Titre: Title:	Apprentissage profond multimodal pour l'estimation de pose d'humains alités
Auteur: Author:	Ghassen Cherni
Date:	2022
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Cherni, G. (2022). Apprentissage profond multimodal pour l'estimation de pose d'humains alités [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/10517/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/10517/
Directeurs de recherche: Advisors:	Lama Séoud, Quentin Cappart, & Philippe Jouvet
Programme: Program:	Génie informatique

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Apprentissage profond multimodal pour l'estimation de pose d'humains alités

GHASSEN CHERNI

Département de génie informatique et génie logiciel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Génie informatique

Août 2022

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

Apprentissage profond multimodal pour l'estimation de pose d'humains alités

présenté par Ghassen CHERNI

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

a été dûment accepté par le jury d'examen constitué de :

Farida CHERIET, présidente
Lama SÉOUD, membre et directrice de recherche
Quentin CAPPART, membre et codirecteur de recherche
Philippe JOUVET, membre et codirecteur de recherche
Guillaume-Alexandre BILODEAU, membre

DÉDICACE

À mes parents, merci pour tout.

REMERCIEMENTS

Ce présent travail est le fruit de deux années de travail qui ont permis de me former grâce à non seulement les cours que j'ai suivis et les longues heures de recherches, mais aussi aux échanges et discussions que j'ai entrepris avec mes collègues et mes professeurs.

Je souhaite ainsi adresser mes remerciements à toutes ces personnes sans lesquelles ce travail n'aurait pas abouti.

Je tiens d'abord à remercier ma directrice de recherche, Lama Seoud, pour son encadrement et son support tout au long de cette période. Ses conseils avisés et sa disponibilité à travers toute cette période ont été d'une grande aide, notamment dans les moments difficiles en pleine pandémie.

Je tiens aussi à remercier mes co-directeurs, Quentin Cappart et Dr. Philippe Jouvet, pour leurs expertises qui m'ont permis d'étoffer le sujet de recherche et de mieux m'y approfondir.

Je souhaite également remercier Philippe Debanné pour ses relectures, corrections et conseils tout au long de la période de rédaction.

Finalement je tiens à remercier ma famille et mes amis sans lesquels je n'aurai pas pu garder le moral haut durant toute cette période.

RÉSUMÉ

Le sujet de ce travail de recherche est la détection de pose de patients alités en utilisant une approche multimodale; des images de différentes modalités, RGB, proche infrarouge (Low-Width Infrared - LWIR) et profondeur, sont combinées à l'aide de méthodes d'apprentissage profond et sont utilisées afin d'estimer la pose de patients. Cette fusion de modalités devra permettre de surmonter les défis spécifiques à la détection de pose de patients dans un contexte hospitalier, c'est-à-dire la possible présence d'occlusions dues à l'utilisation par exemple d'une couverture sur le patient, et la variation de luminosité dans les chambres d'hôpital. L'objectif de cette recherche est de déterminer la combinaison optimale de modalités et la meilleure méthode de fusion qui permettent d'obtenir l'estimation la plus précise des poses des patients tout en ayant une latence permettant un déploiement en temps réel. Ainsi une comparaison de 5 différentes méthodes de fusion multimodale est faite afin de déterminer la plus performante et une comparaison de différentes combinaisons de modalités est aussi réalisée afin de trouver la plus optimale. De plus, on démontre que la méthode choisie est généralisable lorsque déployée dans un contexte un peu différent (ex : une chambre d'un autre hôpital). Ces méthodes sont évaluées sur la base de données publique SLP qui contient des images de sujets alités. Ces conditions sont impératives vu que cette recherche s'inscrit dans un projet plus large qui vise à déployer un système vidéo de monitorage continu de patients au service des soins intensifs pédiatriques du Centre Hospitalier Universitaire Sainte Justine (CHUSJ) afin d'analyser les mouvements des patients en temps réel.

Ainsi nous implémentons 5 différentes méthodes de fusion de modalités, appartenant à trois différentes catégories (fusions en amont, en aval et intermédiaire); DenseFuse, Multimodal Transfer Module (MMTM), Concaténation, Ensemble et Channel-exchange (CE) afin de choisir la meilleure en termes de précision tout en respectant les contraintes de latence. De plus, chaque méthode est implémentée en utilisant différentes combinaisons de modalités afin de trouver la plus optimale.

Les résultats obtenus démontrent que la meilleure méthode de fusion est Channel-exchange en utilisant les modalités de profondeur et LWIR. Cette méthode de fusion avec cette combinaison de modalités nous permet de dépasser les résultats obtenus jusque-là sur la base de données publiques Simultaneously-Collected Multimodal Lying Pose (SLP). En effet, le meilleur résultat sur SLP, publié à ce jour, est de 96.6% tandis que l'on arrive à 97.1% en termes de métrique PCKh@0.5.

Cela s’explique par le fait qu’en fusionnant adéquatement les modalités, on arrive à surmonter les limites de chacune prise individuellement.

Ainsi nous proposons donc un modèle d’estimation de pose de patients alités utilisant l’apprentissage profond, qui permet de surpasser les résultats précédemment obtenus sur la base de données SLP. Nous démontrons aussi que le modèle est généralisable à un contexte différent de celui des données d’entraînement (ex : provenant d’un hôpital différent) ce qui supporte la volonté de déployer un tel modèle d’EPH au service des soins intensifs du CHUSJ pour un monitorage en continu des patients en état critique.

ABSTRACT

In-bed human pose estimation (HPE) is an important step for sleep behavior analysis and for patient monitoring in the intensive care unit. Challenges specific to this context include wide variability in scene illumination (from darkness to bright light), and blanket occlusions. A vision-based multi-modal approach offers an interesting solution to this problem. Work has been done for the creation of an annotated domain-specific dataset for in-bed HPE, containing RGB, long-wavelength infrared (LWIR) and depth images. How to combine these modalities to optimize the in-bed HPE? The present work aims at answering this question by comparing different multimodal deep learning methods. Extensive experiments show the superiority of the Channel-exchange (CE) method over other fusion methods. CE consists in a parameter-free framework that exchanges channels between the modality-specific sub-branches of the neural network, guided by the contributions of individual channels to the overall learning of the model as measured by the magnitude of batch-normalization. The combination of LWIR and depth images using CE reaches a performance of 97.1% using the PCKh@0.5 evaluation metric, outperforming the current benchmark of 96.6% on a domain-specific dataset for in-bed HPE

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

BN Batch normalization

CE Channel-exchange

CNN Convolutional Neural Network

CHUSJ Centre Hospitalier Universitaire Sainte Justine

EPH Estimation de pose humaine

IR Infrarouge

LWIR Low-width infra-red

MMTM Multimodal Transfer Module

PCK Percentage of Correct Key-Points

PCP Percentage of Correct Parts

PDJ Percentage of Detected Joints

ReLU Rectified Linear Unit

SDK Software Development Kit

SI Soins intensifs

SLP Simultaneously-Collected Multimodal Lying Pose

CHAPITRE 1 INTRODUCTION

Le monitorage des patients est une importante et lourde responsabilité du personnel médical hospitalier. Celle-ci est d'autant plus importante en soins intensifs (SI), lorsque le patient est dans un état de santé critique. Il est alors important de surveiller les signes vitaux comme la respiration, la fréquence cardiaque, mais également le profil neurologique du patient évalué partiellement à partir de ses mouvements dans le lit. Actuellement, les infirmières en soins intensifs ont la tâche de monitorer les signes vitaux sur un écran d'ordinateur et de garder un œil direct sur le patient à travers une fenêtre pour évaluer son état de santé général. Cette tache de surveillance humaine est plus que nécessaire, mais contraignante, d'autant plus dans un système de santé en grand manque de main-d'œuvre.

C'est donc dans ce contexte qu'un système de monitorage assisté par ordinateur s'avérerait utile. Un tel système devra présenter des caractéristiques particulières pour pouvoir être déployé dans un contexte hospitaliser. Il devra être de nature non-invasive (pour de ne pas gêner le patient), peu coûteuse (pour être déployé dans toutes les chambres des SI), précise (pour avoir un système fiable) et déployable en temps réel (pour avertir rapidement l'infirmière en cas de problème). L'utilisation de caméras de grade commercial s'avère comme une solution intéressante : à la fois abordables et non-invasives, les images de ces caméras nous permettront d'atteindre nos objectifs de monitorage dépendamment de la précision et de la latence des algorithmes d'aide à la décision déployés.

Dans ce travail de maîtrise, nous nous concentrons sur le monitorage neurologique, c'est-à-dire sur le suivi des mouvements du patient dans son lit d'hôpital. À cette fin, il est nécessaire de reconnaître la pose du patient dans toutes les images de la séquence vidéo.

L'estimation de pose humaine (EPH) est un sujet largement étudié par la communauté scientifique de vision par ordinateur. Elle consiste en la détection des articulations et des membres de sujets humains à partir d'images de caméra. Ce domaine a connu une grande évolution depuis l'introduction de l'apprentissage profond, permettant d'obtenir des résultats jusque-là inatteignables. L'EPH a atteint un stade de maturité technologique avancée, des algorithmes sont maintenant disponibles sous forme de kit de développement (SDK) à l'achat de caméras commerciales du type Kinect ou Intel RealSense. Dans le domaine du divertissement, les performances obtenues sont souvent très satisfaisantes. Toutefois, pour les applications médicales, les travaux de recherche ne sont pas encore assez concluants [1].

L'estimation de pose de sujets humains alités, quant à elle, n'est que très peu étudiée dans la communauté scientifique. Elle présente des défis spécifiques qui nécessitent une approche différente. En effet, les modèles d'EPH classiques sont entraînés sur des images en couleur de personnes faisant des activités quotidiennes, souvent en plein air (sport, travail…) et en pleine journée, tandis que la détection de pose d'humain alité traite le cas de personnes couchées dans un lit, souvent d'hôpital, avec la possibilité d'occlusions visuelles notamment dues à la présence de couverture au-dessus du patient et avec une variabilité en termes d'illumination dans la chambre d'hôpital. De plus, les poses possibles en position couchée sont différentes de celles que l'on retrouve dans un jeu de données en posture debout, ce qui veut dire qu'un modèle entraîné sur des images de personnes faisant des activités régulières de tous les jours ne généralisera pas aussi bien sur des images de sujets alités [2]. Il est donc nécessaire d'entraîner un modèle sur une base de données contenant des images de sujets alités et, afin de relever les défis d'occlusion et de variation de lumière, d'utiliser d'autres modalités qui fonctionnent sous ces conditions, notamment des images de profondeur et infrarouges. Cependant, l'utilisation de différentes modalités entraîne la nécessité de trouver une méthode de fusion adéquate qui permettra de combiner les modalités afin de relever les manquements de chacune d'elles. En effet chaque modalité présente certaines limites; les images couleurs ne sont plus utilisables en cas de présence d'occlusions dues à la présence de couverture ou dans le cas où la chambre dans laquelle se trouve le patient est mal illuminée, les images infrarouges quant à elles permettent de mitiger ces limites, mais la performance de l'inférence peut être significativement affectée par la présence de résidus de chaleur sur le lit (ex : si le patient reste longtemps dans une position donnée, des résidus de chaleurs s'accumulent sur le lit et l'inférence de la pose peut être faussée à cause de cela). Les images de profondeur ne présentent pas ces mêmes inconvénients, mais la performance peut chuter en cas d'occlusion d'un membre du corps par un autre, un exemple étant si le patient dort sur le côté.

Trouver une méthode de fusion de modalités adéquate est donc impératif dans le cadre de ce travail. En apprentissage profond, les méthodes de fusion sont divisibles en trois catégories, dépendamment de la profondeur à laquelle se fait la fusion dans le réseau de neurones utilisé. On parle de fusion en amont (Early fusion), lorsque la fusion est effectuée au début du réseau, fusion intermédiaire (Intermediate fusion) lorsqu'elle est effectuée à un niveau intermédiaire du réseau et finalement, fusion en aval (Late fusion) lorsqu'elle est effectuée à la fin du réseau, juste avant la sortie.

Additionnellement, nous nous intéresserons aux différentes combinaisons de modalités afin de trouver la combinaison optimale dans le cadre de l'EPH en position couchée.

Ce travail de recherche servira de socle pour de futurs travaux qui serviront à déployer un système de monitorage des mouvements des patients aux soins intensifs du CHUSJ.

Ce mémoire est organisé en cinq chapitres. Le premier présente une revue des connaissances présentant les notions nécessaires à la compréhension du projet ainsi que l'avancement et l'état de l'art de chacune des composantes du sujet de recherche. Le second chapitre présente la problématique de recherche et les objectifs spécifiques de ce travail de recherche. Le troisième chapitre présente l'article soumis en juillet 2022 à la revue IEEE Biomedical and Health Informatics sur l'estimation de pose humaine en position alitée à l'aide de l'apprentissage multimodal profond. Le quatrième chapitre présente une discussion générale des travaux réalisés et des résultats obtenus. Enfin, le dernier chapitre conclut sur les points importants du projet et propose des avenues de recherche subséquentes.

CHAPITRE 2 REVUE CRITIQUE DE LA LITTÉRATURE

Dans ce chapitre, nous allons passer en revue les connaissances nécessaires à la compréhension du projet ainsi que l'état de l'art en EPH. Nous allons commencer par introduire les réseaux de neurones, qui constituent la base des systèmes d'EPH modernes. Ensuite nous allons expliquer ce qu'est l'EPH, classique et alité, ainsi que les avancements dans ces domaines. Finalement nous introduirons l'apprentissage profond multimodal et son importance dans l'EPH alité.

2.1 Du neurone au réseau de neurones convolutifs

Les méthodes d'estimation de pose les plus performantes à ce jour se basant toutes sur des réseaux de neurones, nous ne nous attarderons pas sur les méthodes classiques et nous élaborerons plutôt sur les réseaux de neurones et leur application en vision par ordinateur.

2.1.1 Neurone artificiel

Le neurone est le composant le plus basique d'un réseau de neurones, mais aussi le plus important. L'idée du neurone artificiel a commencé à partir d'un modèle du fonctionnement des neurones du cerveau, développé en 1943 par les neurophysiologistes Warren McCulloch et Walter Pitts [2]. D'après ce modèle, un neurone reçoit en entrée des signaux pondérés d'autres neurones et retourne 0 ou 1 dépendamment de si la somme des entrées est au-dessus ou au-dessous d'un certain seuil. Le neurone artificiel moderne se comporte de façon similaire, il reçoit en entrée des signaux d'entrées et ressort la somme pondérée de ces signaux après l'avoir transformé via une fonction d'activation non linéaire telle que ReLU (Rectified Linear Unit) [3]. La Figure 2.1 illustre la structure générale d'un neurone artificiel.

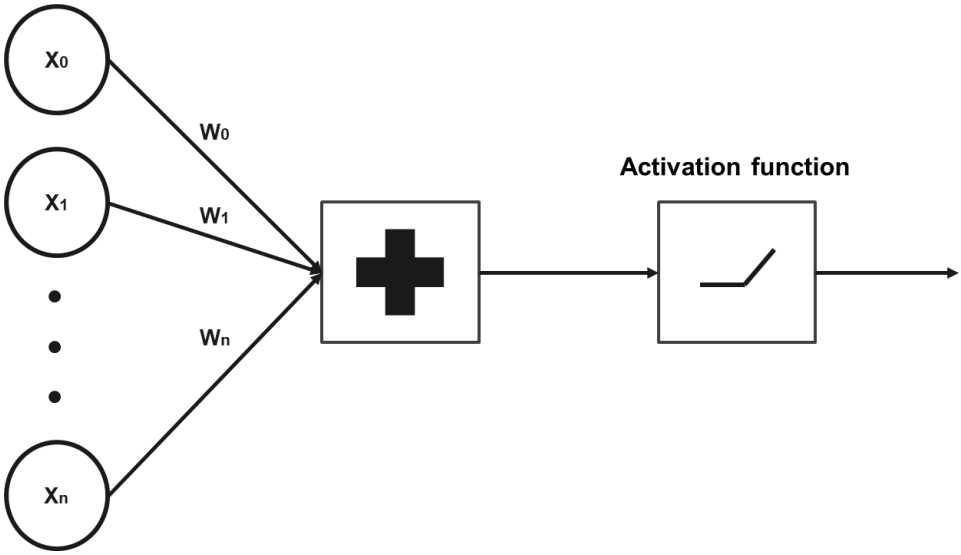


Figure 2.1 Structure d'un neurone artificiel

2.1.2 Réseau de neurones

Plusieurs neurones arrangés en une couche, ou chaque neurone est connecté à tous les neurones de la couche précédente constituent un réseau de neurones, tel qu'on peut le voir à la Figure 2.2. Les couches qui ne sont pas les couches d'entrées et de sorties sont appelées des couches cachées.

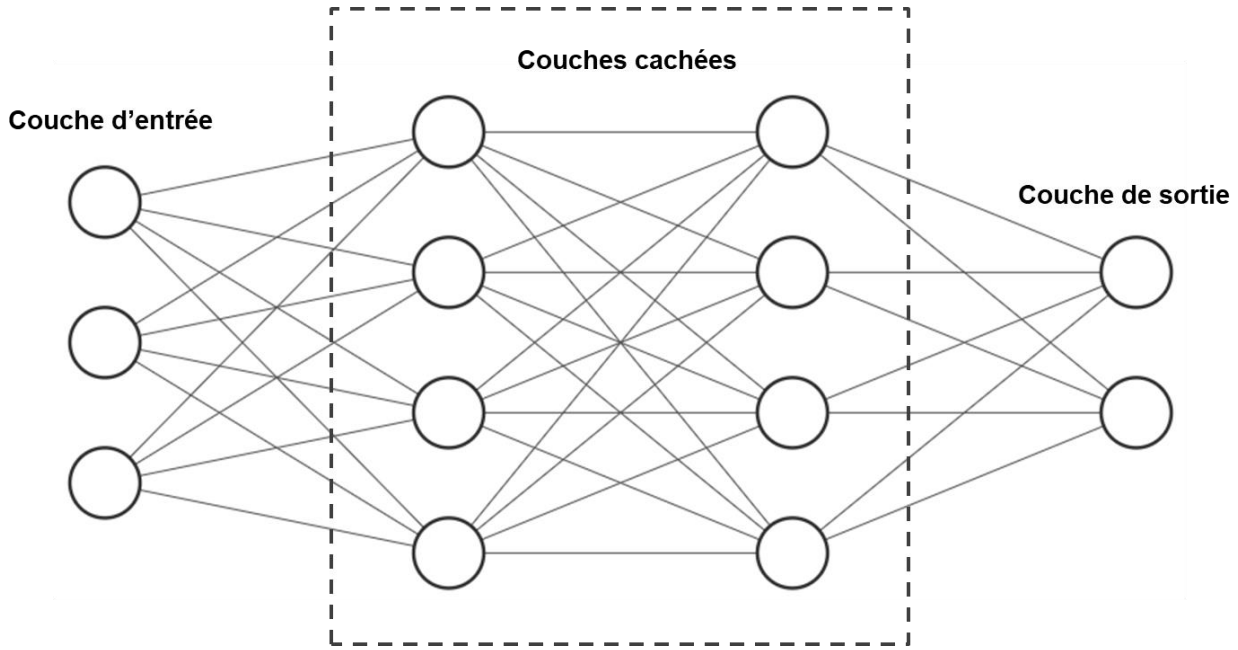


Figure 2.2 Structure d'un réseau de neurones artificiel

Cet arrangement de neurones permet de représenter des fonctions non-linéaires. Les poids des connexions entrants dans chaque neurone sont initialement assignés aléatoirement (échantillonnée d'une distribution normale par exemple) et ensuite appris via un mécanisme de rétropropagation basé sur la descente de gradient. En effet, lorsque les données d'entraînement sont entrées dans le réseau, l'information est propagée à l'avant du réseau via les neurones vers les nœuds de sorties. Une fonction de perte qu'on essaye d'optimiser est ensuite calculée, et la dérivée partielle de la fonction de perte en fonction de chaque poids est déterminée, ces dérivées partielles qu'on appelle gradients sont ensuite utilisées pour modifier la valeur de chaque poids de façon à minimiser la fonction de perte.

Lorsque plusieurs couches de neurones cachées sont empilées, le réseau est appelé un réseau de neurones profond. Empiler plusieurs couches permet de représenter des fonctions de plus en plus complexes, mais introduit un défi supplémentaire; la disparition du gradient. En effet, empiler plusieurs couches fait en sorte que les gradients qui permettent de mettre à jour les poids tendent vers 0, faisant ainsi en sorte que le réseau n'arrive plus à apprendre s'il dépasse une certaine profondeur. En effet, un poids de neurone est mis à jour selon l'équation suivante :

$$w^{nouveau} = w^{ancien} - \eta \frac{dL}{dw}$$

Où $\frac{dL}{dw}$ est la dérivée de la fonction de perte en fonction du poids et η est le coefficient d'apprentissage.

Si $\frac{dL}{dw}$ est très petit, l'apprentissage sera lent, voire impossible, or les gradients des couches initiales sont obtenus, selon la méthode de rétropropagation de gradient, en multipliant les gradients des couches subséquentes, ainsi si un des gradients des couches subséquentes est inférieur à 1, le gradient risque de disparaître.

Les chercheurs ont cependant trouvé une solution à ce problème en introduisant des connexions résiduelles [4] telles que le montre la Figure 2.3. Une connexion résiduelle est une fonction identité qui permet à l'information d'une couche de sauter quelques couches, ainsi, l'information d'une couche l sera additionnée à celle d'une couche $l + t$ ou $t \geq 2$. Les connexions résiduelles permettent aux informations de gradient de mieux circuler dans le réseau et ont permis d'avoir des réseaux substantiellement plus profonds.

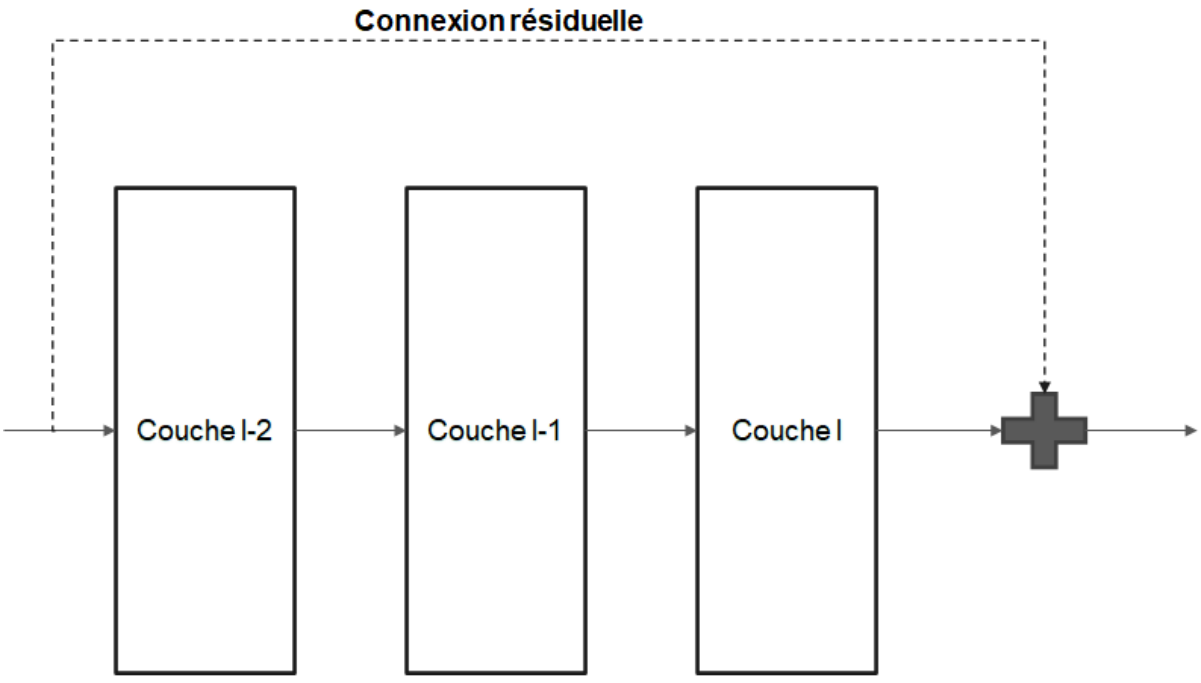


Figure 2.3 Connexion résiduelle

2.1.3 Réseau de neurones convolutif

Les réseaux de neurones denses, c'est-à-dire pour lesquels tous les neurones d'une couche sont connectés à tous les neurones de la couche précédente, ne sont cependant pas adéquats aux taches de traitement d'images. Ils ne permettent pas d'extraire des propriétés spécifiques aux images telles que les contours, les formes, possibilité de présence de symétrie dans l'image et de plus, nécessitent beaucoup plus de paramètres que les réseaux convolutifs.

Un réseau convolutif (CNN) est composé de filtres, ou chaque filtre est convolué sur la sortie de la convolution précédente, permettant ainsi d'extraire des caractéristiques de plus en plus sophistiquées. Les filtres des premières couches du réseau permettent d'extraire des caractéristiques basiques telles que les formes géométriques dans l'image tandis que les filtres des dernières couches permettent d'extraire des caractéristiques et formes plus complexes. Les valeurs dans les filtres de convolutions sont dans ce cas les poids du réseau, qu'il s'agit d'apprendre.

2.2 Estimation de pose humaine

2.2.1 Définition

L'estimation de pose humaine est une application importante de la vision par ordinateur.

L'EPH consiste à restituer, à partir d'images ou de vidéos, une représentation 2D ou 3D du corps humain du ou des sujets présents dans les images [5]. Il existe trois types de modèles utilisés pour la représentation du corps:

- Le modèle basé sur le squelette : C'est le modèle le plus utilisé pour l'EPH. Il consiste à détecter un ensemble de points-clés relatifs aux articulations (coude, genoux…). Ces points-clés sont ensuite connectés entre eux pour créer un modèle squelettique du corps humain. Les coordonnées des points clés peuvent être 2D ou 3D.
- Le modèle basé sur le contour : Ce modèle est utilisé pour l'EPH 2D et consiste à détecter des rectangles englobants autour des membres du corps humain. Ainsi au lieu de représenter le corps par un squelette, chaque membre est représenté par un rectangle définissant le contour de celui-ci.
- Le modèle basé sur le volume : Celui-ci est utilisé pour l'EPH 3D. Il consiste à produire une représentation volumétrique du corps humain. Ainsi, il est différent des modèles précédents dans la mesure où celui-ci encode également l'information relative à la forme du corps humain.

Dans le présent travail, nous nous intéressons à une représentation squelettique 2D du corps humain.

2.2.2 Métriques de performance

Plusieurs métriques sont utilisées pour évaluer la performance d'un modèle d'EPH [6], les plus largement utilisées étant :

Percentage of Correct Parts (PCP): Cette métrique mesure le taux de détection des membres du corps. Un membre est considéré comme étant détecté si la distance entre les deux articulations d'un membre détecté et la vraie position des articulations est inférieure à la moitié de la vraie longueur

de ce membre. PCP n'est plus très répandue comme métrique car elle pénalise les membres les plus petits du corps.

Percentage of Detected Joints (PDJ): Cette métrique a été proposée dans l'objectif de pallier les limitations de PCP. Une articulation est considérée comme correctement détectée si la distance entre l'articulation prédite et la vraie position de cette articulation est inférieure à une certaine fraction du diamètre du torse, c'est-à-dire la distance entre la cote droite et l'épaule gauche de la personne. PDJ@0.2 signifie par exemple que la fraction du torse est fixée à 0.2, donc une articulation est considérée comme étant détectée si la distance entre sa position prédite et sa vraie position est inférieure à 20% du diamètre du torse.

Percentage of Correct Key-point (PCK): Pour cette métrique, une articulation prédite est considérée comme correctement prédite si la distance entre celle-ci et sa vraie position est inférieure à un certain seuil définit comme étant une fraction du côté le plus long de la boîte englobante (bounding box) autour de la personne.

PCKh est une version modifiée de PCK. Le seuil dans PCKh est défini comme étant 50% de la longueur du segment de la tête. La longueur du segment de la tête est définie quant à lui comme étant 60% de la diagonale de la boîte englobante autour de la tête de la personne. PCKh@0.5 signifie que le seuil est défini comme étant 50% de la longueur du segment de la tête.

2.2.3 Modalités d'imagerie utilisées

Une modalité d'image fait référence à la méthode d'acquisition ou l'information qu'encode une image;

- Les images RGB encodent dans leurs trois canaux l'information relative à la présence des couleurs vertes, bleues et rouges pour chaque pixel.
- Les images de profondeur (Depth) encodent quant à elles dans un canal l'information relative à la profondeur, c'est-à-dire la distance séparant un point de la scène de la caméra dans l'axe de la focale.
- Les images infrarouges (IR) encodent dans un canal l'information relative aux ondes infrarouges émises par la scène et captées par les senseurs de la caméra.

Bien que le RGB soit la modalité la plus utilisée pour l'EPH, les modalités de profondeur [1] et d'IR [7] sont aussi utilisées dans certaines applications et circonstances pour les avantages qu'elles présentent par rapport aux images de couleurs.

En effet l'inférence de la pose à partir d'images RGB peut devenir compliquée en cas de faible illumination ou de présence d'occlusion.

L'utilisation d'images IR ou de profondeur permet de pallier ces problèmes, mais présente d'autres défis. En effet, lorsqu'une personne est allongée sur un lit pendant une longue période sans changer de pose, des résidus de chaleurs apparaissent sur le lit, faisant en sorte que même si la personne change de pose, ces résidus induisent en erreur le système d'EPH à partir d'images IR. L'EPH à partir d'images de profondeur présente quant à elle un défi différent; l'estimation de l'information de profondeur d'une personne couchée peut se compliquer s'il y'a occlusion d'un des membres de cette personne par un autre membre (par exemple si la main de la personne est sous sa tête).

Le choix de modalité se fait donc généralement selon les conditions de prise d'image (i.e. illumination, occlusion) et selon le problème spécifique que l'on essaye de résoudre.

2.2.4 Solutions basées sur l'apprentissage profond

L'EPH a connu une évolution drastique depuis l'avènement de l'apprentissage profond et la multiplication des bases de données spécifiques à l'estimation de pose. Depuis, les percées les plus importantes dans ce domaine ont été fortement couplées avec celles de l'apprentissage profond. L'estimation de pose humaine par apprentissage profond peut être classifiée en approches basées sur la régression et en approches basées sur la prédiction de cartes de chaleur (Heatmaps), dépendamment de la façon avec laquelle les points-clés représentant les articulations sont prédits.

L'approche basée sur la régression consiste à directement prédire les coordonnées des points-clés sous forme de vecteur $y = ({y_1}^T,\ {y_2}^T,\ \dots, {y_i}^T),\ i \in \{1, \dots, k\}$ où $y_i = (x, y)$ représente les coordonnées d'une des articulations et k représente le nombre d'articulations que l'on essaye de détecter.

La première implémentation d'un réseau de neurones pour la tâche d'EPH, Deep Pose, se basait sur une approche régressive. Le modèle était basé sur l'architecture AlexNet [9] qui est constituée de 5 couches convolutives, 2 couches denses et un classificateur softmax. Une image d'une personne est passée à ce modèle qui permet d'obtenir en sorties les coordonnées estimées des

articulations. Ces estimations sont ensuite utilisées pour raffiner les résultats de la prédiction en utilisant plusieurs régressions successives. Des modèles subséquents se sont basés sur les architectures R-CNN [10] et Mask R-CNN [11] et les modèles les plus récents sur l'architecture ResNet [5].

L'approche basée sur les cartes de chaleur consiste quant à elle à prédire des cartes de chaleur de la même taille que l'image d'entrée. Il y a autant de cartes de chaleurs qu'il y a de points-clés. Pour une carte, la valeur de sortie au pixel *(x,y)* représente la probabilité de présence du point-clé à cette position dans l'image.

Les premiers modèles d'EPH se basant sur l'apprentissage profond utilisaient des architectures basées sur des réseaux de neurones convolutifs qui réduisent graduellement la résolution spatiale des caractéristiques générées par chaque couche de réseau, puis qui appliquent un suréchantillonnage afin d'obtenir les cartes de chaleur. Cependant la réduction graduelle de la résolution spatiale des cartes de caractéristiques, aux dépens de la sémantique, cause une perte d'information qui ne peut être rétablie de manière précise avec le suréchantillonnage (up-sampling).

Les architectures plus récentes telles que Stacked Hourglass [8] et High-Resolution Net [9] se basent sur la prédiction de cartes de chaleur et ont incorporées des 'Skip layers', inspirés des connexions résiduelles, afin de préserver la haute résolution des caractéristiques, permettant ainsi d'obtenir une représentation sémantique et spatiale plus riche des représentations de caractéristiques.

De plus, les architectures comme Stacked Hourglass utilisent aussi plusieurs étapes de raffinement de résultats tel qu'on peut le voir à la Figure 2.4, ou chaque étape améliore les résultats de l'étape précédente; la première étape prédit des cartes de chaleurs à partir des images en entrées et les étapes qui lui succèdent utilisent non seulement les images en entrée, mais aussi les cartes de chaleurs de l'étape précédente, gagnant ainsi de l'information contextuelle qui permet d'améliorer le résultat.

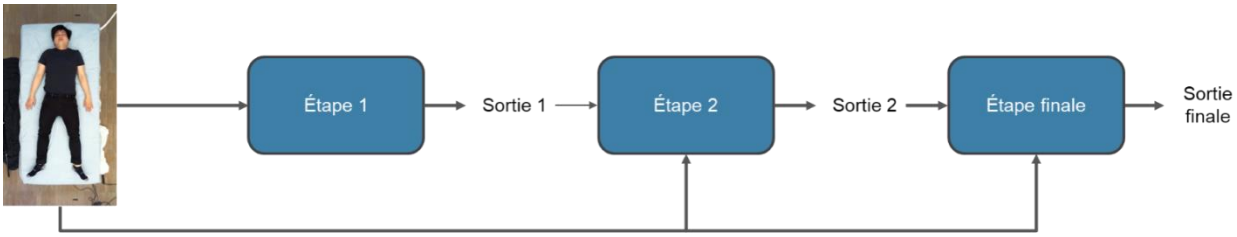


Figure 2.4 Architecture avec étapes de raffinement

2.2.5 Les bases de données

Un autre facteur qui a permis de développer des modèles d'EPH de plus en plus performants à part la conception de modèles efficaces est la création de bases de données spécifiques à l'EPH telle que MPII Human Pose [10] qui contient 25 000 images RGB annotées, COCO [11] qui en contient plus de 200 000 et Human3.6M [12] qui en contient 3 600 000. Ces bases de données ont permis sont d'une grande importance à cause de la difficulté d'annoter ce type de données et du fait qu'elles ont permis de répondre aux besoins des modèles d'apprentissage profond pour une vaste quantité de données pour l'entraînement. Ces bases de données contiennent une grande variété d'images de personnes faisant des activités diverses, mais ne contiennent pas des images de personnes alitées, ce qui limite la généralisation des modèles entraînés sur celles-ci.

2.3 L'estimation de pose de sujets alités

Une des applications plus récentes et de nature plus complexe de l'EPH est l'EPH de sujets alités, c'est-à-dire en position allongée. Le travail de Liu et al. [13] a démontré que l'utilisation de modèle d'EPH classique, c'est-à-dire entraîné sur des poses générales ne permet pas de bien généraliser sur les poses de personnes alitées. Ceci s'explique par le fait que les poses possibles en position alitée diffèrent des poses en position debout telle qu'on peut le voir dans la Figure 2.5. De plus les occlusions visuelles dues à la présence de couverture sur les personnes allongées et la variabilité de l'illumination dans l'environnement du sujet alité rajoutent une couche de difficulté qui n'est pas prise en considération par les modèles d'EPH classiques.

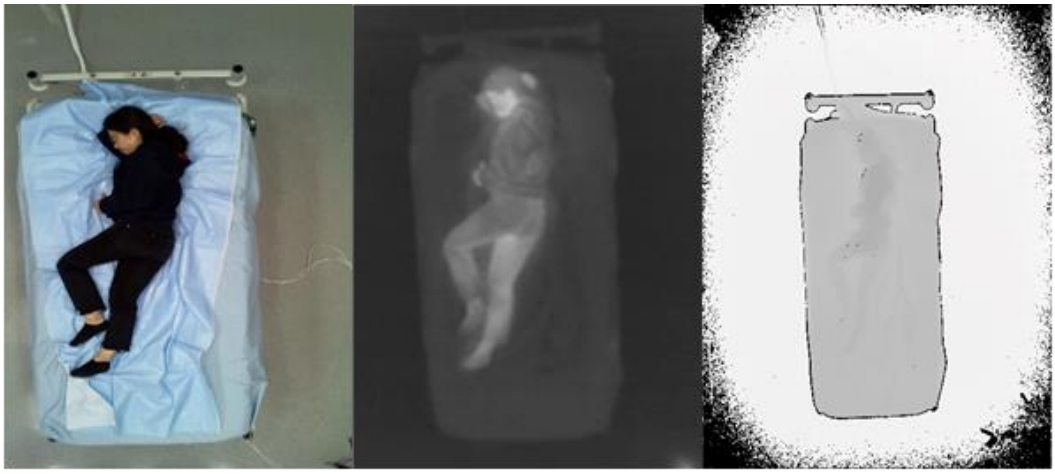


Figure 2.5 Images prises sous différentes modalités

Afin d'adresser ces problèmes spécifiques à l'EPH alité, Liu et al. se sont concentrés sur la collection et l'annotation d'une base de données spécifiques à cette tâche, qu'ils ont appelé Simultaneously-collected multi-modal Lying Pose (SLP), contenant des images de personnes en position alitée sous différentes modalités : RGB, profondeur, proche infrarouge (LWIR) et matelas de pression. Ils ont ensuite entraîné des modèles avec différentes architectures, notamment HRnet et StackedHourGlass, sur chacune des modalités séparément, atteignant la précision la plus haute en utilisant la modalité de profondeur et le modèle StackedHourGlass : un PCKh@0.5 de 96.6%.

Cette base de données a aussi été utilisée par d'autres travaux, notamment Cao et al. [14] qui ont exploité la présence de plusieurs modalités pour concevoir un système robuste capable d'estimer la pose de façon précise en l'absence d'une des modalités en utilisant une méthode d'apprentissage auto-supervisé, atteignant une précision de 94.3% en termes de PCKh@0.5.

D'autres travaux impliquant SLP se sont concentrés sur des tâches plus en amont telles que l'estimation de pose et de formes 3D [15]. Ces travaux se sont aussi basés sur des méthodes du fusion des modalités RGB, IR et profondeur pour inférer la forme des patients couverts par une couverture.

Cependant, aucun des travaux précédemment cités ne s'est concentré à trouver la méthode de fusion multimodale qui optimise l'estimation de pose des patients alités par apprentissage profond.

2.4 Apprentissage profond multimodal

Plusieurs taches qui peuvent être résolues grâce à l'apprentissage profond sont de nature multimodale. Par exemple, l'être humain afin de comprendre des conversations intègre des données visuelles (mouvement des lèvres) et auditives et interagit avec son environnement à l'aide de ses 5 sens, c'est-à-dire qu'il utilise 5 différentes sources d'informations. Ainsi, lorsque plusieurs modalités sont disponibles, n'en utiliser qu'une peut être sous-optimal, par exemple, dans des problèmes de vision par ordinateur tel que la détection d'êtres humains à partir d'images de drone, n'utiliser que des images RGB peut s'avérer problématique lorsque le contraste de couleur dans les images n'est pas grand ou encore lorsqu'il y a une variabilité de lumière (exemple : image prise la nuit), ainsi, rajouter une autre source d'information, notamment des images infrarouges permet de paliers aux faiblesses inhérentes à l'utilisation du RGB, d'où la nécessité de fusionner les modalités disponibles afin d'en tirer le plus d'information. L'apprentissage profond multimodal consiste donc en la création de modèles capable de traiter et connecter des données provenant de différentes sources [17]. Les différentes modalités peuvent être intégrées dans un modèle d'apprentissage profond selon différentes stratégies de fusion, chacune présentant des avantages et des inconvénients. La méthode de fusion optimale doit être capable d'exploiter la complémentarité entre les différentes modalités (inter-modal processing) tout en préservant les caractéristiques spécifiques à chaque modalité (intra-modal processing) [18].

2.4.1 Fusion en amont

La fusion en amont consiste à fusionner les données en une représentation unifiée au début du réseau ou avant de la placer à l'entrée dans le réseau. Un exemple d'une méthode de fusion en amont est de concaténer des images provenant de différentes modalités d'images avant d'envoyer le résultat de cette concaténation en entrée dans le réseau [19].Une autre méthode de fusion en amont nommé DenseFuse [20] combine les différentes modalités à l'aide d'une architecture encodeur-décodeur. Le résultat de cette fusion est ensuite envoyé au réseau de neurones.

La fusion en amont, bien qu'elle permette d'avoir un bon traitement d'information inter-modal, peut cependant causer une perte d'information intra-modale lorsque la fusion est faite, vu que le réseau va extraire les caractéristiques de la représentation des modalités et ne pourra donc pas extraire les caractéristiques de chaque modalité.

2.4.2 Fusion en aval

La fusion en aval consiste à passer chaque modalité par un réseau de neurones qui lui est spécifique, et ensuite combiner les sorties de chaque réseau. Un exemple de fusion en aval est un réseau d'ensemble [21]; les décisions de chaque réseau spécifique à une modalité sont combinées via une opération (ex : maximum ou moyenne) afin d'obtenir la décision finale. Cependant, dans la fusion en aval, les réseaux n'interagissent entre eux qu'après que leurs décisions soient prises, ce qui fait que le traitement inter-modal soit très minime dans cette catégorie de fusion.

2.4.3 Fusion intermédiaire

Dans cette catégorie de fusion, la fusion se fait au niveau des caractéristiques intermédiaires, c'est-à-dire au niveau des couches intermédiaires des réseaux de neurones.

Par exemple, Multimodal Transfer Module (MMTM) [22] est une méthode de fusion intermédiaire qui consiste à implémenter un module entre deux réseaux convolutifs, chacun étant spécifique à une modalité. Ce module permet l'échange d'informations entre les deux réseaux et de recalibrer leurs caractéristiques spécifiques. Une autre méthode de fusion intermédiaire nommée Channel-exchange (CE) [23] consiste quant à elle en l'échange de canaux entre les réseaux parallèles, l'échange étant guidé par le facteur de mise à l'échelle de la normalisation de lot [24] (Batch normalization).

Cette méthode de fusion permet un compromis entre le traitement intra-modal et inter-modal vu qu'avant qu'il n'y ait de fusion, c'est-à-dire au niveau des couches initiales des réseaux de neurones de chaque modalité, les caractéristiques spécifiques à chaque modalité ont le temps d'être traitées, puis, une fois que la fusion est faite au niveau des couches intermédiaires, le modèle apprend à exploiter les complémentarités entre les modalités.

CHAPITRE 3 HYPOTHÈSE ET OBJECTIFS DE RECHERCHE

La revue de littérature a permis de synthétiser l'avancement de la recherche dans le domaine de l'EPH et l'EPH en position alitée ainsi que les enjeux actuels et a donc permis d'orienter nos travaux de recherche. Pour résumer, les réseaux convolutifs de type Stacked Hourglass présentent des performances considérables sur des bases de données de sujets en posture debout. Toutefois, l'EPH en position alitée, notamment dans le cadre du monitorage de patients en lit d'hôpital, bien que semblable dans son objectif à l'EPH, présente des défis qui lui sont spécifiques :

- La différence de poses possibles en position alitée et en position debout
- La présence d'occlusions dues à la couverture au-dessus de la personne allongée
- La variation de luminosité dans la chambre dans laquelle se trouve le patient, dépendamment de si la lumière est allumée ou pas dans la chambre

L'ensemble de ces facteurs font en sorte qu'il est nécessaire d'utiliser des modalités d'images qui permettent de mitiger ces problèmes.

Une base de données contenant des images de patients alités, sous différentes modalités, se nommant Simultaneously-collected multimodal Lying Pose (SLP) est disponible et a été utilisée dans des travaux de recherche visant à utiliser des modalités autres que le RGB. Cependant, la plupart des travaux réalisés sur SLP n'exploitent pas la nature multimodale de la base de données et proposent des modèles unimodaux. De plus, les résultats qu'ils obtiennent ne surpassent pas les résultats obtenus par les créateurs de SLP (96.6% PCKh) en utilisant l'unique modalité de profondeur.

L'hypothèse de cette recherche est qu'une approche d'apprentissage profond multimodale augmenterait la performance d'un modèle d'EPH en position couchée par rapport au meilleur résultat jusque-là obtenu sur la base de données SLP (96.6% PCKh), se basant sur un modèle uni-modal.

Pour vérifier cette hypothèse, deux questions se posent : quelles modalités devraient être combinées? Et comment fusionner optimalement ces modalités?

Les objectifs spécifiques de ce travail sont donc de :

- Trouver la meilleure combinaison de modalités à fusionner

- Réaliser une comparaison exhaustive entre différentes stratégies de fusion multimodale.

CHAPITRE 4 ARTICLE 1: MULTIMODAL DEEP LEARNING FOR IN-BED HUMAN POSE ESTIMATION

Authors. Ghassen Cherni, Quentin Cappart, Philippe Jouvet, and Lama Seoud. Submitted to IEEE - Journal of Biomedical and Health Informatics (JBHI, 18 juillet 2022)

4.1 Introduction

Human pose estimation (HPE) constitutes the basis of a continuous contact-less monitoring of a subject's movements. It is a computer vision task required in several healthcare applications such as the development of intelligent context-aware assistance systems in the operating room [25], the detection of sleep apnea and restless legs syndrome [26], the prevention of pressure ulcers in bedridden patients [27] or the neurological monitoring of critically ill patients (movement analysis for estimation of sedation level, coma, seizures for example) [28]. Using images taken from cameras to estimate the pose of the patients is a non-intrusive and relatively inexpensive method that can be used in addition to visual inspection by nurses or as a replacement to invasive devices such as bracelets. It is also a cheaper solution than the use of pressure mats that are both expensive and less accurate [14]. In order for a camera-based solution to be a viable alternative, it has to be both accurate and low in latency during the inference of the pose. Tremendous work has been done these past years in 2D HPE with deep learning based algorithms achieving high performances [29]. The majority of this work has been done to detect poses of standing subjects under general settings (doing sports or walking for example). In-bed HPE has the same goal of predicting 2D joint coordinates (x,y) in pixel space but it presents specific challenges such as occlusions due to the presence of clothing and/or blankets, highlighting variability in the room and specific lying poses that cannot be reproduced standing.

Some recent work has focused on creating an annotated domain-specific dataset for in-bed HPE containing RGB, LWIR and depth images [14]. Additionally, unimodal models were created to infer poses from these images. However, using a single modality for in-bed pose estimation can be sub-optimal since each modality has drawbacks; indeed, RGB images are not suited when the patient is covered by a blanket or if the light in the room is turned-off. LWIR can mitigate these problems but the performance of the inference can be significantly affected by the presence of heat residues on the bed. As for depth, the performance can be hindered if the patient is sleeping on his

side since some limbs will be occluded by others, making depth information related to the occluded limbs inaccurate.

Based on this context, the main objective of this paper is to determine ”what”,” where” and ”how” to fuse the modalities in order to provide the most accurate in-bed HPE. To this end, we implement and compare different fusion methods using a public multimodal domain-specific dataset. The specific contributions are as follows.

- Finding the best combination of modalities to fuse
- An extensive comparison of different multimodal fusion methods on a public dataset
- A multimodal fusion method that allows to outperform the current benchmark

We find that combining the Depth and LWIR modalities yields the best results, surpassing methods that use a unique modality and methods that combine Depth, LWIR and RGB. Moreover, after comparing the Channel-exchange, Multimodal Transfer Module, Concatenation, DenseFuse and Ensemble modality fusion methods on a domain-specific dataset called Simultaneously-collected multi-modal Lying Pose, we find that CE gives the best result, 97.1%, and allows to surpass the previous best result of 96.9% achieved on this benchmark dataset.

4.2 Related works

In-bed HPE, a sub-branch of HPE, presents similar objectives with additional challenges. Therefore, we first cover the work done in HPE before addressing the specific work done for in-bed pose estimation.

4.2.1 Human pose estimation

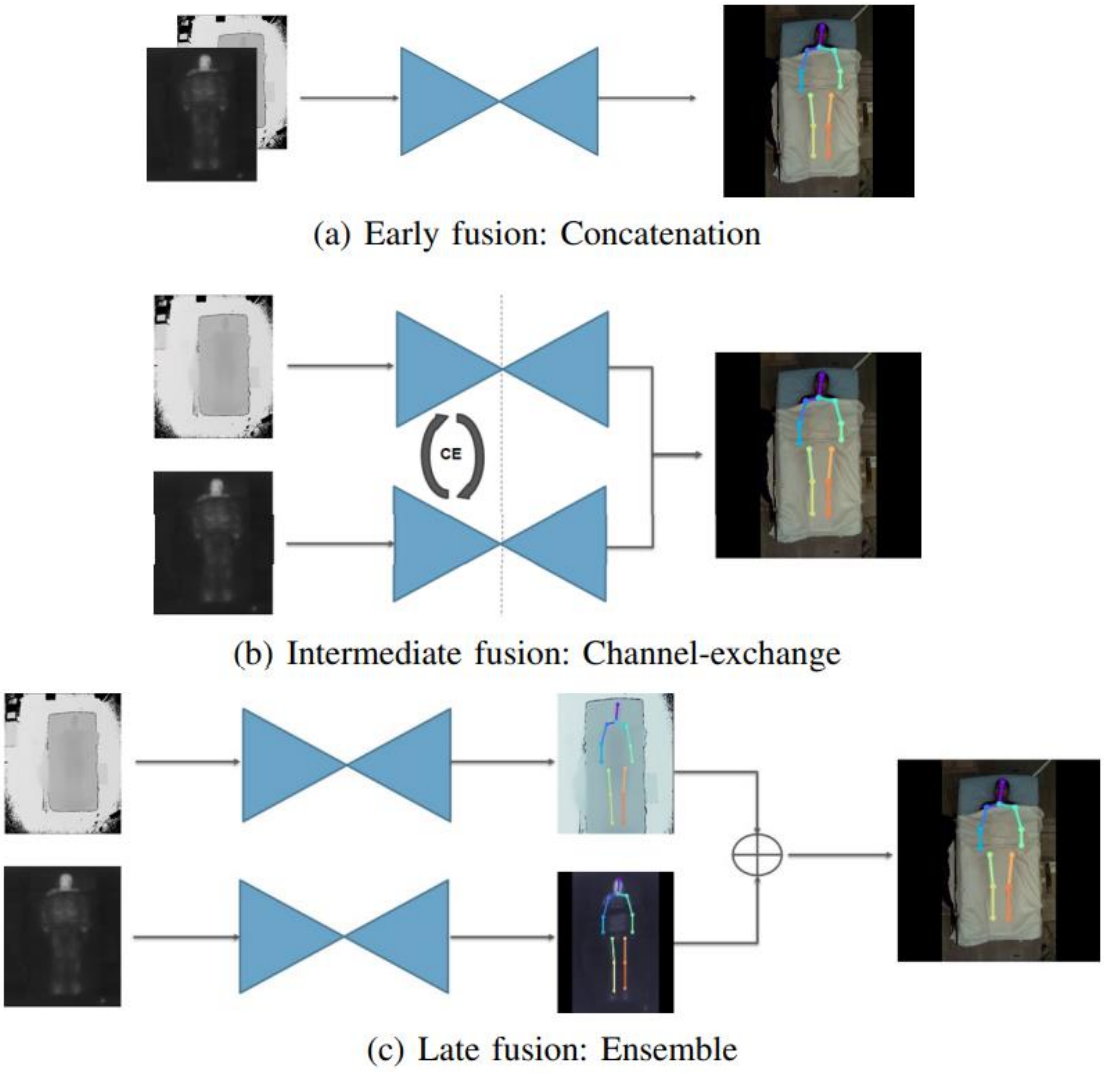
HPE is a widely studied field in computer vision [30], it has known a tremendous evolution with the advent of deep learning [31] and the creation of new HPE-specific datasets. Since then, major breakthroughs in the field were tightly coupled with those of deep learning. HPE can be classified into regression based approaches and heatmap based approaches [29], depending on the way keypoints are predicted. Regression based approaches directly predict the (x, y) coordinates of the keypoints, while heatmap based approaches predict heatmaps, one for each keypoint, where each pixel in the heatmap corresponds to the probability of the presence of the keypoint in that position.

The first pose estimation systems that relied on deep learning [32] used architectures consisting of convolutional neural networks (CNN) [33] that gradually reduced the feature representations, allowing the model to regress towards the joints coordinates.

More recent architectures such as the Stacked Hourglass Network [9] and high-resolution network [23] incorporated skip layers inspired from residual layers [5] in order to preserve high resolution features, allowing for semantically richer and spatially more precise representations.

Multiple architectures also use a multi-stage approach [34], [35], [9], where each stage improves the results of the previous one. The first stage predicts heatmaps from the input images and subsequent stages use, additionally to the input image, the predicted heatmaps from the previous stages, therefore gaining contextual information that is used to improve the pose estimation.

Additionally, the availability of HPE-specific datasets such as MPII Human Pose [11], COCO [12] and Human3.6M [13] allowed to leverage the deep learning models' need of vast amounts of data. These datasets present a variety of images of people doing daily life activities, with their corresponding joints coordinates as annotations. However, these popular datasets do not include images of people in a lying position, which limits the generalization of models trained on these datasets to in-bed poses.



(a) Early fusion: Concatenation

(b) Intermediate fusion: Channel-exchange

(c) Late fusion: Ensemble

Figure 4.1 Examples of fusion methods

4.2.2 In-bed human pose estimation

In-bed HPE is a recent sub-branch of HPE that is beginning to receive more attention by the computer vision community. Work by Liu et al. [36] showed that using a HPE model trained on general purpose pose estimation datasets is sub-optimal when it comes to predicting the pose of in-bed patients due to the difference in pose distributions and the domain-specific challenges such as the presence of occlusion and lighting variability.

To address this limitation, more recent work by Liu et al. [14], [37] focused on collecting an annotated domain-specific dataset for in-bed HPE called Simultaneously-collected multi-modal

Lying Pose (SLP). They trained different models taking each modality separately. The best performance is achieved when considering the depth modality.

The SLP dataset was also used for tasks that leveraged the presence of different modalities to create robust multimodal systems that can work even in the absence of one of the modalities by using a self-supervised framework such as in the work of Cao et al. [15] that slightly improved the PCKh@0.5 score on the SLP dataset from 94.2%, obtained using the LWIR modality, to 94.3% by combining the LWIR and RGB modalities.

Therefore, the current benchmarks on the SLP dataset are 96.6% when using the depth modality and 94.3% when combining the RGB and LWIR modalities.

Other work [16] focused on more upstream tasks such as not only pose but also 3D shape estimation. They also relied on multimodal fusion of depth, IR and RGB images from SLP (by using a cascaded concatenation) to infer the shape of the subject under a blanket.

Nevertheless, none of these previous works focused on finding an optimal multimodal deep learning method to improve the in-bed pose estimation.

4.3 Methods and Experiments

4.3.1 Fusion methods

Multimodal deep learning based fusion methods can be categorized into early, intermediate or late fusion based on whether the fusion is done at the input level, the feature-level or the output level of the network [38], [18]. The challenge in the fusion process is to exploit the complementarity of the modalities (inter-modal processing) while keeping their respective characteristics (intra-modal processing).

Late fusion methods tend to downplay the inter-modal processing while early fusion methods, by aggregating the modalities via a set of operations such as concatenation [39], tend to cause a disruption in the intra-modal processing when the aggregation is done. Intermediate fusion methods therefore provide a reasonable compromise between intra-modal and inter-modal processing.

In this paper, in order to find the most suitable fusion for in-bed HPE, we compare 5 methods from different categories of multi-modal fusion, some examples of which are shown in Figure 4.1:

1) Concatenation: We first implement a simple concatenation of the modalities, which is categorized as an early fusion strategy. Images of different modalities are concatenated in different channels and a single neural network learns the correlation and interactions between the low level features of each modality. Although it simplifies the training, by using one single model to make the predictions, we assume that the model is well suited for all the modalities.

2) DenseFuse: We implement another early fusion method called DenseFuse [20] which differs from the previous one in the fact that images of different modalities are fused via an encoder-decoder architecture into a one-channel image, that is then given as an input to the network. In this method, an encoder and a decoder are trained on one modality, then, the encoder is used on each of the input modalities, and the encoded images are fused via a simple addition before being decoded, thus giving us a fused image. An example is provided in Figure 4.2.

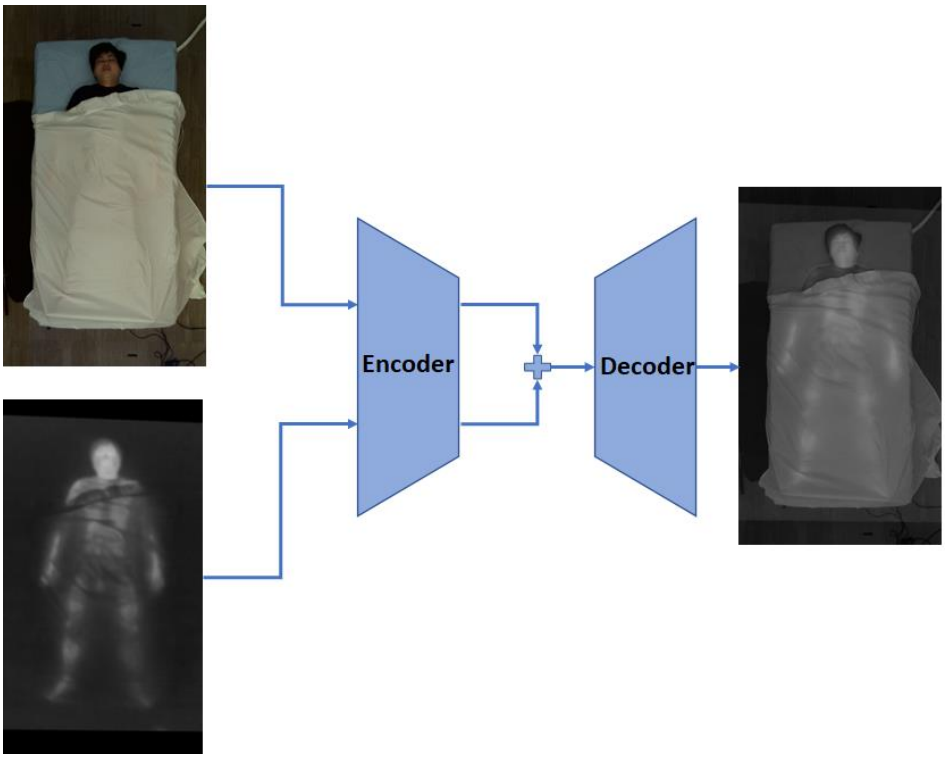


Figure 4.2 Fusion of RGB and IR modalities by using DenseFuse

3) Multimodal Transfer Module: We implement an intermediate fusion method, called Multi-Modal Transfer Module [22] (MMTM). This method consists in implementing a module between parallel CNN streams, each stream being specific to a modality, allowing the parallel networks to transfer knowledge and re-calibrate their channel-wise features. An MMTM at a given layer *i*

between two sub-networks takes their respective features, squeezes them channelwise by calculating the average of the pixels with regards to the channels in order to get channel descriptors $\mu_i^{(1)}$ and $\mu_i^{(2)}$. The channel descriptors of each sub-network are then combined and a joint representation is learned;

$$Z = \boldsymbol{W}\left[\mu_i^{(1)}, \mu_i^{(2)}\right] + \boldsymbol{b}$$

where W and b are the learnable weights and biases respectively.

Excitation signals are then sent back to each sub-network after being learned;

$$E^{(1)} = \boldsymbol{W}^{(1)} Z + \boldsymbol{b}^{(1)}$$

$$E^{(2)} = \boldsymbol{W}^{(2)} Z + \boldsymbol{b}^{(2)}$$

Learning an excitation signal allows the recalibration of features of one modality with features from another. MMTM modules are better added at intermediate and high-level features as explained by Joze et al. [22]. The exact layers to which MMTM modules should be added are selected empirically.

4) Channel Exchange: We implement a method called Channel Exchange [23] (CE), which is method that exchanges channels between sub-networks by replacing channels of one modality's subnetwork that does not contribute to the learning process, with channels from the other modality's sub-network. The contribution of a channel to the learning process of a sub-network is determined via the scaling factor (γ) of the batch-normalization equation;

$$x'_{i,c} = \gamma_{i,c} \frac{x_{i,c} - \gamma_{i,c}}{\sqrt{\sigma_{i,c}^2 + \epsilon}} + \beta_{i,c}$$

where $x_{i,c}$ is the c_{th}channel of the i_{th} layer feature maps, μ and σ represent the mean and standard deviation of the feature maps over the mini-batch and γ and β are the scaling and shifting parameters respectively.

Channels with a close-to-zero scaling factor γ are replaced with channels from other modalities' sub-networks since they lose their influence on the final prediction and replacing them will only enhance the learning.

5) Ensemble: Finally, we implement a simple late fusion method consisting of creating an ensemble model combining the inferences of the modality-specific networks by fusing them. Indeed, each subnetwork predicts a set of heatmaps (one for each joint) containing the pixel-wise probabilities of presence of the joints. The heatmaps of each sub-network are then combined with a max operation.

4.3.2 Dataset and modalities

We evaluate the different methods on the SLP dataset [14] which contains in-bed pose images from 109 participants. Each pose image is captured under different covering conditions; uncovered, covered with a thin blanket (cover1) and with a thick blanket (cover2), and using different imaging modalities (LWIR, Depth, RGB and pressure mat (PM)).

It is noteworthy that in this paper we only consider video based modalities, which means that we do not consider the PM modality.

In order to determine which combination of the three video based modalities provide the best performance in HPE, we tried different modality combinations.

Since the video based modalities were acquired using different sensors, images are of different resolutions and viewpoints. As a preprocessing step, to compensate for these differences, we apply a homographic correction, the matrices of which are provided with the dataset.

SLP was collected under 2 different settings; 102 subjects were lying on a bed in a home setting (danaLab) and 7 patients in a simulated hospital setting (simLab).

For danaLab, the models are trained on the first 90 patients and evaluated on the remaining 12 subjects. Then our model is additionally evaluated on the 7 subjects of simLab.

4.3.3 Model used

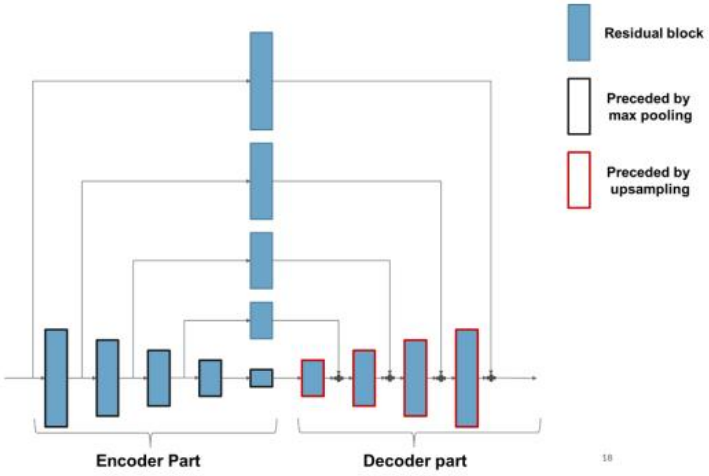
The deep learning model that we used for pose estimation is a Stacked Hourglass (SHG). It consists of multiple stacked CNNs called Hourglasses. Each hourglass has an encoder-decoder structure and consists of multiple residual blocks that apply convolutions, preceded by a max pooling layer.

On the encoder half of the hourglass, successive convolutions and max pooling layers are used to process and extract features down to a very low resolution. Before every block of max pooling and

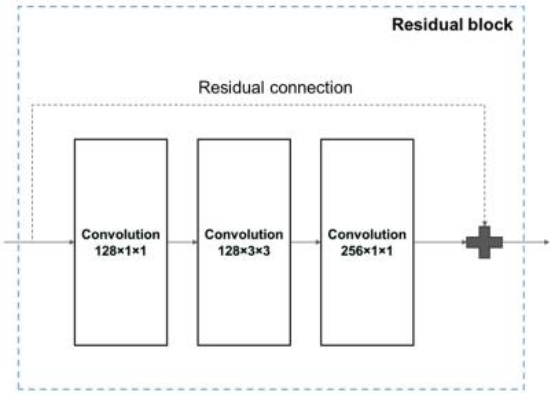
convolutions, the network splits and a skip connection, preserving a high resolution representation, is created.

The decoder side of the Hourglass module applies a sequence of upsampling while combining the resulting upsampled features with the features skipped from the encoder part. This combination allows for a richer semantic representation.

Multiple Hourglass modules can be stacked, and each subsequent Hourglass module refines the predictions of the previous one and the loss is applied at not only the final heatmaps, but also at the intermediate ones. Figure 4.3 presents the overall architecture of the Stacked Hourglass architecture.



(a) Hourglass module



(b) Single block unit in an hourglass

Figure 4.3 Modules in the Stacked Hourglass network

The loss used is a Mean-Squared Error (MSE) that is used to compare the predicted joint heatmap to a groundtruth heatmap consisting of 2D gaussian centered on the joint location. For the CE method, an L1 norm penalty is added to the loss to force the scaling factors of the Batch-Normalization layers to be sparse.

4.3.4 Implementation

The framework that we used for the implementation is Pytorch 1.10. The initial learning rate was set to 1×10^{-3}, and slowly decayed at epochs 70 and 90 by a factor of 0.1 with a total of 100 epochs, which is the same strategy used in the Liu et al. paper [14]. The modules for the MMTM and CE methods were placed at network levels chosen empirically. The hyperparameter lambda (λ) guiding the percentage of channels exchanged in the CE method was chosen following an empirical search strategy and the value that proved to give the best result was $\lambda = 3 \times 10^{-10}$. The models were trained on a single NVIDIA GeForce RTX 3080 GPU. The batch size for all methods was set to 7 for computational considerations.

4.3.5 Performance metric

The metric that was used to compare the different fusion methods and modality combinations is the PCKh@0.5 [11], a widely used HPE metric. PCK stands for Percentage of Correct Key-points. A predicted joint is considered in a correct position if the distance between it and the reference joint position is within a certain threshold, defined as a fraction of the person bounding box. When using PCKh@0.5, the threshold is set to 50% of the head segment length.

4.4 Results and Discussion

4.4.1 Comparison between fusion methods

Table 4.1 shows the results in terms of PCKh@0.5 obtained when considering RGB and LWIR images or depth and LWIR images or depth and RGB images as inputs, with each fusion method. Out of the two early fusion methods, concatenation seems to perform better than DenseFuse, which can be explained by the fact that DenseFuse, by fusing the modalities into one channel, causes a loss of information that does not happen when the modalities are instead concatenated. CE

consistently performs better than the other intermediate fusion method (MMTM) as well as all the other modality fusion methods.

CE method not only surpasses unimodal strategies and other fusion methods that we implemented, but also yields the best PCKh@0.5 results amongst papers that used the SLP dataset; indeed Cao et al. obtained a PCKh0.5 of 94.3% when fusing LWIR and RGB compared to the 95.5% that we obtain when fusing the same modalities. These results show that we reach state-of-the-art results on the SLP dataset and also show that CE method yields the best results across all the different modality combinations.

Another significant advantage of the CE method is its scalability. We can see in Table 4.2 that the number of parameters when using CE does not change much from using a single modality. This is due to the parameter sharing between the sub-networks as explained in the previous section. This means that the CE method, while surpassing the concatenation method and having better results than the unimodal methods, is not heavier in terms of memory.

The inference time of the different fusion methods is also reported in Table 4.1, we can see that using a modality fusion method does not affect the usability of the HPE system in real-time since the highest inference time does not exceed 14 milliseconds. Indeed, the usual frame-per-second (fps) of consumer grade cameras is between 30 and 60 fps, which means that the threshold of time per frame is comprised between 33ms and 16ms per frame, which the CE method, as well as all the other fusion method, respects.

Table 4.1 Comparison between different modality fusion methods, results in PCKh@0.5

Fusion Method	**LWIR+RGB**	**Depth+RGB**	**Depth+LWIR**	**Inference time (ms)**
Concatenation	94.9	96.9	96.9	6.4
DenseFuse	94.4	23.9	94.2	6.3
MMTM	94.1	96.0	94.3	9.6
Channel-exchange	**95.5**	**97.0**	**97.1**	14
Ensemble	93.8	95.2	95.5	9.42

Table 4.2 Number of parameters for each method

Method	**Number of parameters**
Unimodal - LWIR/Depth	12 638 878
Concatenation	12 648 292
DenseFuse	12 645 154
MMTM	25 546 944
Channel-exchange	12 705 194
Ensemble	25 284 032

4.4.2 Optimal modalities combination

The results in Table 4.1 clearly show that combining depth and LWIR gives us the best performance compared to other fusion methods and Table 4.3 shows the superiority of using modality fusion compared to using only one modality. We can also note that even though the presence of occlusion and lighting variations make the use of RGB by itself problematic, fusing it with LWIR or depth can slightly improve the results compared to using these modalities by themselves. But once we combine depth and LWIR, adding RGB becomes trivial, as can be seen in Table 4.3, and causes the predictions to slightly worsen. This can be explained by the fact that RGB images do not contribute much to the learning of the deep neural network. It can even confuse the network when the patient is occluded or when the room in which the patient is present is dark. Therefore fusing RGB images with the other modalities can impede the learning of the network.

Table 4.3 Comparison between modality combinations

Modality	PCK@0.5
LWIR	94.2
LWIR+RGB (CE)	95.5
Depth	96.6
Depth + RGB (CE)	97.0
Depth+LWIR (CE)	**97.1**
Depth+LWIR+RGB (CE)	96.9

Figure 4.4 presents an example of how an appropriate combination of modalities can mitigate the shortcomings of each one. Indeed, we can see in a) that the right arm is positioned behind the head of the patient, while in b), when using the depth modality alone, the right arm is estimated as being extended. This is due to the fact that when one limb is occluded by another, it becomes hard to

estimate the position of the occluded one. The LWIR image doesn't have the same short-coming, and therefore, by using an appropriate fusion method, namely CE, the combination of depth with LWIR allows to mitigate this problem. However, we can see that in the LWIR image, the left arm position is not well estimated which can be caused by the fact that a part the arm is superposed on the body of the patient, causing the heat residues of the body and the arm to mix. We can see that this problem is not present when we use CE for the prediction. CE can therefore overcome the shortcomings of LWIR and depth.

Table 4.5 shows the performance of the 2 best fusion methods compared to unimodal methods under different covering settings; uncovered meaning without any visual occlusion, cover1 meaning that a thin blanket is covering the patient and cover2 meaning that a thick blanket is covering the patient. We can see that the fusion methods consistently outperform the unimodal ones under all of the covering settings, which further proves that using an appropriate modality fusion method not only mitigates the problem of occlusion, but also increases the performance when no occlusion is present. Another observation that can be made from this table is that even though under an uncovered setting concatenation and CE have a similar performance, CE outperforms concatenation when the level of occlusion increases.

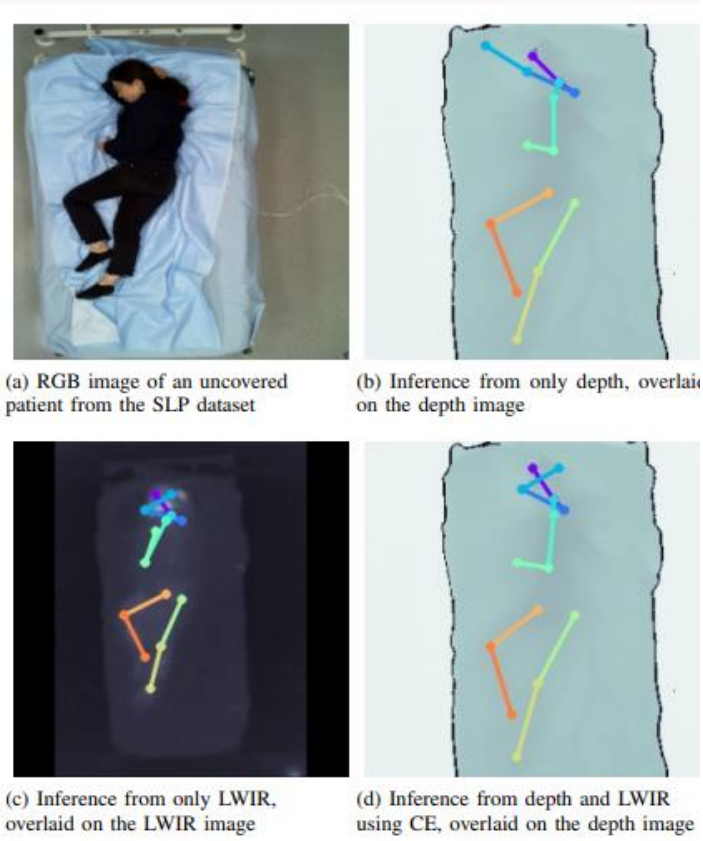


Figure 4.4 Advantage of fusing modalities versus using one modality for pose inference

4.4.3 Generalization to real data

In order to test the generalization performance to images acquired in a hospital setting, we run our models with the two best fusion methods on the simLab part of the SLP dataset. We see in Table 4.4 that CE reaches a PCKh@0.5 of 96.9%, outperforming again the concatenation and unimodal methods. These results show that the improvement of performance that is produced by the modality fusion methods, and more specifically the CE method, is generalizable to other settings.

Table 4.4 Results on simLab data

Method	PCKh@0.5
Unimodal, LWIR	96.4
Unimodal, depth	96.2
Concatenation, depth and LWIR	96.5
Channel-exchange, depth and LWIR	**96.9**

Table 4.5 Performance (PCKh@0.5) in different covering settings

Method	uncovered	cover1	cover2
Unimodal, LWIR	96.1	93.5	92.9
Unimodal, depth	97.8	96.1	95.7
Concatenation, depth and LWIR	**98.2**	96.3	96.2
Channel-exchange, depth and LWIR	98.1	**96.8**	**96.5**

4.5 Conclusion

This work provides for the first time a comparison of different modality combinations and fusion methods to address the problem of in-bed human pose estimation. The results showed that an intermediate fusion method, called Channel-exchange (CE) [23], combining LWIR and depth images gives the best results in terms of accuracy while having a low latency on a public multi-modal dataset (SLP).

We therefore demonstrate that choosing an adequate combination of modalities and an appropriate fusion method can mitigate the challenges of blanket occlusion and lighting variability in scene illumination, improving the performance of in-bed human pose estimation systems. This work can serve as a basis for improvement of upstream tasks such as shape estimation.

4.6 Acknowledgment

We thank Philippe Debanné for his help in proofreading the ´ paper. This research was conducted as part of the activities of the TransMedTech Institute, thanks in part to financial support from the Apogee Canada Research Excellence Fund and the Fonds de recherche du Quebec. This project was also supported in part by L. Seoud’s NSERC Discovery grant.

CHAPITRE 5 DISCUSSION GÉNÉRALE

5.1 Synthèse du travail

Le présent travail visait à trouver une stratégie de fusion de modalités adéquate afin de relever les défis spécifiques à l'EPH de patients alités, c'est-à-dire la possibilité d'occlusion due à la présence de couverture au-dessus des patients et la variation de luminosité dans les chambres dans lesquels ils se trouvent. De plus, cette stratégie devait optimiser l'exploitation de la complémentarité des différentes modalités, c'est-à-dire faire en sorte que les limitations de l'une soient pallier par l'utilisation d'autres modalités; les images RGB perdent toute valeur en cas d'occlusion du patient ou en cas de manque de lumière dans la chambre, tandis que les images LWIR peuvent contenir des résidus de chaleur qui peuvent fausser l'estimation de la pose et les images de profondeur perdent en précision en cas d'occlusion d'une membre du patient par un autre, ce qui arrive souvent dans le cas où le patient est couché sur un côté.

Le projet mené visait donc à mitiger tous ces problèmes en, premièrement, trouvant une méthode de fusion optimale des modalités LWIR et Profondeur uniquement vue que l'utilisation du RGB faussait l'estimation de pose dans les cas mentionnés plus haut. Afin d'arriver à cet objectif, 5 différentes méthodes de fusion de modalités basées sur l'apprentissage profond, appartenant à différentes catégories de fusion (en amant, en aval et intermédiaire) ont été comparées; Concaténation, DenseFuse, Multimodal Transfer Module (MMTM), Channel-Exchange (CE) et Ensemble.

Les résultats ont montré que Concaténation et Channel-Exchange permettent d'obtenir la meilleure performance en termes de PCKh@0.5 (96.9% et 97.1% respectivement), avec un léger avantage pour Channel-Exchange, surtout lorsque le niveau d'occlusion augmente. Le meilleur modèle multimodal dépasse donc le meilleur résultat obtenu dans la littérature sur la base publique SLP (96.6%). De plus CE a un temps d'inférence d'en moyenne 14 millisecondes et a presque le même nombre de paramètres qu'un modèle uni-modal, ce qui veut dire que la méthode est déployable en temps réel et ne prend pas plus d'espace en mémoire qu'un modèle uni-modal. Notre hypothèse de recherche est donc validée.

Il est cependant important d'expliquer que le score élevé en terme de PCKh@0.5 sur la base de données SLP limite l'évaluation des méthodes de fusions. En effet, lorsque le score est élevé (supérieur à 96% dans notre cas), les limitations inhérentes à la base de données font en sorte que l'implémentation d'une méthode plus performante ne résultera pas en une augmentation très importante du score (0.5% dans notre cas). Cela peut être observé par exemple sur la base de données ImageNet, ou la différence entre le modèle le plus performant et le deuxième plus performant est de 0.02%.

De plus ce travail a démontré que cet avantage que procure une bonne stratégie de fusion de modalités subsiste même lorsque le contexte change. En effet, Channel-Exchange permet d'obtenir le meilleur résultat non seulement sur danaLab, sous ensemble de SLP qui contient des images de personnes dans un contexte d'une maison, mais aussi sur simLab, qui lui contient des images de patients dans un contexte hospitalier. Cette aptitude de généralisation est importante dans le cadre de ce projet vu que l'objectif final est de déployer un système d'estimation de pose de patients en position alitée au CHUSJ.

5.2 Limitations de la solution proposée

Dans ce travail, nous avons voulu démontrer qu'avec une fusion de modalité il était possible de dépasser le meilleur score obtenu dans la littérature sur la base de données SLP. Ainsi, pour des fins de comparaison, nous avons considéré le même modèle utilisé dans la littérature et le même ensemble d'hyperparamètres. Bien que dans ces conditions, nous ayons réussi à dépasser le meilleur score, une sélection judicieuse de modèle et d'hyperparamètres permettrait éventuellement d'obtenir encore de meilleurs résultats.

Bien que la performance ait été améliorée sur la base de données SLP et qu'il ait été démontré qu'une méthode de fusion de modalité adéquate permet de mitiger les problèmes reliés à l'estimation de pose en position alitée, il demeure des limitations prévisibles qui peuvent affecter la performance du modèle à l'étape du déploiement;

- Même si le modèle généralise bien sur simLab, il en demeure que simLab et danaLab présentent des similarités, notamment, les caméras sont positionnées aux mêmes angles et les couvertures utilisées sont les mêmes. Ainsi, déployer le modèle dans un nouvel environnement, tel un hôpital où les couvertures sont différentes, où les caméras sont

différentes et/ou placées à des angles différents pourrait causer une baisse de la performance.

- Dans les données d'entraînement, les seules occlusions possibles sont celles de la couverture et des autres membres (ex : bras au-dessous de la tête). Dans un environnement réel (c'est-à-dire non simulé), le patient pourrait avoir au-dessus de lui un plateau de nourriture ou de l'équipement médical (ex : quand le patient se casse la jambe), ce qui peut aussi affecter la performance du modèle.
- Le modèle est entraîné sur des images contenant une seule personne, cependant, dans un environnement hospitalier, il est possible d'avoir plus d'une personne dans une image (ex : Infirmière assistant le patient), or il n'y a pas de mécanisme de gestion de ce cas pour l'instant et l'inférence de pose sera erronée dans une telle situation.
- Une seule architecture CNN a été utilisée (Stacked Hourglass) pour la comparaison entre les différentes méthodes de fusions de modalités. Il est donc possible que CE ne soit pas la solution optimale si une autre architecture est utilisée lors du déploiement.
- Le système actuel nécessite la présence des modalités LWIR et profondeur. Or en déploiement, il est possible qu'une des modalités de soit pas présente due à un bris d'équipement. Ce cas n'est actuellement pas géré.

5.3 Travaux futurs

Une solution qui permettra d'augmenter la robustesse du modèle en temps de déploiement est d'acquérir un jeu de données propre au CHUSJ et de raffiner le modèle avec ces données. Le protocole de collecte de données devrait s'appuyer sur les limites découlant de ce projet, c'est-à-dire faire en sorte à inclure une diversité dans le type d'occlusions possibles dans un contexte hospitalier, et dans les positions des caméras l'une par rapport à l'autre ainsi que leurs champs de vue.

De plus, étant donné qu'un des objectifs futurs est l'EPH d'enfants au CHUSJ, il est nécessaire de bâtir une base de données spécifiques aux enfants alités. Des travaux sont en cours dans notre équipe pour la création d'une base de données particulières aux enfants alités au CHUSJ.

Il est aussi impératif de rajouter un mécanisme de gestion du cas ou plus d'une personne se trouvent dans l'image; une suggestion serait de désactiver l'inférence dans ce cas, vu que l'objectif est de surveiller l'état des patients en l'absence d'infirmière, or si une personne est présente près du patient, il n'est plus indispensable de faire cela.

Un autre mécanisme qu'il faudra rajouter dans le futur est la gestion de l'absence d'une modalité due à, par exemple, un bris d'équipement. Il faut que le système soit capable de prédire la pose des patients en s'appuyant sur une modalité même si l'autre n'est pas présente.

CHAPITRE 6 CONCLUSION

Le travail effectué dans le cadre de cette maîtrise permet de proposer, suite à la comparaison de différentes stratégies de fusion de modalités ainsi que différentes combinaisons de modalités, une architecture de réseau de neurones incorporant une méthode de fusion qui permet d'exploiter la complémentarité entre différentes modalités.

À notre connaissance, ce travail est le premier qui porte sur une comparaison des méthodes de fusion multimodale dans le contexte d'EPH, et bien que le travail traite de sujets couchés, l'EPH de sujets debout pourrait aussi bénéficier d'une approche multimodale pour, par exemple, mitiger les variations d'illuminations entre jour et nuit dans un contexte de surveillance.

L'architecture développée permet de surpasser les résultats obtenus jusque-là sur la base de données SLP et constituera le socle d'un système d'EPH de patients alités au CHUSJ.

Ce travail sera donc intégré dans un projet plus large en entraînant l'architecture qui y est développée sur des données qui seront collectées au CHUSJ et en le déployant ensuite dans un système incluant des caméras afin d'estimer en temps réel la pose des patients.

Ce système devra aussi inclure des mécanismes qui permettront de mitiger les limitations mentionnées précédemment, notamment la gestion des cas ou plusieurs personnes sont présentes dans les images acquises. La qualité de la base de données qui sera créée est aussi très importante pour le projet vu qu'elle permettra, comme mentionné dans la partie Synthèse, d'avoir un modèle robuste aux changements.

RÉFERENCES

[1] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman et A. Blake, «Real-Time Human Pose Recognition in Parts from Single Depth Images,» chez *CVPR*, Colorado, 2011.

[2] G. Jones, «Rapport de stage,» Montréal , 2021.

[3] W. S. McCulloch et W. Pitts, «A logical calculus of the ideas immanent in nervous activity,» *The bulletin of mathematical biophysics* , vol. 5, pp. 115-133, 1943.

[4] A. F. Agarap, «Deep Learning using Rectified Linear Units (ReLU),» *arXiv preprint arXiv*, vol. 1803.08375, 2018.

[5] K. He, X. Zhang, S. Ren et e. al., «Deep residual learning for image recognition,» *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778, 2016.

[6] S. Ostadabbas et R. Josyula, «A Review on Human Pose Estimation,» *arXiv preprint* , vol. 2110.06877, 2021.

[7] T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang et C. Yang, «The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation,» *IEEE Access*, vol. 8, pp. 133330-133348, 2020.

[8] Y. Zhang, C. Fan, Z. Zheng et D. Yang, «Pose estimation at night in infrared images using a lightweight multi-stage attention network,» *Springer, Signal Image and Video Processing*, p. 1757–1765, 2021.

[9] A. Krizhevsky, S. Ilya et G. E. Hinton, «Imagenet classification with deep convolutional neural networks,» chez *Communications of the ACM*, 2017.

[10] R. Girshick, J. Donahue et T. Darrell, «Rich feature hierarchies for accurate object detection and semantic segmentation,» chez *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[11] K. HE, G. Gkioxari et P. Dollar, «Mask r-cnn,» chez *Proceedings of the IEEE international conference on computer vision*, 2017.

[12] A. Newell, K. Yang et J. Deng, «Stacked hourglass networks for human pose estimation,» chez *European conference on computer vision*, 2016.

[13] K. Sun, B. Xiao, D. Liu et J. Wang, «Deep high-resolution representation learning for human pose estimation,» *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5693--5703, 2019.

[14] M. Andriluka, L. Pishchulin, P. Gehler et B. Schiele, «2D Human Pose Estimation: New Benchmark and State of the Art Analysis,» chez *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar et C. L. Zitnick, «Microsoft coco: Common objects in context,» *European conference on computer vision*, pp. 740--755, 2014.

[16] C. Ionescu, D. Papava, V. Olaru et C. Sminchisescu, «Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments,» *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, n° %17, pp. 1325-1339, 2014.

[17] S. Liu, X. Huang, N. Fu, C. Li, Z. Su et S. Ostadabbas, «Simultaneously-collected multimodal lying pose dataset: Towards in-bed human pose monitoring under adverse vision conditions,» *arXiv preprint*, vol. 2008.08735, 2020.

[18] T. Cao, M. A. Armin, S. Denman, L. Petersson et D. Ahmedt-Aristizabal, «In-Bed Human Pose Estimation from Unseen and Privacy-Preserving Image Domains,» *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pp. 1-5, 2022.

[19] Y. Yin, J. P. Robinson et Y. Fu, «Multimodal in-bed pose and shape estimation under the blankets,» *arXiv preprint*, vol. 2012.06735, 2020.

[20] J. Summaira, X. Li, A. M. Shoib, S. Li et J. Abdul, «Recent Advances and Trends in Multimodal Deep Learning: A Review,» *arXiv preprint*, vol. 2105.11087, 2021.

[21] T. Baltrušaitis, C. Ahuja et L.-P. Morency, «Multimodal Machine Learning: A Survey and Taxonomy,» *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, n° %12, pp. 423-443, 2019.

[22] J. Zeng, Y. Tong, Y. Huang, Q. Yan, W. Sun, J. Chen et Y. Wang, «Deep surface normal estimation with hierarchical rgb-d fusion,» chez *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.

[23] H. Li et X.-J. Wu, «DenseFuse: A fusion approach to infrared and visible images,» *IEEE Transactions on Image Processing*, vol. 28, pp. 2614--2623, 2018.

[24] P. R. Kunekar, M. Gupta et B. Agarwal, «Deep Learning with Multi Modal Ensemble Fusion for Epilepsy Diagnosis,» chez *2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)*, 2020.

[25] H. R. V. Joze, A. Shaban, M. L. Iuzzolino et K. Koishida, «MMTM: Multimodal transfer module for CNN fusion,» chez *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[26] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong et J. Huang, «Deep multimodal fusion by channel exchanging,» *Advances in Neural Information Processing Systems*, vol. 33, pp. 4835--4845, 2020.

[27] S. Ioffe et C. Szegedy, «Batch normalization: Accelerating deep network training by reducing internal covariate shift,» chez *International conference on machine learning*, 2015.

[28] V. Srivastav, A. Gangi et N. Padoy, «Self-supervision on unlabelled or data for multi-person 2d/3d human pose estimation,» chez *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020.

[29] A. Oksenberg et D. S. Silverberg, «The effect of body posture on sleep-related breathing disorders: facts and therapeutic implications,» *Sleep medicine reviews*, vol. 2, p. 139–162, 1998.

[30] S. Ostadabbas, R. Yousefi, M. Nourani, M. Faezipour, L. Tamil et M. Q. Pompeo, «A resource-efficient planning for pressure ulcer prevention,» *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, p. 1265–1273, 2012.

[31] C. N. Sessler, M. J. Grap et M. A. E. Ramsay, «Evaluating and monitoring analgesia and sedation in the intensive care unit,» *Critical care*, vol. 12, p. 1–13, 2008.

[32] C. Zheng, W. Wu, T. Yang, S. Zhu, C. Chen, R. Liu, J. Shen, N. Kehtarnavaz et M. Shah, «Deep learning-based human pose estimation: A survey,» *arXiv preprint arXiv:2012.13392*, 2020.

[33] R. Poppe, «Vision-based human motion analysis: An overview,» *Computer vision and image understanding*, vol. 108, p. 4–18, 2007.

[34] A. Toshev et C. Szegedy, «Deeppose: Human pose estimation via deep neural networks,» chez *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[35] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler et B. Schiele, «Deepcut: Joint subset partition and labeling for multi person pose estimation,» chez *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[36] K. Simonyan et A. Zisserman, «Very deep convolutional networks for large-scale image recognition,» *arXiv preprint arXiv:1409.1556,* 2014.

[37] Z. Cao, T. Simon, S.-E. Wei et Y. Sheikh, «Realtime multi-person 2d pose estimation using part affinity fields,» chez *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[38] S.-E. Wei, V. Ramakrishna, T. Kanade et Y. Sheikh, «Convolutional pose machines,» chez *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016.

[39] S. Liu, Y. Yin et S. Ostadabbas, «In-bed pose estimation: Deep learning with shallow dataset,» *IEEE journal of translational engineering in health and medicine,* vol. 7, p. 1–12, 2019.

[40] S. Liu et S. Ostadabbas, «Seeing under the cover: A physics guided learning approach for in-bed pose estimation,» chez *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019.

[41] D. Ramachandram et G. W. Taylor, «Deep Multimodal Learning: A Survey on Recent Advances and Trends,» *IEEE Signal Processing Magazine,* vol. 34, pp. 96-108, 2017.

[42] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee et A. Y. Ng, «Multimodal deep learning,» chez *ICML*, 2011.

[43] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman et A. Blake, «Efficient Human Pose Estimation from Single Depth Images,» *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 35, pp. 2821-2840, 2013.