Résumé :

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

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

L'utilisation de la photopléthysmographie par imagerie (iPPG) pour mesurer les données physiologiques par vidéo se révèle plus pratique et confortable par rapport aux dispositifs intrusifs en contact, susceptibles d'interférer avec le sujet et de modifier son état émotionnel. En plus d'améliorer les performances et la fiabilité grâce à l'intégration de paramètres physiologiques, l'utilisation d'une caméra intégrée dans les appareils numériques courants contribue à réduire les coûts et à rendre l'approche plus accessible.

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

- **En télé-santé :** nous avons introduit une approche novatrice baptisée X-iPPGNet, conçue pour l'estimation sans contact de la fréquence cardiaque à partir d'enregistrements vidéo du visage. Ce système repose sur un réseau spatio-temporel profond, offrant ainsi une solution bout-en-bout optimisée. Le pipeline X-iPPGNet excelle dans la prédiction rapide de la fréquence cardiaque en seulement 2 secondes, éliminant la nécessité d'une extraction séparée du signal iPPG. Cette méthodologie se révèle particulièrement pertinente pour les fréquences cardiaques élevées et fortement fluctuantes. Son potentiel se déploie dans l'estimation sans

contact de signaux physiologiques, ouvrant ainsi la voie à de futures architectures robustes pour des applications en temps réel grâce à son faible nombre de paramètres et à l'utilisation d'un court fragment vidéo. Les résultats expérimentaux dépassent significativement les méthodes actuelles de l'état de l'art, attestant de sa performance sur trois ensembles de données de référence.

La capacité de reconnaître l'état affectif d'une personne et de mesurer ses signaux physiologiques de manière sans contact et à distance s'avère très prometteuse et pertinente dans le contexte du vieillissement et du suivi de l'évolution de maladies invalidantes. L'intégration de cette technologie dans les systèmes d'assistance à la personne, notamment lors des séances de thérapie par réalité virtuelle, offre de nouvelles opportunités pour le traitement des maladies invalidantes. Elle peut être exploitée pour suivre l'évolution de la maladie, évaluer l'efficacité du traitement et fournir un soutien aux patients. L'impact de cette solution est donc significatif, ouvrant de nouvelles perspectives pour l'amélioration de la santé publique et de la qualité de vie des personnes handicapées.

- **En informatique affective :** nous avons proposé un schéma de fusion physio-visuelle pour la reconnaissance de l'état affectif de la personne à partir de vidéos du visage, y compris les émotions et le stress. Cette approche présente plusieurs avantages par rapport aux systèmes existants car elle permet à la fois de surmonter le problème des émotions contrefaites et également améliorer les performances en recueillant en permanence des informations complémentaires sur l'état affectif de la personne. Cela est utile dans le cas d'acquisitions manquantes ou de données corrompues qui peuvent survenir lors de l'utilisation d'une seule modalité dans un environnement bruyant ou dans le cas d'une expression falsifiée.

- **En apprentissage profond :** nous avons proposé en premier lieu un réseau de neurones spatio-temporels basé sur le squelette de l'architecture Xception. Ce réseau repose sur le découplage des canaux de couleur et permet d'extraire des informations supplémentaires de chaque canal séparément. Ensuite, nous avons amélioré l'architecture proposée en intégrant le module Squeeze-Excitation qui joue le rôle d'un mécanisme d'attention. Il vise à modéliser explicitement l'interdépendance entre les canaux de l'image afin de recalibrer les cartes de caractéristiques par canal d'une manière adaptative et efficace en termes de temps de calcul. Par ailleurs, des techniques avancées d'optimisation de l'apprentissage profond ainsi que des stratégies de régularisation sont utilisées pour surmonter les problèmes de surajustement et améliorer la généralisation du modèle à de nouvelles données.

Les travaux présentés dans cette thèse se révèlent prometteurs. Les stratégies et méthodes développées ont été transférées industriellement à l'entreprise messine i-virtual. De plus, les résultats des différentes études ont été valorisés par la publication de trois articles dans les revues Computers in Biology and Medicine (Elsevier) et Biomedical Signal Processing and Control (Elsevier), et trois communications présentées lors de conférences prestigieuses telles que CVPR, Conference on Computer Vision and Pattern Recognition et ICCV, International Conference on Computer Vision et ICPR, International Conference on Pattern Recognition. Les travaux de cette thèse ont également été communiqués dans des journées doctorales du laboratoire LCOMS et les journées du GT ASHM, GDR MACS.