M2IIA : STAGE/ALTERNANCE

Université de Bourgogne, UFR Sciences et Techniques, Département I.E.M.

Rapport de stage/alternance

Auteur/Autrice
Gauthier THOMAS

Détection et Analyse des émotions dans des vidéos : extraction de caractéristiques spécifiques par Vision par Ordinateur et Intelligence Artificielle

Laboratoire Interdisciplinaire Carnot de Bourgogne

Dominique GINHAC
Manon ANSART
Fan YANG SONG
Cyrille MIGNIOT (UB)

2023-24

Table des matières

Remerciements 5
1.1 Introduction 6
1.1.1 Le laboratoire 6
1.1.2 Le sujet 7
1.2 Revue de la littérature 8
1.2.1 Les émotions 8
1.2.2 Les micro et macro expressions 10
1.2.3 Variabilité du rythme cardiaque 11
1.2.4 Segmentation de l'iris 13
1.3 Comparaison des méthodes 15
1.3.1 Variabilité du rythme cardiaque 15
1.3.2 Segmentation de l'iris 17
1.4 Application sur des données réelles 22
1.4.1 L'expérience mise en place 22
1.4.2 Installation de la caméra 22
1.4.3 Difficultées rencontrées 23
1.5 Conclusion 25
1.5.1 À propos du projet 25
1.5.2 À propos de mon expérience personnelle 25

Résumé Ce rapport rend compte du travail effectué lors de ce stage de fin d'étude au laboratoire ICB.

L'objectif du stage est d'extraire des caractéristiques de visages à partir de vidéos afin d'avoir les données nécessaires à la création d'un modèle d'intelligence artificielle capable de détecter les émotions du sujet. Ce modèle aura pour but d'analyser les émotions suscitées par les aliments pour comprendre des comportements et choix alimentaires.

Pour ce faire, une première partie du stage a été consacrée en une revue bibliographique, afin de comprendre les détails du fonctionnement des émotions et des méthodes d'obtention des caractéristiques pertinentes du visage. Une seconde partie a ensuite été consacrée à la mise en place de ces méthodes, notamment pour récupérer les variations du rythme cardiaque, ainsi que les variations de taille de la pupille.

Mots-clés Émotions, IA, Rythme cardiaque, PPG, Segmentation d'iris, Comportement alimentaire

---

Abstract This report describes the work carried out during my final year internship at the ICB laboratory.

The aim of the internship was to extract facial features from videos in order to obtain the data needed to create an artificial intelligence model capable of detecting the subject's emotions. The model will aim to analyze the emotions aroused by food in order to understand eating behaviors and choices.

To achieve this, the first part of the internship was devoted to litterature review to understand the details of how emotions work and the methods for obtaining the relevant facial features. A second part was then devoted to implementing these methods, in particular for recovering variations in heart rate and pupil size.

Keywords Emotions, AI, Heart rate, PPG, Iris segmentation, Eating behavior

---

Remerciements

Ces cinq mois au laboratoire de l'ICB ont été une expérience enrichissante et inoubliable. Cette opportunité m'a permis de m'immerger dans un environnement de recherche dynamique et stimulant, de développer de nouvelles compétences et de consolider mes connaissances théoriques. C'est avec gratitude et beaucoup d'enthousiasme que je rédige ce mémoire de stage, qui témoigne des précieux enseignements et expériences acquis durant cette période.

Je tiens à exprimer ma sincère gratitude à Dominique Ginhac, dont la confiance m'a permis de mener à bien ce stage. Ses conseils ont été essentiels pour m'aider à relever les défis que j'ai rencontrés et ont largement contribué à rendre cette expérience si enrichissante.

Mes remerciements vont également à Manon Ansart et Fan Yang Song, pour leur encadrement et leurs conseils avisés tout au long du stage. Leur aide et leurs précieux conseils m'ont permis d'aborder les difficultés rencontrées pendant le stage avec une plus grande confiance en moi.

Je tiens également à remercier Cyrille Migniot pour son suivi tout au long du stage.

Enfin, je tiens à exprimer ma gratitude à l'ensemble de l'équipe VAI de l'ICB pour son accueil chaleureux ainsi que leurs explications qui m'ont permis de découvrir un peu plus le fonctionnement de la recherche académique.

1.1 Introduction

Le but de ce rapport est de présenter en détail le travail réalisé au cours de ce stage de fin d'études, qui s'est déroulé au sein de l'équipe "Vision and Artificial Intelligence" (VAI) du Laboratoire Interdisciplinaire Carnot de Bourgogne (ICB). Ce rapport mettra en lumière les diverses étapes du projet, les méthodes employées, ainsi que les résultats obtenus.

Dans ce rapport, une présentation détaillée du laboratoire de l'ICB sera faite, en mettant en avant ses missions, son organisation, et ses domaines de recherche. Il sera également expliqué le rôle et les objectifs de l'équipe VAI, en décrivant ses membres, leurs spécialités respectives, ainsi que les projets majeurs auxquels ils contribuent.

Le sujet de stage sera ensuite détaillé, en expliquant l'origine du projet et en précisant le problème auquel il répond. Il sera présenté les défis techniques et scientifiques rencontrés, en expliquant comment ces obstacles se sont présentés et pourquoi ils étaient significatifs pour la réussite du projet.

Ensuite, il sera proposé un résumé détaillé des recherches bibliographiques, en présentant les principales sources consultées et en expliquant leur pertinence pour le projet. Il sera discuté des concepts, théories et méthodes trouvées dans la littérature, qui ont influencé l'approche du sujet.

Par la suite, il sera expliqué les tests et les implémentations que réalisés tout au long du stage. Les différentes étapes de développement seront décrites, en exposant les choix technologiques et les outils. Pour chaque problème rencontré, les méthodes de résolution envisagées et les résultats obtenus suite aux tests seront détaillés. Cette section mettra en lumière le processus d'itération entre la conception, l'expérimentation et l'amélioration des solutions.

Enfin, une conclusion sera faite sur les notions apprises lors de cette expérience, en abordant à la fois les aspects techniques et personnels. Les compétences acquises seront évoquées, et il sera également expliqué comment cette expérience a contribué à approfondir mes connaissances dans le domaine de l'intelligence artificielle ainsi que mon envie de découvrir davantage le monde de la recherche.

1.1.1 Le laboratoire

Le Laboratoire Interdisciplinaire Carnot de Bourgogne (ICB) est une unité mixte de recherche sous la tutelle du CNRS et de l'Université de Bourgogne qui regroupe environ 300 physiciens, chimistes, ingénieurs et techniciens. Implanté en Bourgogne-Franche-Comté, il s'étend sur plusieurs sites dont Dijon, Le Creusot, Chalon-sur-Saône et Belfort.

L'ICB a été créé en 2007 à la suite de la fusion de deux laboratoires et d'une équipe de recherche de l'Université de Bourgogne. En 2017, il s'est encore renforcé avec l'intégration de deux groupes issus de l'Université de Technologie de Belfort-Montbéliard (UTBM). L'ICB collabore étroitement avec des entreprises pour développer des projets de recherche ayant des applications directes dans l'économie régionale et nationale.

Les thèmes de recherche de l'ICB sont centrés sur la physique, la chimie physique, l'ingénierie des matériaux et le numérique. Les principales orientations scientifiques du laboratoire, formant un continuum de la recherche fondamentale aux applications industrielles, sont portées par six départements et quatre plateformes technologiques en continuelle évolution avec l'acquisition de nouveaux équipements toujours plus performants.

Le stage s'est déroulé dans le bâtiment de l'Institut Marey Maison de la Métallurgie (I3M) à Dijon, au sein de l'équipe Vision and Artificial Intelligence (VAI). Les travaux de l'équipe VAI au sein du laboratoire de l'ICB sont axés sur la recherche en vision par ordinateur, en apprentissage profond et vision pour la robotique et intelligence artificielle pour la Physique. L'équipe, dirigée par le Pr. Dominique Ginhac, est constituée de 5 enseignants-chercheurs permanents, dont 3 professeurs et 2 maîtres de conférence. 13 contractuels font partie de l'équipe, dont un post-doctorant, 12 doctorants et plusieurs autres stagiaires de M1 ou M2.

L'encadrement du stage a été effectué par Dominique Ginhac, professeur à l'ICB, Manon Ansart, maîtresse de conférence au LEAD et Fan Yang, professeure au LEAD.

1.1.2 Le sujet

Le sujet du stage se situe au sein du projet ANR aiMotions (Artificial Intelligence for food-related eMotions analysis), qui vise à décrypter dans quelle mesure les émotions suscitées par les aliments sous-tendent les comportements et choix alimentaires. Pour répondre à cet objectif, un modèle d'analyse des émotions faciales utilisant les toutes dernières techniques d'intelligence artificielle (IA) va être développé.

En plus de l'ICB, le Centre des Sciences du Goût et de l'Alimentation (CSGA) participe au projet et a pour but de mettre en place une expérience qui sera utilisée pour réaliser un jeu de données. Ce jeu de données contiendra des séquences vidéos montrant des participants interagir avec des aliments et sera utilisé afin d'entraîner le modèle d'analyse des émotions faciales.

L'objectif du stage au sein de ce projet est de réaliser des revues de littérature scientifique afin de sélectionner les modalités les plus pertinentes à la reconnaissance d'émotions faciales, ainsi que les derniers modèles d'IA montrant les meilleurs résultats pour l'extraction de ces modalités. J'aurai également pour but de participer à la mise en œuvre du dispositif d'acquisition de séquences vidéos qui sera utilisé afin de construire le jeu de données.

1.2 Revue de la littérature

1.2.1 Les émotions

Classification des émotions

Plusieurs modèles existent afin de catégoriser et décrire des émotions. La roue des émotions de Plutchik décrit huit émotions de base et les différentes façons dont elles interagissent entre elles, y compris celles qui sont opposées et celles qui peuvent facilement se transformer en une autre émotion, comme on peut le voir dans la Figure 1.1.

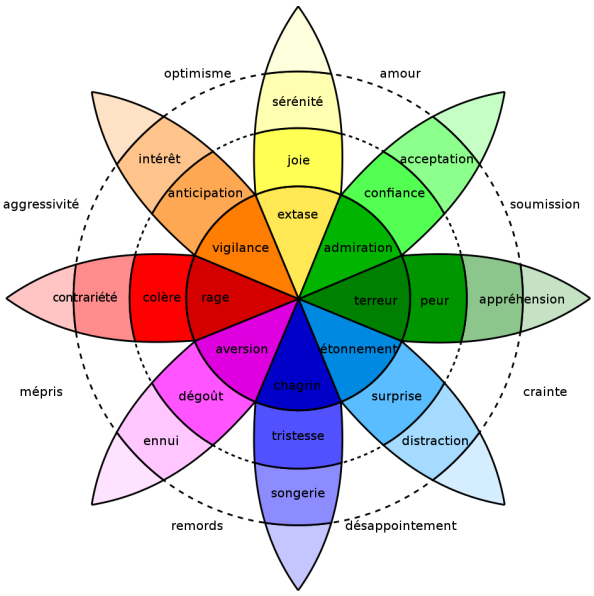

FIGURE 1.1 – Roue des émotions de Plutchik

Il est également possible d'utiliser le Circumplex Model of Affect (Figure 1.2) [Russell, 1980] pour classifier des émotions. Ce modèle réparti les émotions dans un espace circulaire à deux dimensions comprenant les axes d'activation (émotion intense/faible) et de valence (émotion agréable/désagréable). Chaque émotion est une combinaison linéaire de ces deux dimensions. La valence représente l'axe horizontal et l'activation représente l'axe vertical. Bien que la plupart des modèles de détection d'émotions utilisent une discrétisation comme celle proposée par la roue des émotions de Plutchik, certains modèles utilisent ce Circumplex Model of Affect pour définir une émotion.

Déclenchement des émotions

Lorsque l'on doit provoquer des émotions chez les participants, on utilise souvent des images ou des vidéos qui sont conçues pour déclencher des réactions émotionnelles. Par exemple, une image d'un aliment pourri depuis plusieurs semaines provoquera du dégoût chez la personne qui la regarde. En s'appuyant sur ce principe, on peut créer des jeux de données en sélectionnant

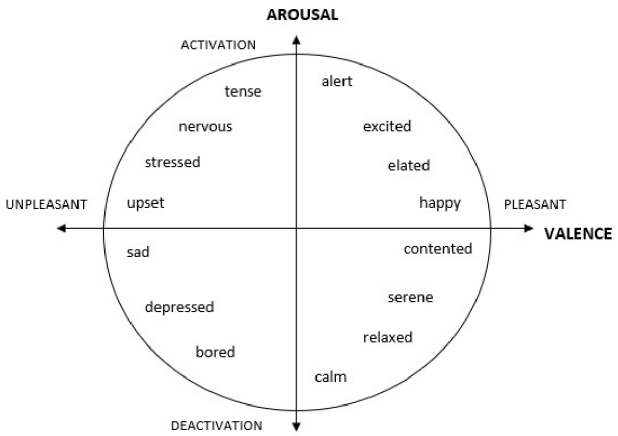

FIGURE 1.2 – Circumplex Model of Affect

des images et des vidéos capables de générer des émotions spécifiques, comme la joie, la peur ou la tristesse.

Le jeu de données IAPS (International Affective Picture System) [Lang and Bradley, 2007] est l'un des jeux de données les plus utilisés pour susciter des émotions à partir d'images. Ce jeu de données contient environ 1 195 images, couvrant un large éventail d'émotions, incluant la joie, la tristesse, la peur, le dégoût, la colère et la surprise.

Détection des émotions

Les macro-expressions (MaE) sont des expressions du visage qui peuvent être volontaires ou non, et qui durent en général entre une demi seconde et quelques secondes. Les macro-expressions couvrent une partie du visage relativement large.

Les micro-expressions (ME) sont involontaires, rapides et peuvent refléter le vrai sentiment d'une personne de façon non intentionnelle, à l'inverse des macro-expressions qui peuvent être simulées [Ekman and Friesen, 1969]. La durée d'une micro-expression est comprise entre quelques dizaines de millisecondes et une demi seconde, et elle ne concerne que de petites surfaces du visage. Cela rend leur détection particulièrement difficile, même pour les experts entraînés à les détecter, et par conséquent rend compliqué la création de gros jeux de données utiles à l'analyse par des réseaux de neurones profonds.

Récemment, la recherche dans le domaine de la détection d'émotion a également fait preuve d'intérêt envers les signaux physiologiques, car ceux-ci trahissent avec certitude des émotions induites par le système nerveux autonome. Ce système nerveux contrôle notamment le rythme cardiaque, depuis lequel il est donc possible d'interpréter ses changements d'états internes [Gouizi et al., 2011].

[Ferdinando et al., 2014] obtiennent une précision de 48% et de 43% pour la mesure de l'excitation et de la valence lors du visionnage de vidéos, en classifiant les émotions dans l'espace excitation-valence discrétisé en 3 catégories. [Shu et al., 2020] atteignent plus de 80% de précision avec une montre connectée utilisant la photopléthysmographie (PPG) pour déterminer si une émotion est positive, neutre ou négative lorsqu'une personne regarde une vidéo.

[Ménard et al., 2015] parviennent à une précision de 89% lors de la classification de 6

émotions primaires à partir du rythme cardiaque. [Nardelli et al., 2015] arrivent à un résultat de 84% de précision pour classifier les émotions dans l'espace excitation-valence lors de l'écoute de musique, d'après la mesure du rythme cardiaque par électrocardiogramme (ECG).

[Guo et al., 2016] classifient 5 émotions à partir de la variation du rythme cardiaque (HRV) calculée par PPG. L'analyse de ce HRV permet de classifier les émotions avec une précision de 71,4% pour les états positifs/négatifs et de 56,9% pour les cinq états. [Benezeth et al., 2018] utilisent aussi la PPG pour mesurer le HRV d'une personne, mais avec de la photopléthysmographie à distance (rPPG) à partir de vidéo du visage d'une personne, de façon similaire à [Li et al., 2019]. Ils analysent ensuite l'état d'excitation par rapport à l'état neutre, et montrent que dans 75% des cas il est possible de distinguer l'état neutre et l'état d'excitation.

D'autres études ont montré qu'il était possible de détecter des émotions grâce aux yeux, car le système nerveux interne du corps humain contrôle la taille de la pupille ou les variations de mouvements des yeux [Picard, 2000].

En utilisant un système d'eye-tracking, [Lanatà et al., 2011] montrent qu'il est possible de distinguer les états neutres et les états d'excitation à partir de la durée de fixation, la direction du regard et de la taille et des mouvements des pupilles.

[Aracena et al., 2015] classent les émotions en trois catégories selon la valence : positif, neutre ou négatif. Lors de leur expérimentation, ils ont pu obtenir 71% de précision en mesurant la taille de la pupille et la position du regard lorsque les sujets regardaient des images du jeu de données IAPS.

D'après ces publications, il est possible de conclure que l'intensité de la réponse pupillaire augmente proportionnellement à la force de la valence émotionnelle, qu'elle soit très négative ou très positive [Oliva and Anikin, 2018].

1.2.2 Les micro et macro expressions

Pour formaliser la classification des expressions, le Facial Action Coding System (FACS) [Ekman and Friesen, 1978] a été créé afin d'avoir un système de notation dédié à l'analyse des mouvements du visage. Dans ce système, la partie du visage où l'on observe le mouvement est désignée comme une Unité d'Action ou Action Unit (AU). Le FACS se compose de 46 AU, dont 12 sont associées à la partie supérieure du visage, 18 à la partie inférieure, et les AU 1 à 7 se réfèrent spécifiquement aux sourcils, au front ou aux paupières. D'après la combinaison des mouvements de ces AU on peut identifier six émotions de base : la colère, le dégoût, la peur, le bonheur, la tristesse et la surprise. Le FACS est à la fois applicable pour la détection des macro-expressions et des micro-expressions.

Les modèles

Afin d'améliorer l'état de l'art en reconnaissance de micro-expressions, qui a toujours une précision relativement faible, le Micro-Expression Grand Challenge (MEGC) propose tous les ans un challenge qui a pour but de repérer ces macro et micro-expressions à partir de longues vidéos. Ces longues vidéos permettent d'avoir une meilleure analyse des micro-expressions faciales dans des scénarios complexes du monde réel, où les micro-expressions sont susceptibles

d'être entrecoupées avec des expressions normales.

Parmi les derniers modèles présentant les meilleurs résultats sur ces jeux de données se trouve Integrating VideoMAE based model and Optical Flow for Micro- and Macro-expression Spotting [Xu et al., 2023], qui a remporté la 1ère place au MEGC de 2023, et propose une méthode de détection basée sur de l'apprentissage profond auto-supervisé qui utilise à la fois les macro et les micro expressions, en s'aidant notamment du flux optique (calculé avec la méthode Main Directional Mean Optical) pour repérer la localisation temporelle des expressions dans les vidéos.

Ensuite, Micro-Expression Spotting with Face Alignment and Optical Flow [Qin et al., 2023] remporte la 2ème place du MEGC de 2023 en alignant les visages, puis en définissant des régions d'intérêt à partir de statistiques des Action Unit. D'après ces statistiques, un flot optique avec la méthode de Farneback sera calculé pour générer des propositions d'expressions.

Efficient Micro-Expression Spotting Based on Main Directional Mean Optical Flow Feature [Yu et al., 2023] a remporté la 3ème place au MEGC 2023 en adoptant une approche similaire à [Qin et al., 2023]. Les vidéos sont alignées sur les visages, puis un flot optique est calculé avec la méthode Main Directional Mean Optical (de la même façon que [Xu et al., 2023]) pour génerer des propositions d'expressions.

Micro-Expression Recognition Enhanced by Macro-Expression from Spatial-Temporal Domain (2021) [Xia and Wang, 2021] propose un modèle de transformation macro-vers-micro pour la reconnaissance des micro-expressions en utilisant deux modèles pré-entraînés sur des données de micro-expressions et de macro-expressions, et propose ensuite une méthode pour aligner les caractéristiques spatiales et temporelles des deux modèles.

CapsuleNet for Micro-Expression Recognition (2019) [Quang et al., 2019] est un modèle basé sur CapsuleNet [Sabour et al., 2017] qui a pour but d'extraire seulement les images avec la plus grande intensité de micro expression dans les vidéos, puis d'entraîner CapsuleNet à classifier ces images.

Les jeux de données

Les tableaux 1.1 et 1.2 présentent respectivement les listes des jeux de données les plus populaires pour l'analyse des macro expressions et des micro expressions.

Lors du MEGC de 2023, les jeux de données recommandés pour la détection de micro-expressions étaient SAMM Long, CAS(ME)^2, SMIC-E-Long, CAS(ME)^3 et 4DME.

1.2.3 Variabilité du rythme cardiaque

Afin d'acquérir le rythme cardiaque d'une manière non intrusive, la photopléthysmographie (PPG) est utilisée car elle permet de détecter les impulsions du volume sanguin cardiovasculaire (BVP) grâce aux variations de la lumière transmise ou réfléchie sur la peau [Allen, 2007]. Il a été montré qu'il était possible de faire cette acquisition en n'utilisant aucune source de lumière additionnelle et en n'utilisant que la lumière ambiante [Verkruysse et al., 2008, Poh et al., 2011]. En utilisant la lumière ambiante comme source de lumière, il nous sera donc possible de faire la mesure du rythme cardiaque grâce à une caméra standard. Cette technique est appelée rPPG (remote PPG) ou iPPG (imaging PPG).

Nom	# Expressions	# Sujets	Résolution	FPS	Type	Émotions	# d'AU
AffectNet [Mollahosseini et al., 2019]	1M (400k labellisées)		$\sim 425 \times 425$		Images RGB	Hap, Sad, Sur, Fea, Ang, Dis, Contempt, Neutral	0
CK+ [Lucey et al., 2010]	593 (327 labellisées)	123	640×480	30	Vidéos RGB	Ang, Dis, Fea, Hap, Sad, Sur, Contempt	46
FER2013 [Goodfellow et al., 2013]	30k		48×48		Image Grayscale	Ang Dis Fea Hap Sad Sur Neutral	0
DISFA [Mavadati et al., 2013]		27	1024×768	20	Vidéos RGB	Hap, Sur, Dis, Sad, Fea	12
RAF-DB [Li et al., 2017]	30k		224×224		Images RGB	Sur, Fear, Dis, Hap, Sad, Ang, Neutral	15
FER+ [Barsoum et al., 2016]	35k		48×48		Images Grayscale	Hap, Sur, Sad, Ang, Dis, Fea, Contempt, Neutral	0
BP4D [Zhang et al., 2014]	2500	41	1040×1392	25	Images en 3D	Ang, Dis, Fea, Hap, Sad, Sur, Embarrassment, Pain	46
MMI Facial Expression Database [Pantic et al., 2005]	2900	75	720×576	24	Vidéos RGB	Ang, Dis, Fea, Hap, Sad, Sur	46

TABLE 1.1 – Liste des jeux de données de macro expressions

Nom	# ME	# Sujets	Résolution	FPS	Type	Émotions	# d'AU
CAS(ME)^3 [Li et al., 2023a]	1109		30				
CAS(ME)^2 [Qu et al., 2018]	57	22	640×480	30	RGB	Pos (8) Neg (21) Sur (9) N/A (19)	28
4DME [Li et al., 2023b]	1068	56	1200x1600 640×480 640×480 640×480	60 60 30 30	4D Grayscale RGB Depth	Pos (34) Neg (127) Sur (30) Rep (6) PS (13) NS (8) RS (3) PR (8) NR (7) N/A (31)	22
SAMM [Davison et al., 2018]	32	159	2040×1088	200	RGB	Hap (24) Ang (20) Sur (13) Dis (8) Fea (7) Sad (3) N/A (84)	46
SAMM-Long [Yap et al., 2020]	147	32	2040×1088	200	RGB		
MMEW [Ben et al., 2021]	300	36	1920×1080	90	RGB	Hap (36) Ang (8) Sur (89) Dis (72) Fea (16) Sad (13) N/A (66)	17
CASME II [Yan et al., 2014]	247	26	640×48	2000	RGB	Hap (33) Sur (25) Dis (60) Rep (27) N/A (102)	11+
CASME [Yan et al., 2013]	195	19	640×480 1280×720	60 60	RGB	Hap (5) Dis (88) Sad (6) Con (3) Fea (2) Ten (28) Sur (20) Rep (40)	12+
SMIC [Li et al., 2013]	164 71 71	16 8 8	640×480 640×480 640×480	100 25 25	RGB NIR RGB	Pos (51) Neg (70) Sur (43) Pos (28) Neg (23) Sur (20) Pos (28) Neg (24) Sur (19)	0
MEVIEW [Husak, 2017]	31	16	720×1280	30	RGB	Hap (4) Ang (1) Sur (9) Dis (1) Fea (2) Unc (7) Con (5)	7

TABLE 1.2 – Liste des jeux de données de micro expressions

Une fois le PPG récupéré depuis le BVP, on peut ensuite analyser ce signal afin d'obtenir la variabilité du rythme cardiaque (HRV), qui est une mesure des variations de la fréquence cardiaque d'un battement à un autre.

Plusieurs méthodes existent pour extraire le signal PPG à partir d'une caméra : les méthodes "classiques" qui effectuent différents traitements du signal après avoir extrait les zones d'intérêt du visage, ainsi que les méthodes utilisant des réseaux de neurones profonds de bout en bout.

Dans [Pai et al., 2018, Pai et al., 2021] les auteurs utilisent la combinaison spatiale et la démodulation de fréquence pour obtenir le HRV à partir de la fréquence instantanée du signal BVP. [Chen and McDuff, 2018] propose un réseau convolutionnel avec un mécanisme

d'attention pour détecter les régions d'intérêt et y récupérer les signaux BVP et respiratoires.

[Li et al., 2019, Gudi et al., 2020, Qiao et al., 2021, Shoushan et al., 2021] capturent un visage et suivent les pixels de la peau du visage dans une région d'intérêt. Un signal unidimensionnel est extrait des valeurs moyennes spatiales de ces pixels au fil du temps avec différentes méthodes (méthode de la chrominance ou méthode POS). Ensuite, différents traitements du signal sont effectués, comme du filtrage par bande ou de la normalisation du signal, et enfin une détection des pics du signal est faite pour extraire le signal BVP.

[Yu et al., 2019] propose deux réseaux spatiotemporel de bout en bout prenant des vidéos de visages brutes. Deux méthodes de modélisation spatio-temporelle sont explorées et comparées : 3DCNN et RNN. À partir de ceci, [Botina-Monsalve et al., 2022] propose une adaptation du réseau 3DCNN de [Yu et al., 2019], se concentrant sur le meilleur compromis entre temps d'inférence et précision.

Les auteurs de [Niu et al., 2020] proposent un système de mesure de rythme cardiaque qui peut gérer les variations de la pose, de l'éclairage et des différences entre les appareils. Ils proposent un réseau de neurones profond de bout en bout qui va analyser des cartes spatio-temporelles formées à partir des zones d'intérêt du visage.

Dans [Kuang et al., 2022], la méthode proposée vise à améliorer la qualité du signal récupéré et réduire la complexité temporelle du réseau. La méthode intègre un bloc d'attention léger appelé 3D Shuffle Attention (3D-SA), qui permet au réseau de capturer les dépendances au niveau des pixels et des canaux, et d'améliorer la précision du réseau. [Zhao et al., 2023] utilisent eux aussi un mécanisme d'attention et proposent une méthode qui combine l'analyse multi-échelle et les réseaux de neurones pour extraire les signaux de pouls à différentes échelles.

Enfin, [Yu et al., 2022] présentent un réseau de bout en bout basé sur l'architecture transformer qui retourne un signal rPPG d'après un flux vidéo.

rPPG-Toolbox [Liu et al., 2023] introduit un framework en Python regroupant de nombreuses méthodes de mesure de rPPG dont une majorité de celles citées précédemment, ainsi que des modèles pré-entrainés sur plusieurs jeux de données populaires.

1.2.4 Segmentation de l'iris

Puisque la taille de l'iris est corrélée avec les changements d'émotions, il est nécessaire d'en extraire ses paramètres. Pour ce faire, une segmentation d'image de l'iris est effectuée, d'après laquelle sera calculé le diamètre de la pupille. La figure 1.3 montre les différentes parties composant un œil, dont l'iris et la pupille. Les jeux de données d'iris les plus populaires sont recensés dans le tableau 1.3.

[Chen et al., 2019] introduisent Dense-Fully Convolutional Network (DFCN), un modèle de segmentation d'iris qui combine les Fully Convolutional Networks (FCNs) avec des blocs denses. L'article souligne l'utilisation des blocs denses, qui sont utilisés pour atténuer les problèmes de surapprentissage et de disparition du gradient, assez présent à cause des tailles relativement petites des jeux de données de segmentation d'iris.

IrisParseNet [Wang et al., 2020] est basé sur de l'apprentissage profond, développé pour améliorer la précision et l'efficacité de la segmentation dans des environnements non coopératifs.

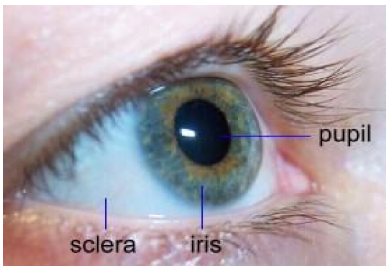

FIGURE 1.3 – Image d'un œil avec l'iris, la pupille et la sclère
SOURCE : [Alhamrouni, 2017]

Cette approche utilise un réseau d'attention multitâche pour prédire simultanément les masques de l'iris, de la pupille et les contours externes de l'iris. Le processus est complété par un post-traitement qui permet de déterminer les paramètres du contour de l'iris, facilitant l'extraction des caractéristiques de l'iris.

Iris R-CNN [Feng et al., 2022] propose un réseau basé sur Mask R-CNN [He et al., 2018] et de son principe de création de propositions de régions, qui seront classifiées pour ensuite identifier et localiser les objets dans une image. Le réseau utilise notamment des ancres en forme de double cercle correspondant aux cercles de la pupille et de l'iris, au lieu des propositions rectangulaires de Mask R-CNN. Iris R-CNN a pour but d'améliorer la précision de la reconnaissance de l'iris en identifiant explicitement les cercles de la pupille et de l'iris dans le réseau, au lieu de se limiter à la simple segmentation. La précision annoncée est très proche de IrisParseNet. Iris R-CNN obtient une meilleure précision sur le jeu de données MICHE (17.9% d'erreur de segmentation contre 29.2% pour IrisParseNet), tandis que IrisParseNet obtient une meilleure précision sur NICE-II (11% contre 12.3% pour Iris R-CNN).

Iris-SAM [Farmanifard and Ross, 2024] parle de l'évolution des techniques de segmentation et de l'impact des modèles de fondation, dont Segment Anything Model (SAM) de Meta. Ces modèles de fondation, grâce à leur capacité à généraliser à partir de grands ensembles de données et à s'adapter facilement à plein de tâches différentes, ont amélioré les performances de certains domaines de segmentation d'images. L'article montre l'intérêt de SAM dans la segmentation de l'iris, une tâche complexe en raison des particularités de chaque œil et des conditions d'acquisition d'image très variable. En fine-tunant SAM avec des jeux de données de segmentation d'iris, le modèle démontre son potentiel pour améliorer la précision de cette segmentation.

Nom	# Images	# Sujets	Résolution	Environnement	Couleur
Casia-Iris V1 [Chinese Academy of Sciences' Institute of Automation, 2010]	756	108	320 × 280	Très contrôlé	NIR
Casia-Iris V3 [Chinese Academy of Sciences' Institute of Automation, 2010]	22 034	700	320 × 280 640 × 480	Peu contrôlé	NIR
Casia-Iris V4 [Chinese Academy of Sciences' Institute of Automation, 2010]	44k réelles 10k virtuelles	1800 réel 1000 virtuels	320 × 280 640 × 480 2352 × 1728	Peu contrôlé	NIR
MMU Iris Database [Multimedia University, 2006]	460	46	320 × 240	Très contrôlé	NIR
IITD v1 [IIT Delhi, 2007]	1120	224	320 × 240	Peu contrôlé	NIR
OpenEDS [Garbin et al., 2019]	12759	152	640 × 480	Peu contrôlé	NIR
LPW [Tonsen et al., 2016]	130 856	22	640 × 480	Non contrôlé	NIR
Q-FIRE [Johnson et al., 2010]	390	195	640 × 480	Non contrôlé	NIR
PupilNet [Fuhl et al., 2016a]	41217		384 × 288	Non contrôlé	NIR
ExCuSe [Fuhl et al., 2015]	17 vidéos			Non contrôlé	NIR
ElSe [Fuhl et al., 2016b]	55712			Non contrôlé	NIR
UBIRIS.v2 [Proenca et al., 2010]	11 102	261	800 × 600	Non contrôlé	Visible
MICHE [De Marsico et al., 2015]	3732	92		Non contrôlé	Visible
NVGaze [Kim et al., 2019]	2.5M réelles 2M virtuelles	35 réels			Visible
RITEyes [Nair et al., 2020]					Visible
U2Eyes [Porta et al., 2019]	6M virtuels	1000			Visible
GI4E [Ponz et al., 2012]	1339	103	800 × 600	Non contrôlé	Visible
BioID [BioID, 2022]	1521	23	384 × 286	Non contrôlé	Visible

TABLE 1.3 – Liste des jeux de données d'iris

1.3 Comparaison des méthodes

Une autre étape importante du travail réalisé a consisté en la création d'un framework en Python, conçu pour regrouper et comparer les différentes méthodes identifiées lors des recherches bibliographiques. Ce framework a été pensé pour être un outil flexible et extensible, servant de base à l'extraction des caractéristiques du visage pour la future création du modèle d'IA. Il permettra non seulement d'intégrer les algorithmes et techniques pertinents, mais aussi de les évaluer de manière cohérente et systématique à la fois avec des jeux de données publiques populaires et avec les futures données du projet.

Afin de créer ce framework, les méthodes explicitées ci-dessous ont été implémentées une à une et ensuite regroupées au sein du même code.

1.3.1 Variabilité du rythme cardiaque

Méthode non supervisée

Les premières tentatives pour extraire le HRV depuis une caméra ont été faites en reproduisant des méthodes non supervisées, ne nécessitant pas de jeux de données. Pour ce faire, Measuring Heart Rate and Heart Rate Variability with Smartphone Camera [Qiao et al., 2021] et HRVCam : robust camera-based measurement of heart rate variability [Pai et al., 2021] sont les plus pertinents afin de récupérer ce HRV dans des conditions similaires à celles où seront faites les acquisitions par la suite, c'est-à-dire en utilisant comme source de données des caméras RGB de qualité moyenne et en faisant l'acquisition du visage dans des environnements non contrôlés.

À la différence de ces deux méthodes, [Qiao et al., 2021] proposent d'extraire du visage une zone commençant au-dessus de la bouche jusqu'en dessous des yeux, afin d'avoir un rec-

tangle qui recouvre le nez et les joues et d'éviter le bruit induit par les mouvements indésirables des yeux et de la bouche. Dans un objectif similaire, [Pai et al., 2021] propose d'extraire 4 zones d'intérêt : les 2 joues et deux zones à gauche et à droite du front.

Ainsi, plusieurs outils ont été testés pour extraire différentes zones d'intérêt du visage, en commençant par les outils fournis par OpenCV pour détecter le visage, avec la méthode de Haar Cascade [Viola and Jones, 2001]. Avec cette méthode, le visage est détecté, mais la zone du visage détectée avec OpenCV n'est pas stable et bouge beaucoup d'une image à une autre, avec parfois des images où la détection est fausse. Cela rend compliqué l'analyse de l'évolution d'une même zone avec précision et induit beaucoup de bruit.

YuNET [Wu et al., 2023], un modèle populaire de détection de visage léger et performant, a été testé et a produit une détection avec moins d'imprécision, mais avec toujours trop de mouvements indésirables sur la zone du visage détectée entre chaque image de la vidéo.

MediaPipe [Lugaresi et al., 2019], un framework simplifiant le développement d'applications de perception, a produit de meilleurs résultat. Ce framework peut notamment effectuer une extraction des repères du visage (Face Landmarks) avec beaucoup plus de précision, car MediaPipe suit les mouvements de 478 repères du visage dans une vidéo. D'autres méthodes d'extraction des Face Landmarks auraient pu être utilisées pour obtenir des résultats similaires, mais MediaPipe s'est montré efficace et précis pour l'extraction de la position du visage tout au long d'une vidéo.

Grâce à ces repères du visage, la zone contenant la peau du visage a pu être extraite avec précision. Sur ce visage, les méthodes suivantes ont pu être testés avec les différentes zones citées précédemment.

Un traitement du signal a ensuite été effectué pour extraire uniquement les fréquences pertinentes pouvant correspondent à un rythme cardiaque, avec un filtre passe bande entre 0.7 et 3.5 Hz, c'est-à-dire entre 42 et 210 BPM. Un filtre de Butterworth a été utilisé pour le filtrage du signal, afin d'éviter le phénomène de Gibbs, qui peut provoquer des distorsions du signal. Les 3 caneaux rouges, vert et bleu des images ont après été combinés avec la méthode de la chrominance [Pai et al., 2021].

Les pics du signal ont ensuite été extraits avec la bibliothèque SciPy et utilisés pour calculer le BPM du sujet.

Les résultats présentés dans la figure 1.4 montrent qu'avec cette méthode, il est possible d'extraire la tendance du rythme cardiaque. Cependant, la précision du rythme cardiaque extrait reste peu élevée.

rPPG-Toolbox

rPPG-Toolbox [Liu et al., 2023] est un framework destiné à standardiser et à faciliter la recherche dans le domaine de la mesure physiologique par caméra pour les signaux cardiaques et pulmonaires. rPPG-Toolbox propose une solution complète qui offre un ensemble d'outils open-source incluant le chargement de jeux de données publiques, des codes de pré-traitement des données, ainsi que les implémentations des modèles supervisés et non-supervisés les plus populaires. De plus, rPPG-Toolbox permet l'évaluation et l'inférence de ces modèles. Cette boîte à outils vise à améliorer la reproductibilité des recherches, à fournir des benchmarks clairs et à offrir une base solide pour la comparaison des méthodes dans ce domaine.

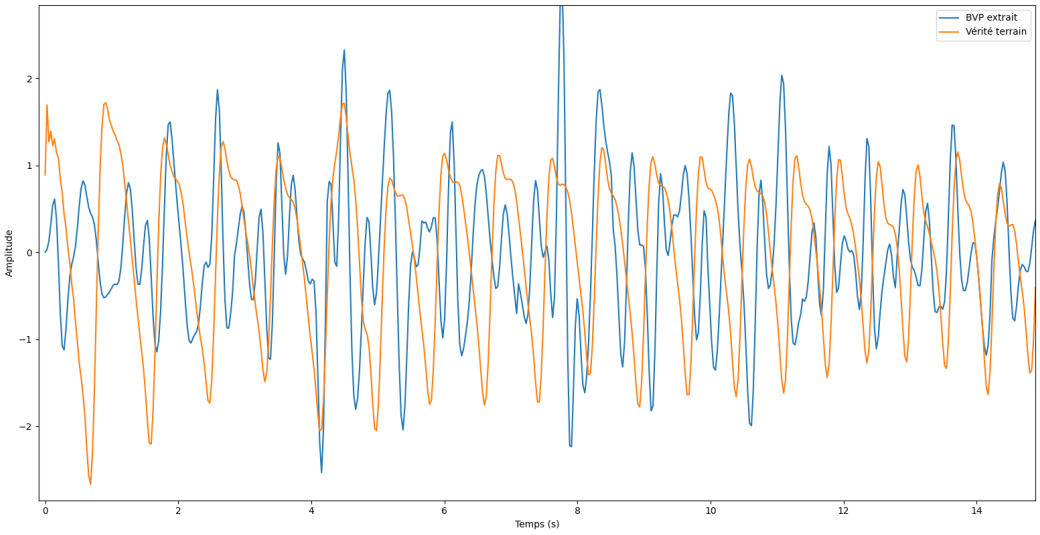

FIGURE 1.4 – Signal BVP extrait après implémentation manuelle des traitements

rPPG-Toolbox ainsi que ses méthodes ont été réutilisés au sein du framework réalisé. Il a été nécessaire de reprendre et modifier certaines parties de rPPG-Toolbox, notamment parce que ce dernier ne permettait pas de faire une inférence sur les modèles sans avoir d'étiquettes associées aux données. Un calcul de chaque métrique étant en effet calculé de façon systématique. rPPG-Toolbox ne permettait également pas d'effectuer une inférence sur une simple vidéo au lieu d'un jeu de données entier.

La figure 1.5 montre le résultat du signal BVP extrait en utilisant une des méthodes non supervisées proposée par rPPG-Toolox. Comparée à la première approche réalisée précédemment, on obtient de meilleurs résultats. En effet, dans la figure 1.4, la synchronisation entre chaque pics (représentant un battement du coeur) extraits et ceux de la vérité terrain est très variable, avec beaucoup de pics incohérents ou non interprétables. À l'inverse, les pics de la figure 1.5 sont réguliers, et presque synchronisés avec la vérité terrain.

pyVHR

Ayant un objectif similaire à rPPG-Toolbox, pyVHR [Boccignone et al., 2022] propose également un framework destiné à faciliter la mesure de signaux rPPG. pyVHR supporte également le chargement de jeux de données populaires, les méthodes classiques non supervisées, mais une seule méthode à partir d'apprentissage (MTTS-CAN).

1.3.2 Segmentation de l'iris

À la différence des mesures rPPG, il n'existe pas de framework regroupant les méthodes de segmentation d'iris les plus populaires. Ainsi, les méthodes les plus pertinentes identifiées lors de la revue bibliographique ont été reprises pour être ajoutées au sein du framework crée pour le projet.

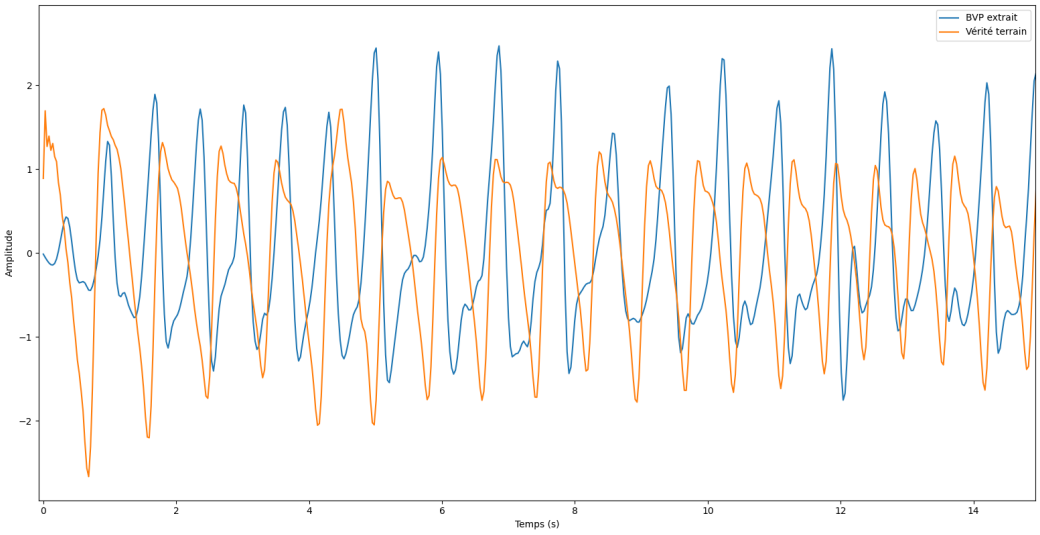

FIGURE 1.5 – Signal BVP extrait avec rPPG-Toolbox en utilisant la méthode POS [Wang et al., 2017]

IrisParseNet

IrisParseNet propose une solution complète avec un code mis à disposition et des modèles pré-entraînés sur des jeux de données comme CASIA.v4, MICHE-I et Ubiris.v2. Cependant, l'implémentation proposée est conçue avec Python 2.7, une version de Python sortie il y a un peu plus de 10 ans, et une version modifiée de Caffe, une bibliothèque Python de réseaux de neurones n'étant plus maintenue depuis 5 ans. Cela a rendu impossible l'exécution code et l'intégration du code dans le projet déjà existant. Ainsi, le code et les modèles pré-entraînés ont été convertis vers des outils plus récents, c'est-à-dire Python 3.11 et PyTorch, l'une des bibliothèques de réseaux de neurones les plus populaires.

Pour convertir les modèles pré-entraînés depuis Caffe vers PyTorch, un script nommé `caffemodel2pytorch` faisant la conversion des poids dans le format `caffemodel` vers le format `pt` a été utilisé. Ensuite, l'architecture du modèle a été reproduite avec PyTorch grâce aux détails présentés dans la publication et ajustée pour que chaque couches du réseau écrites avec PyTorch aient la même taille que celle des poids publiés. Les étapes de pré-traitement et post-traitement ont été de la même façon ré-implémentées depuis les détails de la publication.

La figure 1.6 montre les résultats de la segmentation de IrisParseNet, ainsi que les cercles approximatifs des zones de la pupille et de l'iris calculées avec le post-traitement. L'image de gauche montre le résultat des 3 zones de l'oeil segmenté : l'iris (en vert), la pupille (en rouge), et le contour extérieur de l'iris (en bleu). L'image de droite montre le résultat après post-traitement des cercles estimés pour l'iris et la pupille.

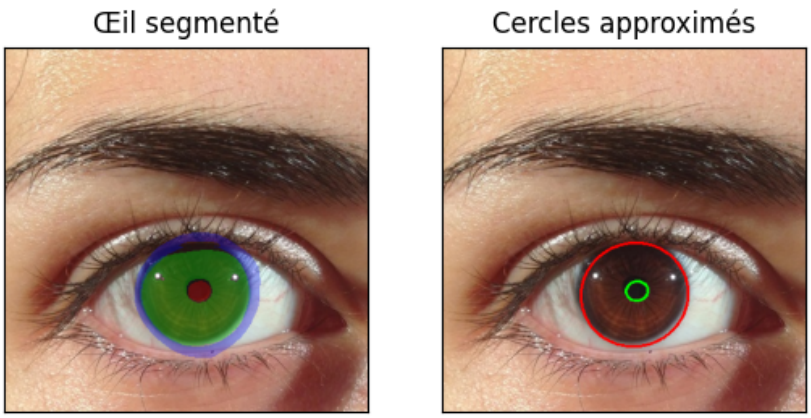

FIGURE 1.6 – Résultats de IrisParseNet

DFCN

DFCN [Chen et al., 2019] montre également des résultats intéressants, comparable avec IrisParseNet et Iris R-CNN. Cependant, les auteurs n'ayant pas publié le code source pour reproduire les résultats, le modèle a été ré-implémenté d'après les informations de la publication.

Le modèle a été implémenté avec PyTorch, en suivant précisément chaque détail donné afin de reproduire la même architecture. Les auteurs de DFCN utilisent les jeux de données UBIRIS.v2, CASIA-Interval-V4 et IITD. Cependant, puisque ces trois jeux de données ne fournissent pas la vérité terrain de la segmentation de l'iris, mais seulement des images d'iris, cette vérité terrain a été récupérée grâce à IRISSEG [Hofbauer et al., 2014], qui fournit des masques de segmentation pour plusieurs jeux de données d'iris populaires. De plus, seul le jeu de données UBIRIS.v2 a été utilisé lors des tests réalisé pendant le stage, puisque les autres jeux de données n'ont pas pu être obtenus au moment des tests.

Seulement 2250 images de UBIRIS.v2 ont été utilisées pour l'entraînement du modèle, car IRISSEG ne fournissait que 2250 masques pour le jeu de données UBIRIS.v2. Cela correspond à environ un tiers des données utilisées pour entrainer le modèler dans la publication de DFCN.

Malgré les efforts pour essayer de faire correspondre le nombre de poids des blocs du réseau de neurones réalisé et le nombre de poids des blocs de la publication, des différences ont été remarquées. En effet, il est annoncé que les Dense Blocks du réseau ont 2 couches de convolutions de taille 3×3, avec 48 noyaux et un pas de 1. La formule pour calculer le nombre de poids d'une couche de convolution est la suivante :

$$\text{Nombre de poids} = (m \times n \times d + 1) \times k \tag{1.1}$$

Avec m la largeur du filtre, n la hauteur du filtre, d le nombre de canaux d'entrée et k le nombre de filtres dans la couche de convolution.

Pour le premier Dense Block, celui-ci a deux couches de convolution de taille 3×3 avec 48 filtres, avec 64 canaux en entrée, ce qui fait donc $(3 \times 3 \times 64 + 1) \times 48 + (3 \times 3 \times (64 + 48) + 1) \times 48 =$

76128 paramètres. Cependant, la publication annonce 161000 paramètres pour ce même Dense Block. Cette incohérence se reproduit pour chaque Dense Block, où le nombre de poids cité dans la publication ne correspond pas au nombre théorique.

Après plusieurs jours de tentatives pour arriver à reproduire le modèle, une demande a été adressée aux auteurs de la publication afin d'obtenir des informations supplémentaires susceptibles de résoudre le problème, mais aucune réponse n'a été reçue.

Avec cette différence dans l'architecture du modèle, les résultats se sont avérés être mauvais. La sortie du modèle ne produit aucun résultat de segmentation, même de façon imprécise, comme le montre la figure 1.7.

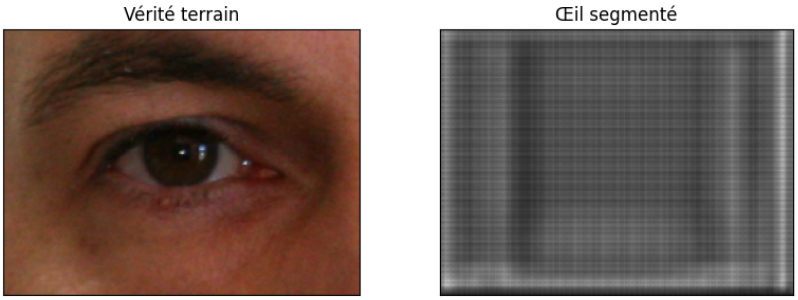

FIGURE 1.7 – Résultats de Iris DFCN

Iris-SAM

En plus des modèles précédents, Iris-SAM [Farmanifard and Ross, 2024] a aussi été testé. La figure 1.8 montre les résultats de la segmentation. La zone de l'iris segmentée est très proche de la vérité terrain, avec seulement un petit peu de bruit sous les cils. Il est possible d'extraire facilement de cette segmentation les paramètres des cercles intérieur et extérieur de l'iris de façon précise.

Un des inconvénients d'Iris-SAM est qu'il faut lors de la segmentation de l'iris non seulement l'image de l'œil, mais aussi une bounding box (c'est-à-dire deux points qui délimitent une zone) pour guider la segmentation. Cela correspond au rectangle vert que l'on peut voir sur la figure 1.8 Cela signifie que les résultats peuvent varier considérablement pour une même image en fonction de la bounding box fournie.

De plus, Iris-SAM a été entraîné uniquement sur des jeux de données composées d'image d'iris dans le domaine de couleur proche infrarouge.

Autres tests

En plus des modèles cités précédemment, d'autres tests ont également été effectués à partir de codes publics. Ces tests ont été effectués avant la ré-implémentation des modèles cités précédemment, pour tenter d'obtenir des premiers résultats rapidement.

PuPal-Beta a paru être une solution simple pour obtenir de premiers résultats de segmentation d'iris, à partir d'une vidéo de visage. Pour obtenir cette segmentation, les yeux ont été

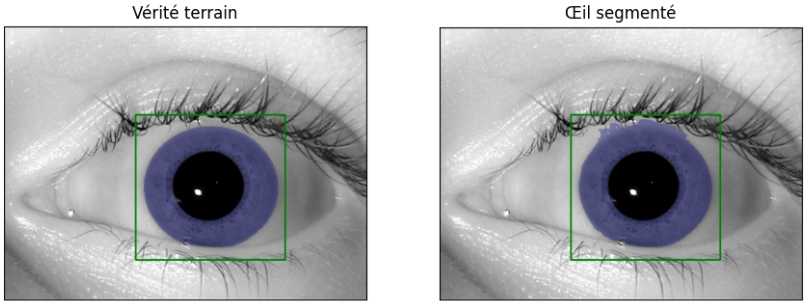

FIGURE 1.8 – Résultats de Iris-SAM

récupérés avec la méthode Haar Cascade, puis une segmentation avec U-net a été appliquée. Malheureusement, aucun résultat n'a pu être reproduit, car le modèle pré-entraîné publié n'est plus disponible à cause d'un lien expiré.

irisSeg propose de récupérer les paramètres de l'iris, c'est-à-dire son rayon minimal et maximal, en utilisant l'algorithme de Daugman [Oad and Ahmad, 2012]. Une bibliothèque Python est proposée, qui n'a pas pu être rendue fonctionnelle à cause de problèmes de dépendances des bibliothèques trop vieilles.

[Eivazi et al., 2019] publient le code source utilisé dans leurs expérimentations. Cependant, le code n'a pas pu être utilisé, car il était nécessaire d'entraîner le modèle avec un jeu de données qui n'a pas encore pu être récupéré.

[Ahmad and Fuller, 2019] proposent une segmentation de l'iris à l'aide de CNN, en s'inspirant d'une architecture proche de Mask R-CNN, comme le fait Iris R-CNN [Feng et al., 2022]. Cependant, le lien de partage des poids du modèle publié n'est plus disponible, et l'entraînement utilise des jeux de données n'ayant pas non plus pu être récupérés.

1.4 Application sur des données réelles

1.4.1 L'expérience mise en place

Afin d'entraîner le modèle d'IA qui sera créée à l'issue du projet, il est nécessaire d'avoir un jeu de données avec des sujets manifestants divers degrés d'émotion, en jouant sur l'intensité et la valence de ces émotions. Ainsi, le CSGA a mis en place un protocole pour réaliser une expérience répondant à ces besoins.

Afin de constituer ce jeu de données, une première expérience pilote a été réalisée par le CSGA. Cette expérience avait pour objectif de recueillir des premiers retours des sujets sur les stimuli choisis, mais en raison de contraintes administratives liées à la protection des données, il n'a pas été possible de filmer ces premières sessions. Une autre expérience sera elle réalisée plus tard, où les données seront cette fois-ci acquises par caméra.

Les stimuli choisis pour déclencher les émotions dans cette expérience sont des odeurs d'arômes alimentaires et des images de nourriture. Le choix des stimuli a une importance particulière, afin de pouvoir couvrir un large spectre émotionnel pour garantir la robustesse du modèle d'IA. Les stimuli doivent non seulement évoquer des émotions fortes, mais aussi des réactions plus nuancées pour enrichir le jeu de données.

1.4.2 Installation de la caméra

Afin de préparer l'acquisition du jeu de données, qui sera réalisé plus tard, l'installation et le développement des premières étapes du logiciel d'acquisition avec la caméra a également été réalisé.

La caméra qui sera utilisée pour les acquisitions est la caméra LUCID Vision Labs Atlas10 ATX081S-CC. Elle est équipée d'un capteur CMOS Sony de 8,1 mégapixels, ayant une résolution de 2840×2840 pixels. Elle est capable d'atteindre un taux d'images par seconde (FPS) de 136.7, ce qui en fait un très bon compromis entre une grande qualité d'image et un grand taux de rafraîchissement, nécessaire pour détecter plus tard les plus petites variations du visage.

Cette caméra utilise une interface 10GigE avec prise en charge du Power over Ethernet (PoE+), permettant à la fois l'alimentation et le transfert de données via un seul câble Ethernet. La caméra est connectée à un injecteur PoE via un câble Ethernet Cat 6a, suivi d'un second câble Ethernet Cat 6a qui relie l'injecteur PoE à une carte réseau de 10 Gbps qui elle est branchée à l'ordinateur faisant les acquisitions.

Pour optimiser la connexion de la caméra, plusieurs modifications de la configuration réseau ont été appliquées. Les Jumbo Frames ont été activées, ce qui permet de transmettre de plus grandes quantités de données par paquet, réduisant le nombre de paquets envoyés. La taille du tampon de réception (receive buffer size) a aussi été augmentée pour mieux gérer les données entrantes. L'auto-négociation, qui permet aux dispositifs de choisir automatiquement la meilleure vitesse de connexion, a également été désactivée pour forcer manuellement une vitesse de connexion à 10 Gpbs.

Une fois l'installation des drivers de la caméra et du SDK permettant d'interagir avec

la caméra effectuée, des tests ont été faits en Python pour évaluer la vitesse d'acquisition. Cependant, les performances se sont avérées assez faibles, avec un taux d'images par seconde limité à 20-30 FPS. Il a donc été décidé d'utiliser la version C++ du SDK pour développer un programme capable de récupérer les images plus efficacement. Pour cela, l'API de FFmpeg en C a été utilisée pour écrire ces images dans un fichier vidéo, avec également ajout d'une légère compression afin d'éviter de trop grosses tailles de fichier (plusieurs Go pour une minute de vidéo).

1.4.3 Difficultées rencontrées

Lors des tentatives de reproduction et d'implémentation des résultats, plusieurs problèmes ont été rencontrés.

Une des principales difficultés a été de faire fonctionner les codes publiés. En effet, parmi les publications fournissant le code utilisé pour mesurer leurs résultats, beaucoup des codes sources n'ont pas été mis à jour depuis plusieurs années, et utilisent d'anciennes versions de Python ou de bibliothèques Pip. Par exemple, il est arrivé régulièrement que certaines bibliothèques ne soient plus disponibles au téléchargement, ou bien que celles-ci ne soient plus compatibles avec les drivers récents de la carte graphique utilisée.

Une fois le code rendu fonctionnel, il s'est également avéré que très peu d'auteurs fournissent le modèle entraîné, ce qui a rendu plus compliqué le test des modèles, car il a aussi été difficile d'acquérir les jeux de données utilisés pour l'entraînement des réseaux de neurones. Beaucoup de jeux de données sont annoncés comme disponibles sur demande à des fins de recherche, mais peu ont répondu aux requêtes pour obtenir les données.

Pour la segmentation d'iris, certaines publications comparent leurs résultats de segmentation auprès de jeux de données populaires, mais il est difficile de déterminer avec certitude quelles méthodes sont les plus performantes. En effet, la majorité des publications ne fournissent pas de code public, ne permettant pas de reproduire exactement les résultats. [Wang et al., 2020] et [Feng et al., 2022] utilisent par exemple des jeux de données publiques (comme UBIRIS.v2 [Proenca et al., 2010] ou MICHE-I [De Marsico et al., 2015]), mais ne publient pas le code utilisé pour implémenter les réseaux présentés dans leurs publications. Ensuite, les comparaisons entre les modèles ne sont pas forcément objectives. Il n'est pas possible de savoir si les données des ensembles d'entraînement ou de test ont été filtrés pour obtenir de meilleurs résultats, ou bien si certaines étiquettes des jeux d'entraînement ont été modifiées. Plusieurs auteurs mentionnent d'ailleurs avoir étiqueté eux-mêmes des jeux de données, sans pour autant publier ce travail. Les étapes de pré-traitement ou de post-traitement sont aussi souvent décrites de manière trop succincte, nécessitant d'interpréter ou de déduire certaines étapes, puisque la publication se concentre souvent davantage sur la méthode proposée en elle-même.

Une autre difficulté rencontrée lors des tests des modèles de segmentation d'iris a été de se procurer les jeux de données populaires d'iris. Malgré plusieurs demandes et relances, seulement les datasets UBIRIS.v2 [Proenca et al., 2010] et MMU Iris Database [Multimedia University, 2006] ont pu être obtenus.

De plus, beaucoup de méthodes se concentrent sur la segmentation de l'iris appliquée à

de l'eye tracking dans des casques de réalité virtuelle ou d'autres dispositifs d'eye tracking, qui utilisent des caméras proche infrarouge.

Grâce au framework rPPG-Toolbox, ces difficultés ont été moins présentes pour la photopléthysmographie à distance que pour la segmentation d'iris, puisque rPPG-Toolbox fournit justement une plateforme où tous les codes sources sont publics, avec le chargement des données, ainsi que les méthodes de pré-traitement et post-traitement.

Lors de l'implémentation du programme en C++, des problèmes ont également été rencontrés lors de l'enregistrement de la vidéo avec FFmpeg. Ces problèmes ont surtout été lié à l'initialisation de l'encodeur avec FFmpeg, où il a été compliqué de trouver les bons paramètres à définir. Des difficultés ont aussi été rencontrées lors de l'envoi des frames à l'encodeur de FFmpeg, où il a fallu coordonner le format de pixel des frames entre celles reçues de la caméra et celles envoyées à l'encodeur, ainsi que la synchronisation du taux de rafraîchissement des deux sources. Le tout a été réalisé en faisant de la programmation multithread pour paralléliser les tâches, afin d'éviter tout blocage lors de la réception des frames et l'encodage puis écriture par FFmpeg, chaque tâche prenant un peu de temps.

1.5 Conclusion

1.5.1 À propos du projet

À l'issue du stage, une revue littéraire des domaines de la segmentation d'iris et de la photopléthysmographie à distance a été réalisée. Les méthodes et modèles les plus pertinents ont été étudiés pour déterminer les façons les plus efficaces d'extraire certaines caractéristiques du visage afin de pouvoir y analyser par la suite des émotions.

Afin d'améliorer la pertinence des benchmarks de segmentation d'iris et de rPPG, il pourra être envisagé par la suite d'implémenter ou de tester de nouvelles méthodes qui n'avaient pas pu être utilisées à cause de jeux de données manquants. Par exemple, les résultats de Iris-SAM n'ont pas été comparés dans la publication avec le jeu de données UBIRIS.v2 ou avec le modèle Iris R-CNN. De la même manière, une fois tous les jeux de données les plus populaires obtenus, les comparaisons des méthodes seront plus pertinentes, puisque celles-ci seront effectuées sur des données plus diversifiées et permettront donc de mieux cerner avantages et inconvénients de chaque méthode.

Iris-SAM annonçait déjà des performances promettantes avec de bonnes capacités de généralisations. Récemment, la version 2 de Segment Anything Model (SAM 2) [Ravi et al., 2024] a été publiée. Cette nouvelle version est supposée apporter de meilleures performances, et permet d'effectuer une segmentation sur une vidéo. Avec une méthode similaire à celle d'Iris-SAM, il serait sûrement possible d'améliorer les résultats d'Iris-SAM, en utilisant également des jeux de données d'iris dans le domaine de couleur du visuel.

1.5.2 À propos de mon expérience personnelle

Ce stage au sein de l'équipe "Vision and Artificial Intelligence" du Laboratoire Interdisciplinaire Carnot de Bourgogne a été une expérience extrêmement enrichissante, tant sur le plan technique que personnel. J'ai eu l'opportunité de pouvoir étudier en détail des problématiques liées à l'intelligence artificielle, ce qui m'a permis de développer mes compétences en matière de recherche, de développement et de résolution de problèmes techniques.

Au-delà des compétences techniques acquises, ce stage a également été une occasion précieuse pour moi de découvrir le monde de la recherche scientifique, un environnement stimulant et exigeant qui m'a beaucoup appris sur la rigueur, la persévérance et l'importance de la collaboration. J'ai pu non seulement mettre en pratique les connaissances théoriques acquises durant mes études, mais aussi explorer de nouvelles approches et technologies qui m'étaient auparavant inconnues.

Cette expérience m'a offert une meilleure compréhension des défis et des enjeux du domaine, tout en m'inspirant à continuer à explorer et à innover dans ce secteur en pleine expansion. Elle m'a également conforté dans mon intérêt pour l'intelligence artificielle et a renforcé mon désir de poursuivre dans cette voie, puisque je commencerai en octobre une thèse sur le même sujet et dans le même laboratoire.

Bibliographie

[Ahmad and Fuller, 2019] Ahmad, S. and Fuller, B. (2019). Unconstrained Iris Segmentation Using Convolutional Neural Networks. In Carneiro, G. and You, S., editors, Computer Vision – ACCV 2018 Workshops, pages 450–466, Cham. Springer International Publishing.

[Alhamrouni, 2017] Alhamrouni, M. (2017). IRIS RECOGNITION BY USING IMAGE PROCESSING TECHNIQUES. PhD thesis.

[Allen, 2007] Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. Physiological Measurement, 28 :R1–39.

[Aracena et al., 2015] Aracena, C., Basterrech, S., Snael, V., and Velasquez, J. (2015). Neural Networks for Emotion Recognition Based on Eye Tracking Data. In 2015 IEEE International Conference on Systems, Man, and Cybernetics, pages 2632–2637, Kowloon Tong, Hong Kong. IEEE.

[Barsoum et al., 2016] Barsoum, E., Zhang, C., Ferrer, C. C., and Zhang, Z. (2016). Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution. arXiv :1608.01041 [cs] version : 2.

[Ben et al., 2021] Ben, X., Ren, Y., Zhang, J., Wang, S.-J., Kpalma, K., Meng, W., and Liu, Y.-J. (2021). Video-based Facial Micro-Expression Analysis : A Survey of Datasets, Features and Algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1–1. arXiv :2201.12728 [cs].

[Benezeth et al., 2018] Benezeth, Y., Li, P., Macwan, R., Nakamura, K., Gomez, R., and Yang, F. (2018). Remote heart rate variability for emotional state monitoring. In 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), pages 153–156, Las Vegas, NV, USA. IEEE.

[BioID, 2022] BioID (2022). The BioID Face Database.

[Boccignone et al., 2022] Boccignone, G., Conte, D., Cuculo, V., D'Amelio, A., Grossi, G., Lanzarotti, R., and Mortara, E. (2022). pyVHR : a Python framework for remote photoplethysmography. PeerJ Computer Science, 8 :e929.

[Botina-Monsalve et al., 2022] Botina-Monsalve, D., Benezeth, Y., and Miteran, J. (2022). RTrPPG : An Ultra Light 3DCNN for Real-Time Remote Photoplethysmography. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 2145–2153, New Orleans, LA, USA. IEEE.

[Chen and McDuff, 2018] Chen, W. and McDuff, D. (2018). DeepPhys : Video-Based Physiological Measurement Using Convolutional Attention Networks. arXiv :1805.07888 [cs].

[Chen et al., 2019] Chen, Y., Wang, W., Zeng, Z., and Wang, Y. (2019). An Adaptive CNNs Technology for Robust Iris Segmentation. IEEE Access, 7 :64517–64532.

[Chinese Academy of Sciences' Institute of Automation, 2010] Chinese Academy of Sciences' Institute of Automation (2010). BIT. Casia iris image database v4.

[Davison et al., 2018] Davison, A. K., Lansley, C., Costen, N., Tan, K., and Yap, M. H. (2018). SAMM : A Spontaneous Micro-Facial Movement Dataset. IEEE Transactions on Affective Computing, 9(1) :116–129. Conference Name : IEEE Transactions on Affective Computing.

[De Marsico et al., 2015] De Marsico, M., Nappi, M., Riccio, D., and Wechsler, H. (2015). Mobile Iris Challenge Evaluation (MICHE)-I, biometric iris dataset and protocols. Pattern Recognition Letters, 57 :17–23.

[Eivazi et al., 2019] Eivazi, S., Santini, T., Keshavarzi, A., Kübler, T., and Mazzei, A. (2019). Improving real-time CNN-based pupil detection through domain-specific data augmentation. In Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, ETRA '19, pages 1–6, New York, NY, USA. Association for Computing Machinery.

[Ekman and Friesen, 1969] Ekman, P. and Friesen, W. V. (1969). Nonverbal Leakage and Clues to Deception.

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

[Farmanifard and Ross, 2024] Farmanifard, P. and Ross, A. (2024). Iris-SAM : Iris Segmentation Using a Foundation Model. arXiv :2402.06497 [cs].

[Feng et al., 2022] Feng, X., Liu, W., Li, J., Meng, Z., Sun, Y., and Feng, C. (2022). Iris R-CNN : Accurate iris segmentation and localization in non-cooperative environment with visible illumination. Pattern Recognition Letters, 155 :151–158.

[Ferdinando et al., 2014] Ferdinando, H., Ye, L., Seppänen, T., and Alasaarela, E. (2014). Emotion Recognition by Heart Rate Variability. Australian Journal of Basic and Applied Science, 8(14) :50–55. Number : 14 Publisher : Petra Christian University.

[Fuhl et al., 2015] Fuhl, W., Kübler, T., Sippel, K., Rosenstiel, W., and Kasneci, E. (2015). ExCuSe : Robust Pupil Detection in Real-World Scenarios. In Azzopardi, G. and Petkov, N., editors, Computer Analysis of Images and Patterns, pages 39–51, Cham. Springer International Publishing.

[Fuhl et al., 2016a] Fuhl, W., Santini, T., Kasneci, G., and Kasneci, E. (2016a). PupilNet : Convolutional Neural Networks for Robust Pupil Detection. arXiv :1601.04902 [cs].

[Fuhl et al., 2016b] Fuhl, W., Santini, T. C., Kübler, T., and Kasneci, E. (2016b). ElSe : ellipse selection for robust pupil detection in real-world environments. In Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications, ETRA '16, pages 123–130, New York, NY, USA. Association for Computing Machinery.

[Garbin et al., 2019] Garbin, S. J., Shen, Y., Schuetz, I., Cavin, R., Hughes, G., and Talathi, S. S. (2019). OpenEDS : Open Eye Dataset. Publication Title : arXiv e-prints ADS Bibcode : 2019arXiv190503702G.

[Goodfellow et al., 2013] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z., and Bengio, Y. (2013). Challenges in Representation Learning : A report on three machine learning contests. arXiv :1307.0414 [cs, stat].

[Gouizi et al., 2011] Gouizi, K., Bereksi Reguig, F., and Maaoui, C. (2011). Emotion recognition from physiological signals. Journal of Medical Engineering & Technology, 35(6-7) :300–307. Publisher : Taylor & Francis _eprint : https ://doi.org/10.3109/03091902.2011.601784.

[Gudi et al., 2020] Gudi, A., Bittner, M., and van Gemert, J. (2020). Real-Time Webcam Heart-Rate and Variability Estimation with Clean Ground Truth for Evaluation. Applied Sciences, 10(23) :8630. Number : 23 Publisher : Multidisciplinary Digital Publishing Institute.

[Guo et al., 2016] Guo, H.-W., Huang, Y.-S., Lin, C.-H., Chien, J.-C., Haraikawa, K., and Shieh, J.-S. (2016). 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), pages 274–277. ISSN : 2471-7819.

[He et al., 2018] He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2018). Mask R-CNN. arXiv :1703.06870 [cs].

[Hofbauer et al., 2014] Hofbauer, H., Alonso-Fernandez, F., Wild, P., Bigun, J., and Uhl, A. (2014). A Ground Truth for Iris Segmentation. In 2014 22nd International Conference on Pattern Recognition, pages 527–532. ISSN : 1051-4651.

[Husak, 2017] Husak, P. (2017). Spotting Facial Micro-Expressions “In the Wild”.

[IIT Delhi, 2007] IIT Delhi (2007). IIT Delhi Iris Database.

[Johnson et al., 2010] Johnson, P. A., Lopez-Meyer, P., Sazonova, N., Hua, F., and Schuckers, S. (2010). Quality in face and iris research ensemble (Q-FIRE). In 2010 Fourth IEEE International Conference on Biometrics : Theory, Applications and Systems (BTAS), pages 1–6.

[Kim et al., 2019] Kim, J., Stengel, M., Majercik, A., De Mello, S., Dunn, D., Laine, S., McGuire, M., and Luebke, D. (2019). NVGaze : An Anatomically-Informed Dataset for Low-Latency, Near-Eye Gaze Estimation. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–12, Glasgow Scotland Uk. ACM.

[Kuang et al., 2022] Kuang, H., Lv, F., Ma, X., and Liu, X. (2022). Efficient Spatiotemporal Attention Network for Remote Heart Rate Variability Analysis. Sensors, 22(3) :1010. Number : 3 Publisher : Multidisciplinary Digital Publishing Institute.

[Lanatà et al., 2011] Lanatà, A., Armato, A., Valenza, G., and Scilingo, E. P. (2011). Eye tracking and pupil size variation as response to affective stimuli : A preliminary study. In 2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops, pages 78–84. ISSN : 2153-1641.

[Lang and Bradley, 2007] Lang, P. and Bradley, M. M. (2007). The International Affective Picture System (IAPS) in the study of emotion and attention. Handbook of emotion elicitation and assessment, 29 :70–73.

[Li et al., 2023a] Li, J., Dong, Z., Lu, S., Wang, S.-J., Yan, W.-J., Ma, Y., Liu, Y., Huang, C., and Fu, X. (2023a). CAS(ME)3 : A Third Generation Facial Spontaneous Micro-Expression Database With Depth Information and High Ecological Validity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3) :2782–2800. Conference Name : IEEE Transactions on Pattern Analysis and Machine Intelligence.

[Li et al., 2019] Li, P., Benezeth, Y., Nakamura, K., Gomez, R., Li, C., and Yang, F. (2019). An Improvement for Video-based Heart Rate Variability Measurement. In 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP), pages 435–439.

[Li et al., 2017] Li, S., Deng, W., and Du, J. (2017). Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2584–2593, Honolulu, HI. IEEE.

[Li et al., 2023b] Li, X., Cheng, S., Li, Y., Behzad, M., Shen, J., Zafeiriou, S., Pantic, M., and Zhao, G. (2023b). 4DME : A Spontaneous 4D Micro-Expression Dataset With Multimodalities. IEEE Transactions on Affective Computing, 14(4) :3031–3047. Conference Name : IEEE Transactions on Affective Computing.

[Li et al., 2013] Li, X., Pfister, T., Huang, X., Zhao, G., and Pietikäinen, M. (2013). A Spontaneous Micro-expression Database : Inducement, collection and baseline. In 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pages 1–6.

[Liu et al., 2023] Liu, X., Narayanswamy, G., Paruchuri, A., Zhang, X., Tang, J., Zhang, Y., Sengupta, S., Patel, S., Wang, Y., and McDuff, D. (2023). rPPG-Toolbox : Deep Remote PPG Toolbox. arXiv :2210.00716 [cs].

[Lucey et al., 2010] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+) : A complete dataset for action unit and emotion-specified expression. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pages 94–101, San Francisco, CA, USA. IEEE.

[Lugaresi et al., 2019] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., and Grundmann, M. (2019). MediaPipe : A Framework for Building Perception Pipelines. arXiv :1906.08172 [cs].

[Mavadati et al., 2013] Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., and Cohn, J. F. (2013). DISFA : A Spontaneous Facial Action Intensity Database. IEEE Transactions on Affective Computing, 4(2) :151–160.

[Mollahosseini et al., 2019] Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2019). AffectNet : A Database for Facial Expression, Valence, and Arousal Computing in the Wild. IEEE Transactions on Affective Computing, 10(1) :18–31. Conference Name : IEEE Transactions on Affective Computing.

[Multimedia University, 2006] Multimedia University (2006). MMU iris dataset.

[Ménard et al., 2015] Ménard, M., Richard, P., Hamdi, H., Dauce, B., and Yamaguchi, Y. (2015). Emotion Recognition based on Heart Rate and Skin Conductance :. In Proceedings of the 2nd International Conference on Physiological Computing Systems, pages 26–32, ESEO, Angers, Loire Valley, France. SCITEPRESS - Science and and Technology Publications.

[Nair et al., 2020] Nair, N., Chaudhary, A. K., Kothari, R. S., Diaz, G. J., Pelz, J. B., and Bailey, R. (2020). RIT-Eyes : realistically rendered eye images for eye-tracking applications. In ACM Symposium on Eye Tracking Research and Applications, ETRA '20 Adjunct, pages 1–3, New York, NY, USA. Association for Computing Machinery.

[Nardelli et al., 2015] Nardelli, M., Valenza, G., Greco, A., Lanata, A., and Scilingo, E. P. (2015). Recognizing Emotions Induced by Affective Sounds through Heart Rate Variability. IEEE Transactions on Affective Computing, 6(4) :385–394.

[Niu et al., 2020] Niu, X., Shan, S., Han, H., and Chen, X. (2020). RhythmNet : End-to-End Heart Rate Estimation From Face via Spatial-Temporal Representation. IEEE Transactions on Image Processing, 29 :2409–2423. Conference Name : IEEE Transactions on Image Processing.

[Oad and Ahmad, 2012] Oad, P. and Ahmad, W. (2012). Iris localization using Daugman's algorithm.

[Oliva and Anikin, 2018] Oliva, M. and Anikin, A. (2018). Pupil dilation reflects the time course of emotion recognition in human vocalizations. Scientific Reports, 8(1) :4871. Publisher : Nature Publishing Group.

[Pai et al., 2018] Pai, A., Veeraraghavan, A., and Sabharwal, A. (2018). CameraHRV : robust measurement of heart rate variability using a camera. In Optical Diagnostics and Sensing XVIII : Toward Point-of-Care Diagnostics, volume 10501, pages 160–168. SPIE.

[Pai et al., 2021] Pai, A., Veeraraghavan, A., and Sabharwal, A. (2021). HRVCam : robust camera-based measurement of heart rate variability. Journal of Biomedical Optics, 26(2) :022707. Publisher : SPIE.

[Pantic et al., 2005] Pantic, M., Valstar, M., Rademaker, R., and Maat, L. (2005). Web-Based Database for Facial Expression Analysis. In 2005 IEEE International Conference on Multimedia and Expo, pages 317–321, Amsterdam, The Netherlands. IEEE.

[Picard, 2000] Picard, R. W. (2000). Affective Computing. MIT Press. Google-Books-ID : GaVncRTcb1gC.

[Poh et al., 2011] Poh, M.-Z., McDuff, D. J., and Picard, R. W. (2011). Advancements in Non-contact, Multiparameter Physiological Measurements Using a Webcam. IEEE Transactions on Biomedical Engineering, 58(1) :7–11.

[Ponz et al., 2012] Ponz, V., Villanueva, A., and Cabeza, R. (2012). Dataset for the evaluation of eye detector for gaze estimation. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, UbiComp '12, pages 681–684, New York, NY, USA. Association for Computing Machinery.

[Porta et al., 2019] Porta, S., Bossavit, B., Cabeza, R., Larumbe-Bergera, A., Garde, G., and Villanueva, A. (2019). U2Eyes : A Binocular Dataset for Eye Tracking and Gaze Estimation. pages 0–0.

[Proenca et al., 2010] Proenca, H., Filipe, S., Santos, R., Oliveira, J., and Alexandre, L. A. (2010). The UBIRIS.v2 : A Database of Visible Wavelength Iris Images Captured On-the-Move and At-a-Distance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(8) :1529–1535.

[Qiao et al., 2021] Qiao, D., Zulkernine, F., Masroor, R., Rasool, R., and Jaffar, N. (2021). Measuring Heart Rate and Heart Rate Variability with Smartphone Camera. In 2021 22nd IEEE International Conference on Mobile Data Management (MDM), pages 248–249, Toronto, ON, Canada. IEEE.

[Qin et al., 2023] Qin, W., Zou, B., Li, X., Wang, W., and Ma, H. (2023). Micro-Expression Spotting with Face Alignment and Optical Flow. In Proceedings of the 31st ACM International Conference on Multimedia, pages 9501–9505, Ottawa ON Canada. ACM.

[Qu et al., 2018] Qu, F., Wang, S.-J., Yan, W.-J., Li, H., Wu, S., and Fu, X. (2018). CAS(ME)^2 : A Database for Spontaneous Macro-Expression and Micro-Expression Spotting and Recognition. IEEE Transactions on Affective Computing, 9(4) :424–436. Conference Name : IEEE Transactions on Affective Computing.

[Quang et al., 2019] Quang, N. V., Chun, J., and Tokuyama, T. (2019). CapsuleNet for Micro-Expression Recognition. In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), pages 1–7.

[Ravi et al., 2024] Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., and Feichtenhofer, C. (2024). SAM 2 : Segment Anything in Images and Videos. arXiv :2408.00714 [cs].

[Russell, 1980] Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6) :1161–1178. Place : US Publisher : American Psychological Association.

[Sabour et al., 2017] Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. Advances in neural information processing systems, 30.

[Shoushan et al., 2021] Shoushan, M. M., Alexander Reyes, B., Rodriguez, A. M., and Woon Chong, J. (2021). Contactless Heart Rate Variability (HRV) Estimation Using a Smartphone During Respiratory Maneuvers and Body Movement. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pages 84–87, Mexico. IEEE.

[Shu et al., 2020] Shu, L., Yu, Y., Chen, W., Hua, H., Li, Q., Jin, J., and Xu, X. (2020). Wearable Emotion Recognition Using Heart Rate Data from a Smart Bracelet. Sensors, 20(3) :718. Number : 3 Publisher : Multidisciplinary Digital Publishing Institute.

[Tonsen et al., 2016] Tonsen, M., Zhang, X., Sugano, Y., and Bulling, A. (2016). Labelled pupils in the wild : a dataset for studying pupil detection in unconstrained environments. In Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications, ETRA '16, pages 139–142, New York, NY, USA. Association for Computing Machinery.

[Verkruysse et al., 2008] Verkruysse, W., Svaasand, L. O., and Nelson, J. S. (2008). Remote plethysmographic imaging using ambient light. Optics Express, 16(26) :21434–21445. Publisher : Optica Publishing Group.

[Viola and Jones, 2001] Viola, P. and Jones, M. (2001). 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, volume 1, pages I–I. ISSN : 1063-6919.

[Wang et al., 2020] Wang, C., Muhammad, J., Wang, Y., He, Z., and Sun, Z. (2020). Towards Complete and Accurate Iris Segmentation Using Deep Multi-Task Attention Network for Non-Cooperative Iris Recognition. IEEE Transactions on Information Forensics and Security, 15 :2944–2959. Conference Name : IEEE Transactions on Information Forensics and Security.

[Wang et al., 2017] Wang, W., den Brinker, A. C., Stuijk, S., and de Haan, G. (2017). Algorithmic Principles of Remote PPG. IEEE Transactions on Biomedical Engineering, 64(7) :1479–1491. Conference Name : IEEE Transactions on Biomedical Engineering.

[Wu et al., 2023] Wu, W., Peng, H., and Yu, S. (2023). YuNet : A Tiny Millisecond-level Face Detector. Machine Intelligence Research, 20(5) :656–665.

[Xia and Wang, 2021] Xia, B. and Wang, S. (2021). Micro-Expression Recognition Enhanced by Macro-Expression from Spatial-Temporal Domain. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, pages 1186–1193, Montreal, Canada. International Joint Conferences on Artificial Intelligence Organization.

[Xu et al., 2023] Xu, K., Chen, K., Sun, L., Lian, Z., Liu, B., Chen, G., Sun, H., Xu, M., and Tao, J. (2023). Integrating VideoMAE based model and Optical Flow for Micro- and Macro-expression Spotting. In Proceedings of the 31st ACM International Conference on Multimedia, MM '23, pages 9576–9580, New York, NY, USA. Association for Computing Machinery.

[Yan et al., 2014] Yan, W.-J., Li, X., Wang, S.-J., Zhao, G., Liu, Y.-J., Chen, Y.-H., and Fu, X. (2014). CASME II : An Improved Spontaneous Micro-Expression Database and the Baseline Evaluation. PLOS ONE, 9(1) :e86041. Publisher : Public Library of Science.

[Yan et al., 2013] Yan, W.-J., Wu, Q., Liu, Y.-J., Wang, S.-J., and Fu, X. (2013). CASME database : A dataset of spontaneous micro-expressions collected from neutralized faces. In 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pages 1–7.

[Yap et al., 2020] Yap, C. H., Kendrick, C., and Yap, M. H. (2020). SAMM Long Videos : A Spontaneous Facial Micro- and Macro-Expressions Dataset. arXiv :1911.01519 [cs].

[Yu et al., 2023] Yu, J., Cai, Z., Du, S., Shen, X., Wang, L., and Gao, F. (2023). Efficient Micro-Expression Spotting Based on Main Directional Mean Optical Flow Feature. In Proceedings of the 31st ACM International Conference on Multimedia, pages 9541–9545, Ottawa ON Canada. ACM.

[Yu et al., 2019] Yu, Z., Li, X., and Zhao, G. (2019). Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks. arXiv :1905.02419 [cs].

[Yu et al., 2022] Yu, Z., Shen, Y., Shi, J., Zhao, H., Torr, P., and Zhao, G. (2022). PhysFormer : Facial Video-based Physiological Measurement with Temporal Difference Transformer. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 4176–4186, New Orleans, LA, USA. IEEE.

[Zhang et al., 2014] Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., Liu, P., and Girard, J. M. (2014). BP4D-Spontaneous : a high-resolution spontaneous 3D dynamic facial expression database. Image and Vision Computing, 32(10) :692–706.

[Zhao et al., 2023] Zhao, C., Wang, H., Chen, H., Shi, W., and Feng, Y. (2023). JAMSNet : A Remote Pulse Extraction Network Based on Joint Attention and Multi-Scale Fusion. IEEE Transactions on Circuits and Systems for Video Technology, 33(6) :2783–2797. Conference Name : IEEE Transactions on Circuits and Systems for Video Technology.