Rapport sur le mémoire de thèse de Monsieur Yassine OUZAR intitulé « Reconnaissance automatique sans-contact de l'état affectif de la personne par fusion physio-visuelle à partir de vidéo du visage »

établi par Frédéric VANDERHAEGEN,
Professeur à l'INSA Hauts-de-France de l'Université Polytechnique Hauts-de-France

A partir de la fusion de données captées sur le visage, les travaux de recherche de Monsieur Yassine OUZAR visent à déterminer les émotions et le stress en utilisant les techniques d'intelligence artificielle, d'informatique affective et du traitement d'image.

Le manuscrit comporte deux parties principales qui sont résumées et commentées par la suite. La première est un état de l'art sur les définitions et les approches de modélisation de facteurs tels que l'émotion et le stress, ainsi que sur leur mesure et interprétation par une instrumentation sans contact. La seconde détaille les contributions de Monsieur OUZAR.

Etat de l'art

Cette partie contient deux chapitres : un consacré à l'émotion et au stress, le second à la mesure de la fréquence cardiaque par caméra. Elle permet de positionner les travaux de Monsieur OUZAR par rapport aux contributions existantes. Dans ces deux chapitres, Monsieur OUZAR montre la complexité de la prise en compte des émotions et du stress du fait de la diversité des définitions, modèles, théories et mesures. Il établit une synthèse pertinente afin de déterminer les orientations qu'il prend dans la seconde partie de son mémoire.

Ce premier chapitre montre la complémentarité de deux facteurs : l'émotion et le stress. L'émotion peut être généralement associée à une réaction à la perception d'un stimulus et le stress à une tension mentale ou émotionnelle due à un stimulus dont les exigences perçues pour le contrôler sont trop importantes. Un doute sur l'interprétation d'une mesure de l'un peut être levé par une mesure de l'autre. Ces deux facteurs peuvent être étudiés sur deux plans : sur le plan comportemental et sur le plan physiologique. La genèse d'émotions ou de stress semble toutefois dépendante de la modalité du stimulus

en sollicitant les sens perceptifs humains ou les expériences des individus. La complexité de leur détection est multifactorielle. Elle dépend de la théorie appliquée, de l'exhaustivité des émotions à prendre en compte, de l'interdépendance entre émotions ou entre émotion et stress, de la modalité de contrôle des émotions ou du stress, du ressenti individuel, de l'interprétation des émotions ou du stress, ou encore de l'acquisition de données pour détecter ou évaluer les émotions ou le stress. Sur ce dernier point, Monsieur OUZAR fait une synthèse des approches de captation d'émotions ou du stress à partir d'expressions faciales et d'autres données physiologiques par des algorithmes d'apprentissage et de classification, et plus particulièrement en mettant l'accent sur les avancées autour de l'apprentissage profond.

Le second chapitre présente les caractéristiques de fonctionnement du cœur et plus spécifiquement de la fréquence cardiaque et de sa variabilité, indicateurs qui sont souvent utilisés comme mesures de l'état de santé d'un patient, mais aussi de stress, de charge de travail ou d'émotion par exemple. Monsieur OUZAR détaille alors les différentes instrumentations possibles pour obtenir ces mesures et se concentre sur les technologies sans contact par caméra basées sur des mesures de variations d'absorption de la lumière sur la peau (i.e., la photopléthysmographie par imagerie, iPPG). Ces technologies permettent une captation non seulement de la fréquence cardiaque mais aussi d'autres signaux vitaux tels que la pression artérielle ou le taux d'oxygène dans le sang et plusieurs algorithmes tels que ceux basés sur l'apprentissage profond peuvent être mis en œuvre pour traiter ces signaux.

Contributions

Cette partie se décompose également en deux chapitrés. Le premier détaille l'architecture choisie pour la mesure de la fréquence cardiaque et le second donne les résultats de deux études pour la détection de l'état émotionnel et du stress en se basant sur des basés de données présentées dans le premier chapitre.

Dans le premier chapitre, Monsieur OUZAR présente différentes bases de données généralement utilisées pour tester des algorithmes d'apprentissage profond. Pour chacune d'entre elles, il précise leur contenu, e.g., le nombre de captures de vidéos d'expressions faciales, la taille ou la résolution de ces vidéos, le nombre d'hommes et de femmes, les fréquences cardiaques ou autres signaux physiologiques au cours de ces captations d'images. Monsieur OUZAR a choisi de modifier le réseau Xception pour définir une architecture de réseau de neurone convolutif spatio-temporel X-iPPGNet pour traiter la couleur des pixels sur lesquels la peau apparaît dans les images de vidéos de 2 secondes. Il inclut une adaptation 3D du processus 2D de convolution séparable en profondeur pour intégrer les caractéristiques de chaque canal de couleur d'images en entrée et traiter les relations spatio-temporelles entre ces caractéristiques. Cette architecture a été testée avec trois bases de données MMSE-HR, UBFC-rPPG et MAHNOB-HCI et comparées avec d'autres approches à partir de différents critères d'erreur d'estimation de la fréquence cardiaque (i.e., écart-type, erreur absolue moyenne, racine carré de l'erreur quadratique moyenne, coefficient de Pearson). Monsieur OUZAR démontre ainsi la faisabilité de mesure de la fréquence cardiaque sans contact. Son architecture donne des résultats meilleurs que d'autres systèmes sur certains critères, et il évalue ces performances sur la base MMSE-HR en fonction d'autres facteurs comme la couleur de peau, le sexe, le mouvement de la tête, la taille de la fenêtre temporelle traitée ou le temps de calcul.

Le dernier chapitre détaille les résultats de deux études sur l'application de l'architecture proposée dans le chapitre précédent pour la reconnaissance de l'état émotionnel et du stress respectivement, et ce en reconstituant les signaux physiologiques associés à l'iPPG avec la méthode MTTS-CAN. Pour la première étude, deux bases de données MAHNOB-HCI, BP4D+ sont utilisées en adaptant l'architecture du chapitre 3 pour estimer la fréquence cardiaque et en intégrant l'interdépendance entre canaux de couleur des images pour réduire les dimensions de la carte spatio-temporelle à traiter et prioriser certains canaux de celle-ci. Cette combinaison permet d'augmenter les performances et de construire un nouveau modèle 3D-SE-XceptionNet basé sur le réseau Xception. Cette nouvelle architecture présente des résultats plus performants par rapport à d'autres contributions concernant la reconnaissance unimodale d'émotions, donne des résultats satisfaisants pour la reconnaissance d'émotion à partir de données physiologiques et des résultats prometteurs en combinant les expressions faciales avec des données physiologiques. Pour la seconde étude sur le stress, la base UBFC-Phys dont la construction et le contenu sont présentés en détail a été utilisée. La méthode MTTS-SCAN est utilisée comme pour la première étude pour identifier les données iPPG et le modèle VGG16 pour extraire les caractéristiques faciales. Le modèle proposé donne de meilleurs résultats que d'autres classifieurs en termes de reconnaissance multimodale du stress à partir de caractéristiques faciales et de données physiologiques.

Conclusion du manuscrit

En conclusion du mémoire, Monsieur OUZAR résume ses contributions concernant 1) l'architecture proposée fusionnant fréquence cardiaque et signaux physiologiques et 2) les résultats prometteurs pour l'estimation de l'état émotionnel et le stress. Il donne ensuite quelques perspectives d'optimisation et d'utilisation de ses recherches.

Commentaires

Dans le mémoire, la problématique temps-réel aurait pu être traitée de manière plus explicite, surtout en termes d'utilisation en ligne du système proposé par Monsieur OUZAR. Par exemple, d'une part on peut penser que plus les régions d'intérêt ROI sont grandes, plus le traitement de pixels est long et d'autre part que plus le sujet bouge, plus la captation et le traitement des mêmes ROI deviennent difficiles. Les conditions expérimentales de la synthèse du tableau 2.1 auraient pu être indiquées à savoir par exemple la posture des sujets testés, l'intensité lumineuse, les caractéristiques des peaux testées, la qualité de l'instrumentation utilisée, la zone de captation sur le corps, etc., conditions dont semblent dépendre la qualité de captation de la fréquence cardiaque. De plus, l'écart d'occurrence d'un battement cardiaque entre une mesure sans contact et l'ECG est-il négligeable ou en d'autres termes, la détection d'un état affectif avec une mesure sans contact peut-elle être décalée avec celle effectuée par ECG ?

L'interdépendance entre la fréquence cardiaque et d'autres facteurs tels que l'attention ou la direction du regard aurait pu également être discutée dans le cadre de l'interprétation de mesures pour le stress ou l'émotion. Par exemple, celles-ci dépendent des capacités perceptives de stimuli. Or une étude récente a montré que l'occurrence de stimuli tels que des alarmes visuelles et sonores synchronisés avec la fréquence cardiaque réduisait la perception desdits stimuli, et que dans ces conditions, certains sujets pouvaient regarder

les zones d'affichage des alarmes sans les percevoir (voir par exemple : Vanderhaegen, F., Wolff, M., Mollard, R., 2022, « A heartbeat-based study of attention in the detection of digital alarms from focused and distributed supervisory control systems », Cognition Technlogy & Work, 25, 119-134). La détection d'un état affectif peut donc être impactée en fonction de l'occurrence d'un stimulus non-perçue alors que certains indicateurs montrent que celui-ci l'est.

Le pourcentage de précision donné dans les tableaux de synthèse comme les tableaux 1.1, 1.2 ou 4.6 nécessiterait un éclairage sur les émotions les mieux reconnues ou identifiées par les systèmes proposés. En effet, certains travaux comme ceux de Bendjoudi combinant les réseaux Xception et VGG16 et utilisant la base EMOTIC (voir par exemple : Bendjoudi, I., Vanderhaegen, F., Hamad, D., Dornaika, F., 2021, « Multi-label, multi-task CNN approach for context-based emotion recognition », Information Fusion, 76, 422-428) semblent montrer que les performances d'identification de telle ou telle émotion dépendent de la configuration architecturale du système de reconnaissance faciale. Ainsi dans le chapitre 4, lorsque Monsieur OUZAR évalue ses modèles pour 4 émotions (joie, embarras, peur et douleur), quelles sont les émotions les mieux identifiées ?

Dans la description des architectures proposées par Monsieur OUZAR, certains points pourraient être clarifiés. Par exemple, pourquoi et comment des transformations géométriques à savoir des rotations, des translations, etc. lors de la préparation des données (voir page 97 du manuscrit) ont été effectuées, sur des extraits de vidéos pour améliorer les performances ? Sur quels critères de performance a-t-on des améliorations et est-ce que cela a nécessité plusieurs essais ? De même, pourquoi dans la structure du réseau 3D-SE-Xception, le flux intermédiaire est-il répété 8 fois, la couche contient 4 neurones et 2 neurones pour quantifier la valence ? Est-ce que plusieurs tests ont été effectués pour déterminer ces caractéristiques ? Dans la section 4.2.2.4, pourquoi avoir choisi la fonction d'entropie croisée binaire et d'entropie croisée catégorielle ? Enfin, pourquoi prendre l'hypothèse dans le chapitre 3 que le modèle « est capable de se concentrer automatiquement sur les zones les plus vascularisées du visage » et « qu'il apprend les caractéristiques spatio-temporelles associées aux changements subtils de couleur sur la région sélectionnée » ?

Conclusions

Monsieur OUZAR a effectué un travail de recherche important sur une thématique difficile d'analyse d'image et de traitement de signaux appliquée à la détection d'émotion et de stress.

Les travaux ont été valorisés dans 3 publications dans des revues de qualité (Computers in Biology and Medecine 2021 en 3^{e} position sur 5 auteurs, et en 2023 en 1ère position sur 4 auteurs ; Biomedical Signal Processing and Control 2021 en 4^{e} position sur 6 auteurs), 3 conférences internationales (en 1er auteur sur 4) et 6 manifestations scientifiques nationales (dont 3 en 1er auteurs : Journées de la SAGIP 2022 ; Colloque Jeunes Chercheurs IFRATH 2021 ; Journées STP GDR MACS 2020).

Il faut noter également que la bibliographie est constituée d'une liste de 432 références bibliographiques, ce qui montre 1) d'une part la diversité et la variabilité des contributions dans l'étude des émotions et du stress ; 2) et d'autre part le haut niveau d'expertise et la capacité de synthèse de Monsieur OUZAR concernant ses domaines de recherche et leur

positionnement dans la communauté scientifique.

Par conséquent, je donne un avis favorable à la soutenance de Monsieur Yassine OUZAR en vue de l'obtention d'un doctorat de l'Université de Lorraine, mention Automatique, Traitement du Signal et des Images, Génie Informatique.

Valenciennes, le 23 Mai 2023

Frédéric VANDERHAEGEN
Professeur des Universités
INSA Hauts-de-France
Université Polytechnique Hauts-de-France
Membre de la 61e section du CNU

Yannick Benezeth
Maître de Conférences HDR
Univ. Bourgogne
Laboratoire ImViA – EA 7535

Rapport sur le mémoire présenté par Yassine Ouzar

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

Monsieur Yassine Ouzar présente dans son mémoire de thèse les travaux de recherche effectués au sein du laboratoire LCOMS de l'Université de Lorraine. Cette thèse a été préparée sous la direction de Choubeila Maaoui et l'encadrement Frédéric Bousefsaf. Ses travaux ont porté sur la reconnaissance des émotions et du stress par analyse d'expressions faciales et de signaux physiologiques estimés par caméra.

Le manuscrit est composé de courts résumés en Anglais et en Français, d'une introduction générale, de deux chapitres d'état de l'art sur l'émotion et le stress et sur la mesure sans contact de la fréquence cardiaque, de 2 chapitres détaillant les contributions et enfin d'une conclusion. Il est rédigé en français (environ 125 pages hors bibliographie), avec une structure claire, ce qui le rend agréable à lire.

Le manuscrit débute donc par une courte **introduction générale** où sont décrits le contexte et les objectifs des travaux. Les travaux portent sur la reconnaissance de l'état affectif d'une personne par analyse d'expressions faciales et par analyse de signaux physiologiques. L'hypothèse du travail est que la combinaison des expressions faciales et des signaux physiologiques (appelée physio-visuelle dans le manuscrit) peut améliorer cette reconnaissance. Une caractéristique intéressante du positionnement des travaux est que le signal physiologique utilisé dans les analyses est estimé à partir du signal vidéo avec une technique appelée iPPG (*Imaging Photoplethysmography*). Cette technologie repose sur l'estimation des fines fluctuations de la couleur de la peau d'une personne associées à la variation du volume sanguin.

Le **chapitre 1** présente un état de l'art sur l'émotion et le stress. La première section se concentre sur les émotions et commence par présenter quelques définitions et théories générales sur les émotions puis décrit les méthodes de reconnaissance automatique. Les descriptions concernent principalement les expressions faciales et les signaux physiologiques. Pour chaque modalité, les méthodes conventionnelles sont tout d'abord décrites suivies d'une présentation des méthodes basées sur l'apprentissage profond. Un bref aperçu des méthodes de reconnaissance multimodale est ensuite donné. Après les émotions, ce chapitre état de l'art poursuit par une présentation du stress avec la même structure, *i.e.* présentation des définitions et des méthodes d'élicitation du stress puis présentation des méthodes de reconnaissance automatique. **Le chapitre 2** présente un nouveau chapitre d'état de l'art présentant cette fois-ci les méthodes de mesure sans contact de la fréquence

cardiaque. On peut noter la présence d'illustrations soignées tout au long de ce chapitre. La présentation commence par le fonctionnement du cœur, et les différentes méthodes de mesure au contact avant d'introduire la technologie iPPG qui sera développé dans ce travail. On comprend alors que cette technologie est récente, offre de nombreux avantages mais également des verrous à adresser. L'état de l'art sur les méthodes conventionnelles d'iPPG est très complet et la chaine de traitement classique est donné avec suffisamment de détails : acquisition des données, détection de la région d'intérêt, extraction du signal iPPG, filtrage et estimation de la fréquence cardiaque. Un tableau synthétique reprend les principales variantes de chacune de ces étapes. Les méthodes basées sur le deep learning sur ensuite brièvement décrites. Il aurait été intéressant de lire plus de discussion sur ces méthodes deep-learning pour comprendre le positionnement des travaux de la thèse. De même, il aurait été intéressant de lire une description des bases de données existantes pour la reconnaissance des émotions, du stress ou pour valider les algorithmes d'iPPG, ainsi que les méthodes d'augmentation de données utilisées dans certains travaux sur le sujet. Cette discussion aurait permis de justifier les choix effectués dans la suite des travaux.

Comme la partie état de l'art, la partie contribution est structurée avec un chapitre portant sur la mesure sans contact de la fréquence cardiaque et un second sur la reconnaissance des émotions et du stress. Cette partie commence donc par **le chapitre 3** portant sur la mesure du rythme cardiaque. Le chapitre commence par une présentation des quatres bases de données utilisées pour entrainer et valider le réseau proposé. Même s'il existe d'autres bases de données publiques qui auraient pu être utilisées, le choix effectué est tout à fait pertinent. Ensuite, la méthode proposée est présentée. La première étape des traitements consiste à détecter et segmenter le visage dans chaque image de la vidéo. Cette étape est basée sur une méthode de l'état de l'art qui n'est pas décrite mais qui semble donner des résultats remarquables. Ensuite, l'architecture deep learning, appelée X-iPPGNet, est présentée. On comprend alors qu'elle est basée sur une architecture Xception modifiée avec par exemple l'utilisation de la convolution séparable en profondeur pour apprendre les caractéristiques spatiales et temporelles de chaque canal de couleur séparément avant d'être fusionné par une convolution ponctuelle (*pointwise*). Le réseau prend donc en entrée un clip vidéo de 2 secondes, avec le visage segmenté, et retourne le rythme cardiaque moyen sur cette durée. Avant de présenter la validation de ce réseau, une étude intéressante donne la distribution des erreurs d'estimation en fonction de la plage de fréquences cardiaques et du type de couleur de peau. Il est possible d'observer alors que les performances sont meilleures dans les plages de fréquences cardiaques les plus courantes (entre 70 et 90 bpm), qui sont majoritaires dans les bases de données utilisées. Une observation similaire est donnée pour les types de couleur de peau. Pour limiter ce biais important et augmenter les performances du système proposé, une augmentation de données spécifiques est proposée en générant des vidéos avec des fréquences cardiaques inférieures à 70 bpm ou supérieures à 90 bpm. Inspirée d'autres travaux de l'état de l'art, la technique d'augmentation de données est basée sur des transformations géométriques et une méthode d'amplification vidéo (*video color magnification*). On aurait aimé lire un peu plus de détails sur les raisons d'utiliser cette stratégie d'augmentation de données et son positionnement par rapport aux autres méthodes de génération de vidéos synthétiques utilisées quelque fois pour les applications iPPG.

Les métriques utilisées pour l'évaluation sont ensuite présentées et les résultats obtenus sur les bases de données MMSE-HR, UBFC-rPPG, MAHNOB-HCI après un entrainement sur la base BP4D+ sont présetés. Un tableau est donné pour chaque base de données avec un ensemble de résultats tirés de 2 articles différents. Cela rend difficile l'évaluation objective des performances relatives de X-iPPGNet par rapport aux autres méthodes car elles n'ont sans doute pas les mêmes étapes de pré-traitement (segmentation du visage…), de post-traitement (filtrage…) ni le même protocole

d'évaluation (séparation des données entrainement/test…) dans le cas des méthodes basées sur le deep-learning. On peut cependant observer que X-iPPGNet donne de très bons résultats sur ces 3 bases de données. Pour aller plus loin dans l'analyse des performances du réseau proposé, on aurait aimé lire une étude des performances « par ablation », par exemple sans l'étape d'augmentation de données ou sans l'utilisation de filtres de convolution séparable en profondeur. Ensuite, une étude intéressante de l'impact de la distribution des fréquences cardiaques est présentée où il est possible d'observer que le modèle a tendance à produire des prédictions orientées vers des valeurs de fréquence cardiaque moyenne, *i.e.* à surestimer les fréquences cardiaques faibles et à sous-estimer les fréquences cardiaques élevées. Des différences de performance notables sont ensuite données entre les hommes et les femmes, les types de peau ou les mouvements de la tête. Peut-être que des tests statistiques vérifiant si ces différences sont significatives ont été réalisées ? Enfin, la capacité de mesurer directement le rythme cardiaque d'X-iPPGNet, dans une stratégie de bout-en-bout (hors segmentation du visage) est présentée comme un avantage par rapport aux méthodes qui estiment le signal iPPG puis le rythme cardiaque. Comme dans les travaux du chapitre suivant, cette caractéristique peut devenir une limitation pour les applications qui doivent analyser le signal iPPG temporel. Il aurait été intéressant de lire un peu plus de discussion sur les avantages d'une stratégie de bout-en-bout.

Le chapitre 4 présente ensuite la reconnaissance de l'état affectif d'une personne par analyse des expressions faciales et par signaux physiologiques. La première partie du chapitre concerne la reconnaissance des émotions. La validation est faite en utilisant MAHNOB-HCI et BP4D+ avec une classification binaire sur le niveau de valence et d'activation pour la première ou une labellisation discrète sur 4 émotions (*i.e.* joie, embarras, peur et douleur) pour la seconde. Cet ensemble d'émotions est intéressant et assez inhabituel. Il aurait été intéressant d'avoir des illustrations des vidéos pour les différentes classes afin de mieux se rendre compte de la difficulté de la tâche de reconnaissance. Le système proposé est donc basé sur l'étude des expressions faciales et sur le signal iPPG. Comme dans la contribution précédente, la base de données a été augmentée pour améliorer les performances du modèle. Cependant, seules les transformations géométriques ont été appliquées sans l'étape d'amplification de la couleur de la vidéo. Pour les expressions faciales, une étape de sélection des visages les plus expressifs a été réalisée. La reconnaissance des expressions faciales est basée sur une modification d'un réseau de Xception avec l'ajout d'un module *Squeeze-Excitation.* Le réseau proposé s'appelle 3D-SE-XceptionNet. L'architecture proposée dans le chapitre 3 permettait d'estimer le rythme cardiaque à partir d'une séquence vidéo de visages. Cependant, il est expliqué dans ce chapitre que la variabilité cardiaque est plus intéressante pour l'estimation de l'état affectif d'une personne et il est donc proposé de ne pas utiliser le réseau X-iPPGNet mais un réseau de l'état de l'art MTTS-CAN qui permet d'estimer un signal iPPG à partir duquel il est possible d'obtenir la variabilité cardiaque. Ensuite, un ensemble de six descripteurs temporels et fréquentiels sont extraits du signal de variabilité cardiaque. Il aurait été intéressant de lire une discussion sur les raisons du choix de ces descripteurs en particulier. De plus, les signaux considérés sont relativement courts par rapport à ce qui se fait dans la littérature puisque seulement 4 secondes de signaux de variabilité cardiaque sont utilisées. Est-ce que la robustesse des descripteurs utilisés sur des signaux si courts a été étudiée ? On pense en particulier au descripteur fréquentiel BF/HF qui utilise l'énergie du signal entre 0,04Hz et 0,15Hz (qui correspond à des signaux de période comprise entre 6 et 25 secondes). Les résultats sont présentés pour la reconnaissance d'expressions faciales avec des performances moyennes de 3D-SE-XceptionNet supérieures aux autres réseaux testés sur BP4D+. Sur MAHNOB-HCI les performances obtenues sont également favorables par rapport aux autres réseaux. Pour la reconnaissance à partir des signaux physiologiques, un simple réseau de neurones à propagation

avant à deux couches a été utilisé avec en entrée le signal iPPG brut, les 6 descripteurs de variabilité cardiaque ou une combinaison des deux. On peut remarquer que le réseau de neurones est donc très petit pour le cas des descripteurs de variabilité cardiaque. Les performances sont comparables entre le signal iPPG brut et les descripteurs de variabilité cardiaque et la performance est plus faible lorsqu'on combine les deux. L'utilisation du signal iPPG brut est original et intéressant mais on se demande s'il n'aurait pas été préférable d'utiliser un réseau plus complexe avec des convolutions ou un réseau récurent pour considérer la dimension temporelle de manière plus précise. Pour MAHNOB-HCI, les performances sont aussi comparées avec une mesure obtenue par un capteur ECG et on observe des performances de reconnaissance un peu inférieures lorsque la mesure est obtenue sans contact mais la différence est relativement faible. La fusion du signal brut et des descripteurs de variabilité cardiaque ne permet pas une meilleure reconnaissance. Il aurait été intéressant de présenter une matrice confusion pour pouvoir comparer précisément la reconnaissance à partir des expressions faciales et des signaux physiologiques. Ensuite, la reconnaissance par combinaison des deux modalités est présentée et il est possible d'observer un gain en performance assez important en combinant les modalités.

Ensuite, une étude est présentée pour évaluer la reconnaissance du stress à partir de la même méthodologie que celle proposée pour la reconnaissance des émotions. Cette étude est basée sur la base de données UBFC-Phys. Contrairement, à la reconnaissance des émotions, deux descripteurs ont été calculées en plus des six utilisés pour la reconnaissance des émotions, un réseau de neurones basé sur VGG a été utilisé au lieu de 3D-SE-XceptionNet et un ensemble de classifieurs différents du réseau de neurones à 2 couches utilisé précédemment a été utilisé. Il aurait été intéressant de chercher à réutiliser les contributions précédentes dans cette dernière étude. On se demande comment est gérée la dimension temporelle pour la reconnaissance des expressions faciales et si seulement 4 secondes de signaux ont été utilisés ou toute la durée de la séquence (environ 3 minutes) ? La comparaison des résultats entre les systèmes unimodaux montrent que les caractéristiques faciales donnent de meilleurs résultats que l'utilisation des signaux physiologiques et que la fusion des caractéristiques faciales avec les signaux physiologiques améliore les performances. Est-ce que la reconnaissance des différents niveaux de stress, comme proposé dans la base de données UBFC-Phys, a été testée ?

Enfin, le manuscrit se termine sur une conclusion qui reprend les objectifs et les différentes contributions présentées dans la thèse avec plusieurs pistes d'améliorations pertinentes.

La méthodologie proposée avec plusieurs contributions intéressantes ainsi que le positionnement original consistant à utiliser une combinaison des expressions faciales et des signaux physiologiques mesurés par analyse vidéo constituent un ensemble de contributions significatives vis-à-vis de la problématique de la reconnaissance des émotions et du stress. Pour conclure, compte tenu des éléments réunis dans le manuscrit soumis, je donne un **avis favorable** pour que Monsieur Yassine Ouzar soutienne sa thèse, en vue de l'obtention du titre de docteur de l'Université de Lorraine.

Fait à Dijon le 30/05/2023,
Yannick Benezeth

UNIVERSITÉ DE BOURGOGNE

ImViA