'approximer est la suivante $FCmax = 208 - 0,7Age$. Pour une personne âgée de 20 ans, on obtient $FCmax = 194bpm$. Selon cette théorie, la fréquence cardiaque maximale des enfants est plus élevée que celle des adultes et elle diminue avec l'âge. La fréquence à laquelle le cœur bat est dictée par les besoins actuels de l'organisme tels que l'augmentation de l'oxygène et l'excrétion du dioxyde de carbone. Les activités susceptibles d'entraîner un changement des battements cardiaques sont, entre autres, l'exercice, la somnolence, l'état émotionnel, le stress, la maladie, les médicaments, la température du corps et la température ambiante, la déshydratation.

2.4 Variabilité de la fréquence cardiaque

La variabilité de la fréquence cardiaque (VFC) est le phénomène physiologique de variation de l'intervalle de temps entre les battements cardiaques consécutifs. Elle dépend principalement de la régulation extrinsèque de la fréquence cardiaque et reflète la capacité du cœur à s'adapter rapidement à des stimuli [309]. Les fluctuations des battements d'un cœur sain devraient être chaotiques selon les définitions mathématiques [310]. Cela s'explique par le fait que le cœur doit être capable de se réadapter rapidement aux changements d'environnement et d'activité physique. L'analyse de la VFC permet d'évaluer la santé cardiaque globale et fournit des indications sur l'état du système nerveux autonome responsable de la régulation de l'activité cardiaque et affective [311]. La variation de la VFC peut contenir des indications d'une maladie actuelle, ou des avertissements sur des maladies cardiaques imminentes. Les indicateurs peuvent être présents à tout moment ou se produire de manière aléatoire, à certains intervalles de la journée [310].

Les modifications de la VFC sont régulées par l'action synergique des deux branches du système nerveux autonome, à savoir le système nerveux sympathique (SNS) et le système nerveux parasympathique (SNP). Une augmentation de l'activité du SNS ou une diminution de l'activité du SNP entraîne une accélération du rythme cardiaque. Inversement, une faible activité du SNS ou une forte activité du SNP entraîne une décélération cardiaque. Cette modification donne un aperçu de l'état et de l'intégrité du système nerveux autonome, qui peut à son tour indiquer un état de stress, de somnolence ou des facteurs connexes. L'analyse de la VFC peut être effectuée à l'aide de méthodes linéaires, qui se divisent en domaine temporel et domaine fréquentiel, et également à l'aide de méthodes non linéaires. Cependant l'analyse temporelle et fréquentielle sont les méthodes les plus populaires pour mesurer la variabilité cardiaque.

2.4.1 Domaine temporel

L'analyse temporelle est considérée comme la mesure la plus simple de la VFC qui est souvent calculée comme l'intervalle de temps entre deux pics de battements cardiaques sur une mesure électrocardiographique (intervalle RR) ou photopléthysmographique (IBI) (voir Figure 2.2). Ces durées temporelles sont abrégées par IBI dans ce manuscrit (Inter-Beat Interval en anglais). Les caractéristiques couramment extraites comprennent la FC moyenne, l'écart type et sa racine carrée qui reflète la variabilité totale pendant la période d'enregistrement. L'analyse dans le domaine temporel fournit en outre d'autres caractéristiques pertinentes qui sont dérivées des IBI, notamment la moyenne quadratique des IBI successifs (RMSSD), la différence entre la FC la plus élevée et celle la plus basse, ou la variation du rythme cardiaque entre le jour et la nuit.

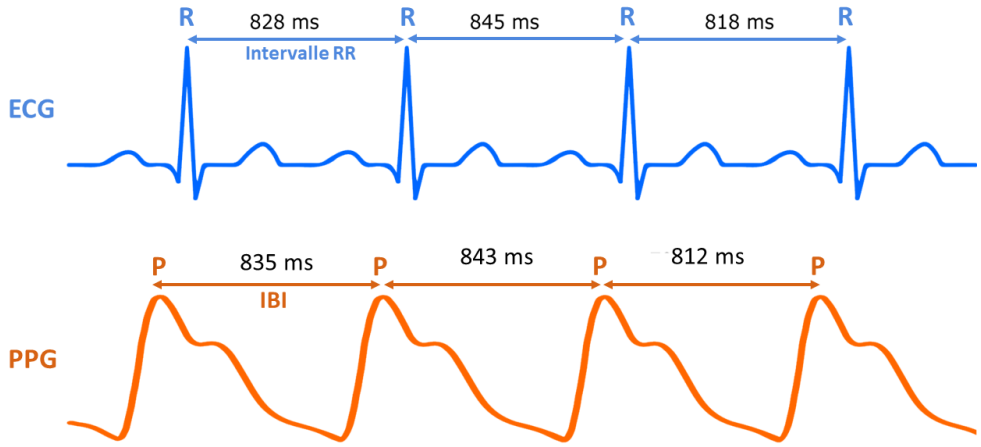


FIGURE 2.2 – La variabilité de la fréquence cardiaque consiste à analyser l'évolution des variations de temps entre chaque intervalle RR pour une mesure ECG ou les IBI pour la PPG.

2.4.2 Domaine fréquentiel

Contrairement au domaine temporel, l'analyse fréquentielle consiste à calculer la densité spectrale de puissance (DSP) pour décomposer la VFC en composantes spectrales spécifiques qui opèrent dans différentes bandes de fréquence. Le calcul de la DSP peut être obtenu par des méthodes paramétriques en utilisant un modèle autorégressif ou non paramétriques à l'aide de la FFT. Les enregistrements à court terme (1-5 minutes) de la VFC permettent de distinguer trois composantes spectrales principales : une composante haute fréquence (HF) (comprise entre 0,15 et 0,40 Hz), une composante basse fréquence (BF) (comprise entre 0,04 et 0,15 Hz) et une composante très basse fréquence (TBF) (comprise entre 0.003 et 0,04 Hz). Ces composantes spectrales sont considérées comme des marqueurs du contrôle parasympathique et sympathique [312]. L'analyse sur 24 heures permet également d'obtenir une quatrième composante, représentée par l'ultra basse fréquence (UBF) ($<$ 0,003 Hz) [313].

2.5 Mesure de l'activité cardiaque

La mesure des signaux physiologiques, notamment la fréquence cardiaque et sa variabilité, est l'un des premiers gestes les plus pratiqués dans la clinique quotidienne [314]. Les signes vitaux sont avant tout des indicateurs critiques qui peuvent informer les professionnels de santé sur le bien-être physique ou psychologique d'une personne. Ils permettent donc le dépistage et le traitement médical initial de plusieurs maladies. Les paramètres physiologiques sont souvent mesurés à l'aide de capteurs invasifs ou non invasifs en contact direct avec le corps humain.

2.5.1 Electrocardiographie

L'électrocardiographie (ECG) est considérée comme la référence en matière de mesure de l'activité cardiaque et pour le diagnostic de pathologies cardiovasculaires. Le principe de l'ECG repose sur la mesure des changements électriques à la surface de la peau provoquées par la dépolarisation des cellules du myocarde à chaque battement cardiaque. Les impulsions électriques sont enregistrées à distance du cœur, à travers la peau, via des électrodes. Cela fait de l'ECG une méthode non invasive pour mesurer l'activité cardiaque. Lorsque les cellules du myocarde sont au repos, il existe une différence de potentiel formée par la différence de concentration des ions positifs et négatifs de part et d'autre de la membrane des cellules du myocarde. L'impulsion électrique se propage de l'électrode polarisée négativement à celle polarisée positivement et génère une déflexion positive sur l'électrocardiogramme. Au contraire, le signal évolue négativement lorsque le vecteur électrique est opposé à la polarisation des électrodes [315].

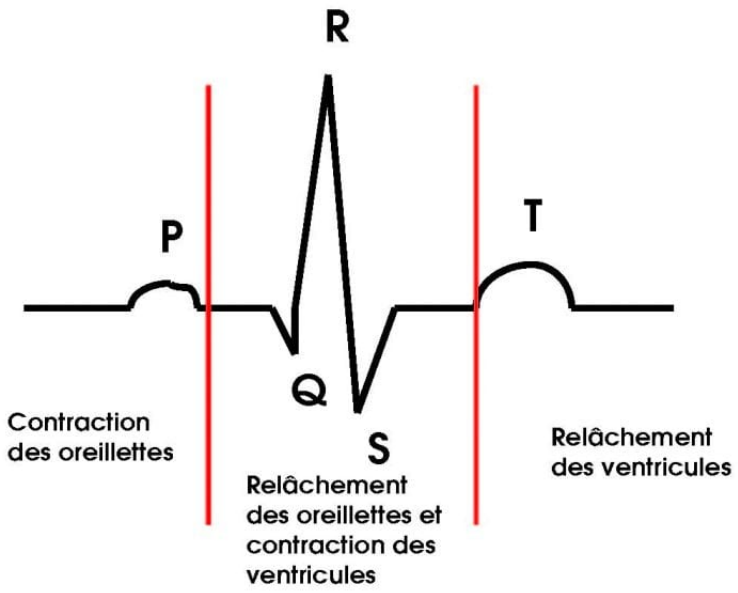


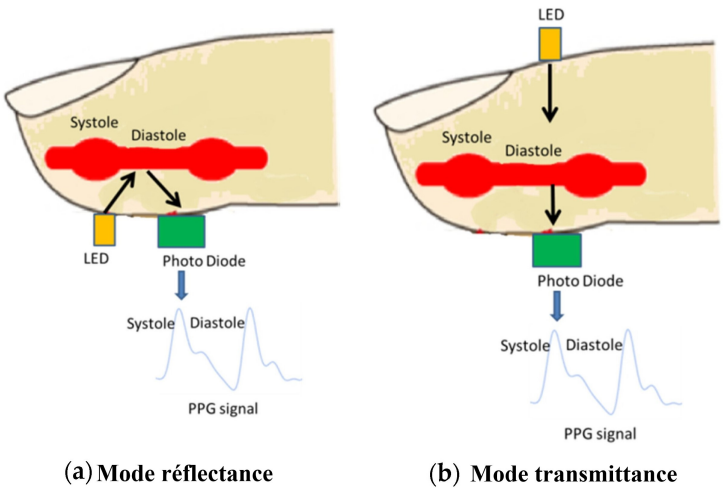
FIGURE 2.3 – Tracé et composition d'un électrocardiogramme.

Un ECG comporte trois composantes principales (Figure 2.3) : le complexe QRS, qui représente la dépolarisation lors de la contraction ventriculaire ; l'onde P, qui représente la dépolarisation lors de la contraction auriculaire ; et l'onde T, qui représente la repolarisation des ventricules. Plusieurs informations peuvent être déduites à partir de la localisation de ces ondes, parmi lesquelles la fréquence cardiaque et sa variabilité qui correspond à l'intervalle de temps entre les ondes R successives.

2.5.2 Photopléthysmographie

La photopléthysmographie (PPG) est une technique émergente et prometteuse qui a remplacé l'Électrocardiographie dans plusieurs applications. Elle consiste en une mesure optique de l'activité cardiaque en observant les variations de volume sanguin dans un tissu biologique de manière non-invasive [316]. La PPG a été introduite pour la première fois en 1937 par Alrick B. Hertzman [317], qui a mesuré les variations de l'absorption de la lumière à travers la peau humaine à l'aide d'une cellule photoélectrique placée sur une région de la peau (le doigt) éclairée par une source lumineuse située au-dessus, comme le montre la figure 2.4. La lumière étant plus fortement absorbée par le sang que par les tissus environnants, les variations du flux sanguin peuvent être détectées par le capteur PPG comme des variations de l'intensité de la lumière. Chaque battement cardiaque est détecté grâce au pic de concentration sanguine qu'il induit et donc, aux variations de la lumière reçue.

Un capteur PPG basique se compose d'un photodétecteur, d'un préamplificateur, d'un filtrage et de diodes électroluminescentes fonctionnant à différentes longueurs d'onde d'émission. Deux méthodologies de mesure de la PPG sont disponibles : le mode de transmittance et le mode de réflectance. Le mode de transmittance est utilisé dans la plupart des dispositifs médicaux et consiste à placer le capteur sur le bout doigt ou sur l'oreille. La source de lumière est placée sur la paroi interne et le photo-détecteur est placé sur la paroi externe, ce dernier va donc capter la lumière résultante de la transmission (Figure 2.4 (b)). En mode réflectance, l'émetteur et le récepteur doivent être placés cote à cote et le récepteur va donc mesurer la lumière réfléchie par la peau (Figure 2.4 (a)).



(a) Mode réflectance (b) Mode transmittance

FIGURE 2.4 – Les deux modes utilisés en photopléthysmographie en contact.

Le signal PPG mesuré par le capteur est proportionnel à la quantité de sang qui circule dans les vaisseaux sanguins et il comprend des composantes variable et continue (voir Figure 2.5). La composante variable (AC) correspond aux variations du volume sanguin en synchronisation avec le rythme cardiaque, tandis que la composante continue (DC) est cadencée par le rythme respiratoire, l'activation du système nerveux sympathique et la thermorégulation. Elle dépend également des signaux optiques réfléchis ou transmis par la peau, le mouvement et le bruit du capteur. Plusieurs paramètres physiologiques peuvent être mesurés à partir du signal PPG comme la fréquence cardiaque, le taux d'oxygène dans le sang ou la pression artérielle.

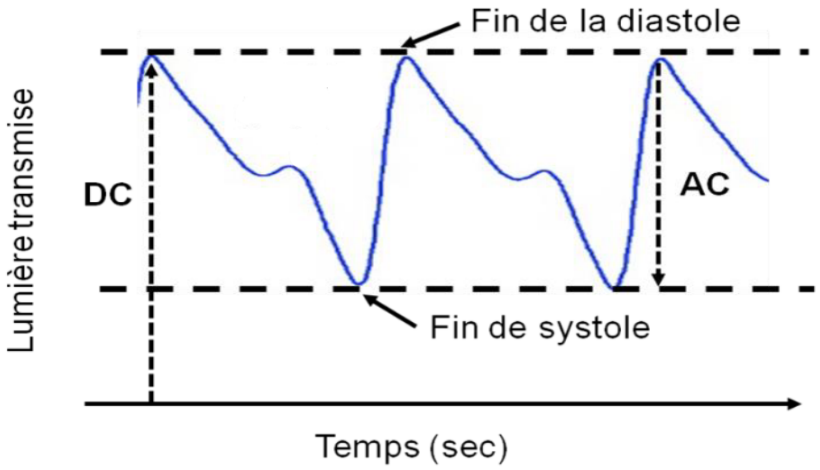


FIGURE 2.5 – Les deux composantes du signal PPG.

2.5.3 Photopléthysmographie par imagerie

Le principe de base de la photopléthysmographie par imagerie (Imaging Photoplethysmography, iPPG), appelée également photopléthysmographie à distance (Remote Photoplethysmography, RPPG), est similaire à la PPG en contact en mode réflectance où la source de lumière et le photodétecteur sont situés l'un à côté de l'autre. La lumière ambiante constitue la source lumineuse remplaçant les diodes électro-luminescentes (Light-Emitting Diode, LED) et une caméra joue le rôle de récepteur au lieu d'une photodiode. Plusieurs types de caméra peuvent être utilisées, allant d'une simple webcam ou d'une caméra d'un smartphone à des caméras plus complexes comme les caméras infra-rouge ou multispectrales. Cependant, afin de réduire le coût et rendre la technologie plus accessible, la majorité des travaux se sont concentrés sur l'utilisation de simples caméras avec une faible fréquence d'échantillonnage permettant un déploiement à très grande échelle, ce type de caméra étant déjà disponible sur presque tous les appareils électroniques que nous utilisons quotidiennement.

La iPPG permet une mesure sans contact de la variation du volume sanguin dans un tissu biologique en observant la quantité de la lumière réfléchie par la surface de la peau (généralement le visage) filmée par une caméra. En pratique, il s'agit d'appliquer une moyenne spatiale pour chaque image contenant les pixels de la peau afin de construire des signaux iPPG bruts qui sont généralement de qualité dégradée [318]. Des techniques de filtrage et de supression des tendances sont par la suite appliquées pour améliorer la qualité du signal. Les paramètres physiologiques sont ensuite estimés dans le domaine temporel ou fréquentiel. Grâce à sa nature sans contact, la méthode iPPG présente de nombreux avantages. Tout d'abord, elle limite les restrictions physiques et réduit le câblage associé à la surveillance des patients. En outre, elle augmente la sécurité des patients et du personnel médical en minimisant le risque de contamination en cas de maladie contagieuse. Actuellement, les mesures sans contact prennent une place de plus en plus importante dans le domaine de la télémédecine et la sécurité. Elles sont ouvertes sur nombreux champs d'application notamment la prévention de maladies cardiovasculaires graves et les soins médicaux, les mesures à domicile, l'automobile et les sciences de sport.

Cette thèse se concentre sur la photopléthysmographie par imagerie qui quantifie uniquement les changements de couleur de la peau dus aux pulsations sanguines. Néanmoins, la photopléthysmographie sans contact présente ses propres défis, en particulier les artefacts de bruit dus aux mouvement et perturbations d'illumination. Ces problèmes ne sont pas triviaux, bien qu'ils soient suffisamment indépendants pour pouvoir être résolus séparément à l'aide d'algorithmes pertinents et intégrés dans un cadre de mesure de la iPPG.

2.5.4 Applications de la photopléthysmographie par imagerie

2.5.4.1 Surveillance médicale

La mesure à distance des signes vitaux est la principale application de la photopléthysmographie par imagerie. Le suivi régulier des signaux physiologiques courants, notamment la fréquence cardiaque, la pression artérielle et la saturation en oxygène, peut prévenir ou indiquer des problèmes cardiaques ou la probabilité d'un accident vasculaire cérébral.

La surveillance par caméra est applicable dans des environnements cliniques et non cliniques. En milieu clinique, l'iPPG peut remplacer les dispositifs en contact intrusifs (tels que la PPG et l'ECG) qui doivent être fixés sur la peau du patient. Cela est particulièrement utile lorsque le contact avec la peau est gênant ou impossible, en particulier pour des patients en gériatrie, pour les nouveau-nés [319], ou en raison de certaines conditions telles que les brûlures, les ulcères cutanés et les maladies contagieuses. L'iPPG peut être utilisée également en milieu non clinique comme un dispositif de télémédecine pour un suivi régulier des signaux physiologiques [320]. Au

quotidien, les signes vitaux peuvent être survéillés pendant les exercices physiques, la prisende repas, le sommeil, ou en phase de repos.

2.5.4.2 Analyse des émotions

Les signaux physiologiques tels que l'activité cardiaque, cérébrale et électrodermale sont des données fiables pour quantifier les émotions car ils varient en fonction de l'intensité de l'état émotionnel ressenti. Ils présentent plusieurs avantages par rapport aux expressions faciales et la parole qui sont facilement contrefaites. Cependant, ils demeurent contraignants dans leurs mise en oeuvre en raison de l'utilisation des dispositifs intrusifs qui peuvent interférer avec les sujets et modifier leur état émotionnel. De plus, la complexité de la mesure et la sensibilité des capteurs limitent fortement leur champs d'application. L'utilisation de l'iPPG dans l'informatique affective et l'analyse des émotions est une piste très prometeuse qui a commencé à attirer l'attention des chercheurs ces dernières années. L'utilisation d'une simple caméra permet à la fois de mesurer l'activité cardiaque, enregistrer les expressions faciales et capturer les mouvements des yeux et la posture, ce qui ouvre la voie au développement de systèmes multimodaux mono-capteur [321].

Cette méthode est particulièrement utile dans les situations où il est difficile de mesurer les signaux physiologiques de manière traditionnelle, par exemple lorsque la personne se trouve dans un environnement bruyant ou lorsqu'il est difficile de placer des capteurs intrusifs sur la peau de la personne. L'aspect sans contact de l'iPPG permet de mesurer les signaux physiologiques pour analyser les émotions dans différents contextes, tels que la recherche en psychologie, la recherche en marketing ou la recherche en neuroscience, pour étudier l'impact des émotions sur le comportement et les décisions des individus. L'iPPG peut également être utilisée dans le domaine de la santé pour surveiller l'état émotionnel des patients et pour aider à la prise en charge de troubles mentaux tels que l'anxiété ou la dépression [322].

2.5.4.3 Surveillance automobile

La surveillance de l'état de santé et affectif des conducteurs permet de prévenir ou éviter plusieurs accidents de la route. Les véhicules automobiles les plus récents sont désormais équipés de caméras embarquées et diverses technologies assistées par la vision par ordinateur sont intégrées pour la détection de la position de la tête, détection des clignements des yeux, suivi du regard, suivi des mouvements de la bouche, analyse des expressions faciales, etc [323, 324]. En plus de ces signaux comportementaux, la surveillance des signes vitaux est désormais possible grâce à l'iPPG. L'analyse des signaux physiologiques et comportementaux permet de déclencher une alarme en cas de détection de comportements dangereux du conducteur ou en cas d'un problème

de santé (fatigue, somnolence, valence et activation émotionnels, stress mental et anxiété.

2.5.4.4 Anti-falsification du visage

Les systèmes de reconnaissance faciale présentent un écueil liè à l'utilisation de faux visages ou de visages artificiels pour permettre l'accès non autorisé au système (appelé usurpation ou falsification de visage). L'usurpation de visage peut être atténuée en utilisant l'iPPG qui est capable de détecter si une personne est réelle ou si elle utilise une image enregistrée, un masque ou tout autre type de falsification. Pour ce faire, l'iPPG peut être utilisée pour mesurer les variations de couleur de la peau sur le visage de la personne pendant qu'elle parle ou respire [325]. Si les variations de couleur sont similaires à celles qui se produiraient naturellement chez une personne réelle, alors l'iPPG peut être utilisée pour confirmer l'authenticité de la personne. Si, en revanche, les variations de couleur ne sont pas similaires à celles qui se produiraient naturellement, cela peut être considéré comme un indicateur de falsification.

2.6 Défis des systèmes iPPG

Bien que les caméras numériques répondent au besoin d'une surveillance plus objective et moins intrusive des signes vitaux, elles ont leurs propres défis comme toute autre méthode de mesure. Il est intéressant d'examiner l'ensemble des facteurs techniques et environnementaux qui pourraient avoir un impact négatif sur les performances des systèmes iPPG. Cela implique le bruit et les spécificités de la caméra, le mouvement, le teint de peau, l'éclairage et l'arrière-plan.

2.6.1 Mouvement

Le mouvement du sujet représente un défi important pour la photopléthysmographie par imagerie. Dans un scénario idéal, lorsque le sujet reste immobile et que la lumière et la position de la caméra restent fixes, le signal iPPG constitué est propre et sans artefacts car il traduit la variation d'intensité des pixels de la peau seulement en plus de bruits des faibles mouvements naturels du corps et du capteur qui peuvent être éliminés par un filtrage passe-bande. Cependant, les mouvements significatifs du sujet modifient les réflexions lumineuses de la peau, ce qui entraînent une distorsion du signal iPPG dont l'amplitude est initialement faible [326].

2.6.2 Conditions de l'éclairage

L'iPPG repose sur la quantification de l'interaction de la lumière avec le sang. L'éclairage joue donc un rôle important dans les performances du système. La source lumineuse doit émettre

une énergie suffisante, car l'amplitude du volume sanguin mesuré est proportionnelle à l'intensité de l'éclairage ; autrement dit, plus l'intensité lumineuse est élevée, plus le signal iPPG extrait est fort [327]. Un éclairage insuffisant peut entraîner une faible amplitude du signal iPPG, car l'énergie de la lumière est trop faible pour pénétrer dans la surface de la peau. Au contraire, une intensité lumineuse très élevée entraîne une augmentation de la réflexion de surface de la peau qui peut saturer les intensités des pixels et conduire à un écrêtage de l'image [322].

2.6.3 Teinte de la peau

Les humains ont un large éventail de teintes de peau, en raison de la quantité variable de mélanine produite par la peau. Les tons de peau foncés contiennent une concentration plus élevée de mélanine que les tons de peau clairs. La mélanine, tout comme l'hémoglobine présente dans le sang, absorbe la lumière. Par conséquent, une concentration plus élevée de mélanine absorbera plus de lumière à l'intérieur des tissus cutanés et moins de lumière retournera vers la caméra. Comme l'iPPG utilise la lumière réfléchie par la peau, moins de réflexion signifie plus faible rapport signal/bruit et des signaux iPPG plus sujets à d'autres sources de corruptions (par exemple, le mouvement) [326]. Un autre problème se pose typiquement pour les femmes qui ont tendance à se maquiller, ce qui empêche une partie de la lumière de pénétrer dans la peau [326] et affaiblit encore le rapport signal/bruit des signaux iPPG.

2.6.4 Bruit de la caméra

Le fonctionnement de l'iPPG repose sur les mêmes principes que le mode réflectance de la photopléthysmographie en contact. L'utilisation d'une simple caméra et la lumière ambiante présente plusieurs avantages : la mesure est sans contact et son coût est faible. La qualité du signal obtenu et donc la fiabilité de la mesure est en contrepartie réduite. La plupart de la lumière incidente sur la surface de la peau est soit réfléchie, soit diffusée. Une infime partie de la lumière est absorbée par l'hémoglobine présente dans le sang, et ce changement d'absorption donne lieu au signal du pouls (signal iPPG). Comme l'intensité du signal iPPG capturé à chaque pixel est extrêmement faible, il est gravement affecté par le bruit de quantification de la caméra. La seule façon de réduire l'effet du bruit de la caméra et d'améliorer le rapport signal/bruit est d'extraire le signal iPPG en faisant la moyenne des intensités des pixels sur une zone cutanée plus large.

2.6.5 Site de mesure

Tout comme la PPG en contact, le site de mesure a également un impact sur les performances. Plus le site de mesure est vascularisé et présente une irrigation sanguine élevée, plus le signal

iPPG est riche plus le rapport signal/bruit est fort. Les sites de mesure les plus courants sont la paume de la main et le visage [328, 329], mais la majorité des travaux ont utilisé le visage entier ou une zone bien définie du visage pour une utilisation plus pratique et des performances élevées car le visage est très vascularisé et il est la partie du corps la plus visible [330].

2.7 Système de mesure de la fréquence cardiaque par iPPG

Un système iPPG basique consiste en une série de traitement d'image et de signal pour extraire les paramètres vitaux de la vidéo. La Figure 2.6 illsutre la structure de base qui comprend l'acquisition des données, la détection de la région d'intérêt, l'extraction du signal iPPG brut, le filtrage et l'extraction des signes vitaux, y compris la fréquence cardiaque.

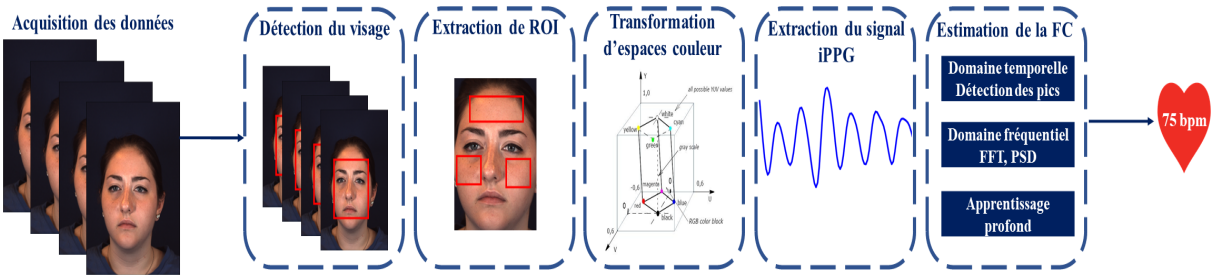


FIGURE 2.6 – La structure de base d'un système iPPG conventionnel pour l'estimation de la fréquence cardiaque à partir des enregistrements vidéos.

2.7.1 Acquisition des données

La première étape consiste à enregistrer une vidéo d'une zone cutanée du corps humain à l'aide d'une caméra qui recueille les photons réfléchis par les tissus éclairés. On distingue plusieurs type de cameras utilisées dans la littérature scientifique :

2.7.1.1 Caméras RGB

Les caméras RGB telles que les caméras de smartphones, les webcams ou les appareils photo numériques ont souvent été utilisées pour mesurer les signaux physiologiques en exploitant deux principes. Le premier principe caractérise la méthode iPPG et repose sur les variations de la couleur de la peau dues à l'activité cardiorespiratoire qui entraîne une variation des valeurs de luminosité dans les séquences d'images [331]. Le deuxième principe caractérise la méthode ballistocardiographie et dépend du mouvement cyclique du corps dû à l'activité cardiorespiratoire [332]. L'évaluation des signes vitaux par caméra RGB semble être une approche prometteuse car

elle est robuste, fiable, sûre, rentable, adaptée à la surveillance à longue distance et à long terme ainsi qu'à la détection simultanée de plusieurs personnes [333].

2.7.1.2 Caméras rapides

Les caméras RGB conventionnelles sont limitées dans la fréquence échantillonnage qui varie généralement entre 25 et 60 fps. Cette fréquence est très faible par rapport à l'ECG et la PPG qui ont une fréquence très supérieure à 60 fps [334]. Cela a un impact direct sur la qualité des signaux iPPG extraits et donc la précision des paramètres physiologiques estimés. Par conséquent, certains travaux ont utilisé des caméras rapides afin d'obtenir le maximum de précision dans l'extraction du signal iPPG [330]. Lors de l'utilisation d'une caméra rapide, le temps d'exposition pour générer chaque image est plus court, ce qui peut affecter la sensibilité des images résultantes dans la capture des changements infimes de la couleur de la peau dus à la pulsation du sang [330, 328]. Bien que les caméras rapides fournissent une meilleure précision par rapport aux caméras conventionnelles, leur coût et leur disponibilité limitent leurs champs d'applications.

2.7.1.3 Caméras thermiques

Une caméra thermique utilise un capteur d'images capable de capter le rayonnement infrarouge à des longueurs d'onde pouvant atteindre 14 m, par opposition aux caméras classiques qui captent la lumière ambiante (longueurs d'onde de 450 nm à 750 nm). Ce type de caméra fonctionne sur le principe que tous les objets dont la température est supérieure au zéro absolu ($-273,15°\text{C}$ ou $459,67°\text{F}$) émettent de l'énergie électromagnétique, également appelée rayonnement infrarouge ou rayonnement thermique [335]. Contrairement aux méthodes basées sur les caméras conventionnelles qui captent le changement de la couleur réfléchis par les tissus éclairés de la peau, les méthodes basées sur l'imagerie thermique extraient les signes vitaux en mesurant la variation de température induite par le passage du sang chaud dans des artères superficielles de la peau. Ce phénomène est modulé par la pulsation cardiaque. Les caméras thermiques permettent une mesure passive de la température et ne nécessitent pas de sources d'éclairage. Elles n'émettent pas d'énergie, ce qui présente des avantages dans des conditions d'illumination variable, notamment lorsque l'illumination ambiante est faible. Cependant, elles présentent encore de nombreuses limitations car elles sont affectées par de petites variations de chaleur et de bruit thermique ambiant dus à la variation de la température de fond en plus des artefacts de mouvement et la rotation de la tête qui sont communes à la plus part des systèmes qui utilisent l'imagerie pour mesurer des signaux physiologiques. Par ailleurs, le coût élevé des caméras thermiques par rapport à d'autres méthodes de mesure physiologique limite leur champ d'utilisation.

2.7.2 Détection de la région d'intérêt

La distribution de la densité des vaisseaux sanguins varie dans les différentes régions du corps et il est important de sélectionner des régions d'intérêt (ROI) efficaces avec des indices photoplethysmographiques riches. Le choix de la ROI a un impact significatif sur la qualité du signal iPPG car l'information iPPG est détectable uniquement au niveau de la peau. Le visage est donc la partie du corps la plus utilisée car il est rarement couvert par des vêtements ou par les cheveux. Les mouvements de la tête et les lèvres ainsi que le clignotement des yeux entrainent néanmoins des variations indésirables affectant la qualité de la mesure du signal iPPG extrait et donc les performances. Pour surmonter ces problèmes, des études antérieures ont utilisé une zone bien définie du visage au lieu du visage entier. Certaines études ont montré que la région du front fonctionne mieux que les autres régions du visages [336, 337], car les signaux iPPG récupérés ont un rapport signal/bruit plus élevé. D'autres travaux se sont concentrès sur les régions inférieures du visage telles que les joues et le nez [338, 339], car le front peut être masqué par les cheveux ou un chapeau. La taille de la région d'intérêt a également un impact direct sur la précision. Plus la taille de la ROI est grande, plus les chances d'extraire un signal iPPG de meilleure qualité sont élevées, mais le temps de calcul est également plus important [340, 341].

L'extraction de la ROI peut être effectuée en détectant la ROI dans toutes les trames ou en identifiant la ROI dans la première trame et en la suivant dans toutes les trames suivantes [342]. Parmi les detecteurs de visage couramment utilisés, on trouve ViolaJones [71], Dlib [72] et MTCNN [73]. Quant au suivi du visage, l'algorithme de Kanade-Lucas-Tomasi est le plus utilisé pour avoir une région de visage stable en cas de mouvement significatif de la tête [343, 344]. La segmentation de la peau a également été utilisée pour segmenter tous les pixels de la peau dans la région du visage [345]. Outre la définition d'une seule ROI, certaines approches utilisent plusieurs petites ROI divisées à partir d'une grande ROI pour former une carte spatio-temporelle [346], ce qui permet d'obtenir des indices iPPG à partir de plusieurs régions du visage et d'atténuer l'impact de l'occlusion et les régions de faible SNR.

2.7.3 Extraction du signal iPPG

A l'exception de PRNet [347], 3D-CNN [332] et 2SR [348], les algorithmes iPPG existants utilisent la moyenne spatiale de toutes les valeurs d'intensité des pixels dans la ROI sélectionnée dans chaque image pour générer des signaux bruts. L'équation 2.1 présente le calcul de la moyenne spatiale, où $i_R(t), i_G(t), i_B(t)$ sont trois signaux sources provenant respectivement des canaux rouge, vert et bleu, I(x,y,t) est la valeur de l'intensité du pixel à l'emplacement de l'image (x,y) au temps t, et |ROI| est la taille de ROI sélectionnée.

L'objectif de la moyenne spatiale est d'atténuer les effets du bruit de quantification de la caméra contenu dans chaque pixel, ce qui améliore le rapport signal/bruit di signal iPPG. Cependant, compte tenu des scénarios pouvant être rencontrés lors de la mesure, les signaux bruts sont sujets à des artefacts de mouvement et à des variations de l'éclairage qui rendent complexe l'extraction des paramètres cardiovasculaires. Par conséquent, un traitement supplémentaire des signaux brutes est nécessaire pour améliorer leur qualité et augmenter leur SNR.

$$i_R(t), i_G(t), i_B(t) \ = \frac{\sum_{x,y \in ROI}^{I(x,y,t)}}{|ROI|} \tag{2.1}$$

2.7.3.1 Filtrage

Un processus de filtrage est généralement effectué sur les signaux bruts pour éliminer les bruits indésirables dus au mouvement et à la variation de l'éclairage en vue d'un bon rapport signal/bruit. Les opérations les plus courantes sont la normalisation temporelle [340], l'interpolation [318, 328], le filtrage passe-bande [318], le filtrage par moyenne mobile [318] et la suppression des tendances [340]. La plage des fréquences cardiaques étant connue [42 bpm, 240 bpm], les filtres sont généralement configurés avec des fréquences de coupure [0,7 Hz, 4,0 Hz] pour éliminer les fréquences en dehors de la plage des fréquenes cardiaques. Le signal iPPG filtré peut être directement utilisé pour l'extraction des fonctions vitales ou il passe encore par d'autres traitements.

2.7.3.2 Réduction de dimensionnalité

Les méthodes de réduction de dimensionnalité sont utilisées pour minimiser la dimensionnalité des signaux bruts afin d'obtenir un signal iPPG plus précis et plus robuste. Le signal iPPG est considéré comme un signal unidimensionnel qui est représenté comme une combinaison linéaire de la somme pondérée des signaux bruts des canaux de couleurs où leurs poids sont difficiles à estimer [349]. Nombreux algorithmes ont été proposés dans la littérature scientifique. Ils peuvent être regroupés en trois grandes catégories [350] :

- La séparation aveugle de sources (Blind Source Separation, BSS) ;
- Méthodes basées sur des modèles ;
- Méthodes basées sur la conception.

La séparation aveugle de sources

La méthode de séparation aveugle de sources (BSS) a été introduite pour la première fois dans [351]. L'idée des algorithmes BSS est de séparer le signal iPPG désiré du bruit et des artefacts en

raison de l'indépendance et de la corrélation statistiques. L'analyse en composantes principales (Principal Component Analysis, PCA) [352] et l'analyse en composantes indépendantes (Independent Component Analysis, ICA) [351] sont des techniques de BSS typiques très utilisées dans le champs du traitement du signal et des images. Cependant, il n'y a pas de consensus sur les avantages des méthodes de BSS en général, ni sur le cadre dans lequel les BSS devraient être appliquées. De même, leur complexité algorithmique les rend moins adaptées pour des mesures en temps réel. Par exemple, Feng et al. [338] ont signalé une augmentation de l'erreur d'estimation de la fréquence cardiaque et un manque de robustesse, respectivement, tandis que Christinaki et al. n'ont signalé que des améliorations subtiles lors de l'utilisation de données multispectrales [353]. Wedekind et al. [354] ont montré que les performances des BSS dépendaient fortement de la qualité des signaux bruts et ces méthodes peuvent, dans certains cas, dégrader la qualité du signal iPPG.

Méthodes basées sur des modèles

Contrairement aux méthodes basées sur la séparation aveugle de source qui ne prennent en considération l'intéraction couleur-peau, les méthodes basées sur les modèles reposent sur des connaissances préalables concernant la réflexion optique de la peau pour améliorer la robustesse. Les algorithmes proposés utilisent des modèles mathématiques impliquant des propriétés physiques de la réflectance de la lumière. De Hann et al. [340] ont proposé la méthode CHROM qui génère un signal iPPG en éliminant le bruit causé par la réflexion de la lumière à l'aide d'un rapport des canaux de couleur normalisés. Ils utilisent une combinaison linéaire des caractéristiques de chrominance pour prendre en compte la réflexion diffuse qui est liée à la pulsation cardiaque et éliminer la composante spéculaire de la lumière. La combinaison linéaire est orthogonale à cette composante. Une autre méthode similaire à CHROM a été proposée par De Hann et al. [355] appeléee PBV. C'est une signature unique sous forme d'un vecteur de changement du volume sanguin définie par le spectre d'absorption de l'hémoglobine et qui récupère le signal iPPG en limitant toutes les variations de couleur à la direction pulsatile. PBV a montré une meilleure robustesse au mouvement que les méthodes BSS et CHROM. Cependant, ses performances étaient légèrement inférieures à celles de CHROM pour les sujets immobiles. Plus tard, Wang et al. [356] ont introduit la méthode POS (Plane Orthogonal-to-Skin) qui est similaire à CHROM mais qui utilise une projection différente orthogonale au ton de la peau. L'idée principale de cet algorithme est de projeter le signal sur un plan orthogonal au ton de la peau normalisé dans le temps, afin d'améliorer la séparation entre les composantes pulsatiles et spéculaires. La méthode POS est considérée comme plus robuste dans les scénarios d'illumination complexes.

Méthodes basées sur la conception

Au lieu de définir un modèle optique, qui nécessite une bonne connaissance des phénomènes physiques complexes, wang et al. [348] ont développé un algorithme appelé rotation du sous-espace spatial (SSR) pour une extraction plus robuste du signal iPPG. Dans cet algorithme, conceptuellement très différent des approches susmentionnées, la distribution des pixels de la peau dans l'espace couleur RGB est utilisée pour créer un sous-espace couleur spatial pour chaque image. La rotation temporelle de ce sous-espace est ensuite mesurée pour estimer le signal iPPG. Cette méthode est plus intéressante par rapport aux autres algorithmes car elle ne nécessite pas de connaissances prioriees liés au teint de la peau ou aux pulsations cardiaque et elle est basée uniquement sur les informations fournies par les données.

2.7.4 Estimation de la fréquence cardiaque

La FC peut être estimée à partir du signal iPPG extrait soit dans le domaine temporel ou dans le domaine fréquentiel. Dans le domaine temporel, la détection des pics est effectuée pour localiser l'instant où le battement de cœur se produit. La FC moyenne est égale à l'inverse de l'intervalle de temps entre deux battements consécutifs (IBI) divisé par 60 pour obtenir la fréquence en battements par minute. Dans le domaine fréquentiel, la fréquence cardiaque correspond à la fréquence du pic maximal de la densité spectrale de puissance du signal iPPG.

2.8 Résumé des travaux existants

En examinant les recherches existantes sur l'estimation sans contact de la fréquence cardiaque à l'aide de l'iPPG, nous pouvons identifier l'existence de deux approches principales selon la manière d'extraire le signal iPPG, soit manuellement en utilisant des méthodes conventionnelles [318, 357], soit automatiquement en utilisant des méthodes d'apprentissage profond [331, 333]. Il existe également une approche bout-en-bout qui permet une estimation de la FC en une seule étape sans passer par l'extraction du signal iPPG. Le Tableau 2.1 résume les travaux de la littérature scientifique avec leurs avantages et inconvénients.

2.8.1 Méthodes conventionnelles

Les premiers travaux sur l'iPPG reposaient sur des approches conventionnelles qui consistent généralement en des étapes de traitement d'image et de signal. Les techniques de traitement d'image sont d'abord appliquées pour localiser les régions de la peau contenant des informations pertinentes sur les changements de couleur subtils associés au flux sanguin. Différents espaces de

couleur et différentes régions d'intérêt ont été exploités pour constituer les signaux iPPG bruts. Verkruysse et al. [358] ont initialement calculé les signaux iPPG bruts à partir du canal vert en utilisant un ensemble de ROIs prédéfinis. Plus tard, plusieurs détecteurs et méthodes de suivi du visage ont été utilisés pour extraire le visage entier ou des sous-régions du visage telles que le front ou les joues [351, 352, 336, 337, 341]. Bousefsaf et al. [345] ont proposé de sélectionner uniquement les pixels d'intérêt à l'aide d'une segmentation de la peau personnalisée, tandis que Tulyakov et al. [346] ont développé une approche permettant de choisir dynamiquement les ROI à l'aide d'un module appelé complément matriciel auto-adaptatif. En outre, différents espaces de couleur ont été étudiés en plus de l'espace RGB standard. Par exemple, la composante u* de l'espace couleur CIE L*u*v* [345] et V du YUV ont été exploitées [359].

Dans la deuxième étape, des algorithmes de traitement du signal sont appliqués pour augmenter le rapport signal/bruit et éliminer le bruit du signal iPPG brut. Parmi les études les plus populaires figurent les méthodes de séparation aveugle des sources, telles que l'analyse en composantes indépendantes [351] et l'analyse en composantes principales [352]. D'autre part, des améliorations supplémentaires ont été obtenues grâce aux approches basées sur des modèles proposées par de Haan et son groupe. Ils ont développé différentes transformations de sous-espace de couleur pour surmonter les artefacts de mouvement et améliorer la qualité du signal iPPG [360, 348, 356].

2.8.2 Méthodes basées sur l'apprentissage profond

Avec le grand succès de l'apprentissage profond et plus particulièrement les réseaux neuronaux convolutifs (CNN) pour les tâches d'imagerie médicale et de vision par ordinateur [371, 372, 373], nombreuses méthodes basées sur l'apprentissage profond pour l'estimation de l'iPPG ont été développées. Selon la revue récente de Ni et al. [374], les méthodes existantes sont construites à l'aide de CNN [331, 370, 375], ou combinent CNN et LSTM pour prendre en compte l'information temporelle [376, 333, 347], ou encore utiliser directement des réseaux de neurones spatio-temporels 3D-CNN pour apprendre simultanément les caractéristiques spatiales et temporelles [369, 377, 332, 170, 378]. Pour citer certains des travaux prometteurs, Chen et McDuff [331] ont proposé un réseau d'attention convolutif nommé DeepPhys, qui consiste en un CNN à deux flux pour extraire la forme d'onde du signal iPPG à partir des vidéos faciales sous un éclairage variable et des mouvements importants de la tête. Ils ont utilisé un modèle d'apparence basé sur un mécanisme d'attention pour trouver les ROI appropriées et pour guider le modèle de représentation du mouvement. Radim et al. [370] ont proposé une architecture de réseau de neurones convolutif en deux étapes, composée respectivement de CNN 2D et de CNN 1D. Le

TABLE 2.1 – Un résumé des approches existantes de l'estimation de la fréquence cardiaque basées sur l'iPPG et leurs avantages et inconvénients.

	Plusieurs étapes		**Une seule étape**
	Conventionnelle	**Apprentissage profond**	
Caméra	Thermique [361]	Thermique [362]	R.G.B [347]
	Monochromatique [363]		
	R.G.B [346, 364]	R.G.B [331, 365]	Données synthetiques [332]
	Cinq bandes [366]		
Pré-traitement	Détection & suivi de ROI [346, 351]	Détection & suivi de ROI [333]	Détection & suivi de ROI [347]
	Transformation de l'espace couleur [345, 359]	Cartes spatio-temporelles [333]	
	Decomposition du signal [318]	Magnification de la video [367]	
Post-traitement	Filtrage [351, 352, 360]	FFT [368]	-
	FFT [340, 318]	Detection des pics[369]	
	Detection des pics [351, 366]	modèle d'apprentissage profond [333, 370]	
Extraction du signal iPPG	Moyenne spatiale [357, 351, 330]	Modèle regressif profond [331, 369]	-
Avantages	Permet l'extraction des caractéristiques de l'onde iPPG	Bonne capacité de généralisation	Bonne capacité de généralisation
		Permet l'extraction des caractéristiques de l'onde iPPG	Facile à déployer
			Fenêtre de temps courte
Inconvénients	Difficile à déployer	Difficile à déployer	Ne permet pas l'extraction des caractéristiques de l'onde iPPG
	Nécessite des étapes de pré-traitement ou de post-traitement	Nécessite des étapes de pré-traitement ou de post-traitement	
	Fenêtre de temps longue	Fenêtre de temps longue	
	Mauvaise capacité de généralisation		

premier extrait le signal iPPG tandis que le second régresse la valeur de fréquence cardiaque. Niu et al. [333] ont généré des cartes spatio-temporelles à partir de plusieurs ROI sur le visage, puis ont entraîné un réseau convolutif combiné avec un réseau récurrent CNN-RNN pour régresser la valeur moyenne de la fréquence cardiaque. Yu et al. [369] ont introduit un réseau spatio-temporel profond (PhysNet) pour extraire les signaux iPPG bruts à partir des vidéos faciales, puis ont mesuré la fréquence cardiaque moyenne et les caractéristiques de la variabilité cardiaque. AutoHR est une contribution récente proposée par Yu et al. [365]. Les auteurs ont utilisé la convolution de différence temporelle combinée avec un module de recherche d'architecture neuronale pour mesurer avec précision le signal iPPG à partir de séquences d'images.

Toutes les méthodes mentionnées ci-dessus sont basées sur plusieurs étapes de traitement. Elles utilisent principalement l'apprentissage profond pour récupérer les signaux iPPG à partir de vidéos faciales, puis la fréquence cardiaque et sa variabilité sont mesurées dans le domaine temporel ou fréquentiel par une détection de pics ou en utilisant la densité spectrale de puissance. Cependant, certains travaux ont adopté des réseaux de neurones profonds de bout en bout pour l'estimation de la fréquence cardiaque sans passer par l'extraction des signaux iPPG. Bousefsaf et al. [332] ont été les premiers à démontrer la possibilité d'estimer la fréquence cardiaque à partir de vidéos de visages sans traitement supplémentaire. Ils ont proposé un CNN 3D entraîné uniquement sur des données synthétiques. Huang et al. [347] ont développé un réseau spatio-temporel à un étage qui combine des modules convolutifs 3D et LSTM pour extraire les caractéristiques spatiales et temporelles et une couche entièrement connectée «dense» pour l'estimation de la valeur de la fréquence cardiaque.

2.9 Conclusion

Dans ce chapitre, nous avons présenté un aperçu général sur la mesure de l'activité cardiaque par photopléthysmographie sans contact. Nous avons abordé dans un premier temps des généralités sur le fonctionnement du coeur et la fréquence cardiaque et sa variabilité ainsi que quelques techniques de mesure existantes. Ensuite, nous avons présenté la photopléthysmographie par imagerie, ses applications et ses défis. Enfin, nous avons passé en revue les diverses approches proposées pour l'estimation de la fréquence cardiaque. Bien que cette analyse ne soit sans doute pas exhaustive, notamment en raison de l'intérêt croissant pour ce domaine de l'analyse des signaux biomédicaux, elle couvre la majorité des méthodes les plus marquantes.

Deuxième partie

Contributions

3

Mesure sans contact de la fréquence cardiaque par caméra

Sommaire

3.1 Introduction

Alors que l'utilisation des signaux physiologiques a longtemps été confinée au domaine médical pour la surveillance de l'état de santé des patients, l'intérêt pour l'utilisation de ces signaux en informatique affective a pris de l'ampleur ces dernières années pour diverses raisons. Par rapport à d'autres modalités comportementales, les signaux physiologiques tels que l'activité cardiaque ou cérébrale sont généralement considérés comme plus fiables et plus objectifs et fournissent une image plus complète de l'expérience émotionnelle. Ils présentent l'avantage de pouvoir capturer des informations liées à l'état émotionnel interne, ce qui donne un aperçu des états émotionnels inconscients dont les gens ne sont pas forcément conscients ou qu'ils ne sont pas en mesure de verbaliser. En outre, les signaux physiologiques peuvent être utilisés pour une surveillance continue, ce qui permet une évaluation plus précise de l'état émotionnel sur des périodes plus longues.

Traditionnellement, des capteurs en contact spéciaux sont nécessaires pour mesurer les signaux physiologiques, par exemple, un ECG ou un oxymètre pour mesurer l'activité cardiaque et un EEG pour mesurer l'activité cérébrale. Les mesures en contact sont peu pratiques et inconfortables, en particulier pour la surveillance à long terme. Les récentes avancées technologiques ont permis le développement de la photoplethysmographie par imagerie, une technique non invasive et sans contact pour mesurer à distance l'activité cardiaque en utilisant une simple camera.

Comme nous l'avons présenté au chapitre 2, la mesure à distance de la fréquence cardiaque à partir des séquences vidéos faciales a suscité une attention particulière ces dernières années. Les recherches présentent des avancées significatives et démontrent que les caméras conventionnelles correspondent à des dispositifs fiables qui peuvent être utilisées pour mesurer un large ensemble de paramètres biomédicaux, y compris la fréquence cardiaque sans aucun contact avec le sujet. De nombreuses méthodes ont été développées au fil du temps afin d'améliorer les performances, en particulier dans des conditions non controlées telles que la variation d'illumination et les mouvements significatifs de la tête.

A l'exception de PRNet [347] et 3D-CNN [332], les approches existantes sont basées sur l'extraction du signal iPPG en utilisant des méthodes conventionnelles ou des méthodes basées sur l'apprentissage profond. La fréquence cardiaque est calculée à partir du signal iPPG dans le domaine temporel ou fréquentiel, comme montré dans la sous-section 2.7.4. La précision dépend de la qualité de la forme d'onde iPPG et de la détection du pic principal. En outre, ces méthodes nécessitent des étapes de pré-traitement et post-traitement supplémentaires (cf. section 2.7). Néanmoins, les bases de données publiques disponibles sont difficiles (la variation d'éclairage, le mouvement et les expressions faciales) et contiennent un grand nombre de signaux de référence

corrompus et de mauvaise qualité [379, 380, 381]. Cela affecte directement la localisation du pic principal et réduit par conséquent la précision de la mesure.

Dans ce chapitre, nous proposons une nouvelle approche bout en bout, à savoir X-iPPGNet, pour l'estimation sans contact de la fréquence cardiaque à partir d'enregistrements vidéo du visage en utilisant un réseau spatio-temporel profond. X-iPPGNet est un pipeline optimisé à une seule étape qui prédit la fréquence cardiaque sur une courte fenêtre temporelle, sans extraction séparée du signal iPPG et sans connaissance préalable.

Cette méthode s'articule autour de trois grandes étapes : la segmentation du visage, augmentation de données et l'estimation de la fréquence cardiaque. La segmentation du visage est d'abord effectuée pour augmenter le rapport signal/bruit et isoler la région d'intérêt contenant le front, les joues et le nez. Cette étape est cruciale car elle permet de s'assurer que seules les variations de couleur et de texture dues aux battements du cœur sont prises en compte, plutôt que les mouvements du visage dans leur ensemble. La deuxième étape consiste en l'augmentation des données. Cette étape vise à augmenter la taille de l'ensemble de données pour améliorer la robustesse du système. Pour cela, on utilise des techniques de génération de données synthétiques en appliquant des transformations géométriques et l'amplification vidéo. Enfin, la troisième étape consiste en l'estimation de la fréquence cardiaque à partir de vidéos du visage à l'aide du modèle X-iPPGNet.

3.2 Bases de données

Le grand succès de l'apprentissage profond pour les tâches de vision par ordinateur est dû principalement à la disponibilité des bases de données massives ainsi qu'à des architectures avancées. Par rapport aux ensembles de données massifs utilisés dans les tâches de vision par ordinateur, tels qu'ImageNet [382] ou Kinetics-700 [383], les bases de données publiques incluant la fréquence cardiaque et les vidéos des participants (telles que MAHNOB-HCI [380], UBFC-rPPG [384], VIPL-HR [381]) sont assez limitées non seulement en termes de volume mais aussi de diversité. Le mouvement de la tête, les expressions faciales, l'occlusion et la couleur de peau correspondent aux principales difficultés qui affectent les performances de la mesure sans contact de la fréquence cardiaque à partir de vidéos du visage. Cependant, les travaux précédents n'ont pas abordé tous ces problèmes en raison de la qualité et de la faible quantité des bases de données susmentionnées.

Dans cette étude, nous avons utilisé quatre ensembles de données publiques pour l'estimation de la fréquence cardiaque afin d'évaluer les performances de la méthode proposée. Nous avons

entraîné notre modèle sur une base de données publique à grande échelle (appelée BP4D+ [379]), tandis que MAHNOB-HCI [380], UBFC-rPPG [384], et MMSE-HR [379] ont été utilisées pour évaluer la capacité de généralisation des approches développées. Nous décrivons brièvement chacune de ces trois bases de données dans les sous-sections suivantes, tandis que nous présentons en détail la base de données BP4D+, car nous sommes, à notre connaissance, la première équipe à l'avoir utilisée dans le contexte de la mesure de fréquence cardiaque à partir de vidéos du visage.

TABLE 3.1 – Résumé des bases de données publiques utilisées dans nos expériences.

Base de données	**Nb de participants**	**Nb de vidéos**	**FPS**	**Ethnicité**	**Tâche/Condition**
MMSE-HR [379]	40	102	25	Latino/Hispanique, Blanc, Africain Americain, Asiatique, et autres	Elicitation de l'émotion
MAHNOB-HCI [380]	27	527	61	Caucasien et Asiatique	Elicitation de l'émotion
UBFC-rPPG [384]	42	42	30	-	Interaction
BP4D+ [379]	140	1400	25	Latino/Hispanique, Blanc, Africain Americain, Asiatique, et autres	Elicitation de l'émotion

3.2.1 MMSE-HR

Le jeu de données MMSE-HR [379] a été collecté pour l'estimation sans contact de la fréquence cardiaque dans des conditions difficiles. Il s'agit de 102 vidéos faciales enregistrées à 25 fps et provenant de 40 sujets (17 hommes et 23 femmes) de diverses origines ethniques présentant diverses couleurs de peau. Les valeurs de vérité terrain correspondent à des moyennes sur les fréquences cardiaques et ont été recueillies simultanément à l'aide d'un capteur BVP en contact fonctionnant à une fréquence d'échantillonnage de 1 KHz.

3.2.2 MAHNOB-HCI

MAHNOB-HCI [380] est une base de données couramment utilisée pour évaluer l'efficacité et la capacité de généralisation des méthodes d'estimation sans contact de la fréquence cardiaque

par caméra. Elle comprend 527 vidéos de 27 sujets (12 hommes et 15 femmes) ainsi que leurs signaux physiologiques correspondants. Toutes les vidéos sont enregistrées à 61 fps avec une résolution de 780 × 580 pixels. Le signal ECG a été utilisé pour calculer les valeurs de vérité terrain de la fréquence cardiaque.

3.2.3 UBFC-rPPG

UBFC-rPPG [384] se compose de 42 vidéos provenant de 42 participants. Les vidéos ont été enregistrées sans compression à l'aide d'une webcam à bas prix (Logitech C920 HD pro) à 30 fps et à une résolution de 640 × 480 pixels. La durée de chaque enregistrement varie entre 50 et 90 secondes. Un oxymètre de pouls au doigt Contec Medical CMS50E est synchronisé avec les enregistrements vidéo afin d'établir le signal PPG de référence.

3.2.4 BP4D+

BP4D+ [379] est une base de données publique à grande échelle principalement dédiée à la reconnaissance multimodale des émotions spontanées basée sur les expressions faciales et les paramètres biomédicaux. Elle comprend plusieurs signaux physiologiques telles que la fréquence cardiaque, la fréquence respiratoire et la pression artérielle. Comparée aux bases de données existantes, BP4D+ est nettement plus importante en termes de quantité de données et de diversité ethnique (y compris les africains, les européens, les asiatiques et les hispaniques/latinos). De plus, les données ont été recueillies dans des conditions difficiles, telles que des mouvements importants de la tête, une plage de fréquence cardiaque très large, d'expressions faciales et d'occlusions.

140 sujets (82 femmes et 58 hommes) ont participé à dix sessions conçues pour susciter différentes émotions. 1400 vidéos RVB d'une durée de 30 secondes à 1 minute ont été enregistrées à 25 fps. La résolution de chaque vidéo est de 1040 × 1392 pixels. La fréquence cardiaque et les autres signaux physiologiques ont été recueillis par un capteur en contact de pression artérielle de type Biopac MP150 [1] fonctionnant à une fréquence d'échantillonnage de 1 kHz.

La Figure 3.1 montre l'histogramme de la distribution de la fréquence cardiaque dans BP4D+. Les valeurs de la fréquence cardiaque varient de 47 à 139 battements par minute (bpm), ce qui couvre presque toute la plage typique de la fréquence cardiaque. L'histogramme forme une distribution inverse-gaussienne car la plupart des adultes en bonne santé et détendus ont une fréquence cardiaque au repos comprise entre 70 et 90 battements par minute. D'autre part, en raison d'un grand nombre de signaux de vérité terrain corrompus (un exemple typique est présenté en Figure 3.2), nous avons recalculé les fréquences cardiaques à partir des signaux

1. http ://www.biopac.com

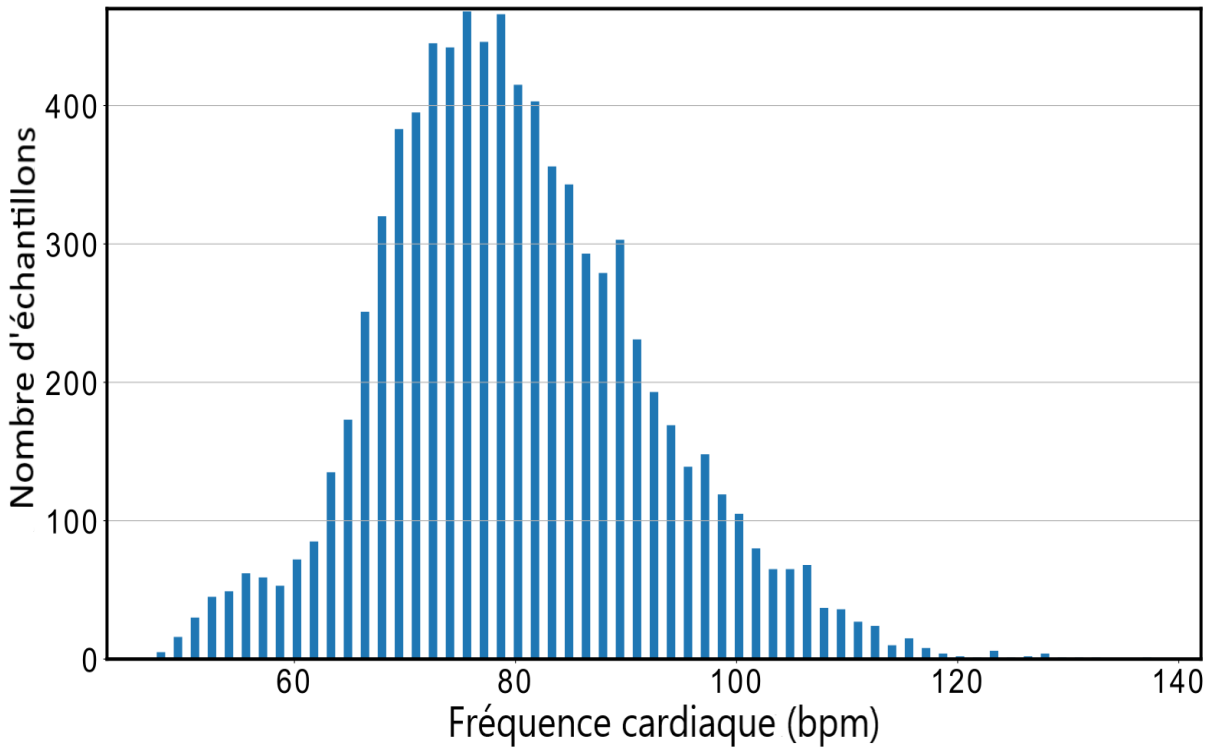


FIGURE 3.1 – Distribution des fréquences cardiaques donnée par la vérité terrain dans la base de données BP4D+.

en contact inclus dans la base de données. Les étapes de mesure de la fréquence cardiaque à partir des signaux bruts de pression artérielle sont illustrées en Figure 3.3. Les signaux sont d'abord rééchantillonnés à la fréquence de la caméra (25 Hz). Ensuite, un algorithme spécifique de suppression de tendances [385] permettant d'atténuer les basses fréquences du signal a été appliqué. Nous avons par la suite appliqué un filtre passe-bande de Butterworth d'ordre 2 avec une fréquence de coupure de 0,75 et 2,5 Hz pour ne garder que la plage des fréquences cardiaques. Enfin, une détection de pics a été appliquée pour calculer la fréquence cardiaque dans le domaine temporel. Nous avons également supprimé les segments videos où les régions du visage sont en dehors de l'image pour fournir à notre réseau uniquement des représentations informatives de données.

3.3 Mesure sans contact bout en bout de la fréquence cardiaque

Notre système général pour l'estimation de la fréquence cardiaque à partir de vidéos faciales est illustré dans la figure 3.4. Nous traitons cette tâche comme un problème de régression à une étape. Le système proposé prend une entrée vidéo de 50 images (correspondant à 2 secondes) et

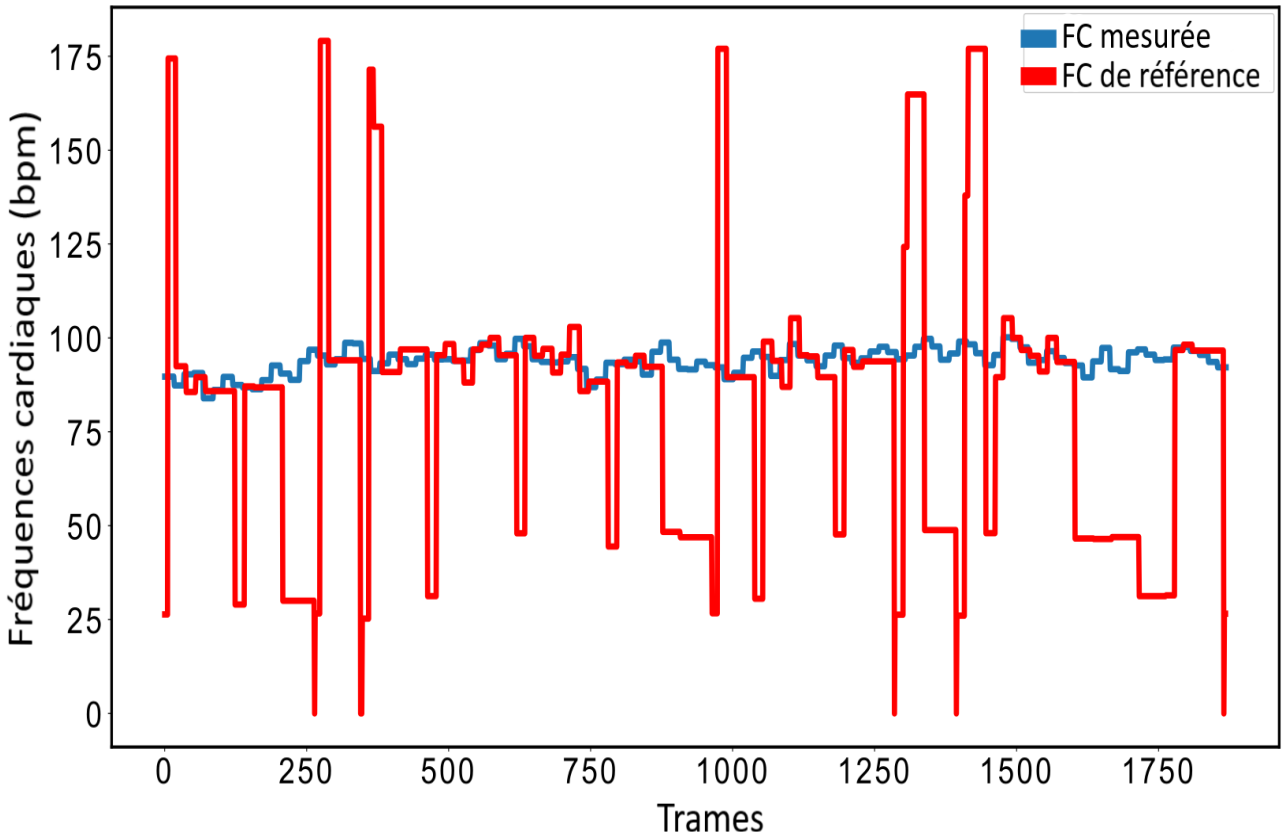


FIGURE 3.2 – Les fréquences cardiaques de référence du participant F005 montrent de fortes incohérences. Courbe rouge : fréquences cardiaques de référence fournies par la base de données ; Courbe bleue : fréquences cardiaques calculées par nos soins à partir du signal PPG en contact fourni dans la base de données.

régresse la valeur de la fréquence cardiaque en sortie. Tout d'abord, une segmentation du visage est effectuée pour éliminer l'arrière-plan et les zones non cutanées [386]. Ensuite, la région du visage est découpée à partir de l'image segmentée en fonction des coordonnées du premier pixel différent de zéro. Enfin, les séquences d'images résultantes sont mises à l'échelle et introduites à un réseau de neurones convolutif spatio-temporel. Nous supposons que le modèle proposé est capable de se concentrer automatiquement sur les zones les plus vascularisées du visage. Nous supposons aussi qu'il apprend les caractéristiques spatio-temporelles associées aux changements subtils de couleur sur la région sélectionnée. Toutes ces étapes sont détaillées dans les sous-sections suivantes.

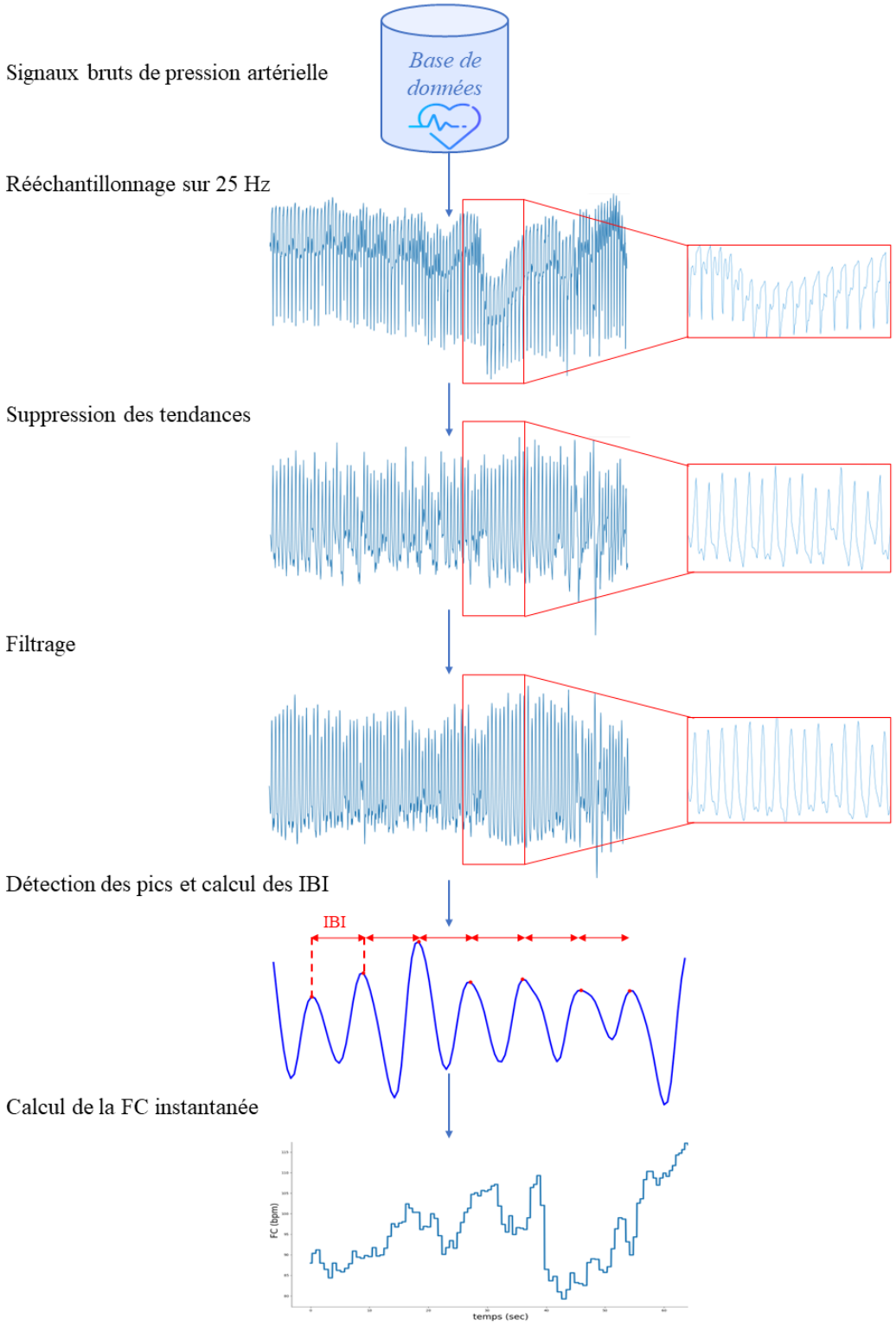


FIGURE 3.3 – Illustration des étapes de mesure de la fréquence cardiaque à partir de signaux bruts fournies dans la base de données BP4D+.

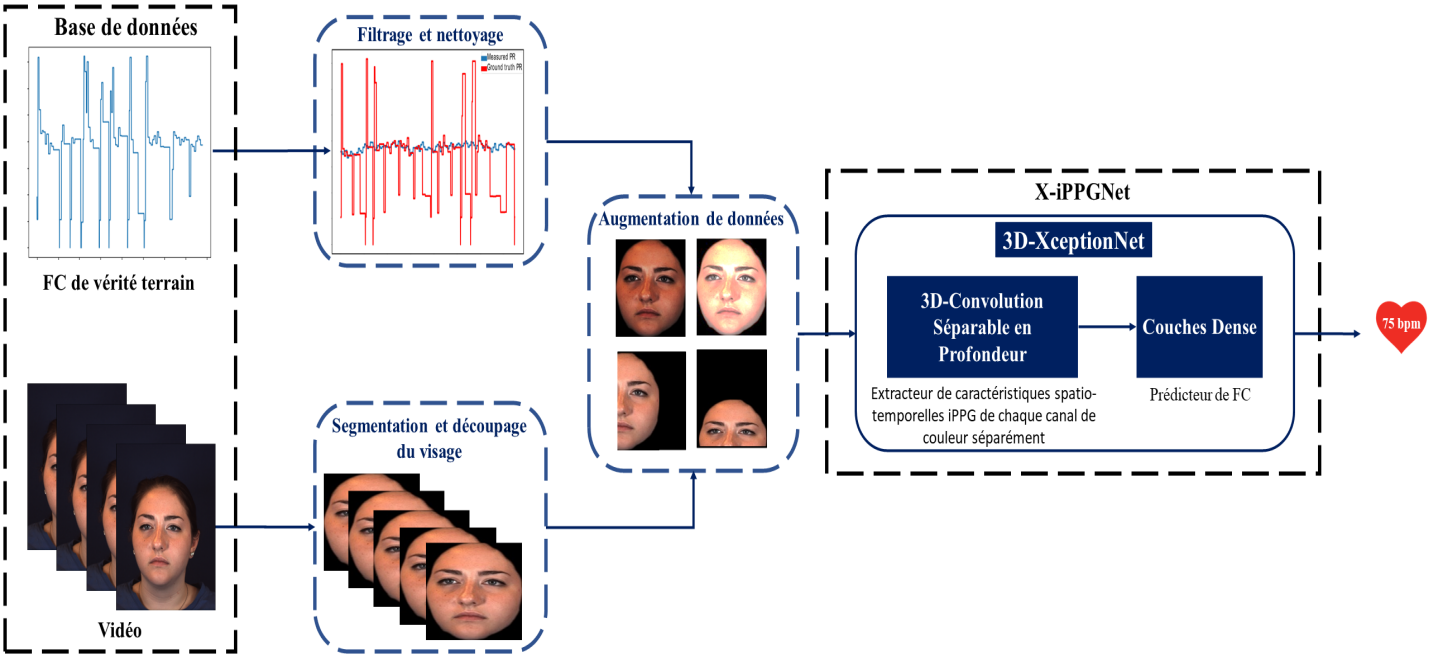


FIGURE 3.4 – Aperçu du système proposé pour l'estimation sans contact de la fréquence cardiaque instantanée. Une segmentation et un découpage du visage sont d'abord effectués sur les vidéos d'entrée pour éliminer les zones ne contenant pas de peau. Les signaux de fréquence cardiaque sont filtrés et nettoyés pour éliminer les données corrompues afin d'entraîner correctement le réseau de neurones (cf. sous-section 3.2.4). Les séquences d'images faciales sont ensuite introduites à un réseau de neurones profond (X-iPPGNet) composé de convolutions séparables en profondeur 3D pour l'extraction des caractéristiques spatiales et temporelles, et de couches denses pour la prédiction de la fréquence cardiaque.

TABLE 3.2 – Nombre d'images ratées selon les algorithmes de détection de visage les plus populaires.

Détecteur de visage	**Nombre d'images ratées**
Viola-Jones [71]	1375
Dlib [72]	227
MTCNN [73]	48
Segmentation de visage [386]	**0**

3.3.1 Segmentation du visage

L'extraction de la région d'intérêt (Region Of Interest, ROI) est la première étape de presque tous les systèmes de mesure de la fréquence cardiaque basés sur la vidéo [351, 333, 387, 369,

347]. Elle vise à maximiser le rapport signal/bruit en éliminant les régions non cutanées qui ne contiennent aucun changement de couleur associé au rythme cardiaque. À notre connaissance, la plupart des systèmes iPPG existants basés sur l'apprentissage profond ont utilisé soit le visage entier [333, 365, 331], soit une zone du visage sélectionnée empiriquement [387, 347]. Plusieurs détecteurs de visages et de points de repères faciaux ont été utilisés pour localiser les ROI. Cependant, ils échouent parfois lorsque les visages présentent des mouvements importants de la tête, des variations de pose, des occultations ou en fonction des expressions faciales. De plus, de nombreux autres éléments affectent la capacité de détection des ROI tels que l'illumination et l'arrière-plan. Nous avons comparé les performances des trois détecteurs de visage les plus populaires utilisés dans le cadre de la mesure de la iPPG, à savoir Viola & Jones [71], Dlib [72], et MTCNN [73].

Le tableau 3.2 illustre le nombre d'images ratées sur le jeu de données MMSE-HR [379]. Les tests sont effectués sur la base MMSE-HR qui contient environ 108 117 images. Les résultats montrent que les trois détecteurs de visage mentionnés ci-dessus sont un peu moins performants dans le cas des conditions non controlées et la détection peut échouer dans certaines images.

FIGURE 3.5 – Exemples montrant la capacité du modèle de segmentation du visage [386] à fonctionner dans des scénarios difficiles. Figures du haut : images brutes, figures du bas : images masquées par le masque de segmentation délivré par l'algorithme.

Pour surmonter les limites des détecteurs de visages, en particulier dans les scénarios non

controlés, nous avons effectué une segmentation de visage à l'aide d'un algorithme récent proposé initialement pour la permutation de visages [386]. Cette méthode est une solution deux en un basée sur un modèle neuronal convolutif. Elle détecte et segmente le visage de manière très fiable et fonctionne idéalement dans toutes les conditions mentionnées ci-dessus sans aucune image ratée. Les visages sont correctement isolés de l'arrière-plan et des occultations avec une grande précision. Des extraits obtenus à partir de plusieurs images de MMSE-HR sont présentés dans la figure 3.5.

3.3.2 X-iPPGNet : Un réseau de neurones de bout en bout pour l'estimation de la fréquence cardiaque par caméra

La plupart des approches basées deep learning pour l'estimation de FC par caméra reposent sur un CNN de type VGG. Les informations temporelles sont traitées à l'aide de réseaux récurrents [347, 333], de convolutions spatio-temporelles [332, 369], ou en incorporant une autre branche temporelle en parallèle [331]. Un CNN de type VGG est une architecture basique qui utilise une pile de convolution standard sans blocs résiduels [89]. Malgré sa simplicité, il est plus enclin au sur-apprentissage. Ses performances sont également inférieures à celles d'autres architectures d'apprentissage profond dans de nombreuses tâches de vision par ordinateur [388]. En outre, la convolution standard prend en compte toutes les informations spatiales et de canal de couleur ensemble. Cependant, des études antérieures ont montré que les canaux de couleur ont des propriétés physiologiques différentes et que l'activité pulsatile varie d'une couleur à l'autre [389]. Bien que le canal vert présente le signal pléthysmographique le plus fort et transporte plus d'informations PPG par rapport aux autres canaux, les canaux rouge et bleu contiennent également des informations pléthysmographiques utiles et complémentaires qui ne doivent pas être négligées [351]. Néanmoins, et à notre connaissance, toutes les approches basées sur l'apprentissage profond ont combiné les canaux rouge, vert et bleu. Cela peut conduire à la perte de caractéristiques utiles entre les canaux, ce qui affecte la précision de la mesure.

Dans ce travail, nous avons conçu un réseau de régression profond de bout en bout basé sur une architecture Xception modifiée [390]. Cette architecture surpasse les autres modèles d'apprentissage profond dans plusieurs tâches de vision par ordinateur [388, 391]. En outre, Elle s'appuie sur la convolution séparable en profondeur (Depthwise Separable Convolution, DSC) au lieu des opérations de convolution standard qui sont très coûteuses en terme de temps de calcul et de besoin en mémoire. Une extension de DSC pour les volumes 3D est utilisée [2] pour apprendre les caractéristiques pertinentes associées au rythme cardiaque de chaque canal de

2. https ://github.com/alexandrosstergiou/keras-DepthwiseConv3D

couleur séparément.

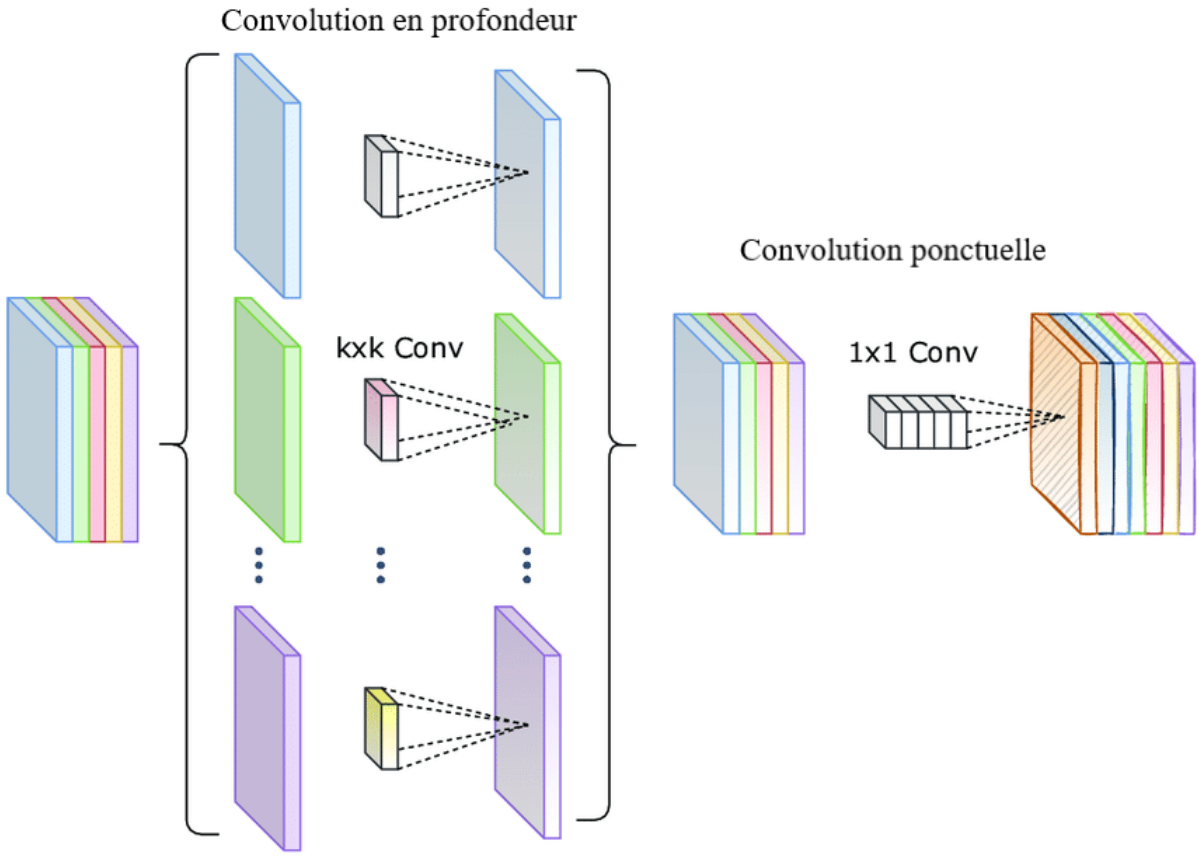


FIGURE 3.6 – Schéma de la convolution séparable en profondeur.

L'idée derrière la DSC est que la profondeur et la dimension spatiale peuvent être découplées dans une couche de convolution (Voir Figure 3.6). Tout d'abord, les dimensions d'intégration de la vidéo sont séparées et une convolution spatio-temporelle indépendante est effectuée pour chaque canal de couleur. Cette opération est appelée convolution en profondeur (Depthwise convolution). Elle vise à extraire les caractéristiques locales de chaque canal de couleur des séquences d'images d'entrée séparément et à capturer les relations temporelles entre les séquences de caractéristiques spatiales. Ensuite, une convolution ponctuelle (Pointwise convolution) est effectuée sur le tenseur convolué pour fusionner les cartes de caractéristiques entre les canaux dans la dimension d'intégration. Cette méthode permet de réduire efficacement le coût de calcul et les besoins en mémoire.

La Figure 3.7 présente l'architecture globale de X-iPPGNet, qui se compose de trois blocs (entrée, intermédiaire et sortie). Elle comprend 36 couches convolutives structurées en 14 modules, tous reliés par des connections résiduelles comme dans l'architecture ResNet, à l'exception du

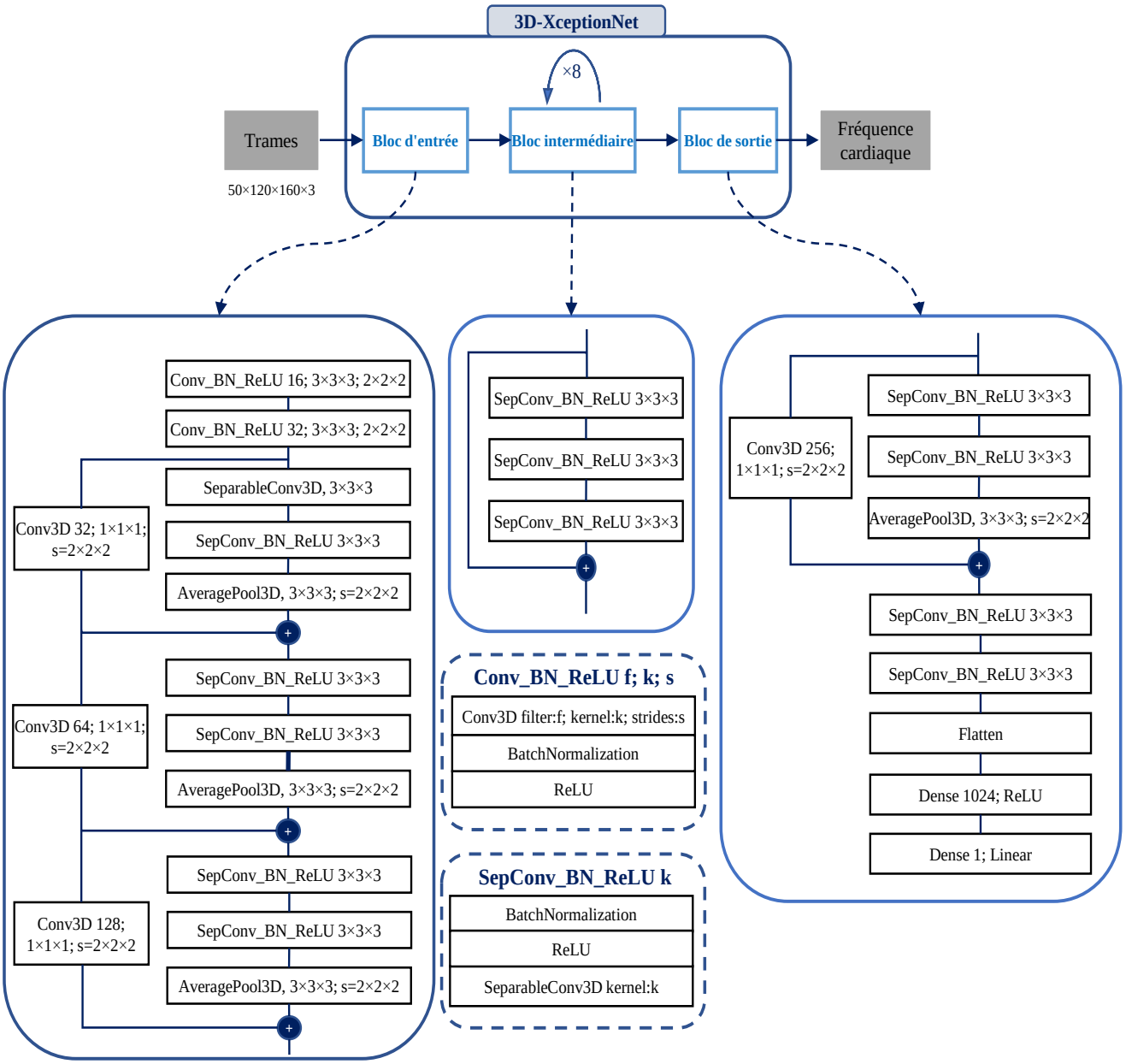


FIGURE 3.7 – Architecture du réseau X-iPPGNet proposé dans ce travail. Elle correspond à une version modifiée du réseau Xception. Les couches de convolution séparable en profondeur 2D sont remplacées par des couches de convolution séparable en profondeur 3D pour capturer les caractéristiques spatiales et temporelles à travers les trames vidéo. Le fragment vidéo d'entrée passe d'abord par le flux d'entrée, puis par le flux intermédiaire qui est répété huit fois, et enfin par le flux de sortie qui se termine par une couche dense avec 1 neurone pour estimer la fréquence cardiaque correspondante.

premier et du dernier module. Le réseau étant très profond, ces connexions résiduelles permettent de réduire l'impact lié aux problèmes de fuite du gradient. Chaque couche convolutive est suivie d'une normalisation par lots (Batch-normalization) pour stabiliser le processus d'apprentissage et accélérer la convergence. Les fonctions d'activation ReLU sont également utilisées pour réaliser une cartographie non linéaire. La sortie de l'extraction des caractéristiques est vectorisée et introduite à deux couches denses de 1024 et 1 neurones, respectivement, pour estimer la valeur de la fréquence cardiaque.

En résumé, l'architecture proposée pour l'estimation sans contact de la fréquence cardiaque est un pipeline à une étape qui prédit la FC moyenne à partir des fragments vidéo de seulement 2 secondes. L'entrée est représentée sous la forme d'un tenseur à 5 dimensions ($N_{bacth} \times Nb_{frames} \times Im_{Height} \times Im_{Weight} \times Channel$), où N_{batch} est la taille du lot ; Nb_{frames} est la longueur du clip vidéo du visage ; Im_{Height}, Im_{Weight} et $Channel$ sont les dimensions de chaque image. La sortie FC est estimée par le dernier neurone via une fonction d'activation linéaire. Elle est donnée en battements par minute.

Nous considérons la prédiction de la fréquence cardiaque comme un problème de régression à une étape. L'entraînement est entièrement supervisé, chaque fragment vidéo de 2 secondes prenant comme étiquette d'entraînement une FC de vérité terrain obtenue avec un dispositif en contact. Au cours de la phase d'apprentissage, le réseau apprend à associer la valeur de FC de vérité terrain à chaque séquence vidéo faciale en construisant une relation de correspondance entre les entrées et les sorties, c'est-à-dire la correspondance d'un tenseur tridimensionnel (données vidéo) à un scalaire unique (fréquence cardiaque). Après la phase d'entraînement, le réseau devrait être capable d'estimer la fréquence cardiaque dans la plage de fréquence sur sur laquelle il a été entraîné.

3.3.2.1 Détails d'implémentation

Le modèle proposé a été implémenté sous Python via l'API Keras et la bibliothèque Tensorflow et a été entraîné avec deux Nvidia Quadro P6000s. Les vidéos ont été découpées en séquences de 50 images (correspondant à 2 secondes). La taille de chaque image est de $160 \times 120 \times 3$ ($ImHeight \times ImWeight \times Channel$).

Inspirés par la procédure d'optimisation SWATS [392], nous démarrons l'apprentissage avec un optimiseur Adam rectifié (Rectified Adam, RAdam) [393] avant de passer à la descente de gradient stochastique (Stochastic Gradient Descent, SGD) [394] lorsque la précision de validation cesse de s'améliorer. Le taux d'apprentissage (Learning rate) est initialement fixé à 10^{-4}, puis diminué à 10^{-6}. Nous entraînons le réseau pendant 25 époques avec une taille de lot (batch-size)

de 64 ($N_{bacth} = 64$) et nous utilisons la fonction de perte (Loss function en anglais) «Erreur quadratique moyenne» (Mean Squared Error, MSE). En outre, une technique d'abandon (Dropout) [395] est appliquée avant la dernière couche dense du réseau (le taux d'abandon est fixé à 40 %). Des stratégies de régularisation L1 et L2 sont également utilisées, ce qui permet de surmonter les problèmes de sur-apprentissage et d'améliorer la capacité de généralisation du modèle à de nouvelles données.

3.3.2.2 Augmentation de données

Le problème courant lors de l'utilisation d'une base de données limitée et/ou déséquilibrée pour l'entrainement d'un réseau de neurones profond est le risque de sur-apprentissage et les mauvaises performances prédictives, notamment pour les échantillons moins représentées dans la base de données.

X-iPPGNet a d'abord été entrainé sur BP4D+ sans augmentation des données. Cependant, plusieurs problèmes qui entravent la précision prédictive de la FC ont attiré notre attention. Ils sont principalement causés par les échantillons de FC fortement déséquilibrés dans la base de données BP4D+, ainsi que par la couleur de la peau des sujets [379]. Par conséquent, l'erreur d'estimation des valeurs de FC élevées et faibles ainsi que le type de couleur de peau avec moins d'échantillons est beaucoup plus élevée par rapport aux FC moyennes (Voir Figure 3.8) et les couleurs de peau bien représentées dans la base de données (Voir Figure 3.9 et Figure 3.10).

L'erreur d'estimation pour les FC moyennes (comprises entre 70 et 90 bpm) est tout à fait acceptable avec une $RMSE = 9.17$ bpm . En revanche, le taux d'erreur passe à 13.61 bpm pour les FC faibles [< 70 bpm] et $RMSE = 16.34$ bpm pour les FC élevées [> 90 bpm]. En ce qui concerne la couleur de peau, l'erreur d'estimation pour la peau claire est $RMSE = 9.72$ bpm, tandis que l'erreur d'estimation est de 9.72 bpm pour les sujets avec une peau foncée. Il est très difficile pour un modèle d'apprentissage profond d'apprendre des caractéristiques pertinentes sur des données mal représentées. Les réseaux de neurones ont tendance à se concentrer sur les cibles ayant un grand nombre d'échantillons.

Pour pallier ce problème, une technique d'augmentation de données a été appliquée pour accroître le volume de données d'apprentissage. Étant donné qu'un plus grand nombre d'échantillons sont disponibles dans la plage des fréquences cardiaques moyennes [70, 90] bpm et moins en dehors de cette plage (voir figure 3.1), nous avons procédé à une triple augmentation de données hors ligne sur les séquences vidéo associées à des valeurs de FC supérieures à 90 bpm ou inférieures à 70 bpm.

En suivant la même stratégie que celle présentée dans [396], nous avons effectué une augmen-

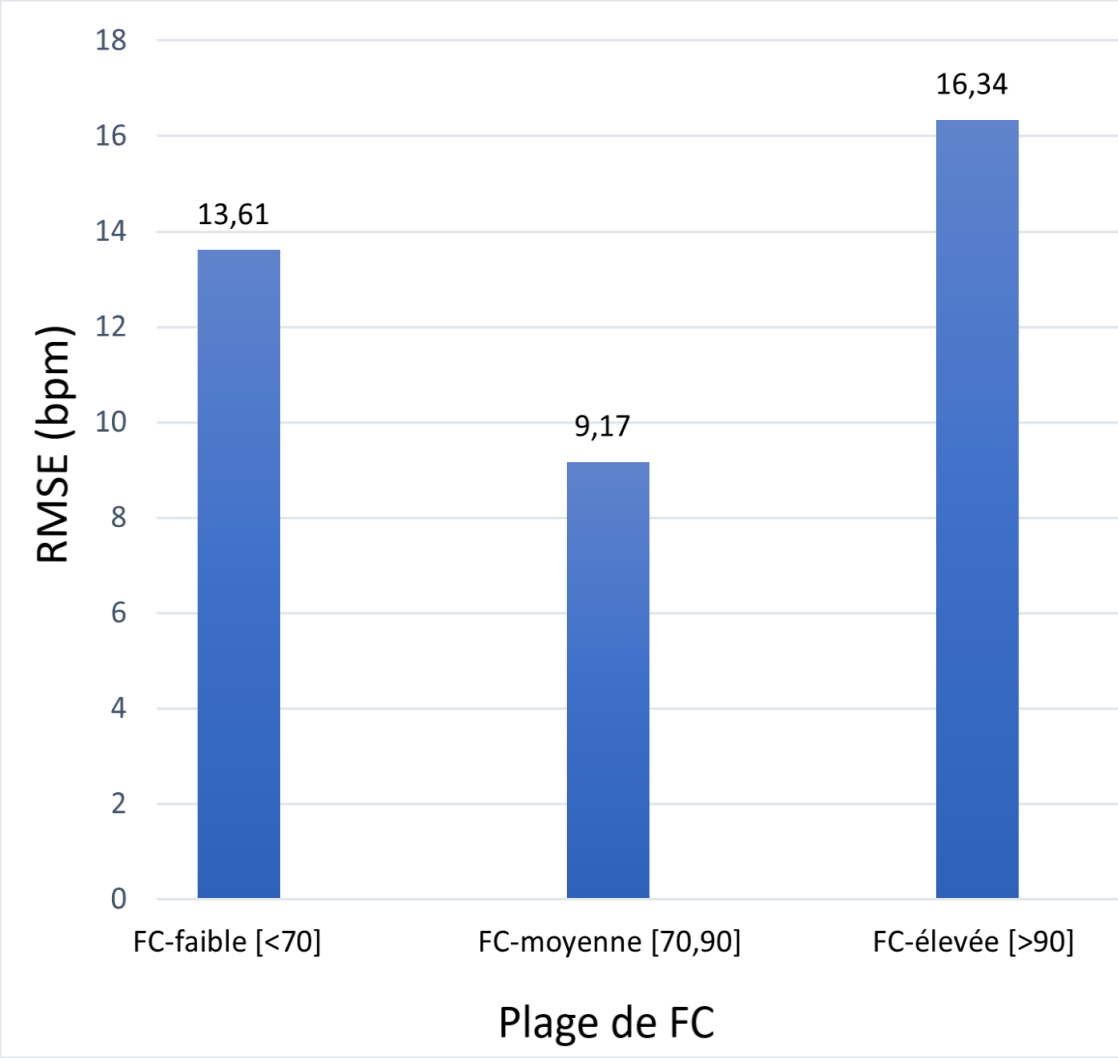


FIGURE 3.8 – Erreur d'estimation de la fréquence cardiaque de X-iPPGNet par plage de fréquence faible [$<$ 70 bpm] ; moyenne [70 bpm, 90 bpm] ; élevée [$>$ 90 bpm].

tation géométrique standard et une amplification de la vidéo (Video magnification en anglais) pour augmenter la taille de l'ensemble d'entraînement et améliorer la robustesse du modèle. Un exemple montrant l'augmentation géométrique et l'amplification vidéo est donné dans la Figure 3.11.

L'augmentation géométrique implique des transformations d'images basiques telles que des rotations aléatoires dans le sens des aiguilles d'une montre et dans le sens horaire et anti-horaire jusqu'à 20 degrés, une mise à l'échelle (dans le sens agrandissement et réduction) jusqu'à 20 %, et un décalage horizontal et vertical de l'image vidéo de 10 % de la largeur et de la hauteur de l'image.

La technique Eulerian Video Magnification (EVM) a été utilisée pour amplifier les subtiles

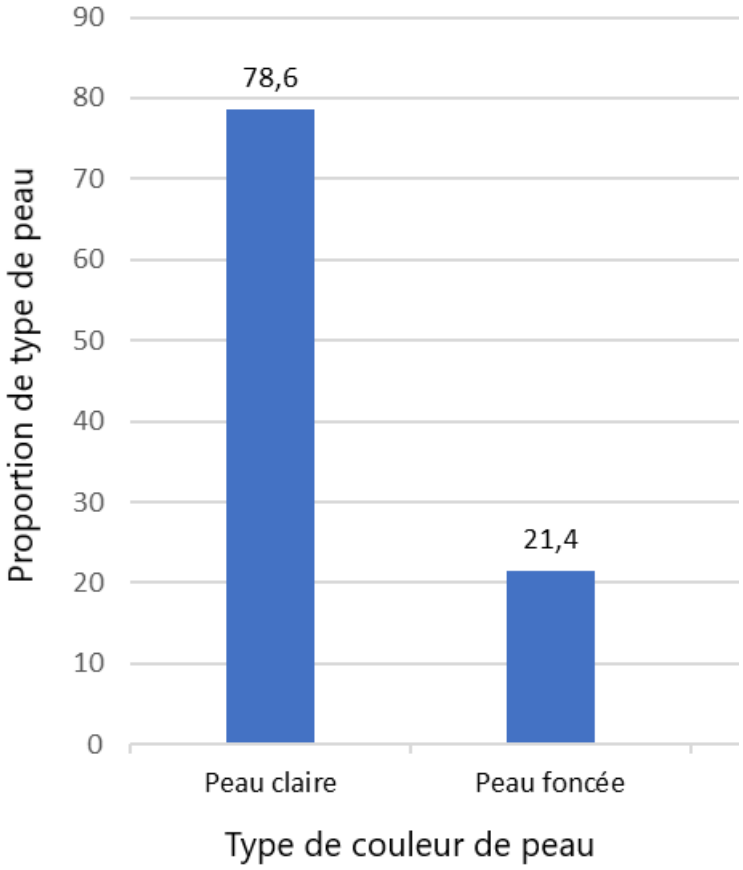


FIGURE 3.9 – Proportion de type de peau dans BP4D+.

fluctuations colorimétriques dues à l'iPPG dans les vidéos. L'intensité de ces fluctuations peut être faible pour les pixels qui couvrent les peaux foncées. La méthode EVM s'est avérée efficace pour l'estimation de la FC [367, 397, 398]. Cette technique prend en entrée une séquence de ROI et applique aux images une décomposition spatiale suivie d'un filtrage temporel. La pyramide laplacienne est utilisée pour la décomposition spatiale, tandis que le filtrage temporel est effectué en appliquant la transformée de Fourier pour chaque pixel. Le facteur d'amplification est fixé à 60 tandis que les fréquences en dehors de la fréquence de coupure (45-240 bpm) sont mises à zéro. Enfin, la transformée de Fourier inverse est appliquée pour reconstruire les images. La vidéo résultante est donc amplifiée et révèle les changements subtils provoqués par le flux sanguins dans les vaisseaux du visage et cachés dans dans la couleur de la peau.

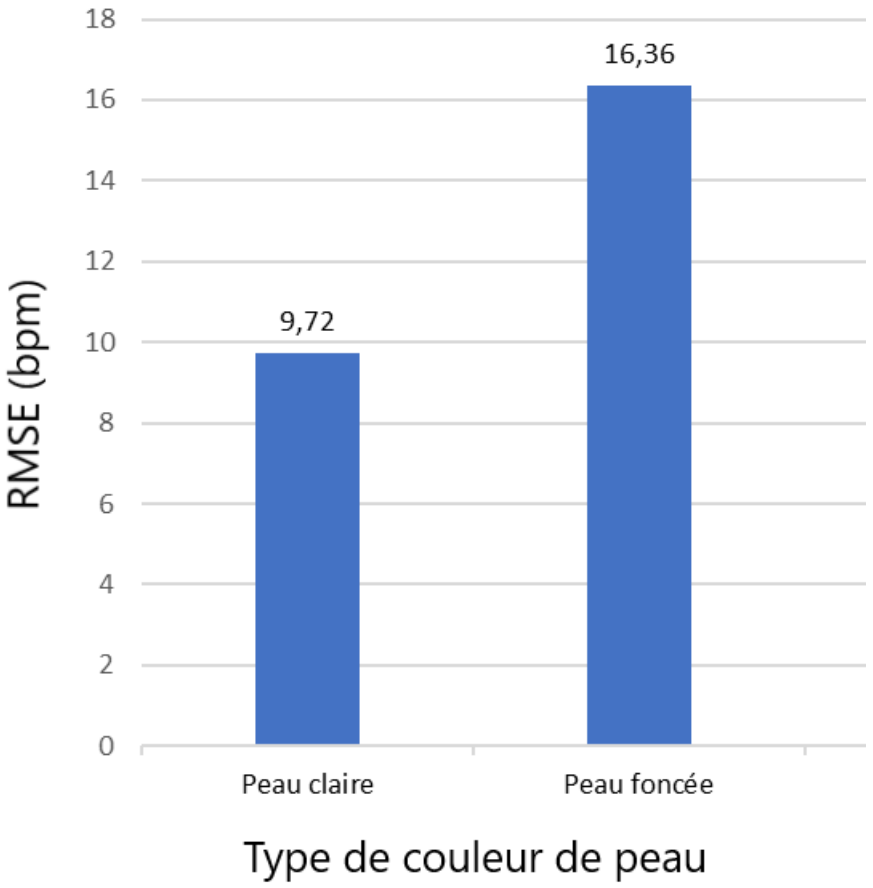


FIGURE 3.10 – L'erreur d'estimation de la fréquence cardiaque de X-iPPGNet par couleur de peau.

3.3.3 Résultats et discussion

Les expériences menées visent à répondre à plusieurs objectifs. Premièrement, nous prouvons la possibilité de mesurer la fréquence cardiaque avec une grande précision sans passer par l'étape d'extraction du signal iPPG couramment utilisée. Deuxièmement, nous fournissons une comparaison des performances avec divers systèmes conventionnels ainsi que d'autres approches d'apprentissage profond récemment proposées pour la mesure sans contact de la fréquence cardiaque à partir de vidéos. Troisièmement, nous démontrons la capacité de généralisation de notre méthode dans des conditions difficiles pour illustrer l'efficacité du modèle proposé.

Afin d'étudier la capacité de généralisation et l'efficacité du réseau X-iPPGNet présenté dans la section 3.3.2, trois bases de données publiques largement utilisées sont employées, à savoir MMSE-HR [379], MAHNOB-HCI [380] et UBFC-rPPG [384]. La base de données MMSE-HR

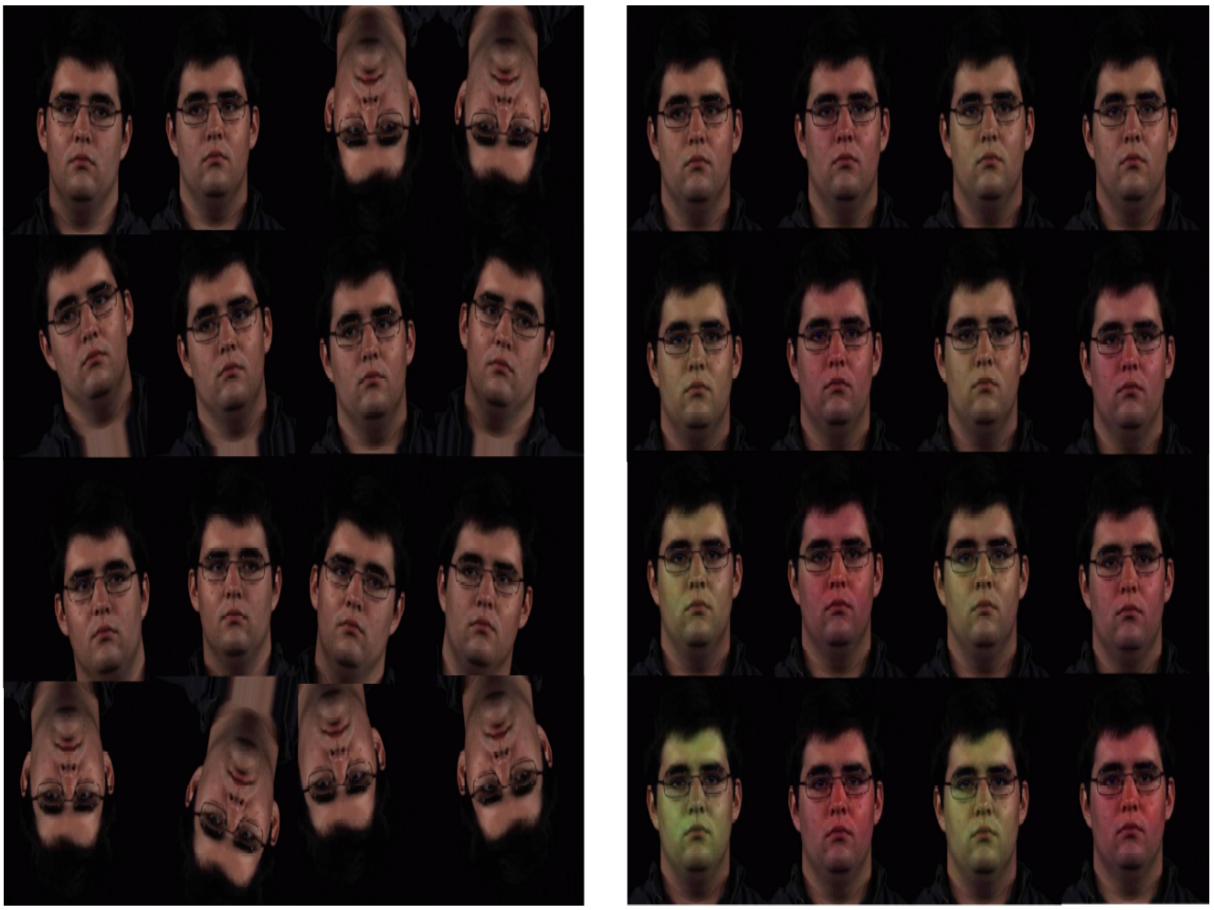


FIGURE 3.11 – Exemples d'augmentation de données avec les transformations géométriques (à gauche) et l'amplification vidéo (à droite).

est directement utilisée pour le test sans traitement supplémentaire car elle a été collectée dans les mêmes conditions que l'ensemble de données d'entraînement. UBFC-rPPG et MAHNOB-HCI sont sous-échantillonnées de respectivement 30 fps et 61 fps à 25 fps afin d'harmoniser les fréquences d'échantillonnage des vidéos d'entraînement et de test. Pour chaque expérience, nous n'utilisons pas les vidéos du même sujet pour l'apprentissage et le test. Nous évaluons et comparons les performances avec d'autres techniques de l'état de l'art en utilisant différentes metriques : l'écart-type (SD), l'erreur absolue moyenne (Mean Absolute Error, MAE, voir l'équation 3.1), la racine carrée de l'erreur quadratique moyenne (Root Mean Square Error, RMSE, voir l'équation 3.2) et le coefficient de corrélation de Pearson (r, voir l'équation 3.3). FC_i et $\widehat{FC_i}$ représentent respectivement la vérité terrain et la fréquence cardiaque estimée.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|FC_i - \widehat{FC_i}| \tag{3.1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(FC_i - \widehat{FC_i}\right)^2} \tag{3.2}$$

$$r = \frac{\sum_{i=1}^{n}(FC_i - \overline{FC_i})(\widehat{FC_i} - \overline{\widehat{FC_i}})}{\sqrt{\sum_{i=1}^{n}(FC_i - \overline{FC_i})^2(\widehat{FC_i} - \overline{\widehat{FC_i}})^2}} \tag{3.3}$$

3.3.3.1 Résultats

Evaluation sur MMSE-HR

Nous évaluons d'abord la capacité de généralisation de X-iPPGNet en entraînant le réseau sur BP4D+ et en le testant sur la base de données MMSE-HR. Le tableau 3.3 présente les comparaisons détaillées avec plusieurs approches de l'état de l'art, y compris les méthodes conventionnelles (Li2014 [399], CHROM [340], SAMC [346]) et les méthodes basées sur l'apprentissage profond (EVM-CNN [367], PhysNet [369], RhythmNet [333] et Auto-HR [365]).

TABLE 3.3 – Résultats de l'estimation de la fréquence cardiaque par notre approche et les méthodes de l'état de l'art sur la base de données MMSE-HR.

Méthode	**SD (bpm)**	**RMSE (bpm)**	***r***
Li2014 [399]	20.02	19.95	0.37
CHROM [340]	14.08	13.97	0.55
SAMC [346]	12.24	11.37	0.71
RhythmNet [333]	6.98	7.33	0.78
PhysNet [369]	12.76	13.25	0.44
AutoHR [365]	5.71	5.87	**0.89**
X-iPPGNet	**5.34**	**5.32**	0.85

Les résultats de la comparaison des performances avec les autres méthodes de l'état de l'art sont tirés de [365]. X-iPPGNet proposé dans cette étude obtient les meilleures performances (SD = 5,34 bpm ; MAE = 4,10 bpm ; RMSE = 5,32 bpm et r = 0,85), surpassant toutes les méthodes concurrentes. Nous observons également un écart significatif de performance entre les approches conventionnelles et celles basées sur l'apprentissage profond. Ces dernières ont obtenu des résultats tout à fait interessants et ont démontré la capacité de généralisation des réseaux de

neurones profonds et à fonctionner aussi bien dans des conditions difficiles que dans des scénarios contrôlés.

Evaluation sur UBFC-rPPG

Dans cette expérience, nous avons suivi la même stratégie que celle présentée dans [347]. 25 vidéos sont sélectionnées aléatoirement pour affiner le modèle pré-entraîné sur BP4D+ et nous réservons les vidéos restantes pour les tests. Étant donné que UBFC-rPPG contient très peu de vidéos de visages (une seule vidéo est enregistrée pour chaque sujet), nous avons utilisé une stratégie de validation croisée indépendante du sujet à 3-fold. La comparaison avec les méthodes de l'état de l'art est tirée de [347] et présentée dans le tableau 3.4. X-iPPGNet obtient de résultats tout à fait pertinents avec une capacité de généralisation remarquable pour de nouvelles données différentes de celles de l'apprentissage. Il convient de noter que nous obtenons les meilleurs SD (6,25 bpm) et RMSE (6,26 bpm) parmi les méthodes existantes et le deuxième meilleur MAE (4.99) derrière la méthode POS [322] avec une petite marge.

TABLE 3.4 – Résultats de l'estimation de la fréquence cardiaque par notre approche et les méthodes de l'état de l'art sur la base de données UBFC-RPPG.

Méthode	**SD (bpm)**	**MAE (bpm)**	**RMSE (bpm)**	***r***
Green [358]	20.2	10.2	20.6	-
ICA [318]	18.6	8.43	18.8	-
CHROM [340]	19.1	10.6	20.3	-
POS [356]	10.4	**4.12**	10.5	-
3DCNN [332]	8.55	5.45	8.64	-
Meta-rPPG [400]	7.12	5.97	7.42	0.53
PRNet [347]	6.45	5.29	7.24	-
X-iPPGNet	**6.25**	4.99	**6.26**	**0.67**

Evaluation sur MAHNOB-HCI

Nous proposons également de vérifier l'efficacité et la capacité de généralisation de notre méthode sur MAHNOB-HCI [380], qui est le jeu de données le plus couramment utilisé pour la

mesure sans contact de la fréquence cardiaque. Le taux de compression élevé et les mouvements spontanés causés par la stimulation émotionnelle rendent l'estimation de la fréquence cardiaque difficile. Nous avons suivi le même protocole que celui utilisé pour l'évaluation sur UBFC-rPPG, c'est-à-dire, la validation croisée indépendant du sujet à 3-fold. Nous avons pris aléatoirement 66 % des vidéos pour affiner le modèle pré-entraîné sur BP4D+ et utilisé les vidéos restantes pour le test.

TABLE 3.5 – Résultats de l'estimation de la fréquence cardiaque par notre approche et les méthodes de l'état de l'art sur la base de données MAHNOB-HCI.

Méthode	**SD (bpm)**	**MAE (bpm)**	**RMSE (bpm)**	r
Poh2011 [318]	13.5	-	13.6	0.36
CHROM [72]	-	13.49	22.36	0.21
Li2014 [399]	6.88	-	7.62	0.81
SAMC [346]	5.81	4.96	6.23	0.83
SynRhythm [401]	10.88	-	11.08	-
DeepPhys [331]	-	4.57	-	-
HR-CNN [370]	-	7.25	9.24	0.51
rPPGNet [377]	7.82	5.51	7.82	0.78
RhythmNet [333]	3.99	-	3.99	0.87
PhysNet [369]	7.84	5.96	7.88	0.76
AutoHR [365]	4.73	3.78	5.10	0.86
PulseGAN [402]	-	4.15	6.53	0.71
X-iPPGNet	**3.93**	**3.17**	**3.93**	**0.88**

Le tableau 3.5 compare les performances de X-iPPGNet avec les techniques existantes, y compris les méthodes conventionnelles et celles basées sur l'apprentissage profond. D'après les résultats, nous pouvons voir que notre méthode donne les meilleures performances pour toutes les metriques (SD = 3,93 ; MAE = 3,17 ; RMSE = 3,93 et r = 0,88), surpassant tous les travaux de l'état de l'art avec une large marge. Il est clair que notre modèle est très performant dans différentes conditions d'acquisition et même dans le cas de vidéos fortement compressées dont le

taux de compression dégrade significativement la qualité des vidéos. Nous remarquons également que les méthodes conventionnelles donnent les plus mauvais résultats par rapport aux systèmes basés sur l'apprentissage profond qui exhibent de meilleures performances dans des conditions non contrôlées.

3.3.3.2 Analyse des performances

Nous fournissons également une analyse supplémentaire pour explorer les critères impactant les performances de mesure tels que la distribution des fréquences cardiaques dans la base de données d'apprentissage, les types de couleur de peau et le sexe, ainsi que les mouvements de la tête. Toutes les expériences ont été menées sur le jeu de données MMSE-HR.

Impact de la distribution des fréquences cardiaques

Afin d'analyser plus en détail l'impact de la distribution de la fréquence cardiaque sur les performances de notre X-iPPGNet, nous traçons le diagramme de Bland-Altman représentant les différences entre la fréquence cardiaque estimée et celle de vérité terrain en fonction de la vérité terrain (voir la Figure 3.12). Le diagramme de Bland-Altman montre que la distribution est plus concentrée entre les limites de concordance 95% (1,96 SD) pour les plages de fréquences basses [$<$ 70 bpm] et moyennes [70 bpm, 90 bpm]. Cependant, les prédictions pour les FC élevées [$>$ 90 bpm] présentent quelques valeurs aberrantes. Nous supposons que cette observation est liée à l'ensemble d'entraînement déséquilibré (voir figure 3.1). Nous supposons aussi que le taux d'erreur augmente de manière significative pour les FC élevées par rapport aux faibles et moyennes fréquences en raison de leurs fluctuations sur la fenêtre temporelle [347].

De plus, le Bland-Altman présente une tendance négative marquée. Le modèle a tendance à surestimer les FC faibles et à sous-estimer les FC élevées. Nous supposons que cette observation est une conséquence directe du déséquilibre de la base de données. Le modèle a tendance à produire des prédictions orientées vers des valeurs de FC moyennes. La différence de FC est donc positive pour les FC faibles et négative pour les FC élevées.

Impact des types de peau et du sexe

MMSE-HR a été choisi pour évaluer la généralisabilité de notre méthode à différentes couleur de peau. Cette base de données est plus diversifiée en termes d'ethnicité par rapport à UBFC-rPPG [384] et MAHNOB-HCI [380], qui sont fortement biaisés vers les peaux claires. En suivant le protocole employé par les auteurs de [403], qui est basé sur l'échelle de Fitzpatrick [404], nous divisons la base de données en 4 catégories selon le type de couleur de peau. En plus des types

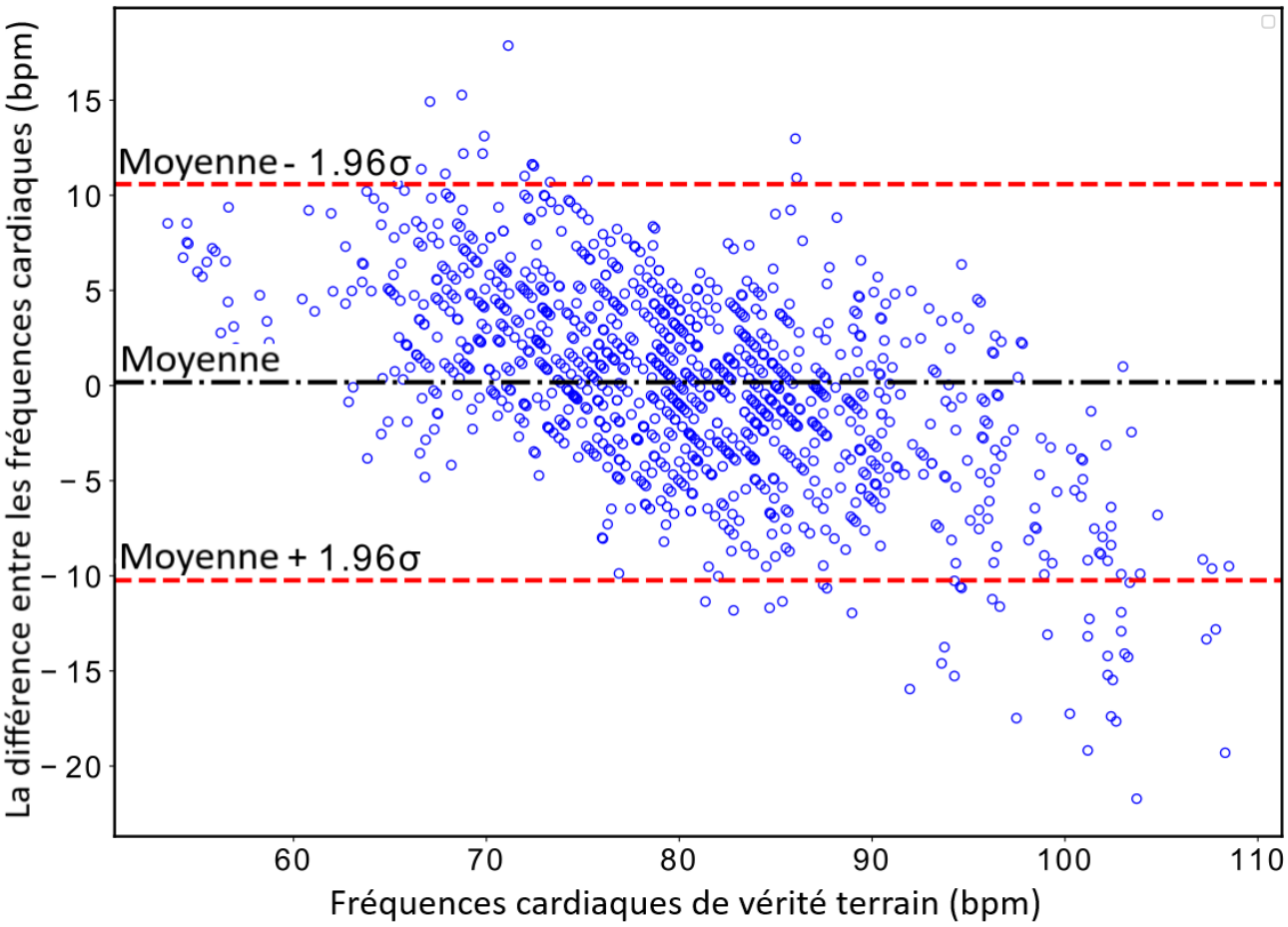
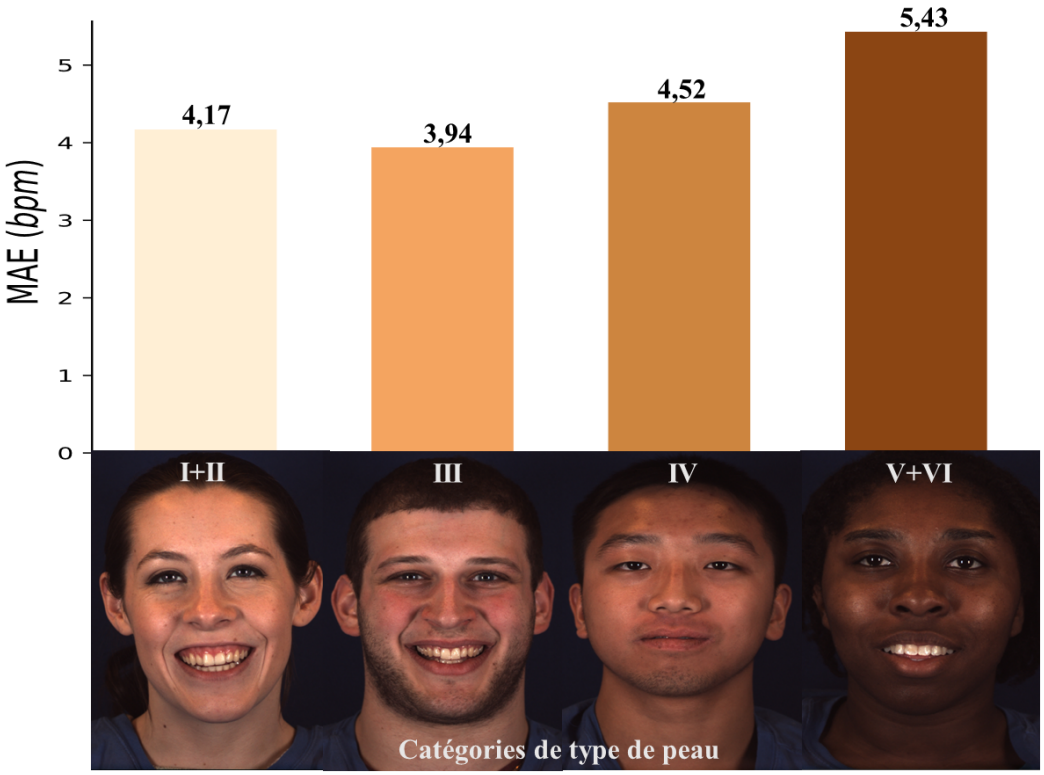


FIGURE 3.12 – Diagramme de Bland-Altman montrant les différences de fréquence cardiaque entre les valeurs de vérité terrain et estimées par rapport aux mesures réelles. Les résultats des analyses sur la base de données MMSE-HR sont ici présentés. Les moyennes sont représentées par des lignes noires en pointillés et les limites de concordance à 95 % (1,96 SD) par des lignes rouges en pointillés.

III et IV de l'échelle de Fitzpatrick, nous avons regroupé les types de peau I + II et V + VI car il y a relativement peu de sujets dans ces catégories.

Les prédictions de X-iPPGNet pour ces différentes couleurs de peau sont raportées dans le tableau 3.6. La technique proposée est très performante pour tous les types de peau et relativement moins pour les peaux les plus foncées (type V et VI), étant donné que les participants ayant des couleurs de peau plus foncées sont sous-représentés dans l'ensemble d'entraînement. Les plus faibles erreurs ont été obtenues pour le type de peau III ($MAE = 3.94, RMSE = 5.18, r = 0.81$), tandis que les peaux foncées (type V et VI) exhibent les mauvais résultats ($MAE = 5.34, RMSE = 6.82, r = 0.4$).

TABLE 3.6 – Erreur de l'estimation de la fréquence cardiaque de notre méthode par type de peau sur le jeu de données MMSE-HR.



Les types de peau Fitzpatrick	I+II	III	IV	V+VI
MAE (bpm)	4.17	3.94	4.52	5.43
RMSE (bpm)	5.31	5.18	5.76	6.82
r	0.87	0.81	0.84	0.40

TABLE 3.7 – Performance de notre méthode par sexe sur la base de données MMSE-HR.

Sexe	Hommes	Femmes
MAE (bpm)	3.74	4.53
RMSE (bpm)	4.76	5.84
r	0.85	0.79

Nous évaluons également l'impact du sexe sur l'estimation de la fréquence cardiaque par analyse vidéo. Les différences de résultats entre les hommes et les femmes sont beaucoup moins marquées que les différences entre les types de peau. Dans l'ensemble, les vidéos d'hommes présentent des erreurs d'estimation légèrement inférieures par rapport aux femmes (Hommes : $MAE = 3.74, RMSE = 4.76, r = 0.85$; Femmes : $MAE = 4.53, RMSE = 5.84, r = 0.79$) (Voir Tableau 3.7. Cela confirme les résultats rapportés dans une étude précédente montrant un taux d'erreur légèrement inférieur pour les hommes que pour les femmes [405].

Impact de mouvement de la tête

L'estimation visuelle de la fréquence cardiaque à l'aide de l'iPPG dans des environnements non contraints reste une tâche difficile. Outre la couleur de la peau et les conditions environnementales, les mouvements de la tête et les expressions faciales doivent également être pris en compte pour construire un système de mesure robuste. Afin d'évaluer comment les mouvements rigides (tels que l'inclinaison de la tête et les changements de posture) et les mouvements non rigides (telles que les expressions faciales) affectent la performance de X-iPPGNet, nous avons calculé l'erreur d'estimation de la FC pour les vidéos dont les sujets sont stables et celles qui incluent des expressions faciales et des mouvements de la tête. Les résultats sont présentés dans le tableau 3.8. Nous observons une faible dégradation de performances pour les mouvements significatifs ($MAE = 4.44, RMSE = 5.74, r = 0.82$) par rapport aux vidéos stables ($MAE = 3.88, RMSE = 4.91, r = 0.86$) mais l'erreur reste tout à fait acceptable. Cela prouve la robustesse de X-iPPGNet dans les scénarios plus complexes.

TABLE 3.8 – Performance de notre méthode sur MMSE-HR dans différentes conditions de mouvement de la tête.

Conditions de mouvement de la tête	**Stable**	**Mouvements significatifs**
MAE (bpm)	3.88	4.44
RMSE (bpm)	4.91	5.74
r	0.86	0.82

Taille de la fenêtre de temps

La taille de la fenêtre de temps est un paramètre important pour l'estimation du rythme cardiaque basée sur la vidéo. Des études antérieures ont rapporté qu'une taille de fenêtre plus longue conduit à de meilleures performances, en particulier lors de l'utilisation d'un filtre passe-bande ou de la densité spectrale de puissance [369, 367]. Néanmoins, cela augmente le temps de calcul, ce qui n'est pas adapté aux applications en temps réel. En effet, il existe un compromis à considérer dans le choix de la taille de la fenêtre. Si la fenêtre temporelle est trop grande, la FC prédite perd de l'information instantanée car nous faisons la moyenne des FC dans le fragment vidéo concerné. Inversement, le fragment vidéo d'entrée peut ne pas contenir un cycle complet de deux battements successifs si la taille de la fenêtre est trop courte, ce qui entraîne une estimation inexacte de la fréquence cardiaque.

TABLE 3.9 – Comparaison de la taille de la fenêtre de temps du fragment vidéo d'entrée de notre système et les méthodes de l'état de l'art.

Méthode	**Taille de la fenêtre de temps**
DeepPhys [331]	30 s
Siamese-rPPG [387]	20 s
CHROM [72]	10 s
POS [356]	10 s
SynRhythm [401]	10 s
RhythmNet [333]	10 s
2SR [348]	6 s
EVM-CNN [367]	4/6/8 s
PhysNet [369]	2/4/8 s
rPPGNet [377]	2 s (64 trames)
PRNet [347]	2 s (60 trames)
3DCNN [332]	2 s (60 trames)
X-iPPGNet	**2 s (50 trames)**

Le tableau 3.9 compare les différentes tailles de fenêtres des méthodes de l'état de l'art, y compris les méthodes conventionnelles et celles basées sur l'apprentissage profond. Toutes les études précédentes utilisent des fenêtres de temps beaucoup plus longues que notre méthode, à l'exception de PRNet [347], 3DCNN [332], et rPPGNet [377] qui utilisent un fragment vidéo de 2 secondes, mais avec un nombre d'images plus élevé.

TABLE 3.10 – Performance et temps de calcul de notre méthode sur MMSE-HR en utilisant différentes tailles de fenêtres de temps.

Taille de fenêtre de temps	**1s**	**2s**	**3s**	**4s**	**6s**
MAE (bpm)	10.21	**4.10**	6.41	7.75	8.13
RMSE (bpm)	12.89	**5.32**	7.98	9.77	10.02
Temps de calcul (ms)	**120**	140	160	180	220

Le tableau 3.10 présente le temps de calcul et la précision en fonction de la taille de fenêtre. Il est clair que l'augmentation de la taille de la fenêtre de temps implique plus d'images d'entrée et plus de paramètres entraînables, ce qui augmente le temps de calcul. Il en va de même pour la précision, où la MAE et la RMSE croissent avec l'augmentation des fenêtres de temps, à l'exception de la fenêtre de 1 seconde qui ne couvre pas l'intervalle de basse fréquence. Pour cette raison, la fenêtre de 2 secondes a été soigneusement sélectionnée pour avoir un cycle cardiaque complet et couvrir toute la plage de fréquence cardiaque. Les temps de calcul des méthodes qui utilisent une fenêtre de 2 secondes sont indiqués dans le tableau 3.11. X-iPPGNet atteint un temps d'inférence de 140 ms derrière PRNet [347], qui fonctionne le plus rapidement parmi les six méthodes. X-iPPGNet est cependant un réseau plus profond et surpasse PRNet en termes de précision.

3.3.3.3 Discussion

Ce travail a été entrepris pour optimiser et améliorer les systèmes basés sur l'iPPG pour l'estimation de la fréquence cardiaque. La plupart des études existantes extraient le signal iPPG en utilisant des approches conventionnelles [318, 340, 357, 399, 348, 356] ou des méthodes basées sur l'apprentissage profond [331, 369, 333, 365]. La fréquence cardiaque est généralement calculée comme l'inverse de la différence de temps entre deux battements consécutifs dans le domaine temporel, ou la fréquence ayant l'énergie du spectre de puissance la plus élevée dans le domaine fréquentiel. Par conséquent, des étapes de traitement supplémentaires telles que la détection des

TABLE 3.11 – Temps de calcul de notre approche par rapport aux méthodes de l'état-de-l'art utilisant une fenêtre de temps de 2 secondes.

Méthode	**Temps de calcul (ms)**
rPPGNet [377]	230
PhysNet [369]	200
3DCNN [332]	155
LCOMS [406]	150
PRNet [347]	**130**
X-iPPGNet	140

pics, la transformation de Fourier rapide ou la densité spectrale de puissance sont nécessaires. En outre, la précision dépend de la qualité de la forme d'onde iPPG et de la précision de la détection des pics principaux. Néanmoins, les bases de données disponibles publiquement intègrent des enregistrements vidéos dont l'extraction du signal iPPG est complexe et fournissent un grand nombre de signaux de références corrompus ou de mauvaise qualité [379, 380, 381]. Cela affecte directement la localisation des pics principaux et, par conséquent, diminue la précision.

Nous avons proposé une approche correspondant à un réseau de neurones entraînable de bout en bout où la FC est directement prédite à partir d'enregistrements vidéo faciaux sans récupération séparée du signal iPPG et sans connaissances préalables. X-iPPGNet fusionne l'extraction du signal iPPG et la prédiction de la FC en une seule étape. Nous nous appuyons sur la capacité des modèles d'apprentissage profond à apprendre et découvrir implicitement des représentations à partir des données. L'apprentissage est entièrement supervisé, chaque fragment vidéo de 2 secondes prenant comme étiquette d'apprentissage une fréquence cardiaque de vérité terrain obtenue avec un dispositif de référence en contact.

Les principaux avantages de l'approche proposée résident dans sa simplicité et son faible temps de calcul. Une courte fenêtre temporelle est utilisée pour estimer la fréquence cardiaque (2 s, 50 frames). La taille de la fenêtre de temps a un impact direct sur les performances. Plus elle est grande, plus l'erreur est élevée, surtout lorsqu'il s'agit de FC élevées et fortement fluctuantes (voir Tableau 3.10). Ceci est dû à la perte d'informations instantanées puisque la fréquence cardiaque est estimée par l'opération de moyennage sur la fenêtre temporelle. De plus, notre modèle est plus adapté à la mesure en temps réel. L'architecture est basée sur une le réseau Xception qui

réduit considérablement le nombre de paramètres et le temps de calcul sans aucune dégradation des performances.

Étant donné que les données constituent le facteur le plus important dans les approches basées sur l'apprentissage profond, X-iPPGNet a été entraîné sur BP4D+ pour fonctionner avec précision dans des scénarios difficiles et permettre un entraînement plus robuste. BP4D+ fournit une grande quantité de données et une diversité ethnique, ainsi que des conditions difficiles. En outre, l'augmentation des données est appliquée pour accroître la quantité d'échantillons sous-représentés pour les hautes et basses fréquences. L'utilisation d'une telle base de données en conjonction avec l'augmentation de données permet l'apprentissage automatique de l'iPPG sans extraction manuelles de caractéristiques et sans connaissances préalabales. De plus, des techniques avancées d'optimisation de l'apprentissage profond ainsi que les stratégies de régularisation ont été utilisées dans notre travail, ce qui permettent de surmonter les problèmes de sur-apprentissage et d'améliorer la généralisation du modèle à de nouvelles données.

Les résultats expérimentaux obtenus montrent l'efficacité de la méthode proposée et prouvent la possibilité de mesurer la FC directement à partir de vidéos du visage sans passer par la récupération du signal iPPG. Les résultats des tests sur trois bases de données de référence surpassent les méthodes existantes et révèlent la capacité de généralisation à de nouvelles données. Nous avons également examiné l'impact de divers facteurs sur les erreurs de prédiction. L'évaluation montre une bonne performance dans des scénarios réels intégrant des mouvements de la tête, l'illumination, la compression vidéo, et différentes couleurs de peau.

La principale limite de notre méthode concerne la façon dont la fréquence cardiaque est mesurée. Bien que l'architecture soit entraînable de bout en bout et supérieure en termes de vitesse et de simplicité, la prédiction de la FC sans passer par l'extraction du signal iPPG ne permet pas l'extraction des caractéristiques de l'onde iPPG qui sont utiles dans les applications médicales [330] ou pour la reconnaissance de l'état affectif [407]. En outre, nous avons identifié plusieurs problèmes qui peuvent être améliorés dans les études futures. Premièrement, la plupart des bases de données accessibles au public sont très limitées en termes de quantité de données [384, 408, 409]. Ce manque de données rend l'entraînement des modèles d'apprentissage profond plus difficile et augmente donc la probabilité de sur-apprentissage et diminue la capacité de généralisation à de nouvelles données. Bien que quelques bases de données à grande échelle soient disponibles [379, 381, 380], elles ne sont pas très diversifiées et sont fortement biaisées vers les tons de peau clairs et les FC moyennes [70, 90]. Cela conduit à un manque de généralisation et à de mauvaises performances pour les échantillons sous-représentés. L'utilisation de données synthétiques [410, 402, 332, 401] ou la combinaison de plusieurs ensembles de données

[411] peut résoudre le problème de la quantité limitée de données tandis que l'application de stratégies avancées d'augmentation des données peut améliorer les performances pour pour les plages d'échantillons sous-représentés.

Deuxièmement, nous avons remarqué un taux élevé de signaux de vérité terrain corrompus et/ou de mauvaises qualités dans les bases de données utilisées. La préparation et le nettoyage des données sont essentiels pour entraîner correctement le réseau et éviter d'obtenir des performances limitées, quelle que soit la robustesse du modèle développé. Enfin, les réseaux existants sont souvent constitués d'un grand nombre de paramètres et nécessitent des temps de calcul élevés, ce qui entrave considérablement leur application sur des appareils aux ressources limitées tels que les téléphones mobiles. Par conséquent, l'étude de modèles de réseaux légers peut améliorer considérablement la vitesse de traitement tout en maintenant des performances similaires.

3.4 Conclusion

Dans ce chapitre, nous avons présenté une nouvelle approche permettant d'estimer la fréquence cardiaque à partir d'enregistrements vidéo du visage en utilisant un réseau spatio-temporel profond. Cette méthode est un moyen efficace de prédire la fréquence cardiaque sans extraction séparée du signal iPPG et sans connaissances préalables. X-iPPGNet est basé sur une architecture de réseau Xception qui s'est avérée efficace pour des tâches de vision par ordinateur en termes de précision, de vitesse de convergence et de faibles temps de calcul.

Nos expériences ont montré l'efficacité de l'architecture proposée, qui atteint une grande précision et surpasse les méthodes existantes sur trois ensembles de données de référence populaires, i.e. MMSE-HR, UBFC-rPPG et MAHNOB-HCI. Les résultats de cette étude ont démontré que la fréquence cardiaque peut être estimée à distance à partir de vidéos du visages sans qu'il soit nécessaire de recourir à l'extraction manuelle du signal iPPG.

4

Reconnaissance physio-visuelle de l'état affectif à partir de vidéos du visage

Sommaire

4.1 Introduction

Dans ce chapitre, deux études basées sur la fusion des expressions faciales et des signaux physiologiques sans contact sont présentées. Dans la première étude, nous proposons un système bimodal pour la reconnaissance de l'état émotionnel, tandis que la seconde porte sur la détection du stress. Les deux études reposent sur la même approche consistant en deux pipelines qui extraient chacun les caractéristiques des expressions faciales et des signaux physiologiques sans contact. Nous évaluons d'abord les performances de chaque modalité séparément, puis fusionons les deux pour estimer l'état émotionnel ou le stress.

4.2 Reconnaissance physio-visuelle des émotions spontanées à partir de vidéos faciales

4.2.1 Bases de données

Les bases de données dédiées à la reconnaissance des émotions peuvent se présenter sous différentes formes, selon la manière dont les données ont été collectées. Elles sont généralement catégorisées selon le format (image, vidéo, texte, audio, signaux physiologiques), le type d'annotation (modèle dimensionnel ou catégoriel) et la méthode d'élicitation (émotions actées ou spontanées). Dans le cadre de cette thèse, nous nous sommes intéressés seulement aux bases de données multimodales et spontanées.

Bien que de nombreuses bases de données multimodales sur les émotions soient disponibles, peu d'entre elles fournissent à la fois les vidéos et les signaux physiologiques correspondants. Les jeux de données existants sont assez limités non seulement en termes de quantité des données mais aussi en diversité. Dans cette étude, nous utilisons deux bases de données volumineuses avec un type d'annotation différent. Nous décrivons brièvement dans les sous-sections suivantes, les deux bases de données utilisées pour la reconnaissance de l'émotion.

4.2.1.1 MAHNOB-HCI

La base de données MAHNOB-HCI présentée dans la section 3.2.2 a été réutilisée dans cette étude. MAHNOB-HCI [380] est une base de données émotionnelle spontanée couramment utilisée dans les recherches sur l'état affectif et la mesure des signaux physiologiques. Elle fournit des données recueillies auprès de 30 participants pour six modalités (signaux EEG, signaux physiologiques, expressions faciales, audio, regard et mouvements du corps). Chaque participant regarde 20 vidéos extraites de films et de sites Web pour induire différentes émotions. La durée

des vidéos est comprise entre 35 et 117 s. Le test SAM (Self Assessment Manikin en anglais) a été utilisé après avoir visualisé chaque vidéo pour évaluer l'activation et la valence perçus sur une échelle discrète de 1 à 9 [35]. Dans notre étude, nous avons divisé l'évaluation de la valence et l'activation en deux classes : un niveau élevé (évaluation 6-9) et un niveau faible (évaluation 1-5). Parmi les 30 participants, les données de 24 d'entre eux ont été utilisées dans nos expériences, tandis que les données du reste des sujets ont été retirées car l'enregistrement était incomplet.

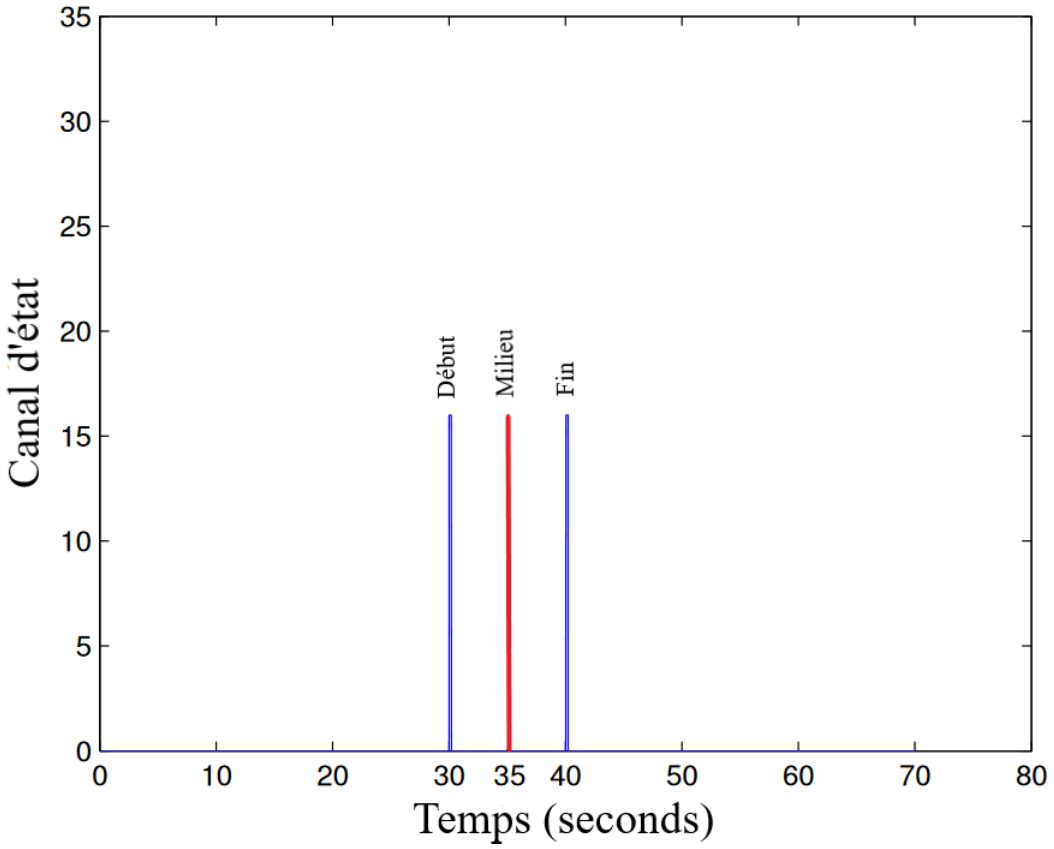


FIGURE 4.1 – Un exemple de canal d'état montrant le début et la fin d'un stimulus. Le stimulus a commencé à exactement 30s et s'est terminé vers 40s.

Les fichiers bdf fournis dans la base de données comprennent 47 canaux dont les signaux ECG sont récupérés du canal n°33, tandis que le canal d'état (n° 47) indique le début et la fin de stimuli (représentés en barre bleu sur la Figure 4.1). Le canal d'état a été utilisé pour extraire les segments les plus expressifs. Nous avons considéré le milieu de la fenêtre de début et de fin des stimuli pour représenter l'APEX émotionnel. Ainsi, 2 secondes ont été prises de la droite et de la gauche de l'APEX émotionnel (milieu, représenté en barre rouge sur la Figure 4.1) afin d'identifier l'émotion.

4.2.1.2 BP4D+

Depuis sa création, BP4D+ [379] a été largement utilisée dans plusieurs travaux liés à l'informatique affective et à la mesure sans contact des signes vitaux [412, 413, 346]. Cependant, nous sommes les premiers à l'avoir utilisée pour la reconnaissance multimodale des émotions à travers la fusion des expressions faciales et les signaux physiologiques sans contact. Comparée à MAHNOB-HCI [380], BP4D+ est plus complexe et diversifiée car elle a été collectée auprès de participants d'origines ethniques différentes présentant des mouvements significatifs (cf. sous-section 3.2.4). De plus, elle est annotée via des étiquettes discrètes et comprend des unités d'action codant les muscles faciaux.

140 sujets (82 femmes et 58 hommes) ont participé à 10 sessions conçues spécialement pour induire les émotions cibles (voir Tableau 4.1). Parmi les 10 tâches, seules quatre émotions sont utilisées dans nos expériences, correspondant à joie, embarras, peur et douleur. Ces catégories d'émotion sont fournies avec des unités d'action codées manuellement permettant d'extraire les images les plus expressives de chaque émotion. Le reste des enregistrements n'a pas été exploité en raison de non étiquetage de données.

TABLE 4.1 – Descriptions des tâches de BP4D+.

Tâche	**Activité**	**Émotion**
T1	Écouter une blague amusante	Joie
T2	Regarder un avatar 3D du participant	Surprise
T3	Appel téléphonique d'urgence	Tristesse
T4	Ressentir une explosion soudaine	Choc
T5	Question vraie ou fausse	Scepticisme
T6	Improviser une chanson	Embarras
T7	Faire une expérience d'une menace physique dans le jeu de fléchettes	Peur
T8	Plonger la main dans l'eau glacée	Douleur
T9	Se plaindre pour une mauvaise performance	Colère
T10	Sentir une mauvaise odeur	Dégoût

4.2.2 Système multimodal pour la reconnaissance des émotions

4.2.2.1 Préparation des données

Tout d'abord, nous avons extrait les trames les plus expressives de chaque vidéo en utilisant les unités d'action et le canal d'état fournis dans BP4D+ [379] et MAHNOB-HCI [380] respectivement. Ensuite, nous avons appliqué le traitement présenté dans la sous-section 3.3.1 pour la segmentation du visage et l'élimination des régions non cutanées. Cette étape permet d'améliorer l'extraction du signal iPPG et ne conserver que les régions d'intérêt pour identifier les expressions faciales et mesurer l'activité cardiaque.

Toutes les images de visages segmentés ont été découpées en fonction des coordonnées des pixels non noirs, puis redimensionnées à $48 \times 48 \times 3$ (hauteur, largeur, nombre de canaux). En outre, une stratégie d'augmentation de données a été appliquée à l'ensemble d'entraînement afin d'accroître le volume de données d'apprentissage. Plusieurs transformations géométriques telles que la rotation, la translation, le zoom et la modification de luminosité de l'image, ont été appliquées de manière aléatoire sur les fragments vidéos (voir sous-section 3.3.2.2). Cela permet d'améliorer les performances du modèle.

4.2.2.2 Réseau de reconnaissance des expressions faciales

Le squelette de l'architecture Xception [390] s'est avéré efficace et a prouvé sa capacité de généralisation pour les tâches de vision par ordinateur ainsi qu'en iPPG d'après les résultats obtenus par X-iPPGNet dans le Chapitre 3. Le modèle proposé a appris le concept iPPG à partir de zéro sans incorporer de connaissances préalables ni passer par l'extraction des signaux iPPG. Les évaluations approfondies ont montré une capacité de généralisation remarquable pour de nouvelles données différentes de celles de l'apprentissage.

L'architecture Xception est un dérivé de l'architecture Inception [91] où les modules d'Inception sont remplacés par des couches de convolution séparables en profondeur et des connexions résiduelles sont intégrées. Cette modification, par rapport à l'architecture Inception, réduit considérablement le temps de calcul et les besoins en mémoire, tout en maintenant des performances similaires [390]. Cependant, la convolution séparable en profondeur effectue une convolution spatiale pour chaque canal de couleur séparément sans tenir compte de la relation entre les différents canaux, tandis que la convolution standard considère toutes les informations spatiales et de canal ensemble sans séparation des canaux de couleur.

L'exploitation de la dépendance des canaux est un moyen important permettant d'améliorer les performances des réseaux de neurones convolutifs. Pour ce faire, nous avons intégré le module

Squeeze-Excitation (SE) [414] dans le réseau Xception. Le bloc SE vise à modéliser explicitement l'interdépendance entre les canaux de l'image, afin de recalibrer les cartes de caractéristiques par canal d'une manière adaptative et efficace en termes de temps de calcul. Cela permet d'augmenter la capacité d'apprentissage du modèle et d'améliorer les performances en mettant en évidence de manière sélective les caractéristiques informatives et en supprimant les informations moins utiles.

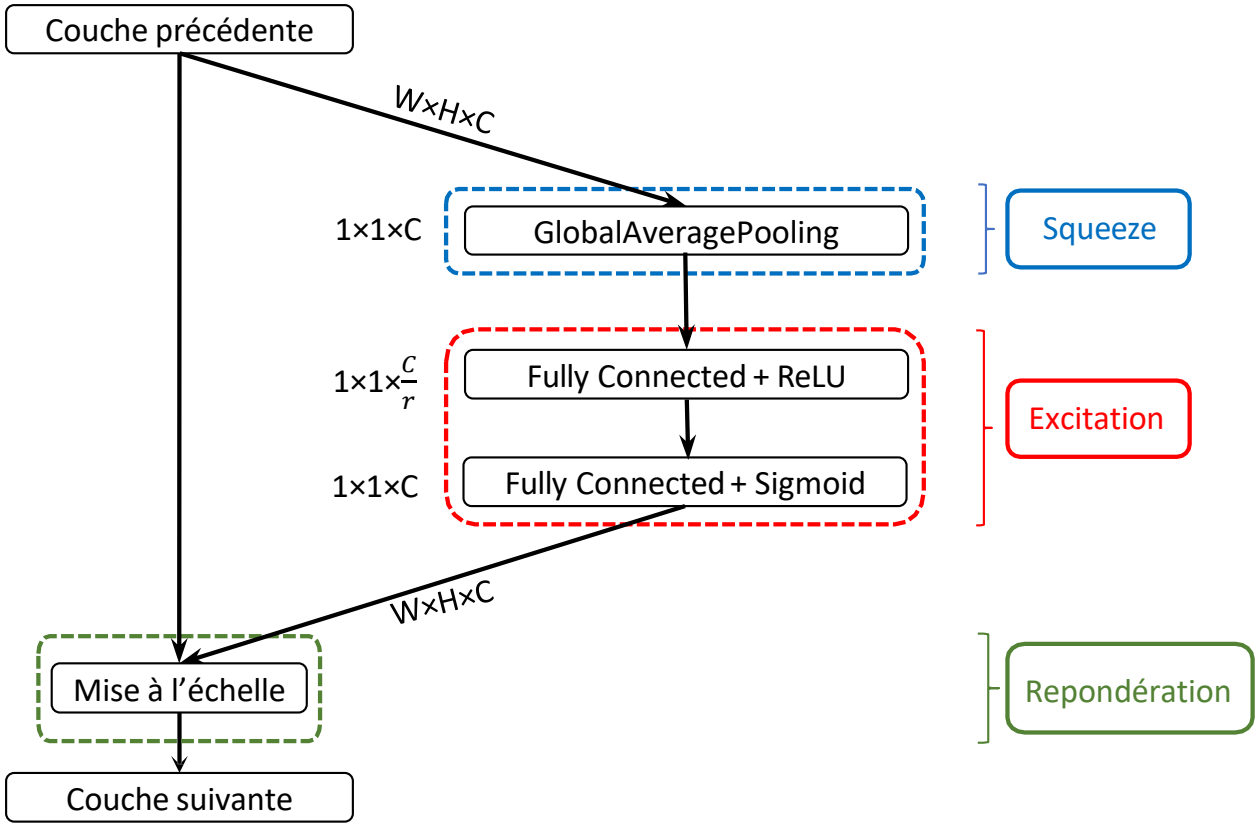


FIGURE 4.2 – Le module Squeeze-Excitation est composé d'une couche pooling par moyenne globale (Global Average Pooling en anglais) en tant qu'opération "Squeeze" et un bloc d'excitation composé de deux couches entièrement connectées (Fully Connected en anglais) qui sont utilisées pour apprendre les poids des caractéristiques. Nous réduisons d'abord la dimension de l'entité avec un paramètre de retrait r, puis nous récupérons la dimension avec le même r dans la prochaine couche entièrement connectée. Après l'opération d'excitation, le bloc SE effectue une mise à l'échelle pour repondérer les couches d'entrée, en multipliant l'élément d'entrée brute par la sortie d'excitation.

La structure du module SE est représentée dans la Figure 4.2. Le bloc SE se compose de deux parties successives : Squeeze et Excitation. L'opération "Squeeze" permet de réduire les dimensions spatiales de la carte de caractéristiques en utilisant une couche de mise en commun

par moyenne globale (GlobalAveragePooling), tandis que l'opération "Excitation" augmente l'importance de certains canaux dans la carte de caractéristiques. Elle se compose de deux couches entièrement connectées qui prennent respectivement les unités ReLU et sigmoïde comme fonction d'activation.

En se basant sur l'architecture de X-iPPGNet basée sur le réseau Xception et les avantages du module SE, nous avons implémenté un nouveau modèle appelé 3D-SE-XceptionNet dédié à la reconnaissance des expressions faciales. La figure 4.3 présente l'architecture globale du réseau proposé qui se compose de trois blocs (entrée, intermédiaire et sortie) comme l'architecture originale du réseau Xception. Cependant, la structure du modèle est simplifiée en réduisant le nombre de couches répétitives de convolution séparables en profondeur. 3D-SE-XceptionNet comprend 15 couches de convolution au lieu de 36 par rapport à la version originale. Ces couches de convolution sont structurées en 14 modules, tous reliés par des connexions résiduelles comme dans l'architecture ResNet [90] sauf le premier et le dernier module. Les blocs SE sont insérés après les connexions résiduelles. La sortie de l'extraction des caractéristiques est aplatie et passée à deux couches denses. La première couche contient 256 neurones tandis que la seconde dépend de l'encodage de la sortie. Quatre neurones sont utilisés pour classifier les quatre catégories d'émotion de BP4D+, tandis que deux neurones sont utilisés dans le cas de MAHNOB-HCI pour coder le niveau de la valence et l'activation. La première couche dense prend la fonction ReLu comme unités cachées, tandis que la seconde prend la fonction d'activation softmax pour prédire la classe d'émotion correspondante ou le niveau de valence/activation.

4.2.2.3 Réseau d'estimation des signaux physiologiques

Dans cette étude, les paramètres physiologiques sont mesurés à distance à l'aide de la méthode de photopléthysmographie par imagerie présentée dans la section 2.5.3. Plusieurs signes vitaux importants peuvent être dérivés de la forme d'onde iPPG, comme la fréquence cardiaque et sa variabilité, la fréquence respiratoire et la pression artérielle. Cependant, parmi ces paramètres physiologiques, seuls le signal iPPG et ses caractéristiques dérivées de la VFC ont été utilisés dans nos expériences. Il a été rapporté dans plusieurs études que la variabilité de la fréquence cardiaque est l'une des caractéristiques physiologiques les plus pertinentes pour l'estimation de l'état affectif d'une personne [415, 170]. Les caractéristiques de la VFC peuvent être dérivées de la variation de l'intervalle de temps entre deux battements cardiaques successifs dans le signal iPPG [416]. Pour cette raison, nous n'utiliserons pas le réseau X-iPPGNet présenté dans le chapitre précédent, car le X-iPPGNet est spécifiquement conçu pour l'estimation de la fréquence cardiaque alors que nous avons désormais besoin d'un modèle d'estimation du signal iPPG pour

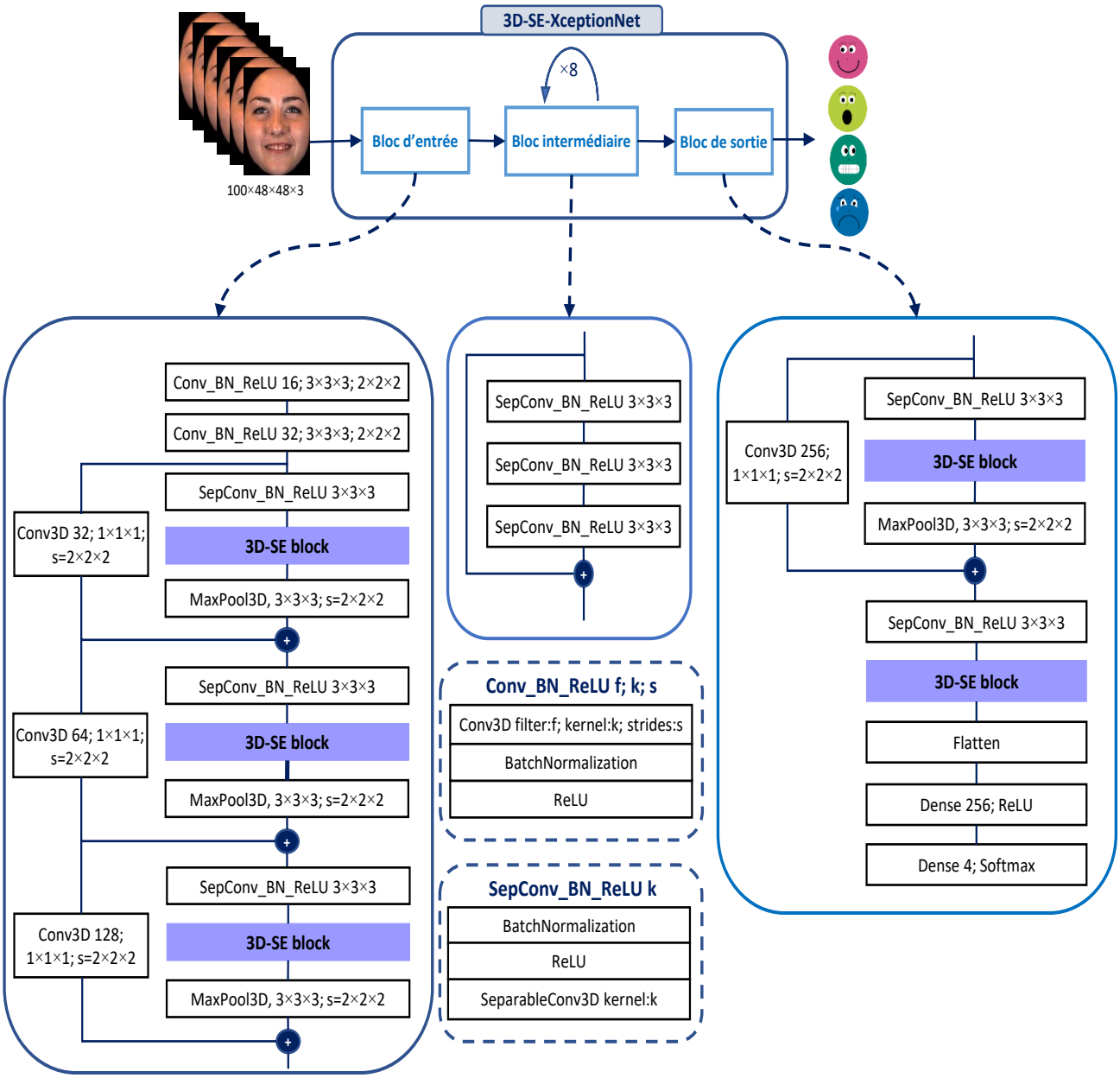


FIGURE 4.3 – La structure du réseau de 3D-SE-XceptionNet correspond à une version modifiée du réseau Xception. Les couches de convolution séparable en profondeur 2D sont remplacées par une convolution séparable en profondeur 3D pour capturer les caractéristiques spatiales et temporelles dans la vidéo. Le bloc SE qui joue le rôle d'un méchanisme d'attention a été intégré pour se focaliser sur les expressions faciales. Le fragment vidéo d'entrée passe d'abord par le flux d'entrée, puis par le flux intermédiaire qui est répété huit fois, et enfin par le flux de sortie qui se termine par une couche dense avec 4 neurones pour classifier les émotions ou 2 neurones pour quantifier le niveau de la valence et l'activation.

en extraire les caractéristiques de la variabilité cardiaque.

Les algorithmes d'extraction de l'iPPG peuvent être divisés en deux catégories : les algorithmes conventionnels [417] qui utilisent des étapes de traitement du signal/image et les approches basées sur l'apprentissage profond [374]. Dans ce travail, nous avons utilisé un réseau d'attention convolutif à décalage séquentiel multitâche (MTTS-CAN, voir Figure 4.4) proposé par Liu et al. [368] pour extraire le signal iPPG. Ce choix est motivé par la précision de l'algorithme et sa vitesse d'inférence ainsi que par le fait que son code source est ouvert au public. MTTS-CAN est l'une des méthodes d'apprentissage profond les plus populaires et les plus récentes, qui offre de bonnes performances en termes de mesure de la fréquence cardiaque et respiratoire. L'architecture du réseau de MTTS-CAN [368] est présentée dans la Figure 4.4. Elle se compose de deux branches en parallèle entraînables conjointement. La première branche est basée sur un modèle de mouvement qui repose sur la différence normalisée entre les images adjacentes comme représentation du mouvement d'entrée et capture des informations temporelles grâce à l'introduction d'un module de décalage temporel. Ce dernier permet l'échange d'informations entre les images voisines et évite l'utilisation des opérations de convolution 3D coûteuses. La deuxième branche consiste en un modèle d'apparence qui permet de guider le modèle de mouvement par le biais d'un mécanisme d'attention. Ce mécanisme d'attention permet la sélection automatique de la région d'intérêt en attribuant des poids plus élevés aux zones de la peau présentant des signaux iPPG plus forts.

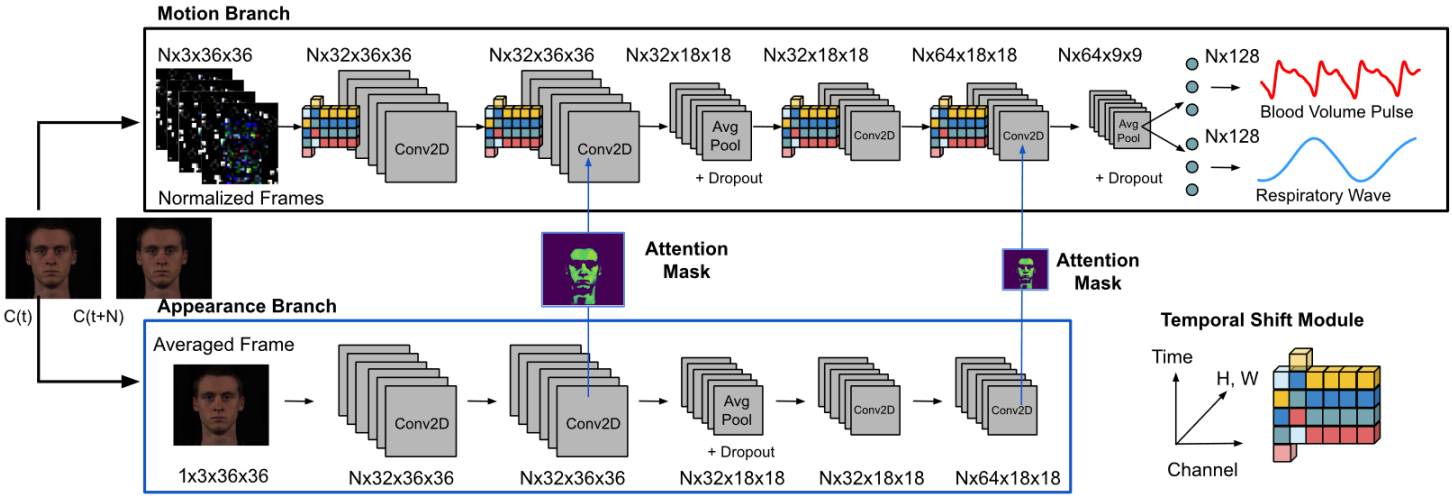


FIGURE 4.4 – L'architecture du réseau MTTS-CAN [368].

Afin de mieux évaluer la qualité du signal iPPG récupéré, nous présentons, dans la figure 4.5, une superposition d'un exemple de signal PPG de vérité terrain enregistré par un capteur en

contact et du signal iPPG prédit par le réseau MTTS-CAN. Le signal iPPG estimé est fortement corrélé avec la vérité terrain et l'emplacement des pics est très proche.

MTTS-CAN correspond à un réseau hybride qui utilise un mécanisme d'attention en conjonction avec des modules de décalage temporel. L'architecture du MTTS-CAN est présentée dans la figure 4.4. Les signaux iPPG récupérés par MTTS-CAN permettent d'extraire les caractéristiques de la VFC à la fois dans le domaine temporel et dans le domaine fréquentiel. Pour l'analyse temporelle et fréquentielle, la détection des pics est effectuée pour localiser l'instant où le battement cardiaque se produit, ce qui permet de calculer les caractéristiques de la VFC.

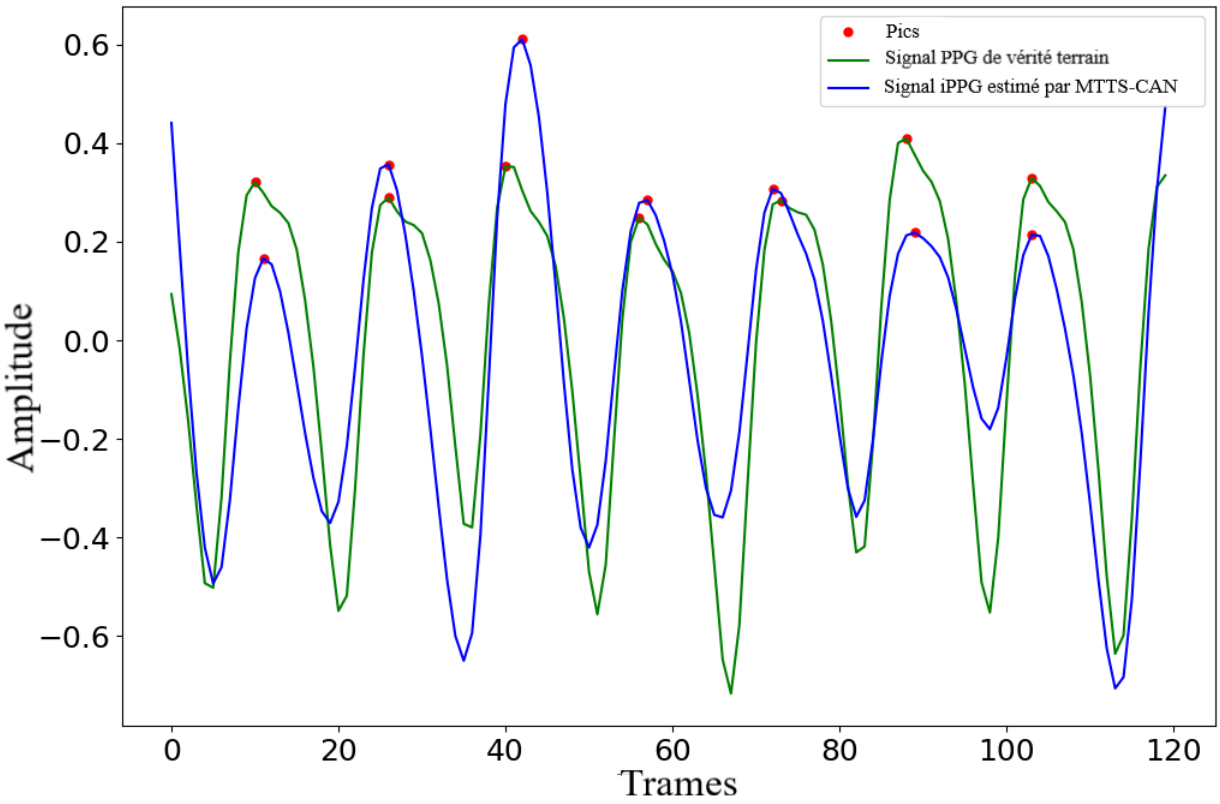


FIGURE 4.5 – Comparaison entre un signal prédit par MTTS-CAN et le signal PPG correspondant de vérité terrain tiré de la base de données BP4D+.

Dans le domaine temporel, la fréquence cardiaque est calculée comme l'inverse des IBI divisé par 60 pour obtenir la fréquence en battements par minute. À partir des variations de la fréquence cardiaque dans la fenêtre sélectionnée, nous avons calculé la moyenne (MeanFC) et l'écart-type (StdFC) de la série de fréquences cardiaques. La moyenne quadratique des différences d'intervalles successifs (RMSSD) est également calculée (voir Equation 4.1). Ce paramètre permet d'évaluer l'activité vagale reflétée dans la variabilité cardiaque [313].

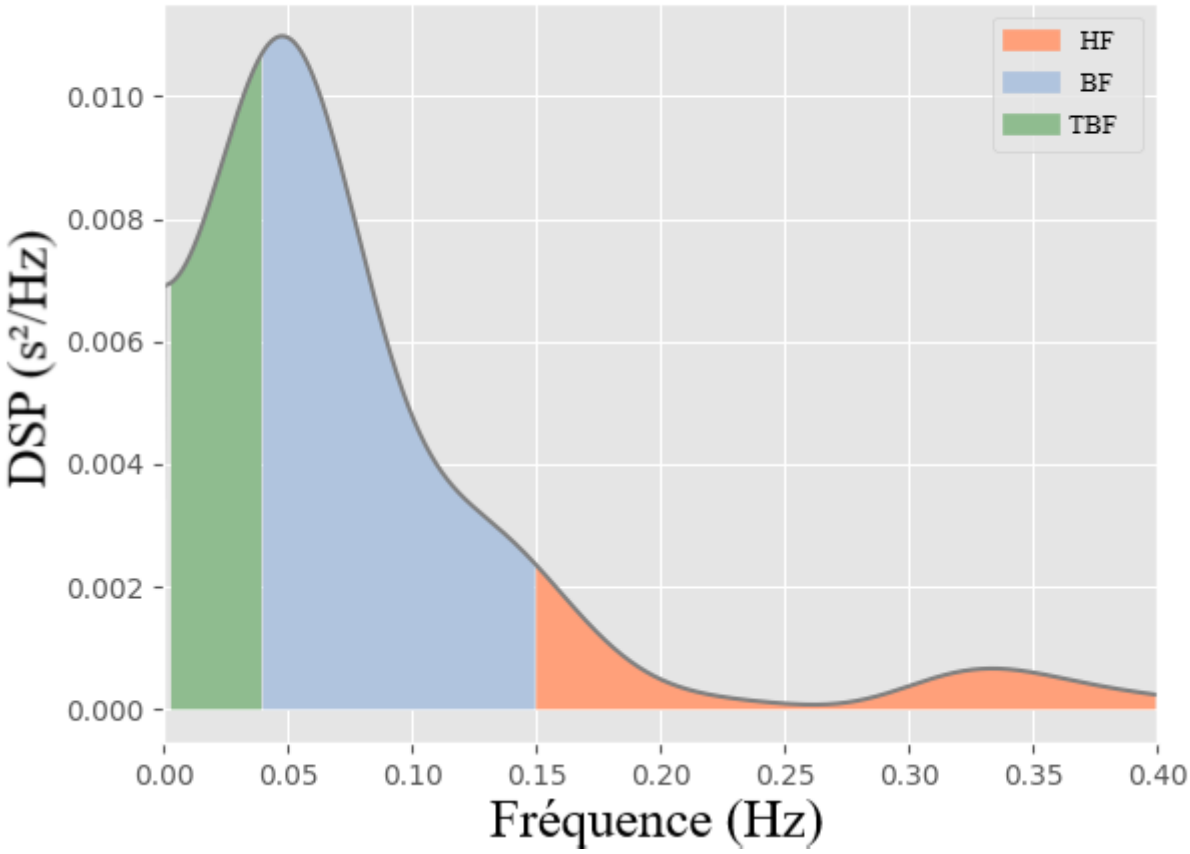


FIGURE 4.6 – Spectre de puissance des intervalles IBI montrant les composantes oscillatoires trés basse fréquence (TBF), basse fréquence (BF) et haute fréquence (HF).

$$RMSSD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(IBI_{i+1} - IBI_i)^2} \tag{4.1}$$

Dans le domaine fréquentiel, les séries IBI ont été interpolées par la méthode de Hermite cubique et les spectres de puissance ont été obtenus en employant la méthode de Welch [418]. La densité spectrale de puissance (DSP) du signal permet d'analyser ses différentes composantes oscillatoires telles que les composantes basse fréquence (BF) et haute fréquence (HF) de la VFC. La composante BF est modulée par l'activité du baroréflexe et contient à la fois l'activité sympathique et parasympathique, tandis que la composante HF reflète la branche parasympathique du SNA [419]. Les puissances BF et HF de la VFC ont été calculées comme l'aire sous la courbe de la DSP dans les plages correspondant à 0,04-0,15 Hz et 0,15-0,4 Hz respectivement (voir Figure

4.6). Nous avons également calculé le rapport BF/HF, qui représente l'équilibre sympatho-vagal [420]. Les composantes de très basse fréquence (TBF) n'ont pas été utilisées dans nos expériences.

4.2.2.4 Détails d'implementation

Le système proposé est mis en œuvre à l'aide des frameworks Keras et tensorflow et exécuté sur deux Nvidia Quadro P6000s. MAHNOB-HCI est sous-échantillonnée de 61 fps à 25 fps afin d'harmoniser la fréquence d'échantillonnage avec BP4D+. La longueur des clip vidéos du visage est fixée à $Nbframes = 100$ images (correspondant à 4 secondes), tandis que la taille de chaque image est de $48 \times 48 \times 3$ ($ImHeight \times ImWidth \times Channel$). La même configuration des hyperparamètres que le Chapitre 3 a été utilisée (cf. sous-section 3.3.2.1). En revanche, nous avons utilisé la fonction de perte d'entropie croisée binaire (Binary Cross Entropy en anglais) avec une fonction d'activation sigmoïde pour l'estimation du niveau de la valence/activation tandis que l'entropie croisée catégorielle (Categorical Cross Entropy en anglais) et la fonction d'activation softmax sont utilisées pour la classification catégorielle des types d'émotions.

4.2.3 Résultats et discussion

Pour illustrer l'efficacité de l'architecture proposée, nous présentons une évaluation expérimentale approfondie et rapportons les résultats sur les deux bases de données BP4D+ [379] et MAHNOB-HCI [380]. Comme chaque base de données est annotée différemment, nous fournissons une évaluation indépendante pour chaque ensemble de données. Une stratégie de validation croisée indépendante du sujet à 3-fold a été utilisée pour évaluer les performances de notre méthode. Trois expériences différentes ont été menées pour classer les émotions en utilisant les expressions faciales uniquement, les modalités physiologiques uniquement et enfin la fusion des expressions faciales et les signaux physiologiques.

4.2.3.1 Reconnaissance unimodale des émotions à partir des expressions faciales

Résultats sur BP4D+

Nous avons évalué d'abord les performances de la méthode proposée avec et sans le bloc Squeeze-Excitation, puis nous l'avons comparé avec cinq réseaux de l'état de l'art à savoir, 3D-VGGNet [89], 3D-ResNet [90], 3D-DenseNet [421], 3D-InceptionNet [91] et 3D-XceptionNet [390]. Le tableau 4.2 présente les résultats en validation croisée 3-fold sur BP4D+. Les différentes architectures de l'état de l'art présentent une faible précision avec un taux de reconnaissance en dessous de 50 %. Cela montre la difficulté de ce jeu de données. Les expériences sur le modèle

TABLE 4.2 – Comparaison de la méthode proposée avec des réseaux de l'état de l'art sur BP4D+ pour la reconnaissance des expressions faciales.

Méthode	**Précision (%)**
3D-DenseNet [421]	37.91
3D-InceptionNet [91]	42.48
3D-ResNet [90]	44.44
3D-VGGNet [89]	49.02
3D-XceptionNet [390]	53.59
3D-SE-XceptionNet	**63.40**

proposé (avec et sans le module Squeeze-Excitation) montre une amélioration de performance par rapport aux réseaux de l'état de l'art. Nous avons obtenu un taux de reconnaissance de 53.59 % en utilisant le réseau 3D-XceptionNet, tandis que 3D-SE-XceptionNet atteint la meilleure précision avec un taux de 63,40%.

La combinaison du réseau Xception et le module SE présente plusieurs avantages. Elle permet d'obtenir des informations plus pertinentes et plus ciblées sur les expressions faciales grâce au module SE qui joue le rôle d'un mécanisme d'attention. Elle permet également d'exploiter les avantages du réseau Xception pour éviter le problème de fuite du gradient grâce aux connexions résiduelles et réduire le coût de calcul et les besoins en mémoire grâce aux convolutions séparables en profondeur.

La matrice de confusion a été également calculée afin de déterminer quelles émotions sont les plus faciles et les plus difficiles à distinguer. La matrice de confusion des résultats de classification de 3D-SE-XceptionNet est présentée dans la figure 4.7. La précision globale du réseau proposé est de 63,4%. Cependant, le taux de reconnaissance est relativement différent pour chaque catégorie d'émotion. La joie et la douleur sont les émotions les plus reconnues avec une précision de 80% et 81% respectivement, tandis que la peur est confondue avec la joie et la douleur. Cela peut s'expliquer en partie par les multiples comportements qui peuvent se produire lors de l'expression de cette émotion.

Résultats sur MAHNOB-HCI

Nous avons également vérifié l'efficacité de notre méthode sur la base de données MAHNOB-

FIGURE 4.7 – Matrice de confusion pour la classification des émotions à partir des expressions faciales.

HCI [380]. Le tableau 4.3 présente le taux de reconnaissance de la valence et l'activation de 3D-SE-XceptionNet par rapport aux méthodes existantes. Les résultats de notre méthode sont obtenus en suivant le même protocole utilisé pour l'évaluation sur BP4D+, i.e. la validation croisée 3-fold. Nous avons effectué une classification binaire sur les niveaux de valence et d'activation, qui sont réparties en deux classes : élevée (score 6-9) et faible (score 1-5). La moyenne de précision des 3 folds pour chaque dimension est présentée dans le tableau 4.3.

Le tableau de comparaison montre une faible précision de la valence par rapport à l'activation pour quasiment tous les travaux. En revanche, nous avons obtenu de bonnes performances pour les deux dimensions notamment pour l'activation où nous avons eu le meilleur taux de reconnaissance de 74 %. Quant à la valence, le taux de reconnaissance est également bon et vient juste après la méthode de Fang et al. [423] avec une faible marge.

TABLE 4.3 – Comparaison de la méthode proposée avec des méthodes de l'état de l'art sur MAHNOB-HCI pour la reconnaissance de la valence et l'arousal à travers les expressions faciales.

Méthode	**Valence (%)**	**Arousal (%)**
Huang et al. [185]	50.57	53.64
Wang et al. [422]	51.01	64.45
Zhong et al. [194]	54.06	56.47
Koelstra et al. [184]	64	67.5
Fang et al. [423]	**67.97**	66.73
3D-SE-XceptionNet	66	**74**

4.2.3.2 Reconnaissance des émotions à partir des signaux physiologiques

La classification des émotions à partir des signaux physiologiques sans contact est effectuée en utilisant les signaux iPPG et les caractéristiques de la variabilité caradiaque (IVFC). Ces dernières sont extraites à partir des signaux iPPG estimés par MTTS-CAN dans le domaine temporel en utilisant la détection des pics et dans le domaine fréquentiel à l'aide de la densité spectrale de puissance (voir sous-section 4.2.2.3). De nombreuses caractéristiques peuvent être estimées à partir du signal temporel iPPG ou du spectrogramme de la VFC. Cependant, nous n'avons utilisé que six caractéristiques de la variabilité de la fréquence cardiaque, à savoir : la moyenne de la fréquence cardiaque (MeanFC), l'écart-type de la fréquence cardiaque (StdFC), la moyenne quadratique des différences d'intervalles successifs (RMSSD), la composante haute fréquence (HF), la composante basse fréquence (BF) et le rapport BF/HF.

Résultats sur BP4D+

Trois expériences ont été menées pour la reconnaissance des émotions à partir de données physiologiques. D'abord, les signaux iPPG et les caractéristiques de la variabilité de la fréquence cardiaque sans contact (IVFC) sont utilisés séparément pour classer les émotions. Ensuite, nous les avons fusionné pour voir quelle approche donne la meilleure précision.

Inspiré par le travail de Fabiano et al. [424], un réseau de neurones à propagation avant (Feedforward Neural Network en anglais) a été utilisé dans nos expériences. Il se compose de deux couches. La couche d'entrée comporte le même nombre de neurones que la longueur du

TABLE 4.4 – Comparaison de la précision de la reconnaissance des émotions à partir de signaux physiologiques sur BP4D+.

Caractéristique	**Précision (%)**
iPPG	55.33
iVFC	53.59
iPPG + iVFC	44.64

signal d'entrée (100 pour la modalité iPPG, 6 pour la iVFC), tandis que la couche de sortie comprend le même nombre de neurones que le nombre de classes d'émotion à prédire (4 neurones correspondant à 4 catégories d'émotions). La fonction d'activation de la couche d'entrée est ReLU, tandis que la fonction d'activation softmax est utilisée pour la couche de sortie.

Le tableau 4.4 illustre la précision de la reconnaissance en utilisant les signaux iPPG et les caractéristiques de la iVFC séparément et après la fusion. Les signaux physiologiques donnent une faible précision par rapport aux expressions faciales, qu'ils soient utilisés séparément ou combinés. En outre, les performances sont meilleures en utilisant seulement les signaux iPPG par rapport à la seule utilisation de la iVFC. Cela peut être justifié par la courte durée des signaux IPPG utilisés pour l'extraction des caractéristiques de la iVFC ainsi que par la qualité des signaux iPPG qui est sujet au bruit et aux artefacts dus aux mouvements et aux conditions d'éclairage. Cela a donc un impact sur la précision des caractéristiques de la iVFC. D'autre part, la fusion des signaux iPPG et de la IVFC présente des performances plus faibles. Cela peut être lié au manque de corrélation entre le signal iPPG et les caractéristiques de la iVFC. Il convient de noter que ces résultats ne peuvent être comparés à aucune méthode de l'état de l'art car nous sommes les premiers a avoir exploité la base de données BP4D+ dans dans le contexte de la reconnaissance des émotions à partir des signaux physiologiques sans contact.

Résultats sur MAHNOB-HCI

Nous avons effectué la même expérience menée sur MAHNOB-HCIen utilisant une taille de fenêtre de temps plus longue (correspondant à 10 s) pour voir son impact sur les performances. Nous avons également comparé la précision de l'estimation entre les signaux ECG et iPPG et leurs caractéristiques VFC dérivées pour démontrer la faisabilité de remplacer les mesures de références en contact par les signaux capturés par caméra. Les caractéristiques de la variabilité cardiaque en contact (CVFC) et sans contact (iVFC) sont extraites de la même façon dans le

TABLE 4.5 – Comparaison de la précision de la valence et l'arousal en utilisant les signaux physiologiques sur MAHNOB-HCI.

Caractéristique	**Valence (%)**	**Arousal (%)**
ECG	80	76
iPPG	71	67.5
cVFC	**86**	**78**
iVFC	78	71
ECG + cVFC	51	47
iPPG + iVFC	49	46

domaine temporel et fréquentiel à partir des signaux temporels et des spectogrammes des signaux ECG et iPPG, respectivement (voir sous-section 4.2.2.3).

Le tableau 4.5 fournit la précision de reconnaissance de la valence et l'activation en utilisant les caractéristiques en contact (ECG et cVFC) et sans contact (iPPG et iVFC) ainsi que la fusion entre elles. Il est tout à fait logique que le taux de reconnaissance des signaux en contact surpasse celui des signaux estimés par caméra. Néanmoins, la différence n'est pas significative et peut être améliorée dans des travaux futurs. D'autre part, la précision des caractéristiques VFC (cVFC et iVFC) dépasse celle des signaux ECG et iPPG contrairement à ce qui a été obtenu dans les expériences sur BP4D+. Cela est lié à la taille de la fenêtre temporelle du signal caradiaque utilisé pour extraire les caractéristiques de la VFC car une durée de fenêtre courte affecte négativement la précision [425]. En revanche, la taille de fenêtre temporelle du signal iPPG utilisée a été déterminée en prenant en compte la durée des segments les plus expressifs qui est comprise entre 4s et 9s pour BP4D+ et environ 10s pour MAHNOB-HCI. Pour cette raison, nous ne pouvons pas augmenter davantage la taille de fenêtre temporelle du signal. D'autre part, le même résultat que BP4D+ a été obtenu lorsque nous combinons les signaux cardiaques (iPPG et ECG) et les caractéristiques de la VFC (iVFC et cVFC). La fusion de ces deux modalités n'est pas pertinente et donne un faible taux de reconnaissance que ce soit pour les signaux en contact ou sans contact.

4.2.3.3 Reconnaissance multimodale des émotions

L'architecture de notre système de reconnaissance multimodale des émotions est présentée dans la figure 4.8. Le modèle proposé se compose de deux pipelines permettant d'extraire les caractéristiques de chaque modalité à partir des flux vidéo (voir sous-section 4.2.2.2 et 4.2.2.3). Les vidéos sont introduites au réseau dédié aux expressions faciales (3D-SE-XceptionNet) et au réseau dédié aux signaux physiologiques (MTTS-CAN). Le premier pipeline permet d'extraire le vecteur de caractéristiques après la couche d'aplatissement (Flatten en anglais) (voir figure 4.8) en utilisant les poids pré-entraînés de notre modèle 3D-SE-XceptionNet, tandis que le second pipeline renvoie le signal iPPG récupéré par le réseau MTTS-CAN ainsi que les caractéristiques VFC calculées à partir du signal iPPG. Ainsi, trois expériences ont été menées pour la reconnaissance multimodale des émotions. Tout d'abord, les caractéristiques des expressions faciales sont fusionnées uniquement avec le signal iPPG, puis uniquement avec le vecteur de la iVFC. Enfin, toutes les modalités sont fusionnées. Le vecteur de résultat de la concaténation est ensuite passé à deux couches denses avec 256 et 4 neurones respectivement pour la base de données BP4D+, tandis que la dernière couche de sortie lorsque nous utilisons MAHNOB-HCI intègre 2 neurones seulement pour la classification des deux niveaux de la valence ou l'activation. La première couche dense prend les unités linéaires rectifiées ReLU comme fonction d'activation, tandis que la seconde prend la fonction d'activation softmax/sigmoïde pour prédire la classe d'émotion correspondante ou le niveau de valence/activation.

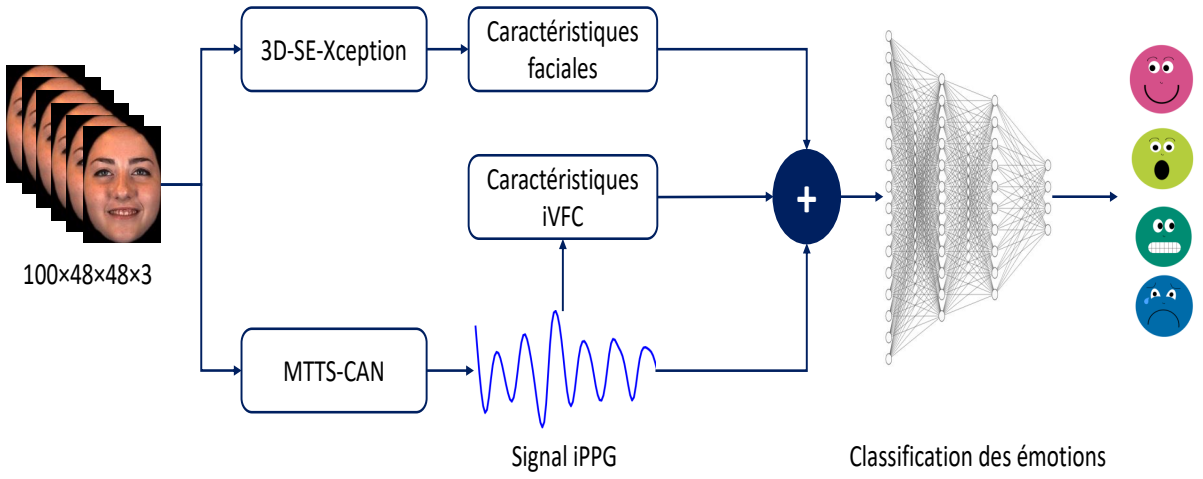


FIGURE 4.8 – Système proposé pour la reconnaissance multimodale des émotions basé sur les expressions faciales, signal iPPG et les caractéristiques de la VFC.

Résultats sur BP4D+

La précision de la reconnaissance pour chaque expérience est indiquée dans le tableau 4.6. Les résultats montrent que la fusion des caractéristiques des expressions faciales avec les paramètres physiologiques améliore les performances par rapport à l'approche unimodale utilisant les expressions faciales ou les données physiologiques séparément. Cela confirme les résultats obtenus par les études antérieures où la précision de la fusion des deux modalités dépasse celle obtenue par les systèmes unimodaux, et la précision des expressions faciales est toujours meilleure par rapport aux signaux physiologiques [426, 427]. En outre, le manque de corrélation entre le signal iPPG et les caractéristiques iVFC a un impact négatif sur la précision, qu'il s'agisse de fusionner uniquement ces deux modalités ou de les fusionner avec les expressions faciales.

TABLE 4.6 – Résultats comparatifs de la fusion des expressions faciales et des signaux physiologiques sur BP4D+.

Caractéristique	**Précision (%)**
Expressions faciales + iPPG + iVFC	67.97
Expressions faciales + iVFC	70.59
Expressions faciales + iPPG	**71.90**

La figure 4.9 représente la matrice de confusion pour le système multimodale de reconnaissance des émotions basé sur la fusion des expressions faciales et les caractéristiques de la iVFC, et sur la fusion des expressions faciales et du signal iPPG. Les performances globales du réseau proposé sont 70,59% et 71,90%, respectivement. Par rapport à l'utilisation des expressions faciales uniquement, la fusion avec les signaux physiologiques a amélioré de manière significative la précision des émotions mal classées. Par exemple, la précision de la peur a doublé, passant de 32% à 64% pour chaque schéma de fusion.

Résultats sur MAHNOB-HCI

Nous avons refait la même expérience menée sur la base de données MAHNOB-HCI. Le tableau 4.7 illustre la précision des trois schémas de fusion pour la reconnaissance de la valence et l'activation. Les résultats confirment ce que nous avons obtenu avec la base de données BP4D+. La fusion des expressions faciales avec les signaux physiologiques à amélioré le taux de reconnaissance de la valence et l'activation. La fusion des expressions faciales et les caractéristiques

(a) Expressions faciales + iVFC.

(b) Expressions faciales + iPPG.

FIGURE 4.9 – Matrice de confusion pour la classification multimodale des émotions à travers les expressions faciales fusionnées aux caractéristiques iVFC et aux signaux iPPG. Les expériences ont été menées sur BP4D+.

de la iVFC a atteint le meilleur taux de reconnaissance (Valence : 86 % ; Activation : 81 %) avec une légère superiorité par rapport à la fusion des expressions faciales et le signal iPPG. Notons la dégradation de performances sur le schéma de fusion combinant les expressions faciales, les signaux iPPG et les caractéristiques de la iVFC. Cette observation suit ce qui a été présenté précédemment (section 4.2.3.3).

TABLE 4.7 – Résultats comparatifs de la fusion des expressions faciales et des signaux physiologiques sur MAHNOB-HCI.

Caractéristique	**Valence (%)**	**Arousal (%)**
Expressions faciales + iVFC	86	81
Expressions faciales + iPPG	79	76
Expressions faciales + iPPG + iVFC	61	59

4.2.4 Conclusion

Les expressions faciales et les signaux physiologiques présentent des niveaux de précision différents, chacun ayant ses propres avantages et faiblesses. Bien que les expressions faciales soient visibles et faciles à classer par rapport aux indices physiologiques, l'incorporation de modalités

physiologiques peut fournir des informations complémentaires et améliorer encore la précision notamment dans le cas des émotions contrefaites ou inexprimées.

Dans cette étude, nous avons développé un système multimodal qui fusionne les expressions faciales et les signaux physiologiques sans contact. Un nouveau réseau de neurones spatio-temporel a été proposé combinant le module Squeeze-Excitation et l'architecture Xception. Deux paramètres physiologiques ont été sélectionnés, i.e. le signal iPPG et les caractéristiques iVFC. Contrairement aux travaux existants, les indices physiologiques ont été mesurées de manière sans contact à l'aide de la photopléthysmographie par imagerie. Ainsi, une seule source d'entrée, les vidéos du visage, a été utilisée pour extraire les caractéristiques de chaque modalité.

La fusion des deux modalités a amélioré considérablement la précision sur les deux bases de données BP4D+ et MAHNOB-HCI. Les résultats comparatifs obtenus s'intègrent parfaitement aux systèmes multimodaux existants qui utilisent différentes sources de données d'entrée, démontrant ainsi la possibilité d'utiliser uniquement des vidéos du visage pour identifier les émotions à l'aide d'indices physiologiques et visuelles.

4.3 Reconnaissance physio-visuelle du stress à partir des vidéos faciales

4.3.1 Base de données

La base de données UBFC-Phys [428] a été explorée dans cette étude. Il s'agit d'un ensemble de données multimodal publique dédié aux études psycho-physiologiques. Il fournit des données recueillies auprès de 56 participants étudiants en psychologie de premier cycle, dont 46 femmes et 10 hommes, tous âgés de 19 à 38 ans (avec un âge moyen de 21,8 ans). Les participants ont été soumis à une expérience de stress divisée en trois étapes : une tâche de repos T1, une tâche d'expression orale T2 et une tâche d'arithmétique T3 (T2 et T3 étant les tâches stressantes), au cours desquelles les participants étaient filmés et portaient un bracelet permettant de mesurer leurs signaux de pouls (BVP) et d'activité électrodermale (Voir Figure 4.10).

Un formulaire permettant de calculer le niveau de stress est présenté aux participants avant et après l'expérience. Pour chaque participant, trois vidéos (une vidéo par tâche) d'une durée de 3 minutes ont été enregistrées à une fréquence de 35 fps et avec une résolution de 1024×1024 pixels. Les signaux BVP et EDA pour chaque tâche ainsi que leurs scores d'anxiété calculés avant et après l'expérience sont disponibles au public [3].

3. https ://ieee-dataport.org/open-access/ubfc-phys-2

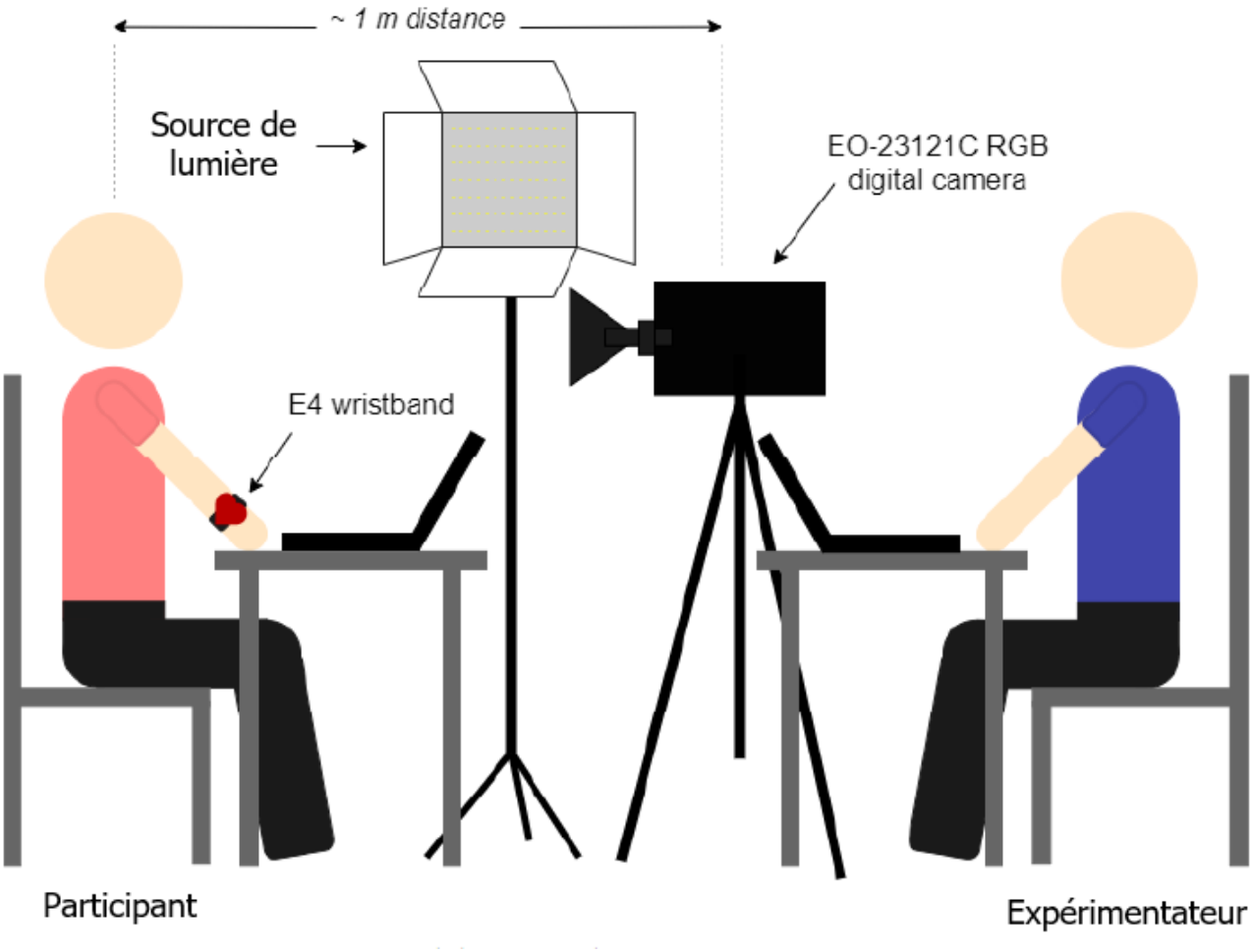


FIGURE 4.10 – Schéma du protocol expérimental développé pour construire la base de données UBFC-Phys [428].

4.3.2 Système multimodal pour la reconnaissance du stress

4.3.2.1 Préparation des données

Les expériences menées comprennent trois types de caractéristiques :

(a) Caractéristiques physiologiques mesurées en contact ;

(b) Caractéristiques physiologiques mesurées à partir des vidéos du visage ;

(c) Caractéristiques faciales.

Les caractéristiques physiologiques en contact sont colllectées à l'aide d'un bracelet. Elles comprennent les signaux BVP et les caractéristiques de la variabilité cardiaque en contact (CVFC). Les caractéristiques physiologiques mesurées à partir des vidéos faciales comprennent le signal iPPG et les caractéristiques de la variabilité cardiaque sans contact (iVFC). A cela s'ajoute

l'exploitation des caractéristiques faciales extraites des enregistrements vidéo à l'aide d'apprentissage par transfert. Les caractéristiques de la variabilité cardiaque en contact (cVFC) et sans contact (iVFC) sont extraites à partir des signaux BVP et iPPG, respectivement (voir sous-section 4.2.2.3).

4.3.2.2 Caractéristiques physiologiques en contact

Tout d'abord, les signaux BVP de vérité terrain sont rééchantillonnés à la fréquence d'échantillonnage de la caméra (35 Fps) afin d'harmoniser les fréquences des signaux en contact et sans contact. Ensuite, la supression des tendances est effectuée en utilisant une approche de lissage [385]. Après cela, nous avons appliqué un filtre passe-bande de Butterworth d'ordre 2 avec une fréquence de coupure de 0,75 et 2,5 Hz pour ne garder que les caractéristiques cardiaques.

À partir des signaux BVP filtrés, 8 caractéristiques de variabilité cardiaque (cVFC) ont été calculées dans le domaine temporel et fréquentiel. La fréquence cardiaque maximale (MaxFC) et minimale (MinFC) ont été calculé en plus des 6 caractéristiques extraites dans la sous-section 4.2.2.3, i.e. la moyenne de la fréquence cardiaque (MeanFC), l'écart-type de la fréquence cardiaque (StdFC), la moyenne quadratique des différences d'intervalles successifs (RMSSD), la composante haute fréquence (HF), la composante basse fréquence (BF) et le rapport BF/HF.

4.3.2.3 Caractéristiques physiologiques estimées à partir de vidéos du visage

La méthode MTTS-CAN présentée dans la sous-section 4.2.2.3 a été utilisée pour l'extraction du signal iPPG. Les mêmes types caractéristiques de la variabilité cardiaque (iVFC) sont ensuite mésurées à partir du signal iPPG après un simple filtrage et supression de tendance (voir la section 4.3.2.2).

4.3.2.4 Caractéristiques faciales

Les modèles d'apprentissage profond ont dépassé les algorithmes d'apprentissage automatique traditionnels dans les tâches de vision par ordinateur. Cependant, une grande quantité de données est nécessaire pour entraîner correctement le modèle afin d'obtenir des performances élevées. En raison de la limitation des données, nous avons examiné l'approche de l'apprentissage par transfert comme une alternative viable et pour l'extraction automatique de caractéristiques faciales.

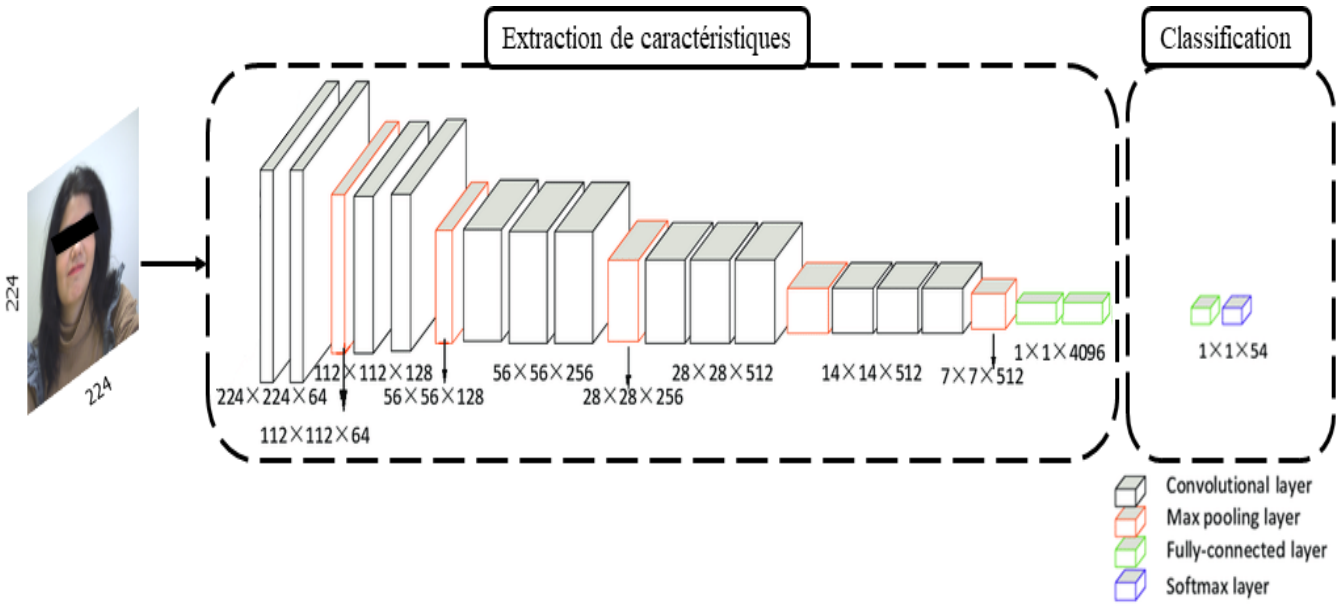


FIGURE 4.11 – L'architecture VGG16 montrant le bloc d'extraction de caractéristiques et de classification [429].

Un modèle VGG16 pré-entraîné [430] est adopté comme extracteur de caractéristiques faciales. Le modèle VGG16 est très populaire et s'est avéré très efficace pour obtenir une grande précision de reconnaissance dans nombreuses tâches de vision par ordinateur [431]. Le réseau se compose d'un bloc d'extraction de caractéristiques basé sur des couches de convolution et d'un bloc de classification constitué de couches denses (Voir Figure 4.11). Le bloc d'extraction de caractéristiques est figé, tandis que le bloc de classification est modifié en remplaçant le nombre de neurones et la fonction d'activation de la dernière couche dense pour qu'elle soit compatibles avec la classification stress/non stress. Ensuite, le réseau est affiné avec les données d'UBFC-Phys dédiées à la reconnaissance du stress.

4.3.3 Résultats et discussion

Les expériences réalisées reposent sur les mêmes spécifications présentées dans l'article original du jeu de données UBFC-Phys [428]. A l'aide du matériel supplémentaire II fourni avec l'article [4], nous avons retiré des données correspondant à des tâches où des problèmes techniques sont survenus lors de l'enregistrement. Parmi les 168 tâches (3 tâches × 56 participants) fournies dans la base de données, nous avons gardé que 101 tâches correspondant à des données non corrompues.

Nous avons utilisé une stratégie de validation croisée indépendante du sujet à 7-fold sur les

4. https ://ieeexplore.ieee.org/document/9346017/media

modalités séparées et fusionnées. Les données d'un participant ne sont jamais utilisées à la fois pour l'entraînement et le test. Nous avons créé aléatoirement 7 folds en utilisant 85 % des données pour l'apprentissage et les 15 % restants pour le test. La précision moyenne pour chaque fold est indiquée dans les tableaux 4.8 et 4.9.

Trois différentes expériences ont été menées pour la détection de l'état de stress :

(a) en utilisant les modalités physiologiques (contact et sans contact) uniquement ;

(b) en utilisant les caractéristiques faciales uniquement ;

(c) en fusionnant les signaux physiologiques et les caractéristiques faciales.

Nous proposons un classification binaire non stress / stress dans cette étude. La tâche T1 représente l'état non-stress, tandis que les tâches T2 et T3 représentent l'état de stress.

4.3.3.1 Reconnaissance du stress à partir de signaux physiologiques

Nous avons utilisé les algorithmes d'apprentissage automatique classiques et les idées proposées dans l'article original de l'ensemble de données UBFC-Phys pour comparer les performances de différents modèles [428]. Cinq classifieurs ont été considérés : SVM avec un noyau polynomial (SVM Poly), SVM avec un noyau gaussien (SVM RBF), les forêts aléatoires (Random Forest, RF), bayésien naïf (Naive Bayes, NB) et les k plus proches voisins (k-Nearest Neighbours, KNN). Nous avons effectué la même validation croisée à 7-fold sur les cinq algorithmes. Chaque classifieur a été entraîné avec 85 % des signaux, et les 15 % restants ont été utilisés pour le test.

Le tableau 4.8 fournit le taux de reconnaissance en utilisant les caractéristiques physiologiques en contact (BVP et cVFC) et sans contact (iPPG et iVFC). Dans cette expérience, le meilleur résultat a été obtenu en utilisant les caractéristiques physiologiques calculées à partir du signal BVP en contact. Les caractéristiques cVFC ont atteint la précision la plus élevée avec 78,16 %, suivies par les signaux BVP avec une précision de 72,61 %. Les meilleures performances des caractéristiques physiologiques en contact ont été obtenues par le classifieur bayésien naïf. En comparant les résultats obtenus, nous constatons que la précision de la reconnaissance de l'état de stress obtenu par le signal BVP est supérieure à celle du signal iPPG. Une observation similaire peut être faite pour les caractéristiques VFC avec et sans contact. Ces observations sont conformes à ce qui a été rapporté dans des études antérieures [432] mais en contradiction avec les résultats présentés dans l'article qui accompagne la publication de UBFC-Phys [428]. Les auteurs ont signalé des précisions plus élevées avec les signaux physiologiques mesurés à partir de la vidéo qu'avec les caractéristiques physiologiques en contact. Nous supposons que la précision de chaque modalité dépend du type de classifieur et de ses paramètres, ainsi que de la

TABLE 4.8 – Résultats de la classification des états de stress et de non-stress basée sur les signaux physiologiques

Caractéristique	**Classifieur**	**Précision (%)**
BVP	SVM RBF Kernel	69.72
	SVM Poly Kernel	58.58
	NB	**72.61**
	RF	66.96
	KNN	44.22
iPPG	SVM RBF Kernel	57.81
	SVM Poly Kernel	57.81
	NB	61.82
	RF	**62.40**
	KNN	59.96
cVFC	SVM RBF Kernel	72.74
	SVM Poly Kernel	74.55
	NB	**78.16**
	RF	58.58
	KNN	73.64
iVFC	SVM RBF Kernel	58.58
	SVM Poly Kernel	57.61
	NB	56.92
	RF	58.58
	KNN	**72.22**

méthode d'extraction du signal iPPG. Dans leurs expériences [428], une méthode conventionnelle composée de plusieurs étapes de traitement du signal et de l'image a été utilisée. Ici, nous avons choisis d'adopter une nouvelle approche d'apprentissage profond de bout en bout qui extrait la

forme d'onde iPPG automatiquement sans aucune étape supplémentaire de prétraitement [368].

4.3.3.2 Reconnaissance du stress à partir des caractéristiques faciales

Une stratégie d'apprentissage par transfert est adoptée dans cette expérience pour exploiter les connaissances pré-aprises dans le domaine de la reconnaissance d'objets et les transférer à la reconnaissance du stress. Les couches denses supérieures sont remplacées et ajustées avec des données de la base UBFC-Phys. Le système proposé est illustré dans la figure 4.12.

Les trames de l'ensemble des vidéos sont redimensionnées à (224×224×3) et ensuite introduite au modèle VGG16 pré-entraîné [430]. VGG16 est initialement entraîné sur le jeu de données ImageNet pour la reconnaissance des objets. La sortie des caractéristiques de VGG16 (avant les couches denses) est extraite et ensuite vectorisée à l'aide de GlobalMaxPooling1D pour avoir le vecteur de caractéristiques faciales. Ce vecteur est transmis à une couche LSTM pour tenir compte la dimension temporelle. Enfin, il est acheminé vers une couche dense composée de 2 neurones pour effectuer la classification. Pour cela, une fonction d'activation sigmoïde est appliquée à la couche dense, permettant ainsi d'effectuer la classification binaire du stress (stress/non-stress).

Les résultats présentés dans le tableau 4.9 montrent que la reconnaissance de l'état de stress basée sur les caractéristiques faciales surpasse les caractéristiques physiologiques mesurées avec ou sans contact. Cela est cohérent avec les résultats obtenus dans des études précédentes où la précision de la reconnaissance des affects/émotions à l'aide de caractéristiques visuelles (par exemple, les expressions faciales) surpasse les modalités physiologiques [432].

4.3.3.3 Reconnaissance multimodale du stress à partir des caractéristiques faciales et les signaux physiologiques

La figure 4.13 présente l'architecture globale du système multimodal proposé pour la reconnaissance du stress. Elle se compose de deux étages pour extraire les caractéristiques de chaque modalité à partir des enregistrements vidéo du visage. Chaque vidéo d'UBFC-Phys est introduite dans le réseau d'extraction des caractéristiques faciales et au réseau d'extraction des paramètres physiologiques (MTTS-CAN). Le premier pipeline extrait le vecteur de caractéristiques après la couche d'aplatissement en utilisant les poids pré-entraînés de VGG16 (voir figure 4.12), tandis que le second pipeline renvoie le signal iPPG récupéré par le réseau MTTS-CAN [368] et les caractéristiques iVFC.

Nous avons mené deux expériences sur notre système multimodal de reconnaissance du stress. La première est basée sur la combinaison des caractéristiques faciales et des caractéristiques iVFC. La deuxième combine les caractéristiques faciales et le signal iPPG. Le vecteur résultant

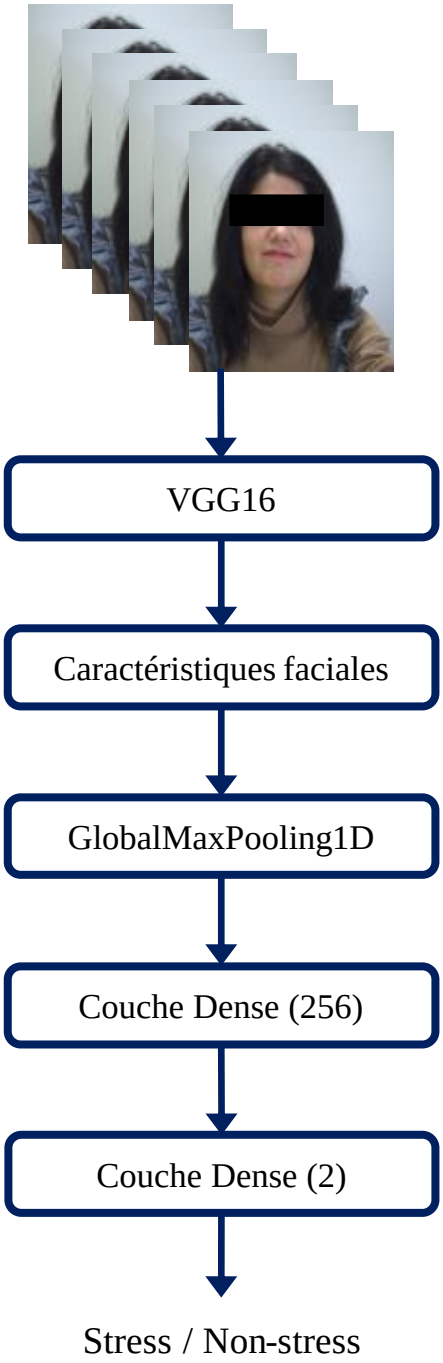


FIGURE 4.12 – Système de reconnaissance de l'état de stress basé sur les caractéristiques faciales.

de la concaténation des deux modalités est passé dans deux couches denses de 256 et 2 neurones respectivement. La première couche prend les unités ReLU comme unités cachées, tandis que la seconde utilise la fonction d'activation sigmoïde pour prédire la classe de stress correspondante, i.e. l'état de stress ou de non-stress.

La précision de la fusion des caractéristiques faciales avec les caractéristiques iVFC et avec

TABLE 4.9 – Résultats de la classification des états de stress et de non stress en utilisant les caractéristiques faciales uniquement et en les fusionnant avec les signaux physiologiques sans contact.

Caractéristique	**Précision (%)**
Caractéristiques faciales	82.48
Caractéristiques faciales + iPPG	83.12
Caractéristiques faciales + iVFC	**91.07**

les signaux iPPG sont présentées dans le tableau 4.9. La combinaison des caractéristiques faciales et les caractéristiques iVFC améliore considérablement le taux de reconnaissance et fournit une meilleure précision (91,07 %) que l'utilisation des caractéristiques faciales (82.48 %) ou iVFC (72.22 %) séparément. La fusion des caractéristiques faciales et des signaux iPPG améliore légèrement les performances, obtenant une précision de 83,12%.

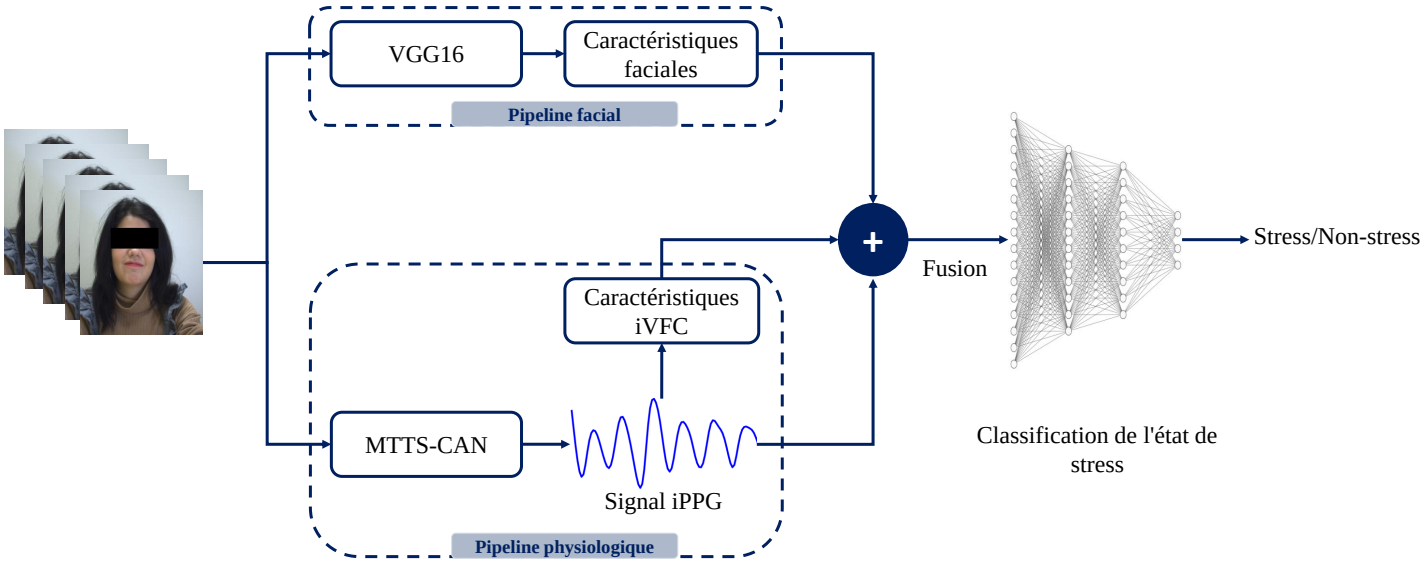


FIGURE 4.13 – Aperçu du système proposé pour la reconnaissance multimodale de l'état de stress à l'aide des caractéristiques faciales, des signaux iPPG et des caractéristiques iVFC. Il se compose de deux pipelines. Le premier (pipeline facial sur la figure) extrait les caractéristiques faciales à l'aide d'un réseau VGG16 pré-entraîné, tandis que le second (pipeline physiologique sur la figure) extrait les signaux physiologiques à l'aide du réseau MTTS-CAN. Ce dernier récupère le signal iPPG, à partir duquel les caractéristiques de la VFC peuvent être estimées. Les caractéristiques extraites de chaque modalité sont ensuite fusionnées et introduites à un réseau de neurones à propagation avant pour la classification stress/non stress.

4.3.4 Conclusion

Nous avons proposé dans cette section une approche multimodale pour la reconnaissance de l'état de stress basée sur la fusion des signaux physiologiques et des caractéristiques faciales. Contrairement aux travaux existants, les modalités utilisées sont receuillies à partir d'enregistrements vidéo du visage seulement. Les paramètres physiologiques sont mesurés à distance à l'aide de la technique iPPG, tandis que les caractéristiques faciales sont extraites par apprentissage par transfert. De cette manière, une seule source d'entrée a été utilisée pour extraire les caractéristiques de chaque modalité.

Les résultats comparatifs entre les systèmes unimodaux ont montré que les caractéristiques faciales sont plus pertinentes et permettent d'obtenir un plus haut niveau de précision. De plus, la fusion des caractéristiques faciales avec les signaux physiologiques a fourni une estimation plus précise, indiquant l'efficacité de l'approche multimodale.

4.4 Conclusion

Nous avons présenté au cours de ce chapitre un système bimodal basé sur la fusion physio-visuelle ainsi que les résultats expérimentaux pour la reconnaissance automatique de l'état affectif (émotions / stress). L'architecture proposée comprend deux pipelines, chacun dédié à l'extraction des caractéristiques de chaque modalité. Le pipeline visuel s'appuie sur un nouveau modèle neuronal profond (3D-SE-XceptionNet) qui combine l'architecture Xception avec le module Squeeze-Excitation. Quant au pipeline physiologique, nous avons adopté une méthode robuste et légère pour l'extraction du signal iPPG et afin d'estimer les caractéristiques de la variabilité cardiaque [368]. Une fusion des caractéristiques est ensuite effectuée pour prédir les états émotionnels ou le stress.

Nous avons dans un premier temps évalué les performances de l'approche proposée pour la reconnaissance des émotions. Les expérimentations ont été réalisées sur des données relatives à deux modèles d'annotation différents : l'annotation de la base de données BP4D+[379] suit une représentation catégorielle, tandis queque l'annotation de MANHOB-HCI [380] repose sur l'approche dimensionnelle et fournit des niveaux de valence et d'activation. Les résultats comparatifs révèlent une amélioration non négligeable de la précision après la fusion des expressions faciales et l'un des paramètres physiologiques (iPPG ou iVFC). En revanche, la combinaison des signaux iPPG et des caractérisques de la iVFC a entraîné une mauvaise précision, que nous associons à l'absence de corrélation entre les deux indices physiologiques. D'autre part, le taux de reconnaissance pour les systèmes unimodaux varie selon la base de données étudiée. Les expressions

faciales ont surpassé les signaux physiologiques sur BP4D+ tandis que les signaux physiologiques ont obtenu la meilleure précision par rapport aux expressions faciales sur la base de données MAHNOB-HCI. Nous associons cela aux défis présents dans la base de données BP4D+ qui sont liés principalement aux mouvements et à la couleur de peau des sujets. Cela a un impact direct sur la qualité des signaux iPPG extraits affectant ainsi la précision de reconnaissance en utilisant les signaux physiologiques. Une observation inverse se dégage de l'étude de la base de données MAHNOB-HCI. Les sujets sont installés de manière stable et leur couleur de peau est fortement biaisée vers les tons claires. Par conséquent, les signaux iPPG et les caractéristiques iVFC dérivées sont mieux extimés et les performances de reconnaissance sont de facto meilleures.

Dans un second temps, nous avons appliqué la même approche à la reconnaissance de l'état de stress. A la différence de la première étude, nous nous sommes appuyés sur une méthode d'apprentissage par transfert pour réduire les efforts de développement et surmonter le problème du manque de données. Les conclusions après analyse des résultats expérimentaux sont identiques à celles qui découlent de l'étude sur la reconnaissance des émotions. La fusion des caractéristiques faciales et des signaux physiologiques a amélioré la précision et permet d'atteindre le meilleur taux.

Les résultats obtenus dans ce chapitre montre l'efficacité de la fusion physio-visuelle par rapport aux approches unimodales. Nous avons également démontré la faisabilité d'utiliser une seule source de données, i.e. la caméra, pour extraire et fusionner des signaux de natures différentes, à savoir les caractéristiques faciales et les paramètres physiologiques.

Conclusion et perspectives

Partant de l'hypothèse initiale que les humains peuvent naturellement interprèter leurs émotions simplement en s'écoutant ou en regardant leur visage, les premiers travaux de la littérature scientifique sur la reconnaissance automatique de l'état affectif reposaient sur l'utilisation des expressions faciales et de la parole séparément. Au fil du temps, un large éventail d'algorithmes ont été proposés et d'autres modalités ont été explorées, telles que les signaux physiologiques et la gestuelle. Cependant, les expressions faciales ont été plus étudiées en raison de leur visibilité et de leur rôle majeur dans les interactions sociales. D'autre part, les signaux physiologiques sont aussi intéressants car ils offrent un potentiel pour une reconnaissance plus précise contrairement aux expressions faciales qui sont facilement contrefaites et plus affectées par les différences sociales et culturelles.

Dans les dernières années, le domaine de la reconnaissance unimodale d'émotions a atteint un stade de saturation conduisant à l'émergence de la fusion multimodale. L'analyse des émotions à partir d'une seule modalité peut être limitée car elle ne prend pas en compte les autres indices émotionnels qui peuvent être présents. Par conséquent, l'utilisation de plusieurs modalités est souvent recommandée pour améliorer la précision de la reconnaissance des émotions. En effet, la majorité de travaux de l'état de l'art se sont focalisés sur la fusion de deux ou plusieurs modalités afin d'améliorer les performances. Différentes combinaisons ont été étudiées mais la fusion des expressions faciales avec les signaux physiologiques est plus efficace en termes de précision et de fiabilité et elle permet d'exploiter les avantages de chaque modalité notamment pour surmonter le problème des émotions contrefaites. Dans ce sens, nous avons étudié dans cette thèse la fusion physio-visuelle pour la reconnaissance automatique de l'état affectif de la personne, y compris les émotions spontanées et le stress. A la différence des schémas de fusion existants, nous nous sommes basés sur une approche sans contact et mono-capteur en utilisant uniquement une seule source de données. Les caractéristiques visuelles et physiologiques sont extraites directement à partir des vidéos du visage. La mesure des données physiologiques par vidéo à l'aide de l'iPPG est plus pratique et confortable par rapport aux dispositifs en contact intrusifs qui peuvent interférer

avec le sujet et modifier son état émotionnel. Mis à part l'amélioration des performances et la fiabilité grâce à l'intégration des paramètres physiologiques, l'utilisation d'une caméra qui est intégrée dans tous les appareils numériques utilisés dans la vie quotidienne permet de réduire le coût et de rendre les approches plus accessibles.

Les contributions de cette thèse peuvent être classées en trois domaines : l'informatique affective, télé-santé, et l'apprentissage profond.

- En informatique affective : nous avons proposé un schéma de fusion physio-visuelle pour la reconnaissance de l'état affectif de la personne à partir de vidéos du visage, y compris les émotions et le stress. Cette approche présente plusieurs avantages par rapport aux systèmes existants car elle permet à la fois de surmonter le problème des émotions contrefaites et également améliorer les performances en recueillant en permanence des informations complémentaires sur l'état affectif de la personne. Cela est utile dans le cas d'acquisitions manquantes ou de données corrompues qui peuvent survenir lors de l'utilisation d'une seule modalité dans un environnement bruyant ou dans le cas d'une expression falsifiée.
- En télé-santé : nous avons développé une nouvelle approche bout-en-bout (X-iPPGNet) pour l'estimation sans contact de la fréquence cardiaque à partir d'enregistrements vidéo du visage en utilisant un réseau spatio-temporel profond. X-iPPGNet est un pipeline optimisé qui prédit la fréquence cardiaque en une courte durée (2 secondes) et sans extraction séparée du signal iPPG. Cela est particulièrement pertinent dans le cas des fréquences cardiaques élevées et fortement fluctuantes. Cette recherche ouvre plusieurs perspectives dans l'estimation des signaux physiologiques sans contact et peut servir de base à des futures architectures robustes dans des applications en temps réel car elle nécessite un nombre réduit de paramètres et un court fragment vidéo. De plus, les résultats expérimentaux surpassent de manière significative toutes les méthodes actuelles de l'état de l'art sur trois ensembles de données de référence.
- En apprentissage profond : nous avons proposé en premier lieu un réseau de neurones spatio-temporels basé sur le squelette de l'architecture Xception. Ce réseau repose sur le découplage des canaux de couleur et premet d'extraire des informations supplémentaires de chaque canal séparémment. Ensuite, nous avons amélioré l'architecture proposée en intégrant le module Squeeze-Excitation qui joue le role d'un mécanisme d'attention. Il vise à modéliser explicitement l'interdépendance entre les canaux de l'image afin de recalibrer les cartes de caractéristiques par canal d'une manière adaptative et efficace en termes de temps de calcul. Par ailleurs, des techniques avancées d'optimisation de l'apprentissage profond ainsi que des stratégies de régularisation sont utilisées pour surmonter les problèmes de

surajustement et améliorer la généralisation du modèle à de nouvelles données.

Les résultats des différentes études présentées dans ce travail sont prometteurs. Il est néanmoins important d'identifier les limites avant de proposer des améliorations et des perspectives de travaux futurs.

La principale limite de la première étude sur la mesure sans contact de la fréquence cardiaque réside dans l'aspect de mesure bout-en-bout qui ne permet pas d'extraire les caractéristiques de la variabilité cardiaque qui sont généralement dérivées à partir du signal iPPG. En revanche, une approche de transfer learning peut être effectuée pour estimer les caractéristiques cardiaques en utilisant les poids pré-entrainés de X-iPPGNet. Par ailleurs, l'amélioration des performances de la méthode proposée doit passer par la résolution du problème de données qui sont soit limitées soit corrompues. Il est donc nécessaire d'accroître la quantité de données d'apprentissage soit en utilisant des techniques avancées d'augmentation de données ou la génération de données synthétiques, soit en combinant plusieurs bases de données. En ce qui concerne les données corrompues qui sont dues à des problèmes techniques lors de l'acquisition, le pré-traitement et le filtrage des données est indispensable pour entraîner le modèle correctement.

Les résultats préliminaires de la deuxième étude sur la reconnaissance de l'état affectif par une fusion physio-visuelle sont très encourageants et démontrent la possibilité de prédire l'état émotionnel et du stress de la personne en combinant les caractéristiques faciales et physiologiques mesurées via la caméra uniquement. Cependant, une analyse approfondie est nécessaire pour examiner et explorer les différents facteurs impactant les performances afin de les améliorer. En termes de méthode, les résultats rapportés dans cette thèse ont été obtenus par la fusion des caractéristiques haut niveau uniquement. Toutefois, des schémas de fusion de caractéristiques à plusieurs niveaux peuvent être testés afin d'évaluer la relation entre le niveau de caractéristiques et les performances. D'autre part, le pipeline visuel pour l'extraction des caractéristiques faciales peut être comparé avec d'autres approches basées sur le transfer learning en utilisant des réseaux tels que OpenFace et VGGFace. En termes de facteurs environnementaux, nous proposons d'étudier l'impact du mouvement, de l'occultation et des conditions d'éclairage sur la précision.

Nous envisageons également d'optimiser l'architecture et les hyper-paramètres de notre modèle afin qu'il soit adapté pour des applications temps réel.

Une suite logique de ce projet de thèse consisterait à exploiter d'autres modalités pour développer un système affectif complet intégrant tous les signaux contenus dans la vidéo tels que la parole, le regard, la posture, la gestuelle, la température de la peau ainsi que d'autres signaux vitaux mesurables à distance via la caméra. Une autre perspective de ce travail serait d'optimiser l'architecture et les hyper-paramètres de notre modèle afin qu'il soit adapté pour des applica-

tions où les ressources sont limitées. Le développement d'une interface homme-machine peut être envisagé comme finalisation de ce projet.

Table des figures

Liste des tableaux

Glossaire

AC Composante variable du volume sanguin. 44

AU Action Units. 9

BF Basse fréquence. 41, 103, 107, 115, 131

BoW Sac de mots visuels. 15

bpm Battements par minute. 39

BSS Séparation aveugle de sources. 52

BVP Volume sanguin. 113

CNN Réseau de neurones convolutif. 17, 19, 24

cVFC Variabilité de la fréquence cardiaque en contact. 108, 114

DC Composante continue du volume sanguin. 44

DNN Réseau de neurones profond. 15

DSC Convolution séparable en profondeur. 71

DSP Densité spectrale de puissance. 41, 103

DT Arbres de décision. 22

ECG Électrocardiographie. 11, 21, 38

EDA Activité électrodermale. 11

EEG Électroencéphalographie. 11, 20

EMD Décomposition en modes empiriques. 20, 22

EMG Électromyographie. 12, 35

EOG Électrooculographie. 24, 35

EVM Magnification vidéo eulérienne. 76

FACS Facial Action Coding System. 9

FC Fréquence cardiaque. 11, 38

FCN Réseau entièrement connecté. 23

FFT Transformation de Fourier rapide. 20

FMI Fonction de mode intrinsèque. 20

fps Nombre de trames par seconde. 115

GRU Réseau de neurones récurrents à portes. 19

HF Haute fréquence. 41, 103, 107, 115, 131

HOG Histogramme de gradient orienté. 14

IBI Intervalle de temps entre deux pics de battements cardiaques. 40, 103, 131

ICA Analyse en composantes indépendantes. 53

iPPG Photopléthysmographie par imagerie. xx, 108

iVFC Variabilité de la fréquence cardiaque mesurée par l'iPPG. 107, 108

kNN k plus proches voisins. 22, 117

LBP Motifs binaires locaux. 14

LDA Analyse discriminante linéaire. 22

LED Diodes électro-luminescente. 44

LFPC Coefficients de puissance log-fréquence à court terme. 34

LSTM Eéseau de neurones récurrent à mémoire court et long terme. 19

MAE Erreur aboslue moyenne. 79

maxFC Fréquence cardiaque maximale. 115

meanFC Moyenne des fréquences cardiaques. 102, 107, 115

MFCC Coefficients cepstraux de fréquence Mel. 33

minFC Fréquence cardiaque minimale. 115

MSE Erreur quadratique moyenne. 75

NB Modèle de Bayes naïf. 22, 117

PCA Analyse en composantes principales. 53

POS Plan orthogonal à la peau. 53

PPG Photopléthysmographie. 11, 21, 38

RAdam Optimiseur Adam réctifié. 74

RESP Respiration. 22

RF Forêts aléatoires. 117

RMSE Racine carrée de l'erreur quadratique moyenne. 79

RMSSD Moyenne quadratique des IBI successifs. 40, 102, 107, 115

RNN Réseau de neurones récurrent. 19

ROI Région d'intérêt. 69

rPPG remote photopléthysmographie. 44

SAM Mannequin d'auto-évaluation. 8

SAN Nœud sino-atrial. 38

SD Ecart-type. 79

SE Squeeze-Excitation. 98

SGD Descente de gradient stochastique. 74

SIFT Transformation de caractéristiques visuelles invariantes à l'échelle. 14

SNA Système nerveux autonome. 2, 38

SNP Système nerveux parasympathique. 40

SNS Système nerveux sympathique. 40

SSR Rotation du sous-espace spatial. 54

stdFC Ecart-type des fréquences cardiaques. 102, 107, 115

STFT Transformée de Fourier fenêtrée. 20

SVM Séparateur à vaste marge. 15, 20, 22, 117

TBF Très basse fréquence. 41, 103, 131

UBF Ultra basse fréquence. 41

VAS Visual Analog Scale. 8

VFC Variabilité de la fréquence cardiaque. 11, 40, 103, 107

Bibliographie

[1] M. Egger, M. Ley, and S. Hanke, "Emotion recognition from physiological signal analysis : A review," *Electronic Notes in Theoretical Computer Science*, vol. 343, pp. 35–55, 2019. The proceedings of AmI, the 2018 European Conference on Ambient Intelligence.

[2] P. R. J. Kleinginna and A. M. Kleinginna, "A categorized list of emotion definitions, with suggestions for a consensual definition," *Motivation and Emotion*, vol. 5, pp. 345–379, 1981.

[3] D. Behm, J. Whittle, D. Button, and K. Power, "Intermuscle differences in activation," *Muscle nerve*, vol. 25, pp. 236–43, 02 2002.

[4] P. Ekman, "Basic emotions," *Handbook of cognition and emotion*, vol. 98, no. 45-60, p. 16.

[5] R. Picard, "Affective computing," 1995.

[6] M. J. (), *Dictionnaire Larousse français - anglais, anglais - français = = [Larousse French - English, English - French dictionary] / Larousse ; [coordination générale pour cette édition Janice McNeillie]*. Paris : Larousse, [nouvelle édition 2007] ed., impr. 2007.

[7] M. W. Eysenck, A. W. Ellis, E. B. Hunt, and P. N. Johnson-Laird, "The blackwell dictionary of cognitive psychology," 1990.

[8] W. B. Cannon, "The james-lange theory of emotions : A critical examination and an alternative theory," *The American Journal of Psychology*, vol. 39, no. 1/4, pp. 106–124, 1927.

[9] W. James, "What is an emotion ?," *Mind*, vol. 9, no. 34, pp. 188–205, 1884.

[10] S. Schachter and J. E. Singer, "Cognitive, social, and physiological determinants of emotional state.," *Psychological review*, vol. 69, pp. 379–99, 1962.

[11] M. Cabanac de Lafregeyre, "What is emotion ?," *Behavioural processes*, vol. 60, pp. 69–83, 12 2002.

[12] J. Kumar and J. Kumar, "Machine learning approach to classify emotions using gsr," *Advanced Research in Electrical and Electronic Engineering*, vol. 2, no. 12, pp. 72–76, 2015.

[13] F. M. P. del Arco, M. T. M. Valdivia, L. A. U. López, and R. Mitkov, "Improved emotion recognition in spanish social media through incorporation of lexical knowledge," *Future Gener. Comput. Syst.*, vol. 110, pp. 1000–1008, 2020.

[14] R. Descartes and S. H. Voss, "The passions of the soul," 1989.

[15] B. de Spinoza, "Ethics part iii . on the origin and nature of the emotions," 2001.

[16] C. DARWIN, "The expression of the emotions in man and animals,"

[17] P. Bard, "A diencephalic mechanism for the expression of rage with special reference to the sympathetic nervous system," *American Journal of Physiology-Legacy Content*, vol. 84, no. 3, pp. 490–515, 1928.

[18] R. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence : analysis of affective physiological state," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[19] D. Västfjäll, "Emotion induction through music : A review of the musical mood induction procedure," *Musicae Scientiae*, vol. 5, pp. 173 – 211, 2001.

[20] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition & Emotion*, vol. 9, pp. 87–108, 1995.

[21] M. K. Uhrig, N. Trautmann, U. Baumgärtner, R.-D. Treede, F. Henrich, W. Hiller, and S. Marschall, "Emotion elicitation : A comparison of pictures and films," *Frontiers in psychology*, vol. 7, p. 180, 2016.

[22] J. H. Janssen, P. Tacken, J. de Vries, E. L. van den Broek, J. H. Westerink, P. Haselager, and W. A. IJsselsteijn, "Machines outperform laypersons in recognizing emotions elicited by autobiographical recollection," *Human–Computer Interaction*, vol. 28, no. 6, pp. 479–517, 2013.

[23] D. Adolph and B. M. Pause, "Different time course of emotion regulation towards odors and pictures : Are odors more potent than pictures ?," *Biological psychology*, vol. 91, no. 1, pp. 65–73, 2012.

[24] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, "K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations," *Scientific Data*, vol. 7, no. 1, pp. 1–16, 2020.

[25] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on intelligent transportation systems*, vol. 6, no. 2, pp. 156–166, 2005.

[26] H. Shahid, A. Butt, S. Aziz, M. U. Khan, and S. Z. H. Naqvi, "Emotion recognition system featuring a fusion of electrocardiogram and photoplethysmogram features," in *2020 14th International Conference on Open Source Systems and Technologies (ICOSST)*, pp. 1–6, IEEE, 2020.

[27] C. E. Izard, "Basic emotions, natural kinds, emotion schemas, and a new paradigm," *Perspectives on Psychological Science*, vol. 2, pp. 260 – 280, 2007.

[28] W. M. Wundt, *Grundriss der psychologie*. A. Kröner, 1913.

[29] A. Mehrabian, "Comparison of the pad and panas as models for describing emotions and for differentiating anxiety from depression," *Journal of Psychopathology and Behavioral Assessment*, vol. 19, pp. 331–357, 1997.

[30] F. Abdat, "Reconnaissance automatique des émotions par données multimodales : expressions faciales et signaux physiologiques," *Université de Metz, France*, 2010.

[31] S. Tomkins, *Affect imagery consciousness : Volume I : The positive affects*. Springer publishing company, 1962.

[32] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of emotion*, pp. 3–33, Elsevier, 1980.

[33] W. contributors, "Fichier :Plutchik-wheel fr.svg — Wikipédia," 11 2013.

[34] A. Mehrabian, "Communication without words," 1968.

[35] M. M. Bradley and P. J. Lang, "Measuring emotion : the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[36] N. Crichton, "Visual analogue scale (vas)," *J Clin Nurs*, vol. 10, no. 5, pp. 706–6, 2001.

[37] C. Bertheaux, *Le rôle de l'émotion dans un processus de conception sensorielle*. PhD thesis, Université de Lyon, 2020.

[38] R. Adolphs, D. Tranel, S. Hamann, A. W. Young, A. J. Calder, E. A. Phelps, A. Anderson, G. P. Lee, and A. R. Damasio, "Recognition of facial emotion in nine individuals with bilateral amygdala damage," *Neuropsychologia*, vol. 37, pp. 1111–1117, 1999.

[39] B. de Gelder, "Why bodies ? twelve reasons for including bodily expressions in affective neuroscience," *Philosophical Transactions of the Royal Society B*, vol. 364, pp. 3475–3484, 2009.

[40] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[41] P. Ekman, W. V. Friesen, and P. C. Ellsworth, "Emotion in the human face : Guidelines for research and an integration of findings," 1972.

[42] P. Ekman, "Emotions revealed : recognizing faces and feelings to improve communication and emotional life. new york," *NY : Times books*, 2003.

[43] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition : History, trends, and affect-related applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

[44] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.

[45] B. Bhushan, "Study of facial micro-expressions in psychology," in *Understanding facial expressions in communication*, pp. 265–286, Springer, 2015.

[46] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu, "How fast are the leaked facial expressions : The duration of micro-expressions," *Journal of Nonverbal Behavior*, vol. 37, no. 4, pp. 217–230, 2013.

[47] P. Eckman, "Telling lies : Clues to deceit in the marketplace, politics, and marriage," 1985.

[48] S. Porter and L. Ten Brinke, "Reading between the lies : Identifying concealed and falsified emotions in universal facial expressions," *Psychological science*, vol. 19, no. 5, pp. 508–514, 2008.

[49] M. Frank, M. Herbasz, K. Sinuk, A. Keller, and C. Nolan, "I see how you feel : Training laypeople and professionals to recognize fleeting emotions," in *The Annual Meeting of the International Communication Association. Sheraton New York, New York City*, pp. 1–35, 2009.

[50] S. Nag, A. K. Bhunia, A. Konwer, and P. P. Roy, "Facial micro-expression spotting and recognition using time contrasted feature with visual memory," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2022–2026, IEEE, 2019.

[51] R. W. Levenson, "Emotion and the autonomic nervous system : A prospectus for research on autonomic specificity.," 1988.

[52] J. Wagner, J. Kim, and E. André, "From physiological signals to emotions : Implementing and comparing selected methods for feature extraction and classification," *2005 IEEE International Conference on Multimedia and Expo*, pp. 940–943, 2005.

[53] N. Simonazzi, *Reconnaissance d'états émotionnels à partir des interactions avec un smartphone : Conception des méthodes et outils pour le domaine de la relation client.* PhD thesis, Bordeaux, 2021.

[54] A. Horvers, N. Tombeng, T. Bosse, A. W. Lazonder, and I. Molenaar, "Detecting emotions through electrodermal activity in learning contexts : A systematic review," *Sensors*, vol. 21, no. 23, p. 7869, 2021.

[55] Y. Benezeth, P. Li, R. Macwan, K. Nakamura, R. Gomez, and F. Yang, "Remote heart rate variability for emotional state monitoring," in *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pp. 153–156, 2018.

[56] A. Woyczyk, S. Rasche, and S. Zaunseder, "Impact of sympathetic activation in imaging photoplethysmography," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[57] M. B. Akçay and K. Oğuz, "Speech emotion recognition : Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[58] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition : a review," *International Journal of Speech Technology*, vol. 21, pp. 93–120, 2018.

[59] C. Whissell, "The dictionary of affect in language," 1989.

[60] R. Plutchik, "Emotion, a psychoevolutionary synthesis," 1980.

[61] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods : Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[62] A. Austermann, N. Esau, L. Kleinjohann, and B. Kleinjohann, "Prosody based emotion recognition for mexi," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1138–1144, 2005.

[63] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar, "Combining prosodic lexical and cepstral systems for deceptive speech detection," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, pp. I–I, 2006.

[64] K. R. Scherer, "Vocal affect expression : a review and a model for future research.," *Psychological bulletin*, vol. 99, no. 2, p. 143, 1986.

[65] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, "Emotion recognition using facial expressions," *Procedia Computer Science*, vol. 108, pp. 1175–1184, 2017. International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.

[66] K. Venkataramanan and H. R. Rajamohan, "Emotion recognition from speech," *arXiv preprint arXiv :1912.10458*, 2019.

[67] F. Noroozi, C. A. Corneanu, D. Kamińska, T. Sapiński, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE transactions on affective computing*, vol. 12, no. 2, pp. 505–523, 2018.

[68] J. Kumari, R. Rajesh, and K. Pooja, "Facial expression recognition : A survey," *Procedia Computer Science*, vol. 58, pp. 486–491, 2015. Second International Symposium on Computer Vision and the Internet (VisionNet'15).

[69] S. Li and W. Deng, "Deep facial expression recognition : A survey," *IEEE transactions on affective computing*, 2020.

[70] P. V. Rouast, M. T. P. Adam, and R. Chiong, "Deep learning for human affect recognition : Insights and new developments," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 524–543, 2021.

[71] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, pp. I–I, 2001.

[72] D. E. King, "Dlib-ml : A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, 2009.

[73] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, pp. 1499–1503, 2016.

[74] C. Shan, S. Gong, and P. McOwan, "Robust facial expression recognition using local binary patterns," in *IEEE International Conference on Image Processing 2005*, vol. 2, pp. II–370, 2005.

[75] L. Greche, M. Akil, R. Kachouri, and N. Es-Sbai, "A new pipeline for the recognition of universal expressions of multiple faces in a video sequence," *Journal of Real-Time Image Processing*, vol. 17, pp. 1389–1402, 2020.

[76] K. Verma and A. Khunteta, "Facial expression recognition using gabor filter and multi-layer artificial neural network," *2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC)*, pp. 1–5, 2017.

[77] S. Berretti, B. Ben Amor, M. Daoudi, and A. del Bimbo, "3D facial expression recognition using SIFT descriptors of automatically detected keypoints," *The Visual Computer*, vol. 27, pp. 1021–1036, June 2011.

[78] R. T. Ionescu and C. Grozea, "Local learning to improve bag of visual words model for facial expression recognition," 2013.

[79] Y. Wang, H. Yu, B. Stevens, and H. Liu, "Dynamic facial expression recognition using local patch and lbp-top," in *2015 8th International Conference on Human System Interaction (HSI)*, pp. 362–367, 2015.

[80] M. Abdulrahman and A. Eleyan, "Facial expression recognition using support vector machines," in *2015 23nd Signal Processing and Communications Applications Conference (SIU)*, pp. 276–279, 2015.

[81] K. Tiwari and M. Patel, "Facial expression recognition using random forest classifier," in *International Conference on Artificial Intelligence : Advances and Applications 2019*, pp. 121–130, Springer, 2020.

[82] G. Ramkumar and E. Logashanmugam, "An effectual facial expression recognition using hmm," in *2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*, pp. 12–15, 2016.

[83] P. Carcagnì, M. Del Coco, M. Leo, and C. Distante, "Facial expression recognition and histograms of oriented gradients : a comprehensive study," *SpringerPlus*, vol. 4, no. 1, pp. 1–25, 2015.

[84] A. Hernandez-Matamoros, A. Bonarini, E. Escamilla-Hernandez, M. Nakano-Miyatake, and H. Perez-Meana, "Facial expression recognition with automatic segmentation of face regions using a fuzzy based classification approach," *Knowledge-Based Systems*, vol. 110, pp. 1–14, 2016.

[85] S. K. A. Kamarol, M. H. Jaward, H. Kälviäinen, J. Parkkinen, and R. Parthiban, "Joint facial expression recognition and intensity estimation based on weighted votes of image sequences," *Pattern Recognition Letters*, vol. 92, pp. 25–32, 2017.

[86] J. R. Lee, L. Wang, and A. Wong, "Emotionnet nano : An efficient deep convolutional neural network design for real-time facial expression recognition," *Frontiers in Artificial Intelligence*, vol. 3, 2021.

[87] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1749–1756, 2014.

[88] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[89] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv :1409.1556*, 2014.

[90] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[91] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[92] T. Chang, G. Wen, Y. Hu, and J. Ma, "Facial expression recognition based on complexity perception classification algorithm," *arXiv preprint arXiv :1803.00185*, 2018.

[93] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter conference on applications of computer vision (WACV)*, pp. 1–10, IEEE, 2016.

[94] H. Hardjadinata, R. S. Oetama, and I. Prasetiawan, "Facial expression recognition using xception and densenet architecture," in *2021 6th International Conference on New Media Studies (CONMEDIA)*, pp. 60–65, 2021.

[95] H.-D. Nguyen, S. Yeom, G.-S. Lee, H.-J. Yang, I.-S. Na, and S.-H. Kim, "Facial emotion recognition using an ensemble of multi-level convolutional neural networks," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 33, no. 11, p. 1940015, 2019.

[96] O. Arriaga, M. Valdenegro-Toro, and P.-G. Plöger, "Real-time convolutional neural networks for emotion and gender classification," *ArXiv*, vol. abs/1710.07557, 2019.

[97] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[98] J. Shao and Y. Qian, "Three convolutional neural network models for facial expression recognition in the wild," *Neurocomputing*, vol. 355, pp. 82–92, 2019.

[99] Y. Fan, V. O. Li, and J. C. Lam, "Facial expression recognition with deeply-supervised attention network," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 1057–1071, 2022.

[100] W. Sato, T. Kochiyama, and S. Uono, "Spatiotemporal neural network dynamics for the processing of dynamic facial expressions," *Scientific reports*, vol. 5, no. 1, pp. 1–13, 2015.

[101] J. Haddad, O. Lézoray, and P. Hamel, "3d-cnn for facial emotion recognition in videos," in *International symposium on visual computing*, pp. 298–309, Springer, 2020.

[102] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, "Hybrid deep neural networks for face emotion recognition," *Pattern Recognition Letters*, vol. 115, pp. 101–106, 2018.

[103] T.-H. S. Li, P.-H. Kuo, T.-N. Tsai, and P.-C. Luan, "Cnn and lstm based facial expression analysis model for a humanoid robot," *IEEE Access*, vol. 7, pp. 93998–94011, 2019.

[104] K. Kang and X. Ma, "Convolutional gate recurrent unit for video facial expression recognition in the wild," in *2019 Chinese Control Conference (CCC)*, pp. 7623–7628, IEEE, 2019.

[105] S. Ouellet, "Real-time emotion recognition for gaming using deep convolutional network features," *arXiv preprint arXiv :1408.3750*, 2014.

[106] J. Li and E. Y. Lam, "Facial expression recognition using deep neural networks," in *2015 IEEE International Conference on Imaging Systems and Techniques (IST)*, pp. 1–6, IEEE, 2015.

[107] M. Nasri, M. A. Hmani, A. Mtibaa, D. Petrovska-Delacretaz, M. B. Slima, and A. B. Hamida, "Face emotion recognition from static image based on convolution neural networks," in *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pp. 1–6, IEEE, 2020.

[108] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3d convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 30–40, 2017.

[109] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1805–1812, 2014.

[110] D. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel convolutional neural network," in *2015 international joint conference on neural networks (IJCNN)*, pp. 1–8, IEEE, 2015.

[111] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2168–2177, 2018.

[112] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion : Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, p. 3046, 2021.

[113] P. Giannopoulos, I. Perikos, and I. Hatzilygeroudis, "Deep learning approaches for facial emotion recognition : A case study on fer-2013," *Advances in Hybridization of Intelligent Methods : Models, Systems and Applications*, pp. 1–16, 2018.

[114] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE international conference on computer vision*, pp. 2983–2991, 2015.

[115] D. Liu, X. Ouyang, S. Xu, P. Zhou, K. He, and S. Wen, "Saanet : Siamese action-units attention network for improving dynamic facial expression recognition," *Neurocomputing*, vol. 413, pp. 145–157, 2020.

[116] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, Y. Zong, and N. Sun, "Multi-clue fusion for emotion recognition in the wild," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 458–463, 2016.

[117] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 569–576, 2017.

[118] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks

for emotion recognition in video," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pp. 467–474, 2015.

[119] D. H. Kim, W. J. Baddar, J. Jang, and Y. M. Ro, "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 10, pp. 223–236, 2019.

[120] K. Kang and X. Ma, "Convolutional gate recurrent unit for video facial expression recognition in the wild," in *2019 Chinese Control Conference (CCC)*, pp. 7623–7628, 2019.

[121] ZhaoLixin, "A facial expression recognition method using two-stream convolutional networks in natural scenes," *Journal of Information Processing Systems*, vol. 17, no. 2, pp. 399–410.

[122] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2983–2991, 2015.

[123] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, 2018.

[124] S. Wioleta, "Using physiological signals for emotion recognition," in *2013 6th international conference on human system interactions (HSI)*, pp. 556–561, IEEE, 2013.

[125] X. Niu, L. Chen, H. Xie, Q. Chen, and H. Li, "Emotion pattern recognition using physiological signals," *Sensors & Transducers*, vol. 172, no. 6, p. 147, 2014.

[126] S. Huynh, S. Kim, J. Ko, R. K. Balan, and Y. Lee, "Engagemon : Multi-modal engagement sensing for mobile games," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 2, no. 1, pp. 1–27, 2018.

[127] M. Murugappan, R. Nagarajan, and S. Yaacob, "Combining spatial filtering and wavelet transform for classifying human emotions using eeg signals," *Journal of Medical and Biological Engineering*, vol. 31, no. 1, pp. 45–51, 2011.

[128] Z. Mohammadi, J. Frounchi, and M. Amiri, "Wavelet-based emotion recognition system using eeg signal," *Neural Computing and Applications*, vol. 28, no. 8, pp. 1985–1990, 2017.

[129] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from brain signals using hybrid adaptive filtering and higher order crossings analysis," *IEEE Transactions on affective computing*, vol. 1, no. 2, pp. 81–97, 2010.

[130] K. Ishino and M. Hagiwara, "A feeling estimation system using a simple electroencephalograph," in *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483)*, vol. 5, pp. 4204–4209, IEEE, 2003.

[131] K. Schaaff, "Eeg-based emotion recognition," *Universitat Karlsruhe (TH)*, 2008.

[132] G. Chanel, J. J. Kierkels, M. Soleymani, and T. Pun, "Short-term emotion assessment in a recall paradigm," *International Journal of Human-Computer Studies*, vol. 67, no. 8, pp. 607–627, 2009.

[133] N. Zhuang, Y. Zeng, L. Tong, C. Zhang, H. Zhang, and B. Yan, "Emotion recognition from eeg signals using multidimensional information in emd domain," *BioMed research international*, vol. 2017, 2017.

[134] A. Samara, M. L. R. Menezes, and L. Galway, "Feature extraction for emotion recognition and modelling using neurophysiological data," in *2016 15th international conference on ubiquitous computing and communications and 2016 international symposium on cyberspace and security (IUCC-CSS)*, pp. 138–144, IEEE, 2016.

[135] A. M. Bhatti, M. Majid, S. M. Anwar, and B. Khan, "Human emotion recognition and analysis in response to audio music using brain signals," *Computers in Human Behavior*, vol. 65, pp. 267–275, 2016.

[136] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, "Ecg pattern analysis for emotion detection," *IEEE Transactions on affective computing*, vol. 3, no. 1, pp. 102–115, 2011.

[137] H.-W. Guo, Y.-S. Huang, C.-H. Lin, J.-C. Chien, K. Haraikawa, and J.-S. Shieh, "Heart rate variability signal features for emotion recognition by using principal component analysis and support vectors machine," in *2016 IEEE 16th international conference on bioinformatics and bioengineering (BIBE)*, pp. 274–277, IEEE, 2016.

[138] S. N. M. S. Ismail, N. A. A. Aziz, and S. Z. Ibrahim, "A comparison of emotion recognition system using electrocardiogram (ecg) and photoplethysmogram (ppg)," *Journal of King Saud University-Computer and Information Sciences*, 2022.

[139] B. Cheng and G. Liu, "Emotion recognition from surface emg signal using wavelet transform and neural network," in *2008 2nd International Conference on Bioinformatics and Biomedical Engineering*, pp. 1363–1366, IEEE, 2008.

[140] C.-K. Wu, P.-C. Chung, and C.-J. Wang, "Representative segment-based emotion analysis and classification with automatic respiration signal segmentation," *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 482–495, 2012.

[141] P. Gong, H. T. Ma, and Y. Wang, "Emotion recognition based on the multiple physiological signals," in *2016 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, pp. 140–143, IEEE, 2016.

[142] Z. Guendil, Z. Lachiri, C. Maaoui, and A. Pruski, "Emotion recognition from physiological signals using fusion of wavelet based features," in *2015 7th International Conference on Modelling, Identification and Control (ICMIC)*, (Sousse, Tunisia), IEEE, Dec. 2015.

[143] C. L. Lisetti and F. Nasoz, "Using noninvasive wearable computers to recognize human emotions from physiological signals," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 11, pp. 1–16, 2004.

[144] L. Xun and G. Zheng, "Ecg signal feature selection for emotion recognition," *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 11, no. 3, pp. 1363–1370, 2013.

[145] A. Goshvarpour and A. Goshvarpour, "Poincaré's section analysis for ppg-based automatic emotion recognition," *Chaos, Solitons & Fractals*, vol. 114, pp. 400–407, 2018.

[146] H. Ferdinando, T. Seppänen, and E. Alasaarela, "Comparing features from ecg pattern and hrv analysis for emotion recognition system," in *2016 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pp. 1–6, IEEE, 2016.

[147] R. Rakshit, V. R. Reddy, and P. Deshpande, "Emotion detection and recognition using hrv features derived from photoplethysmogram signals," in *Proceedings of the 2nd workshop on Emotion Representations and Modelling for Companion Systems*, pp. 1–6, 2016.

[148] A. Sepúlveda, F. Castillo, C. Palma, and M. Rodriguez-Fernandez, "Emotion recognition from ecg signals using wavelet scattering and machine learning," *Applied Sciences*, vol. 11, no. 11, p. 4945, 2021.

[149] S. Jerritta, M. Murugappan, K. Wan, and S. Yaacob, "Electrocardiogram-based emotion recognition system using empirical mode decomposition and discrete fourier transform," *Expert Systems*, vol. 2, no. 31, pp. 110–120, 2014.

[150] J. Wagner, "Augsburg biosignal toolbox (aubt)," *University of Augsburg*, 2005.

[151] M. Soleymani, F. Villaro-Dixon, T. Pun, and G. Chanel, "Toolbox for emotional feature extraction from physiological signals (teap)," *Frontiers in ICT*, vol. 4, p. 1, 2017.

[152] S. Zhao, A. Gholaminejad, G. Ding, Y. Gao, J. Han, and K. Keutzer, "Personalized emotion recognition by personality-aware high-order learning of physiological signals," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 1s, pp. 1–18, 2019.

[153] J. A. Dominguez-Jimenez, K. C. Campo-Landines, J. C. M. Santos, E. J. Delahoz, and S. H. C. Ortiz, "A machine learning model for emotion recognition from physiological signals," *Biomed. Signal Process. Control.*, vol. 55, 2020.

[154] M. B. H. Wiem and Z. Lachiri, "Emotion classification in arousal valence model using mahnob-hci database," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 3, 2017.

[155] W. Lin, C. Li, and S. Sun, "Deep convolutional neural network for emotion recognition using eeg and peripheral physiological signal," in *International conference on image and graphics*, pp. 385–394, Springer, 2017.

[156] N. Liu, Y. Fang, L. Li, L. Hou, F. Yang, and Y. Guo, "Multiple feature fusion for automatic emotion recognition using eeg signals," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 896–900, IEEE, 2018.

[157] Y.-H. Kwon, S.-B. Shin, and S.-D. Kim, "Electroencephalography based fusion two-dimensional (2d)-convolution neural networks (cnn) model for emotion recognition system," *Sensors*, vol. 18, no. 5, p. 1383, 2018.

[158] E. S. Salama, R. A. El-Khoribi, M. E. Shoman, and M. A. W. Shalaby, "Eeg-based emotion recognition using 3d convolutional neural networks," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 8, 2018.

[159] Y. Wang, Z. Huang, B. McCane, and P. Neo, "Emotionet : A 3-d convolutional neural network for eeg-based emotion recognition," in *2018 international joint conference on neural networks (IJCNN)*, pp. 1–7, IEEE, 2018.

[160] Y. Yang, Q. Wu, M. Qiu, Y. Wang, and X. Chen, "Emotion recognition from multi-channel eeg through parallel convolutional recurrent neural network," in *2018 international joint conference on neural networks (IJCNN)*, pp. 1–7, IEEE, 2018.

[161] X. Li, D. Song, P. Zhang, G. Yu, Y. Hou, and B. Hu, "Emotion recognition from multi-channel eeg data through convolutional recurrent neural network," in *2016 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pp. 352–359, IEEE, 2016.

[162] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words : Transformers for image recognition at scale," *arXiv preprint arXiv :2010.11929*, 2020.

[163] A. Arjun, A. S. Rajpoot, and M. R. Panicker, "Introducing attention mechanism for eeg signals : Emotion recognition with vision transformers," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 5723–5726, IEEE, 2021.

[164] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-Gonzalez, E. Abdulhay, and N. Arunkumar, "Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos)," *IEEE Access*, vol. 7, pp. 57–67, 2018.

[165] R. Harper and J. Southern, "A bayesian deep learning framework for end-to-end prediction of emotion from heartbeat," *IEEE transactions on affective computing*, 2020.

[166] S. Siddharth, T.-P. Jung, and T. J. Sejnowski, "Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing," *IEEE Transactions on Affective Computing*, 2019.

[167] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine : theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006.

[168] S.-W. Wang and S.-N. Yu, "Emotion recognition based on photoplethysmography using resnet and bilstm networks," in *2021 International Conference on e-Health and Bioengineering (EHB)*, pp. 1–4, IEEE, 2021.

[169] P. Sarkar and A. Etemad, "Self-supervised ecg representation learning for emotion recognition," *IEEE Transactions on Affective Computing*, 2020.

[170] T. Luguev, D. Seuß, and J.-U. Garbas, "Deep learning based affective sensing with remote photoplethysmography," in *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–4, 2020.

[171] Q. Zhang, X. Chen, Q. Zhan, T. Yang, and S. Xia, "Respiration-based emotion recognition with deep learning," *Computers in Industry*, vol. 92, pp. 84–90, 2017.

[172] N. Ganapathy, Y. R. Veeranki, H. Kumar, and R. Swaminathan, "Emotion recognition using electrodermal activity signals and multiscale deep convolutional neural network," *Journal of Medical Systems*, vol. 45, no. 4, pp. 1–10, 2021.

[173] M.-S. Lee, Y. R. Cho, Y. K. Lee, D. S. Pae, M. T. Lim, and T. K. Kang, "Ppg and emg based emotion recognition using convolutional neural network.," in *ICINCO (1)*, pp. 595–600, 2019.

[174] B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire, and V. Chandran, "Automatic emotion recognition using temporal multimodal deep learning," *IEEE Access*, vol. 8, pp. 225463–225474, 2020.

[175] S. VIJAYAKUMAR, R. FLYNN, P. CORCORAN, and N. MURRAY, "Cnn-based emotion recognition from multimodal peripheral physiological signals,"

[176] W. Tao, C. Li, R. Song, J. Cheng, Y. Liu, F. Wan, and X. Chen, "Eeg-based emotion recognition via channel-wise attention and self attention," *IEEE Transactions on Affective Computing*, 2020.

[177] S. Tripathi, S. Acharya, R. Sharma, S. Mittal, and S. Bhattacharya, "Using deep and convolutional neural networks for accurate emotion classification on deap data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, pp. 4746–4752, 2017.

[178] W. Liu, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal deep learning," in *Neural Information Processing : 23rd International Conference, ICONIP 2016, Kyoto, Japan, October 16–21, 2016, Proceedings, Part II 23*, pp. 521–529, Springer, 2016.

[179] Z. Jia, Y. Lin, J. Wang, Z. Feng, X. Xie, and C. Chen, "Hetemotionnet : two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1047–1056, 2021.

[180] P. Kawde and G. K. Verma, "Multimodal affect recognition in v-a-d space using deep learning," *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, pp. 890–895, 2017.

[181] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal residual lstm network," in *Proceedings of the 27th ACM international conference on multimedia*, pp. 176–183, 2019.

[182] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques : A tutorial and review," *Inf. Fusion*, vol. 59, pp. 103–126, 2020.

[183] M. K. Abadi, J. Staiano, A. Cappelletti, M. Zancanaro, and N. Sebe, "Multimodal engagement classification for affective cinema," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 411–416, IEEE, 2013.

[184] S. Koelstra and I. Patras, "Fusion of facial expressions and eeg for implicit affective tagging," *Image and Vision Computing*, vol. 31, no. 2, pp. 164–174, 2013.

[185] X. Huang, J. Kortelainen, G. Zhao, X. Li, A. Moilanen, T. Seppänen, and M. Pietikäinen, "Multimodal emotion analysis from facial expressions and electroencephalogram," *Computer Vision and Image Understanding*, vol. 147, pp. 114–124, 2016.

[186] Y. Huang, J. Yang, S. Liu, and J. Pan, "Combining facial expressions and electroencephalography to enhance emotion recognition," *Future Internet*, vol. 11, no. 5, p. 105, 2019.

[187] D. Li, Z. Wang, C. Wang, S. Liu, W. Chi, E. Dong, X. Song, Q. Gao, and Y. Song, "The fusion of electroencephalography and facial expression for continuous emotion recognition," *IEEE Access*, vol. 7, pp. 155724–155736, 2019.

[188] Y. Yang, Q. Gao, Y. Song, X. Song, Z. Mao, and J. Liu, "Investigating of deaf emotion cognition pattern by eeg and facial expression combination," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 2, pp. 589–599, 2021.

[189] S. Tivatansakul and M. Ohkura, "Emotion recognition using ecg signals with local pattern description methods," *International Journal of Affective Engineering*, vol. 15, no. 2, pp. 51–61, 2016.

[190] J. Yan, B. Wang, and R. Liang, "A novel bimodal emotion database from physiological signals and facial expression," *IEICE TRANSACTIONS on Information and Systems*, vol. 101, no. 7, pp. 1976–1979, 2018.

[191] W. Yu, S. Ding, Z. Yue, and S. Yang, "Emotion recognition from facial expressions and contactless heart rate using knowledge graph," in *2020 IEEE International Conference on Knowledge Graph (ICKG)*, pp. 64–69, IEEE, 2020.

[192] G. Du, S. Long, and H. Yuan, "Non-contact emotion recognition combining heart rate and facial expression for interactive gaming environments," *IEEE Access*, vol. 8, pp. 11896–11906, 2020.

[193] F. Abdat, C. Maaoui, and A. Pruski, "Bimodal system for emotion recognition from facial expressions and physiological signals using feature-level fusion," in *2011 UKSim 5th European Symposium on Computer Modeling and Simulation*, pp. 24–29, IEEE, 2011.

[194] B. Zhong, Z. Qin, S. Yang, J. Chen, N. Mudrick, M. Taub, R. Azevedo, and E. Lobaton, "Emotion recognition with facial expressions and physiological signals," in *2017 IEEE symposium series on computational intelligence (SSCI)*, pp. 1–8, IEEE, 2017.

[195] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE Access*, vol. 8, pp. 168865–168878, 2020.

[196] M. Li, L. Xie, Z. Lv, J. Li, and Z. Wang, "Multistep deep system for multimodal emotion detection with invalid data in the internet of things," *IEEE Access*, vol. 8, pp. 187208–187221, 2020.

[197] Q. Zhu, G. Lu, and J. Yan, "Valence-arousal model based emotion recognition using eeg, peripheral physiological signals and facial expression," in *Proceedings of the 4th International Conference on Machine Learning and Soft Computing*, pp. 81–85, 2020.

[198] N. Saffaryazdi, S. T. Wasim, K. Dileep, A. F. Nia, S. Nanayakkara, E. Broadbent, and M. Billinghurst, "Using facial micro-expressions in combination with eeg and physiological signals for emotion recognition," *Frontiers in Psychology*, p. 3486, 2022.

[199] M. Hoque and R. W. Picard, "Acted vs. natural frustration and delight : Many people smile in natural frustration," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pp. 354–359, IEEE, 2011.

[200] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of selected topics in signal processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[201] M. Hao, W.-H. Cao, Z.-T. Liu, M. Wu, and P. Xiao, "Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features," *Neurocomputing*, vol. 391, pp. 42–51, 2020.

[202] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9, 2016.

[203] C. M. A. Ilyas, R. Nunes, K. Nasrollahi, M. Rehm, and T. B. Moeslund, "Deep emotion recognition through upper body movements and facial expression.," in *VISIGRAPP (5 : VISAPP)*, pp. 669–679, 2021.

[204] K. Yang, H. Xu, and K. Gao, "Cm-bert : Cross-modal bert for text-audio sentiment analysis," *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[205] A. Bakhshi and S. Chalup, "Multimodal emotion recognition based on speech and physiological signals using deep neural networks," in *International Conference on Pattern Recognition*, pp. 289–300, Springer, 2021.

[206] D. Deng, Y. Zhou, J. Pi, and B. E. Shi, "Multimodal utterance-level affect analysis using visual, audio and text features," *arXiv preprint arXiv :1805.00625*, 2018.

[207] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9, IEEE, 2016.

[208] C. Sandi, "Stress and cognition," *Wiley Interdisciplinary Reviews : Cognitive Science*, vol. 4, no. 3, pp. 245–261, 2013.

[209] D. Rai, K. Kosidou, M. Lundberg, R. Araya, G. Lewis, and C. Magnusson, "Psychological distress and risk of long-term disability : population-based longitudinal study," *J Epidemiol Community Health*, vol. 66, no. 7, pp. 586–592, 2012.

[210] Y. S. Can, B. Arnrich, and C. Ersoy, "Stress detection in daily life scenarios using smart phones and wearable sensors : A survey," *Journal of biomedical informatics*, vol. 92, p. 103139, 2019.

[211] H. Selye, *The stress of life*. 1956.

[212] R. S. Lazarus and S. Folkman, *Stress, appraisal, and coping*. Springer publishing company, 1984.

[213] B. S. McEwen, "The neurobiology of stress : from serendipity to clinical relevance," *Brain research*, vol. 886, no. 1-2, pp. 172–189, 2000.

[214] S. Folkman, *The Oxford handbook of stress, health, and coping*. Oxford University Press, 2011.

[215] L. E. Bourne Jr and R. A. Yaroush, "Stress and cognition : A cognitive psychological perspective," tech. rep., 2003.

[216] T. L. Quick, "Healthy work : Stress, productivity, and the reconstruction of working life," *National Productivity Review*, vol. 9, no. 4, pp. 475–479, 1990.

[217] J. Siegrist, "Effort-reward imbalance at work and cardiovascular diseases," *International journal of occupational medicine and environmental health*, vol. 23, no. 3, p. 279, 2010.

[218] A. R. Jensen and W. D. Rohwer Jr, "The stroop color-word test : a review," *Acta psychologica*, vol. 25, pp. 36–93, 1966.

[219] U. Lundberg, R. Kadefors, B. Melin, G. Palmerud, P. Hassmén, M. Engström, and I. Elfsberg Dohns, "Psychophysiological stress and emg activity of the trapezius muscle," *International journal of behavioral medicine*, vol. 1, no. 4, pp. 354–370, 1994.

[220] M. Willmann, C. Langlet, J.-P. Hainaut, and B. Bolmont, "The time course of autonomic parameters and muscle tension during recovery following a moderate cognitive stressor : Dependency on trait anxiety level," *International Journal of Psychophysiology*, vol. 84, no. 1, pp. 51–58, 2012.

[221] F. Teixeira-Silva, G. B. Prado, L. C. G. Ribeiro, and J. R. Leite, "The anxiogenic video-recorded stroop color–word test : Psychological and physiological alterations and effects of diazepam," *Physiology & behavior*, vol. 82, no. 2-3, pp. 215–230, 2004.

[222] G. Schneider, D. Jacobs, R. Gevirtz, and D. O'connor, "Cardiovascular haemodynamic response to repeated mental stress in normotensive subjects at genetic risk of hypertension : evidence of enhanced reactivity, blunted adaptation, and delayed recovery," *Journal of human hypertension*, vol. 17, no. 12, pp. 829–840, 2003.

[223] E. A. Hines, "A standard stimulus for measuring vasomotor reactions : its application in the study of hypertension," in *Mayo Clin Proc*, vol. 7, pp. 332–335, 1932.

[224] W. Lovallo, "The cold pressor test and autonomic function : a review and integration," *Psychophysiology*, vol. 12, no. 3, pp. 268–282, 1975.

[225] N. Sharma, A. Dhall, T. Gedeon, and R. Goecke, "Thermal spatio-temporal data for stress recognition," *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, pp. 1–12, 2014.

[226] P. Zimmermann, P. Gomez, B. Danuser, and S. Schär, "Extending usability : putting affect into the user-experience," *Proceedings of NordiCHI'06*, pp. 27–32, 2006.

[227] C. Z. Wei, "Stress emotion recognition based on rsp and emg signals," in *Advanced Materials Research*, vol. 709, pp. 827–831, Trans Tech Publ, 2013.

[228] R. M. Sabour, Y. Benezeth, P. De Oliveira, J. Chappe, and F. Yang, "Ubfc-phys : A multimodal database for psychophysiological studies of social stress," *IEEE Transactions on Affective Computing*, 2021.

[229] M. Wang and K. J. Saudino, "Emotion regulation and stress," *Journal of Adult Development*, vol. 18, no. 2, pp. 95–103, 2011.

[230] M. Zubair and C. Yoon, "Multilevel mental stress detection using ultra-short pulse rate variability series," *Biomedical Signal Processing and Control*, vol. 57, p. 101736, 2020.

[231] R. Costin, C. Rotariu, and A. Pasarica, "Mental stress detection using heart rate variability and morphologic variability of eeg signals," in *2012 International Conference and Exposition on Electrical and Power Engineering*, pp. 591–596, IEEE, 2012.

[232] F. Bousefsaf, C. Maaoui, and A. Pruski, "Remote assessment of the heart rate variability to detect mental stress," in *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, pp. 348–351, IEEE, 2013.

[233] D. McDuff, I. Nishidate, K. Nakano, H. Haneishi, Y. Aoki, C. Tanabe, K. Niizeki, and Y. Aizu, "Non-contact imaging of peripheral hemodynamics during cognitive and psychological stressors," *Scientific Reports*, vol. 10, no. 1, pp. 1–13, 2020.

[234] K. Peternel, M. Pogačnik, R. Tavčar, and A. Kos, "A presence-based context-aware chronic stress recognition system," *Sensors*, vol. 12, no. 11, pp. 15888–15906, 2012.

[235] Y. Shi, N. Ruiz, R. Taib, E. Choi, and F. Chen, "Galvanic skin response (gsr) as an index of cognitive load," in *CHI'07 extended abstracts on Human factors in computing systems*, pp. 2651–2656, 2007.

[236] Y. Liu and S. Du, "Psychological stress level detection based on electrodermal activity," *Behavioural brain research*, vol. 341, pp. 50–53, 2018.

[237] S. K. Panigrahy, S. Jena, and A. Turuk, "Study and analysis of human stress detection using galvanic skin response (gsr) sensor in wired and wireless environments," *Research Journal of Pharmacy and Technology*, vol. 10, p. 545, 01 2017.

[238] R. Katmah, F. Al-Shargie, U. Tariq, F. Babiloni, F. Al-Mughairbi, and H. Al-Nashash, "A review on mental stress assessment methods using eeg signals," *Sensors*, vol. 21, no. 15, p. 5043, 2021.

[239] A. Alberdi, A. Aztiria, and A. Basarab, "Towards an automatic early stress recognition system for office environments based on multimodal measurements : A review," *Journal of biomedical informatics*, vol. 59, pp. 49–75, 2016.

[240] G. Giannakakis, D. Grigoriadis, and M. Tsiknakis, "Detection of stress/anxiety state from eeg features during video watching," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6034–6037, IEEE, 2015.

[241] F. Al-Shargie, T. B. Tang, N. Badruddin, and M. Kiguchi, "Mental stress quantification using eeg signals," in *International conference for innovation in biomedical engineering and life sciences*, pp. 15–19, Springer, 2015.

[242] G. Jun and K. G. Smitha, "Eeg based stress level identification," in *2016 IEEE international conference on systems, man, and cybernetics (SMC)*, pp. 003270–003274, IEEE, 2016.

[243] T. K. Calibo, J. A. Blanco, and S. L. Firebaugh, "Cognitive stress recognition," in *2013 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pp. 1471–1475, IEEE, 2013.

[244] V. Vanitha and P. Krishnan, "Real time stress detection system based on eeg signals," 2017.

[245] P. Karthikeyan, M. Murugappan, and S. Yaacob, "A study on mental arithmetic task based human stress level classification using discrete wavelet transform," in *2012 IEEE Conference on Sustainable Utilization and Development in Engineering and Technology (STUDENT)*, pp. 77–81, IEEE, 2012.

[246] D. McDuff, S. Gontarek, and R. Picard, "Remote measurement of cognitive stress via heart rate variability," in *2014 36th annual international conference of the IEEE engineering in medicine and biology society*, pp. 2957–2960, IEEE, 2014.

[247] G. Tanev, D. B. Saadi, K. Hoppe, and H. B. Sorensen, "Classification of acute stress using linear and non-linear heart rate variability analysis derived from sternal ecg," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3386–3389, IEEE, 2014.

[248] N. Keshan, P. Parimi, and I. Bichindaritz, "Machine learning for stress detection from ecg signals in automobile drivers," in *2015 IEEE International conference on big data (Big Data)*, pp. 2661–2669, IEEE, 2015.

[249] G. Giannakakis, M. Pediaditis, D. Manousos, E. Kazantzaki, F. Chiarugi, P. G. Simos, K. Marias, and M. Tsiknakis, "Stress and anxiety detection using facial cues from videos," *Biomedical Signal Processing and Control*, vol. 31, pp. 89–101, 2017.

[250] S. M. U. Saeed, S. M. Anwar, and M. Majid, "Quantification of human stress using commercially available single channel eeg headset," *IEICE Transactions on Information and Systems*, vol. 100, no. 9, pp. 2241–2244, 2017.

[251] F. Al-Shargie, T. B. Tang, N. Badruddin, and M. Kiguchi, "Mental stress quantification using eeg signals," in *International Conference for Innovation in Biomedical Engineering and Life Sciences : ICIBEL2015, 6-8 December 2015, Putrajaya, Malaysia 1*, pp. 15–19, Springer, 2016.

[252] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert, "Discriminating stress from cognitive load using a wearable eda device," *IEEE Transactions on information technology in biomedicine*, vol. 14, no. 2, pp. 410–417, 2009.

[253] P. Ren, A. Barreto, Y. Gao, and M. Adjouadi, "Affective assessment by digital processing of the pupil diameter," *IEEE Transactions on Affective computing*, vol. 4, no. 1, pp. 2–14, 2012.

[254] S. K. Panigrahy, S. K. Jena, and A. K. Turuk, "Study and analysis of human stress detection using galvanic skin response (gsr) sensor in wired and wireless environments," *Research Journal of Pharmacy and Technology*, vol. 10, no. 2, pp. 545–550, 2017.

[255] J. Zhai and A. Barreto, "Stress detection in computer users based on digital signal processing of noninvasive physiological variables," in *2006 international conference of the IEEE engineering in medicine and biology society*, pp. 1355–1358, IEEE, 2006.

[256] J. R. M. Fernández and L. Anishchenko, "Mental stress detection using bioradar respiratory signals," *Biomedical signal processing and control*, vol. 43, pp. 244–249, 2018.

[257] P. Karthikeyan, M. Murugappan, and S. Yaacob, "Emg signal based human stress level classification using wavelet packet transform," in *Trends in Intelligent Robotics, Automation, and Manufacturing :*

First International Conference, IRAM 2012, Kuala Lumpur, Malaysia, November 28-30, 2012. Proceedings, pp. 236–243, Springer, 2012.

[258] J. S. Lerner, R. E. Dahl, A. R. Hariri, and S. E. Taylor, "Facial expressions of emotion reveal neuroendocrine and cardiovascular stress responses," *Biological psychiatry*, vol. 61, no. 2, pp. 253–260, 2007.

[259] D. F. Dinges, R. L. Rider, J. Dorrian, E. L. McGlinchey, N. L. Rogers, Z. Cizman, S. Goldenstein, C. Vogler, S. Venkataraman, and D. N. Metaxas, "Optical computer recognition of facial expressions associated with stress induced by performance demands.," *Aviation, space, and environmental medicine*, vol. 76 6 Suppl, pp. B172–82, 2005.

[260] J. Zhang, X. Mei, H. Liu, S. Yuan, and T. Qian, "Detecting negative emotional stress based on facial expression in real time," *2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP)*, pp. 430–434, 2019.

[261] Y.-S. Og and W. hyun Cho, "Stress detection system for emotional labor based on deep learning facial expression recognition," 2021.

[262] J. Almeida and F. Rodrigues, "Facial expression recognition system for stress detection with deep learning.," in *ICEIS (1)*, pp. 256–263, 2021.

[263] W. T. Chew, S. C. Chong, T. S. Ong, and L. Y. Chong, "Facial expression recognition via enhanced stress convolution neural network for stress detection," *IAENG International Journal of Computer Science*, vol. 49, no. 3, pp. 1–10, 2022.

[264] C. Viegas, S.-H. Lau, R. Maxion, and A. Hauptmann, "Towards independent stress detection : A dependent model using facial action units," in *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, pp. 1–6, IEEE, 2018.

[265] F. Bevilacqua, H. Engström, and P. Backlund, "Automated analysis of facial cues from videos as a potential method for differentiating stress and boredom of players in games.," *International Journal of Computer Games Technology*, 2018.

[266] H. Gao, A. Yüce, and J.-P. Thiran, "Detecting emotional stress from facial expressions for driving safety," in *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 5961–5965, IEEE, 2014.

[267] M. Soury and L. Devillers, "Stress detection from audio on multiple window analysis size in a public speaking task," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 529–533, IEEE, 2013.

[268] H. Lu, D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury, "Stresssense : Detecting stress in unconstrained acoustic environments using smartphones," in *Proceedings of the 2012 ACM conference on ubiquitous computing*, pp. 351–360, 2012.

[269] O. Simantiraki, G. Giannakakis, A. Pampouchidou, and M. Tsiknakis, "Stress detection from speech

using spectral slope measurements," in *Pervasive Computing Paradigms for Mental Health : Selected Papers from MindCare 2016, Fabulous 2016, and IIoT 2015 3*, pp. 41–50, Springer, 2018.

[270] M. Pedrotti, M. A. Mirzaei, A. Tedesco, J.-R. Chardonnet, F. Mérienne, S. Benedetto, and T. Baccino, "Automatic stress classification with pupil diameter analysis," *International Journal of Human-Computer Interaction*, vol. 30, no. 3, pp. 220–236, 2014.

[271] S. Baltaci and D. Gokcay, "Stress detection in human–computer interaction : Fusion of pupil dilation and facial temperature features," *International Journal of Human–Computer Interaction*, vol. 32, no. 12, pp. 956–966, 2016.

[272] J. Aigrain, S. Dubuisson, M. Detyniecki, and M. Chetouani, "Person-specific behavioural features for automatic stress detection," in *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, vol. 3, pp. 1–6, IEEE, 2015.

[273] G. Giannakakis, D. Manousos, V. Chaniotakis, and M. Tsiknakis, "Evaluation of head pose features for stress detection and classification," in *2018 IEEE EMBS international conference on biomedical & health informatics (BHI)*, pp. 406–409, IEEE, 2018.

[274] B. D. Womack and J. H. Hansen, "N-channel hidden markov models for combined stressed speech classification and recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 668–677, 1999.

[275] V. P. Patil, K. K. Nayak, and M. Saxena, "Voice stress detection," *International Journal of Electrical, Electronics and Computer Engineering*, vol. 2, no. 2, pp. 148–154, 2013.

[276] M. Van Puyvelde, X. Neyt, F. McGlone, and N. Pattyn, "Voice stress analysis : a new framework for voice and effort in human performance," *Frontiers in psychology*, vol. 9, p. 1994, 2018.

[277] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on psychological stress detection using biosignals," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 440–460, 2019.

[278] R. Fernandez and R. W. Picard, "Modeling drivers' speech under stress," *Speech communication*, vol. 40, no. 1-2, pp. 145–159, 2003.

[279] X. Yao, T. Jitsuhiro, C. Miyajima, N. Kitaoka, and K. Takeda, "Physical characteristics of vocal folds during speech under stress," *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4609–4612, 2012.

[280] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[281] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," in *Ninth international conference on spoken language processing*, 2006.

[282] H. Han, K. Byun, and H.-G. Kang, "A deep learning-based stress detection algorithm with speech signal," in *proceedings of the 2018 workshop on audio-visual scene understanding for immersive multimedia*, pp. 11–15, 2018.

[283] D. Carneiro, J. C. Castillo, P. Novais, A. Fernández-Caballero, and J. Neves, "Multimodal behavioral analysis for non-invasive stress detection," *Expert Systems with Applications*, vol. 39, no. 18, pp. 13376–13389, 2012.

[284] A. de Santos Sierra, C. S. Avila, J. G. Casanova, G. B. Del Pozo, and V. J. Vera, "Two stress detection schemes based on physiological signals for real-time applications," in *2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 364–367, IEEE, 2010.

[285] V. Sandulescu, S. Andrews, D. Ellis, N. Bellotto, and O. M. Mozos, "Stress detection using wearable physiological sensors," in *International work-conference on the interplay between natural and artificial computation*, pp. 526–532, Springer, 2015.

[286] L. Xia, A. S. Malik, and A. R. Subhani, "A physiological signal-based method for early mental-stress detection," in *Cyber-Enabled Intelligence*, pp. 259–289, Taylor & Francis, 2019.

[287] H. Kurniawan, A. V. Maslov, and M. Pechenizkiy, "Stress detection from speech and galvanic skin response signals," in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, pp. 209–214, IEEE, 2013.

[288] A. Anusha, J. Jose, S. Preejith, J. Jayaraj, and S. Mohanasankar, "Physiological signal based work stress detection using unobtrusive sensors," *Biomedical Physics & Engineering Express*, vol. 4, no. 6, p. 065001, 2018.

[289] J. Zhang, H. Yin, J. Zhang, G. Yang, J. Qin, and L. He, "Real-time mental stress detection using multimodality expressions with a deep learning framework," *Frontiers in Neuroscience*, vol. 16, 2022.

[290] D. Giakoumis, A. Drosou, P. Cipresso, D. Tzovaras, G. Hassapis, A. Gaggioli, and G. Riva, "Using activity-related behavioural features towards more effective automatic stress detection," 2012.

[291] J. Zhai and A. Barreto, "Stress detection in computer users through non-invasive monitoring of physiological signals," *Blood*, vol. 5, no. 0, 2008.

[292] O. M. Mozos, V. Sandulescu, S. Andrews, D. Ellis, N. Bellotto, R. Dobrescu, and J. M. Ferrandez, "Stress detection using wearable physiological and sociometric sensors," *International journal of neural systems*, vol. 27, no. 02, p. 1650041, 2017.

[293] A. Bhatti, B. Behinaein, P. Hungler, and A. Etemad, "Attx : Attentive cross-connections for fusion of wearable signals in emotion recognition," *arXiv preprint arXiv :2206.04625*, 2022.

[294] R. Walambe, P. Nayak, A. Bhardwaj, and K. Kotecha, "Employing multimodal machine learning for stress detection," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–12, 2021.

[295] P. Bobade and M. Vani, "Stress detection with machine learning and deep learning using multimodal physiological data," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 51–57, IEEE, 2020.

[296] S. Koldijk, M. A. Neerincx, and W. Kraaij, "Detecting work stress in offices by combining unobtrusive sensors," *IEEE Transactions on affective computing*, vol. 9, no. 2, pp. 227–239, 2016.

[297] C.-P. Bara, M. Papakostas, and R. Mihalcea, "A deep learning approach towards multimodal stress detection.," in *AffCon@ AAAI*, pp. 67–81, 2020.

[298] L.-l. Chen, Y. Zhao, P.-f. Ye, J. Zhang, and J.-z. Zou, "Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers," *Expert Systems with Applications*, vol. 85, pp. 279–291, 2017.

[299] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "Continuous stress detection using a wrist device : in laboratory and real life," in *proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing : Adjunct*, pp. 1185–1193, 2016.

[300] M. A. B. S. Akhonda, S. M. F. Islam, A. S. Khan, F. Ahmed, and M. M. Rahman, "Stress detection of computer user in office like working environment using neural network," in *2014 17th International Conference on Computer and Information Technology (ICCIT)*, pp. 174–179, IEEE, 2014.

[301] E. Maier, U. Reimer, E. Laurenzi, M. Ridinger, and T. Ulmer, "A mobile solution for stress recognition and prevention," in *Proc. Int'l Conf. Health Informatics (HealthInf)*, pp. 428–433, 2014.

[302] A. Sano and R. W. Picard, "Stress recognition using wearable sensors and mobile phones," in *2013 Humaine association conference on affective computing and intelligent interaction*, pp. 671–676, IEEE, 2013.

[303] D. S. Lee, T. W. Chong, and B. G. Lee, "Stress events detection of driver by wearable glove system," *IEEE Sensors Journal*, vol. 17, no. 1, pp. 194–204, 2016.

[304] L. M. Vizer, L. Zhou, and A. Sears, "Automated stress detection using keystroke and linguistic features : An exploratory study," *International Journal of Human-Computer Studies*, vol. 67, no. 10, pp. 870–886, 2009.

[305] W. Liao, W. Zhang, Z. Zhu, and Q. Ji, "A real-time human stress monitoring system using dynamic bayesian network," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)-workshops*, pp. 70–70, IEEE, 2005.

[306] A. Barreto, J. Zhai, and M. Adjouadi, "Non-intrusive physiological monitoring for automated stress detection in human-computer interaction," in *International Workshop on Human-Computer Interaction*, pp. 29–38, Springer, 2007.

[307] V. C. Scanlon and T. Sanders, *Essentials of anatomy and physiology.* FA Davis, 2018.

[308] H. Tanaka, K. D. Monahan, and D. R. Seals, "Age-predicted maximal heart rate revisited," *Journal of the american college of cardiology*, vol. 37, no. 1, pp. 153–156, 2001.

[309] U. Rajendra Acharya, K. Paul Joseph, N. Kannathal, C. M. Lim, and J. S. Suri, "Heart rate variability : a review," *Medical and biological engineering and computing*, vol. 44, no. 12, pp. 1031–1051, 2006.

[310] A. L. Goldberger, "Is the normal heartbeat chaotic or homeostatic ?," *Physiology*, vol. 6, no. 2, pp. 87–91, 1991.

[311] K. Bilchick and R. Berger, "Heart rate variability.," *Journal of Cardiovascular Electrophysiology*, vol. 17, pp. 691–694, June 2006.

[312] T. F. o. t. E. S. o. C. t. N. A. S. o. P. Electrophysiology, "Heart rate variability : standards of measurement, physiological interpretation, and clinical use," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

[313] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in public health*, p. 258, 2017.

[314] C. Lockwood, T. Conroy-Hiller, and T. Page, "Vital signs," *JBI reports*, vol. 2, no. 6, pp. 207–230, 2004.

[315] F. Bousefsaf, *Mesure sans contact de l'activité cardiaque par analyse du flux vidéo issu d'une caméra numérique : extraction de paramètres physiologiques et application à l'estimation du stress.* PhD thesis, 2014.

[316] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological measurement*, vol. 28, no. 3, pp. R1–R39, 2007.

[317] A. Hertzmann, "Observations on the finger volume pulse recorded photo-electrically," *Am J Physiol*, vol. 119, pp. 334–335, 1937.

[318] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, pp. 7–11, Jan. 2011.

[319] A. C. Kevat, D. V. Bullen, P. G. Davis, and C. O. F. Kamlin, "A systematic review of novel technology for monitoring infant and newborn heart rate," *Acta Paediatrica*, vol. 106, no. 5, pp. 710–720, 2017.

[320] F. Zhao, M. Li, Y. Qian, and J. Z. Tsien, "Remote measurements of heart and respiration rates for telemedicine," *PloS one*, vol. 8, no. 10, p. e71384, 2013.

[321] Y. Ouzar, F. Bousefsaf, D. Djeldjli, and C. Maaoui, "Video-based multimodal spontaneous emotion recognition using facial expressions and physiological signals," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2459–2468, 2022.

[322] W. Wang, "Robust and automatic remote photoplethysmography," 2017.

[323] M. M. Trivedi, T. Gandhi, and J. McCall, "Looking-in and looking-out of a vehicle : Computer-vision-based enhanced vehicle safety," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 1, pp. 108–120, 2007.

[324] L. M. Bergasa, J. M. Buenaposada, J. Nuevo, P. Jimenez, and L. Baumela, "Analysing driver's attention level using computer vision," in *2008 11th International IEEE Conference on Intelligent Transportation Systems*, pp. 1149–1154, IEEE, 2008.

[325] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei, and G. Zhao, "Deep learning for face anti-spoofing : A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[326] E. M. Nowara, *Towards Robust Imaging Photoplethysmography in Unconstrained Settings.* PhD thesis, Rice University, 2021.

[327] A. Tohma, M. Nishikawa, T. Hashimoto, Y. Yamazaki, and G. Sun, "Evaluation of remote photoplethysmography measurement conditions toward telemedicine applications," *Sensors*, vol. 21, no. 24, p. 8357, 2021.

[328] Y. Sun, S. Hu, V. Azorin-Peris, R. Kalawsky, and S. E. Greenwald, "Noncontact imaging photoplethysmography to effectively access pulse rate variability," *Journal of biomedical optics*, vol. 18, no. 6, p. 061205, 2012.

[329] D. Shao, Y. Yang, C. Liu, F. Tsow, H. Yu, and N. Tao, "Noncontact monitoring breathing pattern, exhalation flow rate and pulse transit time," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 11, pp. 2760–2767, 2014.

[330] D. Djeldjli, F. Bousefsaf, C. Maaoui, F. Bereksi-Reguig, and A. Pruski, "Remote estimation of pulse wave features related to arterial stiffness and blood pressure using a camera," *Biomedical Signal Processing and Control*, vol. 64, p. 102242, 2021.

[331] W. Chen and D. McDuff, "Deepphys : Video-based physiological measurement using convolutional attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 349–365, 2018.

[332] F. Bousefsaf, A. Pruski, and C. Maaoui, "3d convolutional neural networks for remote pulse rate measurement and mapping from facial video," *Applied Sciences*, vol. 9, p. 4364, 10 2019.

[333] X. Niu, S. Shan, H. Han, and X. Chen, "Rhythmnet : End-to-end heart rate estimation from face via spatial-temporal representation," *IEEE Transactions on Image Processing*, vol. 29, pp. 2409–2423, 2019.

[334] P. Li, *Pulse rate variability measurement with camera-based photoplethysmography.* PhD thesis, Université Bourgogne Franche-Comté, 2021.

[335] N. Bouzida, A. Bendada, and X. P. Maldague, "Visualization of body thermoregulation by infrared imaging," *Journal of Thermal Biology*, vol. 34, no. 3, pp. 120–126, 2009.

[336] T. Blöcher, J. Schneider, M. Schinle, and W. Stork, "An online ppgi approach for camera based heart rate monitoring using beat-to-beat detection," in *2017 IEEE Sensors Applications Symposium (SAS)*, pp. 1–6, 2017.

[337] M. Kumar, A. Veeraraghavan, and A. Sabharwal, "Distanceppg : Robust non-contact vital signs monitoring using a camera," *Biomedical optics express*, vol. 6, no. 5, pp. 1565–1588, 2015.

[338] L. Feng, L.-M. Po, X. Xu, Y. Li, and R. Ma, "Motion-resistant remote imaging photoplethysmography based on the optical properties of skin," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 5, pp. 879–891, 2014.

[339] G. Lempe, S. Zaunseder, T. Wirthgen, S. Zipser, and H. Malberg, "Roi selection for remote photoplethysmography," in *Bildverarbeitung für die Medizin 2013*, pp. 99–103, Springer, 2013.

[340] G. Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," *IEEE transactions on bio-medical engineering*, vol. 60, 06 2013.

[341] S. Kwon, J. Kim, D. Lee, and K. Park, "Roi analysis for remote photoplethysmography on facial video," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4938–4941, IEEE, 2015.

[342] V. Selvaraju, N. Spicher, J. Wang, N. Ganapathy, J. M. Warnecke, S. Leonhardt, R. Swaminathan, and T. M. Deserno, "Continuous monitoring of vital signs using cameras : A systematic review," *Sensors*, vol. 22, no. 11, p. 4097, 2022.

[343] K. B. Jaiswal and T. Meenpal, "Continuous pulse rate monitoring from facial video using rppg," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–5, IEEE, 2020.

[344] R. Song, S. Zhang, J. Cheng, C. Li, and X. Chen, "New insights on super-high resolution for video-based heart rate estimation with a semi-blind source separation method," *Computers in biology and medicine*, vol. 116, p. 103535, 2020.

[345] F. Bousefsaf, C. Maaoui, and A. Pruski, "Automatic Selection of Webcam Photoplethysmographic Pixels Based on Lightness Criteria," *Journal of Medical and Biological Engineering*, vol. 37, pp. 374–385, June 2017.

[346] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, and N. Sebe, "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2396–2404, 2016.

[347] B. Huang, C.-L. Lin, W. Chen, C.-F. Juang, and X. Wu, "A novel one-stage framework for visual pulse rate estimation using deep neural networks," *Biomedical Signal Processing and Control*, vol. 66, p. 102387, 2021.

[348] W. Wang, S. Stuijk, and G. De Haan, "A novel algorithm for remote photoplethysmography : Spatial subspace rotation," *IEEE transactions on biomedical engineering*, vol. 63, no. 9, pp. 1974–1984, 2015.

[349] D. Djeldjli, F. Bousefsaf, C. Maaoui, and F. Bereksi-Reguig, "Imaging photoplethysmography : Signal waveform analysis," in *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems : Technology and Applications (IDAACS)*, vol. 2, pp. 830–834, IEEE, 2019.

[350] R. Sinhal, K. Singh, and M. Raghuwanshi, "An overview of remote photoplethysmography methods for vital sign monitoring," in *Computer Vision and Machine Intelligence in Medical Image Analysis : International Symposium, ISCMM 2019*, pp. 21–31, Springer, 2020.

[351] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation.," *Optics express*, vol. 18, no. 10, pp. 10762–10774, 2010.

[352] M. Lewandowska, J. Rumiński, T. Kocejko, and J. Nowak, "Measuring pulse rate with a webcam — a non-contact method for evaluating cardiac activity," *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pp. 405–410, 2011.

[353] E. Christinaki, G. Giannakakis, F. Chiarugi, M. Pediaditis, G. Iatraki, D. Manousos, K. Marias, and M. Tsiknakis, "Comparison of blind source separation algorithms for optical heart rate monitoring," in *2014 4th International Conference on Wireless Mobile Communication and Healthcare-Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)*, pp. 339–342, IEEE, 2014.

[354] D. Wedekind, A. Trumpp, F. Gaetjen, S. Rasche, K. Matschke, H. Malberg, and S. Zaunseder, "Assessment of blind source separation techniques for video-based cardiac pulse extraction," *Journal of biomedical optics*, vol. 22, no. 3, p. 035002, 2017.

[355] G. De Haan and A. Van Leest, "Improved motion robustness of remote-ppg by using the blood volume pulse signature," *Physiological measurement*, vol. 35, no. 9, p. 1913, 2014.

[356] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote ppg," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2016.

[357] F. Bousefsaf, C. Maaoui, and A. Pruski, "Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 568–574, 2013.

[358] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light.," *Optics express*, vol. 16, no. 26, pp. 21434–21445, 2008.

[359] J. Rumiński, "Reliability of pulse measurements in videoplethysmography," *Metrology and Measurement Systems*, vol. 23, no. 3, 2016.

[360] G. De Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878–2886, 2013.

[361] S. Bennett, T. N. El Harake, R. Goubran, and F. Knoefel, "Adaptive eulerian video processing of thermal video : An experimental analysis," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 10, pp. 2516–2524, 2017.

[362] D.-Y. Chen, H.-S. Zou, and A.-T. Hsieh, "Thermal image based remote heart rate measurement on dynamic subjects using deep learning," in *2020 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, pp. 1–2, 2020.

[363] K. Humphreys, T. Ward, and C. Markham, "Noncontact simultaneous dual wavelength photoplethysmography : a further step toward noncontact pulse oximetry," *Review of scientific instruments*, vol. 78, no. 4, p. 044304, 2007.

[364] W. Wang, A. D. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote ppg," *IEEE Transactions on Biomedical Engineering*, vol. 64, pp. 1479–1491, 2017.

[365] Z. Yu, X. Li, X. Niu, J. Shi, and G. Zhao, "Autohr : A strong end-to-end baseline for remote heart rate measurement with neural searching," *IEEE Signal Processing Letters*, vol. 27, pp. 1245–1249, 2020.

[366] D. J. McDuff, S. Gontarek, and R. W. Picard, "Improvements in remote cardiopulmonary measurement using a five band digital camera," *IEEE Transactions on Biomedical Engineering*, vol. 61, pp. 2593–2601, 2014.

[367] Y. Qiu, Y. Liu, J. Arteaga-Falconi, H. Dong, and A. E. Saddik, "Evm-cnn : Real-time contactless heart rate estimation from facial video," *IEEE Transactions on Multimedia*, vol. 21, pp. 1778–1787, 2019.

[368] X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-task temporal shift attention networks for on-device contactless vitals measurement," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19400–19411, 2020.

[369] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," *arXiv preprint arXiv :1905.02419*, 2019.

[370] R. Špetlík, V. Franc, and J. Matas, "Visual heart rate estimation with convolutional neural network,"

[371] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on mri," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019. Special Issue : Deep Learning in Medical Physics.

[372] E. Goceri and N. Goceri, "Deep learning in medical image analysis : Recent advances and future trends," pp. 305–310, 07 2017.

[373] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision : A brief review," *Computational intelligence and neuroscience*, vol. 2018, 2018.

[374] A. Ni, A. Azarang, and N. Kehtarnavaz, "A review of deep learning-based contactless heart rate measurement methods," *Sensors*, vol. 21, p. 3719, 05 2021.

[375] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, "Deep ppg : Large-scale heart rate estimation with convolutional neural networks," *Sensors*, vol. 19, no. 14, 2019.

[376] E. Lee, E. Chen, and C.-Y. Lee, "Meta-rppg : Remote heart rate estimation using a transductive meta-learner," in *ECCV*, 2020.

[377] Z. Yu, W. Peng, X. Li, X. Hong, and G. Zhao, "Remote heart rate measurement from highly compressed facial videos : an end-to-end deep learning solution with video enhancement," 2019.

[378] O. Perepelkina, M. Artemyev, M. Churikova, and M. Grinenko, "Hearttrack : Convolutional neural network for remote video-based heart rate monitoring," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1163–1171, 2020.

[379] Z. Zhang, J. Girard, Y. Wu, X. Zhang, P. Liu, U. A. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. Cohn, Q. Ji, and L. Yin, "Multimodal spontaneous emotion corpus for human behavior analysis," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3438–3446, 2016.

[380] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[381] X. Niu, H. Han, S. Shan, and X. Chen, "Vipl-hr : A multi-modal database for pulse estimation from less-constrained face video," in *ACCV*, 2018.

[382] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet : A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[383] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," *CoRR*, vol. abs/1705.06950, 2017.

[384] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognition Letters*, vol. 124, pp. 82–90, 2019. Award Winning Papers from the 23rd International Conference on Pattern Recognition (ICPR).

[385] M. P. Tarvainen, P. O. Ranta-aho, and P. A. Karjalainen, "An advanced detrending method with application to hrv analysis," *IEEE Transactions on Biomedical Engineering*, vol. 49, pp. 172–175, 2002.

[386] Y. Nirkin, I. Masi, A. T. Tran, T. Hassner, and G. G. Medioni, "On face segmentation, face swapping, and face perception," *CoRR*, vol. abs/1704.06729, 2017.

[387] Y.-Y. Tsou, Y.-A. Lee, C.-T. Hsu, and S.-H. Chang, "Siamese-rppg network : Remote photoplethysmography signal estimation from face videos," in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, SAC '20, (New York, NY, USA), p. 2066–2073, Association for Computing Machinery, 2020.

[388] P. Zhao, C. Li, M. M. Rahaman, H. Yang, T. Jiang, and M. Grzegorzek, "A comparison of deep learning classification methods on small-scale image data set : from convolutional neural networks to visual transformers," *arXiv preprint arXiv :2107.07699*, 2021.

[389] A. V. Moço, S. Stuijk, and G. de Haan, "Motion robust ppg-imaging through color channel mapping," *Biomedical optics express*, vol. 7, no. 5, pp. 1737–1754, 2016.

[390] F. Chollet, "Xception : Deep learning with depthwise separable convolutions," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807, 2017.

[391] K. Shaheed, A. Mao, I. Qureshi, M. Kumar, S. Hussain, I. Ullah, and X. Zhang, "Ds-cnn : A pre-trained xception model based on depth-wise separable convolutional neural network for finger vein recognition," *Expert Syst. Appl.*, vol. 191, apr 2022.

[392] N. Keskar and R. Socher, "Improving generalization performance by switching from adam to sgd," *ArXiv*, vol. abs/1712.07628, 2017.

[393] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*, April 2020.

[394] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv :1609.04747*, 2016.

[395] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout : A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[396] E. M. Nowara, D. McDuff, and A. Veeraraghavan, "Combining magnification and measurement for non-contact cardiac monitoring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3810–3819, June 2021.

[397] N. Miljković and D. Trifunović, "Pulse rate assessment : Eulerian video magnification vs. electrocardiography recordings," in *12th Symposium on Neural Network Applications in Electrical Engineering (NEUREL)*, pp. 17–20, IEEE, 2014.

[398] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM transactions on graphics (TOG)*, vol. 31, no. 4, pp. 1–8, 2012.

[399] X. Li, J. Chen, G. Zhao, and M. Pietikäinen, "Remote heart rate measurement from face videos under realistic situations," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4264–4271, 2014.

[400] E. Lee, E. Chen, and C.-Y. Lee, "Meta-rppg : Remote heart rate estimation using a transductive meta-learner," in *European Conference on Computer Vision*, pp. 392–409, Springer, 2020.

[401] X. Niu, H. Han, S. Shan, and X. Chen, "Synrhythm : Learning a deep heart rate estimator from general to specific," in *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 3580–3585, 2018.

[402] R. Song, H. Chen, J. Cheng, C. Li, Y. Liu, and X. Chen, "Pulsegan : Learning to generate realistic pulse waveforms in remote photoplethysmography," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, pp. 1373–1384, 2021.

[403] X. Liu, Z. Jiang, J. Fromm, X. Xu, S. N. Patel, and D. McDuff, "Metaphys : Unsupervised few-shot adaptation for non-contact physiological measurement," *ArXiv*, vol. abs/2010.01773, 2020.

[404] T. Fitzpatrick, "The validity and practicality of sun-reactive skin types i through vi.," *Archives of dermatology*, vol. 124 6, pp. 869–71, 1988.

[405] E. Nowara, D. McDuff, and A. Veeraraghavan, "A meta-analysis of the impact of skin type and gender on non-contact photoplethysmography measurements," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1148–1155, 2020.

[406] Y. Ouzar, D. Djeldjli, F. Bousefsaf, and C. Maaoui, "Lcoms lab's approach to the vision for vitals (v4v) challenge," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 2750–2754, October 2021.

[407] Y. Ouzar, F. Bousefsaf, D. Djeldjli, and C. Maaoui, "Video-based multimodal spontaneous emotion recognition using facial expressions and physiological signals," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2460–2469, 2022.

[408] G. Heusch, A. Anjos, and S. Marcel, "A reproducible study on remote heart rate measurement," *arXiv preprint arXiv :1709.00962*, 2017.

[409] R. Stricker, S. Müller, and H.-M. Groß, "Non-contact video-based pulse rate measurement on a mobile service robot," *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pp. 1056–1062, 2014.

[410] D. McDuff, X. Liu, J. Hernandez, E. Wood, and T. Baltrusaitis, "Synthetic data for multi-parameter camera-based physiological sensing," *arXiv preprint arXiv :2110.04902*, 2021.

[411] B. L. Hill, X. Liu, and D. McDuff, "Beat-to-beat cardiac pulse rate measurement from video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 2739–2742, October 2021.

[412] W. Li, F. Abtahi, Z. Zhu, and L. Yin, "Eac-net : Deep nets with enhancing and cropping for facial action unit detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 11, pp. 2583–2596, 2018.

[413] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[414] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[415] Y. Benezeth, P. Li, R. Macwan, K. Nakamura, R. Gomez, and F. Yang, "Remote heart rate variability for emotional state monitoring," in *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 153–156, IEEE, 2018.

[416] M. Fingar and P. Podrzaj, "Feasibility of assessing ultra-short-term pulse rate variability from video recordings," *PeerJ*, vol. 8, 2020.

[417] S. Zaunseder, A. Trumpp, D. Wedekind, and H. Malberg, "Cardiovascular assessment by imaging photoplethysmography – a review," *Biomedical Engineering / Biomedizinische Technik*, vol. 63, pp. 617 – 634, 2018.

[418] P. Welch, "The use of fast fourier transform for the estimation of power spectra : a method based on time averaging over short, modified periodograms," *IEEE Transactions on audio and electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[419] B. Appelhans and L. Luecken, "Heart rate variability as an index of regulated emotional responding," *Review of General Psychology*, vol. 10, pp. 229–240, Sept. 2006.

[420] R. L. Burr, "Interpretation of normalized spectral heart rate variability indices in sleep research : a critical review.," *Sleep*, vol. 30 7, pp. 913–9, 2007.

[421] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[422] Z. Wang, J. Wang, S. Wang, and Q. Ji, "Sequence-based bias analysis of spontaneous facial expression databases," in *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 1–6, IEEE, 2014.

[423] Y. Fang, R. Rong, and J. Huang, "Hierarchical fusion of visual and physiological signals for emotion recognition," *Multidimensional Systems and Signal Processing*, vol. 32, pp. 1103–1121, 2021.

[424] D. Fabiano and S. Canavan, "Emotion recognition using fused physiological signals," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 42–48, 2019.

[425] N. Bourdillon, L. Schmitt, S. Yazdani, J.-M. Vesin, and G. P. Millet, "Minimal window duration for accurate hrv recording in athletes," *Frontiers in neuroscience*, vol. 11, p. 456, 2017.

[426] Y. Huang, J. Yang, P. Liao, and J. Pan, "Fusion of facial expressions and eeg for multimodal emotion recognition," *Computational Intelligence and Neuroscience*, vol. 2017, 2017.

[427] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE Access*, vol. 8, pp. 168865–168878, 2020.

[428] R. Meziati Sabour, Y. Benezeth, P. De Oliveira, J. Chappe, and F. Yang, "Ubfc-phys : A multimodal database for psychophysiological studies of social stress," *IEEE Transactions on Affective Computing*, pp. 1–1, 2021.

[429] Z. Pei, H. Xu, Y. Zhang, M. Guo, and Y.-H. Yang, "Face recognition via deep learning using data augmentation based on orthogonal experiments," *Electronics*, vol. 8, no. 10, p. 1088, 2019.

[430] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv :1409.1556*, 2014.

[431] A. K. Dubey and V. Jain, "Automatic facial recognition using vgg16 based transfer learning model," *Journal of Information and Optimization Sciences*, vol. 41, no. 7, pp. 1589–1596, 2020.

[432] Y. Ouzar, F. Bousefsaf, D. Djeldjli, and C. Maaoui, "Video-based multimodal spontaneous emotion recognition using facial expressions and physiological signals," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2460–2469, June 2022.