Contributions à l'étude par vision du mouvement humain : analyse du comportement et assistance au déplacement

Cyrille Migniot

Mémoire pour l'obtention de l'Habilitation à Diriger des Recherches

Ecole doctorale n° 37

Sciences Physiques pour l'Ingénieur et Microtechniques

par

Cyrille MIGNIOT

Contributions à l'étude par vision du mouvement humain : analyse du comportement et assistance au déplacement

Soutenu le 20 octobre 2023 devant le jury composé de :

Saïda BOUAKAZ	Professeure à l'Université Claude Bernard Lyon 1, LIRIS - UMR 5205	Rapporteuse
Nicolas PASSAT	Professeur à l'Université de Reims Champagne Ardenne, CReSTIC - EA 3804	Rapporteur
Hazem WANNOUS	Professeur à l'Institut Mines-Telecom de Lille, CRIStAL - UMR 9189	Rapporteur
Frédéric LERASLE	Professeur à l'Université Paul Sabatier, Toulouse, LAAS - UPR8001	Examinateur
Albert DIPANDA	Professeur à l'Université de Bourgogne, Dijon, ImViA - EA 7535	Examinateur

Remerciements

Depuis mon recrutement j'ai eu le bonheur de travailler et de côtoyer de nombreuses personnes. C'est par ces riches rencontres que j'ai pu construire mon parcours et ma vie professionnelle.

Je voudrais dans un premier temps remercier les membres de mon jury : Saïda Bouakaz, Nicolas Passat, Hazem Wannous et Frédéric Lerasle, pour avoir accepté de relire mon manuscrit et d'analyser mon parcours. Leurs retours et conseils avisés me permettront de construire le futur de ma recherche et de mon enseignement.

Je voudrais bien évidemment remercier l'ensemble des membres du laboratoire Le2I puis du laboratoire ImViA pour m'avoir accueilli et m'avoir donné de si bonnes conditions de travail. Je pense en particulier aux personnes avec qui j'ai eu l'honneur de partager l'encadrement d'une thèse ou le suivi d'un projet : Albert Dipanda, Julien Dubois, Fan Yang, Olivier Aubreton, Cédric Demonceaux et Dominique Ginhac. J'ai tant appris à vos côtés. Je pense aussi aux doctorants, post-doctorants et ingénieurs, avec qui j'ai partagé des idées et de bons moments, ainsi qu'aux membres du personnel administratif et technique qui ont tant fait pour me faciliter la tâche.

Un grand merci également aux collègues des laboratoires avec qui j'ai pu réaliser des collaborations fructueuses, mais aussi aux collègues enseignants de l'UFR Sciences et Techniques ainsi que de l'ESIREM qui m'ont entouré et conseillé pour construire un enseignement qui m'est propre tout en étant le plus cohérent et intéressant pour les élèves.

Une grande pensée à ma famille qui me pousse sans cesse. Pour finir je veux remercier mon épouse Béatrice pour son soutien sans faille et sa bienveillance de tous les instants.

Abréviations

CL	Convolutional layer → couche convulotionnelle
CNN	Convolutional Neural Network
CPC	Complete Point Cloud → nuage de points complet
DL	Deep Learning
EOA	Electronic Orientation Assistance
HMM	Hiden Markov Model
HOG	Histogram of Oriented Gradients
HOOF	Histogram of Oriented Optical Flow
HRIR	Head-Related Impulse Response
k-NN	k Nearest Neighbour → algorithme des k plus proches voisins
LBP	Local Binary Pattern
LBP-TOP	Local Binary Pattern on three orthogonal planes
LOSO	Leave One Subject Out
LSTM	Long Short-Term Memory
LTP	Local Temporal Pattern
ME	Micro-Expression
ML	Machine Learning
RGB-D	Red Green Blue Depth → les trois canaux classiques de la couleur et la profondeur
YOLO	You Only Look Once → algorithme de détection d'objets

Table des matières

Part I

Activités de recherche

Chapitre 1

Introduction générale

L'image est un support riche et passionnant. Les traitements qui lui sont appliqués sont si divers qu'ils impliquent des paradigmes excessivement éloignés. L'image peut être le résultat désiré qu'il est possible d'améliorer, de restaurer ou même de synthétiser, ce qui implique des traitements locaux, gérant des modifications au niveau du pixel. Mais une image peut aussi être le miroir du monde et par son intermédiaire nous pouvons le comprendre et l'analyser. Une acquisition bien réalisée peut révéler au mieux les caractéristiques permettant de répondre aux questions posées.

Une image contient une quantité colossale d'information. Encore faut-il savoir les trouver car elles sont noyées par une multitude d'informations redondantes ou perturbatrices. Le phénomène s'accentue encore avec une augmentation de la dimensionnalité, si ce sont des vidéos qui sont traitées ou des images 3D. Ce n'est pas seulement de l'information supplémentaire, c'est aussi plus de connexions. Des rapprochements selon le temps ou l'espace sont rajoutés qui induisent une continuité la plupart du temps et parfois une discontinuité riche en interprétation. Si le traitement d'images introduit une multitude de thématiques de recherche, c'est son enrichissement à travers les vidéos et la 3D, ainsi que les contraintes et adaptations qui en découlent, qui ont guidé mes recherches récentes.

Si le gros de ma recherche se concentre sur le traitement de l'image, j'ai taché de m'y atteler non pas selon une vue fondamentale mais dans une conscience du processus complet : les données sont acquises, puis traitées pour obtenir la réponse à un problème concret. L'application pratique d'une méthode est primordiale car elle fixe à la fois les contraintes qu'il faut établir, les a-priori à intégrer et les hypothèses à poser. Je me suis pour ma part concentré sur du monitoring de l'humain. Au contraire des traitements sur les images médicales, permet -tant d'aider à un diagnostic à partir d'une acquisition très spécifique sous forte contrainte, il s'agit ici, à partir d'une acquisition non invasive, d'extraire des indices concernant le comportement de la personne filmée. La perte en robustesse, si elle existe, est compensée par la facilité d'utilisation de la méthode et sa mise à la disposition du grand public.
L'application peut rester médicale : le suivi de la rééducation des personnes portant une prothèse et l'assistance à la mobilité des personnes mal-voyantes dans mon cas. Mais le traitement doit être proche du système d'acquisition; soit pour permettre une embarquabilité du système ou un traitement temps réel, soit pour permettre une acquisition non invasive et bon marché. Par exemple les caméras de type Kinect, ou les modèles qui ont suivi comme les realsense, sont communes et accessibles à tous tout en apportant une information supplémentaire riche. Il en est de même pour les caméras rapides telles la GO2400C de Stemmer (utilisée dans nos études) permettant de ne pas perdre à l'échantillonnage les mouvements rapides.

En résumé

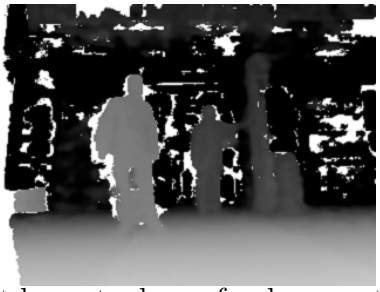
- Traitement des images enrichi par la vidéo et la profondeur.
- Recherches centrées autour du monitoring de l'humain.
- Méthodes non invasives, légères et rapides.

1.1 Données 3D

Jouer sur les données 3D, c'est avoir une représentation plus fidèle du monde plutôt qu'une projection sur un plan défini par l'acquisition. Cette représentation permet de séparer facilement plusieurs plans d'étude de la scène mais est aussi nécessaire pour décrire des formes dont la projection réduit la descriptivité.
Ma recherche ne se concentre pas sur l'acquisition de la profondeur (ni le système d'acquisition ni l'estimation de la profondeur) mais à la façon d'utiliser cette donnée pour en tirer une information haut niveau sur la scène. Même en restant sur les systèmes d'acquisition classiques, le support en sortie, qui sera l'entrée de mes méthodes, peut se présenter sous plusieurs formes. Chacune possède ses spécificités propres à bien prendre en considération dans les méthodologies les utilisant.

J'ai étudié la 3D selon trois niveaux : la carte de profondeur, le nuage de points et le nuage de points complet. La carte de profondeur correspond à une image où chaque pixel est lié à une valeur correspondant à la profondeur, c'est à dire la distance entre l'objet et la caméra. Comme dans une image, chaque pixel possède un voisinage régulier. Ainsi la grande majorité des méthodes de traitement des images couleur peuvent être appliquées aux carte de profondeur. Néanmoins la distance entre deux pixels ne correspond pas à une distance physique relative à la scène.

La carte de profondeur



La carte de profondeur produit une image où chaque pixel est lié à une distance (la distance de l'élément à la caméra). Pour la visualisation de cette carte, les valeurs les plus claires correspondent aux éléments proches et les valeurs les plus sombres aux éléments éloignés. La quantification de cette distance dépend du capteur.

Des données RGB-D sont des données où les données couleurs et la carte de profondeur sont alignées (après calibration des capteurs). Ainsi la valeur associée à un pixel est de dimension 4 : les trois canaux couleur et l'information de profondeur.

Le nuage de point est la représentation des données sur un repère 3D cartésien représentant l'espace monde. Chaque point possède une position en 3 coordonnées et peut être associé à une couleur. Cette fois les distances correspondent à celles du monde réel. Mais le voisinage n'est plus régulier : certaines zones possèdent une forte densité de points et d'autre non. Les algorithmes utilisés sont alors souvent très différents de ceux du traitement d'image.

Un nuage de points ne représente que ce que la caméra peut voir. Ainsi si une personne fait face à la caméra, le nuage de points donnera des information sur sa vue de face mais pas sur sa

Le nuage de points

Un nuage de points est une représentation dans le monde 3D. Chaque pixel capturé correspond à un point. Cette représentation permet de conserver les distances.
Avec une capture RGB-D, il est possible d'associer une couleur à chaque point.

Le nuage de points complet (CPC)

Un nuage de points complet est une représentation à 360° de la scène. La connaissance de la surface des éléments de la scène ne dépend pas du point de vue de l'acquisition.
Cette représentation nécessite une acquisition selon plusieurs points de vue.

vue de dos. De précieuses informations manquent alors, parfois les plus descriptives. Le nuage de points complet est la fusion des nuages de points obtenus par plusieurs caméras permettant d'obtenir une représentation 3D de la scène sur 360°.

1.2 Problématiques

Au delà de l'application pratique, qui rajoute des contraintes et des hypothèses décorrélées des capacité théoriques d'un algorithme et de son principe fondamental, la nature de l'objectif recherché est déterminant. Les méthodes sont définies par ce qu'elles sont capables de fournir avant de s'adapter à une situation. Au niveau de ce manuscrit, j'introduirai trois thématiques : la classification, le détection et la représentation.

Classification — Un échantillon est extrait à partir des données et testé. L'objectif est, parmi un ensemble fixe défini de classes, de trouver celle qui correspond le plus au contenu testé.

Pour la classification, les données sont ciblées. La méthode est centrée sur la discrimination entre les classes.

Détection — L'image est prise dans son ensemble et il faut à la fois dénombrer et localiser les instances d'une classe. Sur une vidéo la localisation peut être temporelle. La taille de l'élément peut aussi être recherchée.

Ici l'applicatif est important pour régler le compromis à faire entre les besoins de bien limiter la détection à la classe dédiée et de réduire le nombre d'occurrences non détectées.

Représentation — L'image ou la vidéo est une donnée complexe. L'idée ici est de générer une autre donnée tout aussi complexe mais représentative d'un aspect précis.

Le tout est de transférer l'information vers un espace plus représentatif de ce qui nous intéresse.

1.3 Projets en cours ou réalisés

La réalisation et le suivi d'un travail de recherche sont guidés par le projet auquel il est rattaché. Dans ce manuscrit, je présenterai la recherche que j'ai effectuée depuis mon recrutement autour de quatre projets principaux.

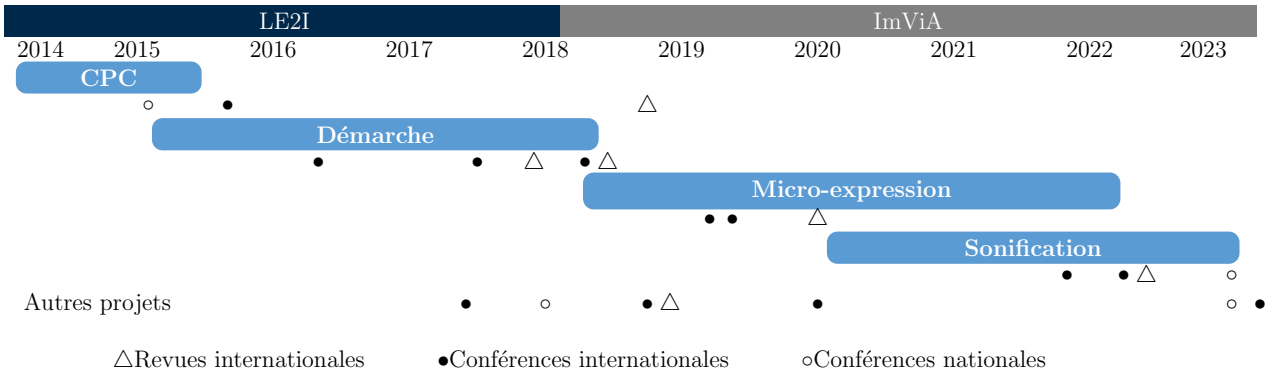


Figure 1.1: Projets et publications depuis mon recrutement.

1.3.1 Détection de personnes dans un CPC

Acquisition	Vue à 360° prise à partir de 3 Kinects.
Objectif	Détecter une personne dans une scène 3D.
Contributions	• Définition d'un descripteur 3D de la forme d'une personne. • Estimation conjointe de la position et de l'orientation de la personne. • Adaptation 3D des HOG à partir d'une projection sur des polyèdres réguliers. • Détection de une ou plusieurs personnes dans une scène malgré un fort recouvrement.
Encadrements	Thèse de Kyis Essmaeel et 4 stages et projets d'études.
Financement	MESRI.
Publications	[MTAP2019] et [VISAPP2016].

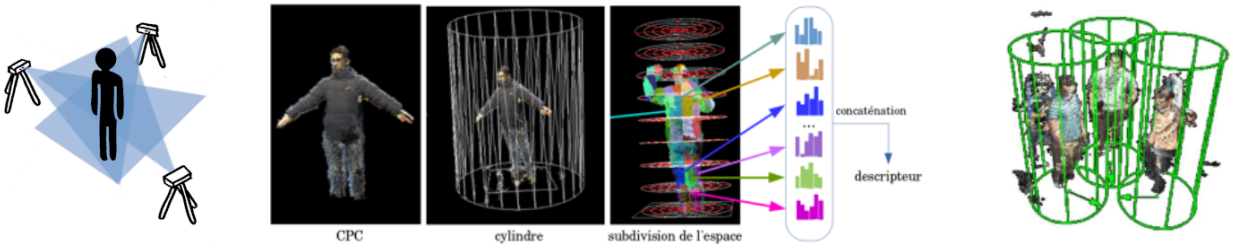


Figure 1.2: À gauche l'acquisition du CPC; au centre les étapes du calcul du descripteur et à droite les résultats de la détection.

L'objectif de ce travail est d'établir un système d'acquisition 3D non invasif pour la détection de personnes robuste à partir de données 3D.

Nous avons construit une plate-forme d'acquisition composée de plusieurs caméras RGB-D (Kinect) permettant d'estimer une vue à 360° de la scène (CPC). Pour une détection robuste à partir de ces données, nous avons introduit un nouveau descripteur basé sur les HOG et adapté à l'environnement 3D et aux caractéristiques géométriques d'une personne. La fenêtre de détection est un cylindre découpé en blocs de façon régulière selon les coordonnées cylindriques. La quantification des normales du nuage de points est réalisée par projection sur un polyèdre régulier. Chaque face correspond à une valeur de l'histogramme.
Nous avons testé ce descripteur associé à un classifieur SVM par balayage de la scène pour la détection des personnes. Les expériences réalisées démontrent à la fois la très grande efficacité du descripteur mais aussi la supériorité du CPC par rapport à un nuage de points classique, principalement au niveau de la robustesse. La classification proposée permet également d'estimer l'orientation de la personne.

1.3.2 Analyse de la démarche

Acquisition	Vue de face de plusieurs cycles de marche par une caméra Kinect.
Objectif	Reconnaître une démarche problématique.
Contributions	• Classification des poses clés à partir d'un descripteur original de présence de données 3D dans un cylindre. • Classification d'une démarche saine à partir des matrices de covariance de l'évolution des angles des articulations de la jambe. • Classification d'une démarche saine à partir de LSTM.
Encadrements	Thèse de Margarita Khokhlova et 2 stages et projets d'études.
Financement	Conseil Régional de Bourgogne (JCE).
Publications	[AIM2019], [MTAP2018], [SITIS2018], [VISAPP2018] et [SITIS2016].

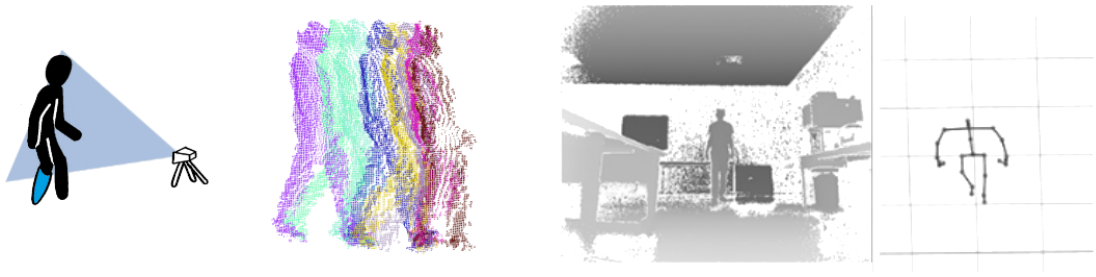


Figure 1.3: À gauche le système d'acquisition; au centre les nuages de points pour chaque étape de la marche; à droite une acquisition de la profondeur et le squelette qui en est extrait.

L'analyse de la démarche clinique est généralement subjective, étant effectuée par des cliniciens observant la démarche des patients. Des alternatives à une telle analyse sont les systèmes basés sur les marqueurs et les systèmes basés sur les plates-formes au sol. Cependant, cette analyse standard de la marche nécessite des laboratoires spécialisés, des équipements coûteux et de longs délais d'installation et de post-traitement. Il y a eu de nombreuses tentatives dans la recherche pour proposer une alternative basée sur la vision par ordinateur pour l'analyse de la démarche. Avec l'apparition de caméras 3D bon marché, le problème de l'évaluation qualitative de la démarche a été réexaminé. Les chercheurs ont réalisé le potentiel des dispositifs de

caméras 3D pour les applications d'analyse de mouvement. Cependant, malgré des progrès très encourageants dans les technologies de détection 3D, leur utilisation réelle dans l'application clinique reste rare.
Ce projet propose des modèles et des techniques pour l'évaluation du mouvement à l'aide d'un capteur Kinect. En particulier, nous avons étudié la possibilité d'utiliser différentes données fournies par une caméra RGB-D pour l'analyse du mouvement et de la posture. Les principales contributions sont les suivantes. Nous avons proposé un descripteur de posture basé sur la répartition dans l'espace d'un nuage de points 3D. Le descripteur conçu peut classer les postures humaines statiques à partir des données 3D. Nous avons construit un système d'acquisition pour l'analyse de la marche basée un réseau récurrent LSTM. Enfin, nous avons proposé une approche de détection de démarche anormale basée sur les données du squelette (position des articulations). Nous avons démontré que notre outil d'analyse de la marche fonctionne bien sur une collection de données que nous avons générée ainsi que sur des données de l'état de l'art. Notre méthode d'évaluation de la démarche montre des avancées significatives dans le domaine. Notre approche nécessite un équipement limité et est prête à être utilisée pour l'évaluation de la démarche en conditions réelles.

1.3.3 Micro-expressions

Acquisition	Vidéo rapide couleur de personnes vue de face.
Objectif	Reconnaître une micro-expression.
Contributions	• Reconnaissance par LBP-TOP unifié dans le temps. • Étude de la profondeur du CNN pour la reconnaissance. • Étude de la dimension du flot optique en entrée d'un CNN pour la reconnaissance. • Évaluation du spoting par les méthodes classiques ainsi que de l'association spotting/reconnaissance. • Introduction d'un paradigme plus simple mais pertinent pour le spotting.
Encadrements	Thèse de Reda Belaiche et 4 stages et projets d'études.
Financement	MESRI.
Publications	[ApSc2020], [ICIAP2019] et [SITIS2019].

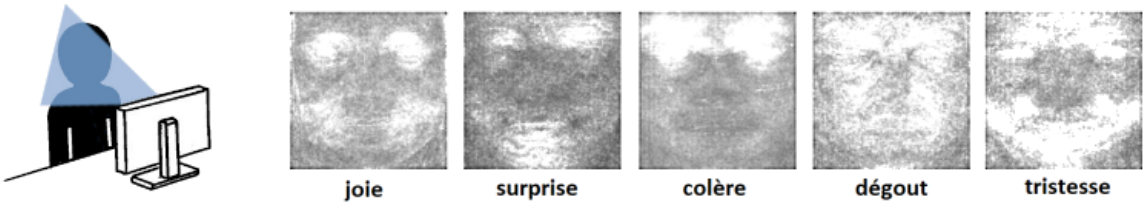


Figure 1.4: À gauche le mode d'acquisition; à droite le flot optique moyen sur les micro-expressions relatives à chacune des émotions.

Les technologies de l'interaction homme-machine se concentrent de plus en plus sur l'être humain, que ce soit sur son identité ou bien sur son état physique et mental. Des progrès conséquents ont été réalisés depuis quelques décennies. Cependant l'étude des pensées et des

émotions reste encore un domaine peu développé, mais qui a commencé à grandement gagner en intérêt. L'une des pistes les plus étudiées dans la recherche autour des émotions humaines est l'apparition et le décodage des expressions faciales. L'homme est instinctivement conscient des macro-expressions faciales et un nombre conséquent d'études se concentrent sur elles. Il existe cependant un autre type d'expression faciale dont la majorité des êtres humains ne sont pas conscient : les micro-expressions du visage. Celles-ci sont caractérisées par leur courte durée et leur très faible intensité. La communauté scientifique en vision par ordinateur étudie depuis quelques années la possibilité de reconnaître automatiquement les micro-expressions à l'aide de cameras rapides et de programmes informatiques. Il s'agit néanmoins d'un problème difficile en raison de la nature de ces micro-expressions. Les récents progrès du machine learning permettent d'adopter des méthodes nouvelles et efficaces pour résoudre diverses tâches de vision par ordinateur applicables à la reconnaissance de micro-expressions.
Basé sur les dernières avancées techniques en machine learning, l'objectif de ce projet est d'établir un système de reconnaissance des émotions sans contact et temps réel, permettant de répondre à des performances en robustesse et flexibilité, faible coût, usage simple et capacité à être embarqué à moindre coût énergétique.
Notre première méthode, basée sur des descripteurs (comme le Local Binary Patterns on Three orthogonal Planes - LBP_TOP) vise à unifier la représentation temporelle du descripteur. Les meilleures performances ont été obtenues sur les représentations les plus courtes, démontrant que c'est bien le mouvement et non son évolution qui est descriptif.
Puis nous nous sommes intéressés aux architectures de réseaux de neurones spécialement adaptées à la classification de micro-expressions. Nous avons étudié la possibilité de réduire les besoins en mémoire et en calculs nécessaires pour classifier des micro-expressions en faisant le minimum de concessions possible sur l'efficacité du système. Nous avons observé que la profondeur du réseau pouvait être réduite sans perte notoire et que le mouvement selon l'axe vertical, le plus descriptif, pouvait se suffire efficacement.
Nous avons étudié l'association du spotting (localisation des micro-expressions) et de la reconnaissance. Le problème étant particulièrement complexe, nous avons proposé un nouveau paradigme où le début de l'expression (ou onset) est connu, ce qui est cohérent avec les applications envisagées où l'émotion est provoquée par un stimulus. Les résultats obtenus montrent qu'une solution basée sur une estimation statistique est plus efficace pour le problème dans son ensemble qu'une recherche temporelle de l'apex (l'instant de plus forte intensité de l'expression).

1.3.4 Sonification d'une scène 3D

Acquisition	Vidéo RGB-D d'un système embarqué sur des lunettes.
Objectif	Sonifier la scène 3D pour aider à la perception et au déplacement des personnes malvoyantes.
Contributions	• Proposition d'un système de substitution visio-auditif temps réel. • Sonification de la présence de personnes dans la scène. • Aide au déplacement dans des locaux à partir de marqueurs fixes. • Aide au déplacement en extérieur dans un milieu urbain.
Encadrement	Thèse de Florian Scalvini.
Financement	Projet envergure Région.
Publications	[ICASSP2022], [SITIS2022] et [Frontiers2023]

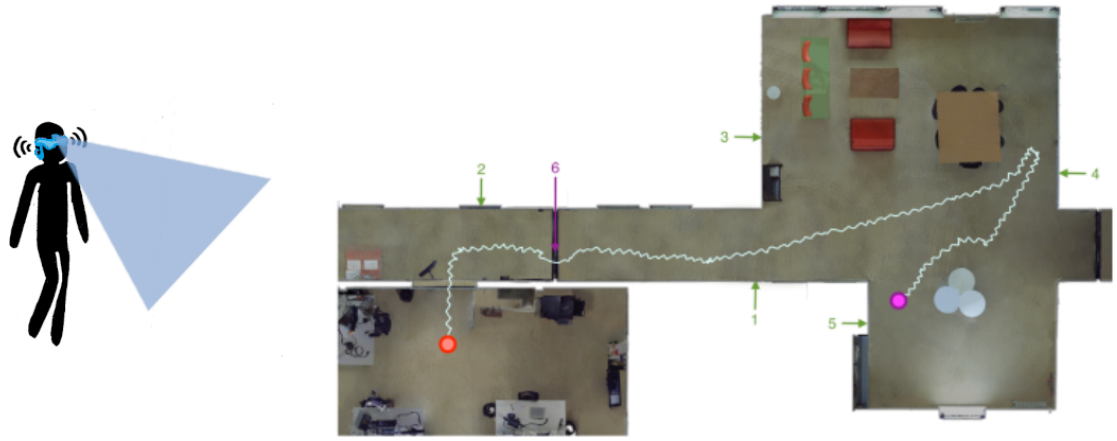


Figure 1.5: Parcours (en blanc) d'une personne avec notre système et la recherche des STag (1 à 6). La personne se dirige naturellement vers la destination (5) en évitant les obstacles (tables) et en passant par la porte (6).

Pour ce projet, nous avons proposé une méthode de substitution visuo-auditive pour aider les personnes malvoyantes à comprendre les scènes et à atteindre une destination dans un environnement intérieur. La sonification se réalise de cette façon. Pour chaque image capturée, les coordonnées d'un pixel à sonifier sont transposées en coordonnées sphériques sur une sphère de 2 mètres de rayon centrée sur la caméra. Nous avons ensuite utilisé l'ensemble des données HRIR (Head-Related Impulse Response) enregistrées dans une chambre anéchoïque pour spatialiser un son monophonique bref (33ms) de 440 Hz avec un fondu en cosinus de 5ms. Pour chaque position de pixel possible, nous avons précalculé sa spatialisation HRIR sur la base des réponses d'entrée de la position azimutale correspondante dans l'ensemble de données HRIR. L'amplitude du son est modulée en fonction de la distance qui sépare la cible de l'utilisateur en utilisant une loi de carré inverse.

Nous avons proposé une première approche se concentrant sur la localisation de personnes dans le voisinage de l'utilisateur afin de faciliter les interactions. Étant donné qu'un traitement temps réel et une faible latence sont nécessaires dans ce contexte pour la sécurité de l'utilisateur, nous avons proposé un système embarqué. Le traitement est basé sur un réseau de neurones convolutif léger pour effectuer une localisation 2D de personne efficace. Cette mesure est enrichie de l'information correspondant à la distance où se trouve la personne (profondeur) et est ensuite transcrite en un signal audio stéréophonique. Nous avons réalisé une implémentation basée GPU qui permet d'atteindre un traitement en temps réel à 23 images/s sur un flux vidéo de 640×480. Nous avons démontré expérimentalement que cette méthode permet une localisation précise en temps réel.
La seconde approche proposée produit un service de navigation dans un bâtiment pour les personnes malvoyantes afin de les aider à atteindre leur destination d'intérêt via le chemin le plus court. Des étiquettes (marqueurs STag) sont placées à des endroits pertinents et sont reconnus par le système porté par l'utilisateur. Ensuite, le système guide l'utilisateur vers ces points les uns après les autres jusqu'à ce qu'il atteigne sa destination. Le système sonifie à la fois la direction à prendre (vers le marqueur détecté) mais aussi la présence d'obstacle sur le chemin. L'utilisateur se dirige vers son objectif de façon assez naturelle en évitant les obstacles et en ouvrant les portes. Le procédé a été étendu à la navigation en extérieur dans un milieu urbain.

1.3.5 Autres projets

- Interaction d'une main filmée par une caméra RGB-D avec un objet virtuel ([IVC2019]).
- Calibration d'une caméra thermique à partir d'un damier refroidi ([QIRT2020]).
- Endoscopie 3D pour la recherche de défaut en aéronautique (projet rapid DGA en collaboration avec la société EFER - [COMPAS2023]).
- Estimation d'éléments pérennes dans des scènes 3D acquises à partir de caméras RGB-D embarquées sur des voitures dans le contexte de la conduite autonome (en collaboration avec la société Huawei - [IROS2023]).

Diagnostic basé sur le mouvement		Interprétation de scène 3D	
Micro-expression	Analyse de la démarche	Détection de personnes	Sonification d'une scène 3D
Vidéo rapide	Vidéo de profondeur	CPC	Vidéo RGB-D par un système embarqué
Classification et spotting de ME.	Reconnaissance de pose de la marche et reconnaissance de démarches anormales.	Classification d'un CPC et détection dans une scène 3D.	Sonification de personnes et aide au déplacement dans une scène.
Section 3.1	Section 3.2	Section 4.1	Section 4.2
[ApSc2020] [ICIAP2019] [SITIS2019]	[AIM2019] [MTAP2018] [SITIS2018] [SITIS2016]	[MTAP2019] [ICASSP2022] [VISAPP2016]	[Frontiers2023] [SITIS2022]

Table 1.1: Positionnement des thématiques traitées

1.4 Composition du manuscrit

Le chapitre 2 contient un état de l'art des descripteurs relatifs aux mouvements 3D des personnes. Les descripteurs permettent d'extraire l'identité profonde d'une classe. Nous cherchons ici à mettre en valeur comment la 3D et le mouvement sont utilisés pour décrire le comportement d'une personne. Les méthodes d'apprentissage profond pourtant prédominantes sur ces dernières années seront peu traitées car plus liées à un principe de classification qu'à une spécialisation à une classe. Le chapitre comprend une réflexion sur l'évolution du domaine ces dernières années et présente des pistes sur son évolution future.

Le chapitre 3 traite de l'influence du mouvement sur l'analyse du comportement humain. Plus spécifiquement nous focalisons le propos sur l'influence de deux caractéristiques du mouvement : les mouvements subtils avec une application à la reconnaissance de micro-expressions (section 3.1) puis les mouvements cycliques avec une application à l'étude de la marche (section 3.2). Il s'agit pour les deux cas de problèmes très complexes du fait de la faible variation inter-classe.

Le chapitre 4 propose une étude sur l'aide à l'interprétation par une personne d'un environnement 3D. Dans un premier temps, nous resterons sur une personne comme centre de l'attention. La section 4.1 traite en effet de la détection d'une personne dans une scène 3D. Ensuite la personne devient acteur de l'interprétation et émettrice du mouvement. La section 4.2 développe l'assistance à une personne pour naviguer dans un environnement 3D.

Je tirerai dans la section 5 un bilan de ces travaux ainsi que sur mon parcours au sein de mon laboratoire. Puis je développerai mon projet de recherche pour les années à venir dans le chapitre 6.

Enfin dans la partie II, je récapitulerai les informations relatives à mon parcours en recherche et en enseignement.

Chapitre 2

Description du mouvement 3D des personnes

Avec l'arrivée de caméras 3D à faible coût et les efforts continus dans le traitement avancé des nuages de points, la perception 3D a gagné en importance dans le domaine de la vision. Les images de profondeur fournissent des surfaces naturelles qui peuvent être exploitées pour capturer les caractéristiques géométriques de la scène observée à partir d'un descripteur géométrique. Par rapport aux données couleur conventionnelles, les informations de profondeur dans les données RGB-D permettent de s'adapter à différentes conditions d'éclairage, à un point de vue différent, d'éliminer le bruit de fond et de simplifier les variations de mouvement intra-classe. Par conséquent, en général, les descripteurs basés sur les données RGB-D sont plus performants, ou au moins plus robustes, que les descripteurs basés sur les données RGB [1].

Les informations extraites des nuages de points 3D sont principalement constituées de la forme, de la couleur (ou de l'intensité) et de la relation spatiale entre les points du nuage. Les descripteurs de forme sont les descripteurs 3D les plus populaires pour les nuages de points, principalement les descripteurs basés sur les normales (Figure 2.1). Les normales sont les vecteurs perpendiculaires aux surfaces du nuage de points. Elles fournissent la plupart des informations pertinentes sur la forme et la structure d'un objet en 3D. Il existe de nombreuses méthodes pour les estimer, mais la plus simple consiste à trouver la normale d'un plan tangent à la surface, qui est un ajustement planaire des moindres carrés.

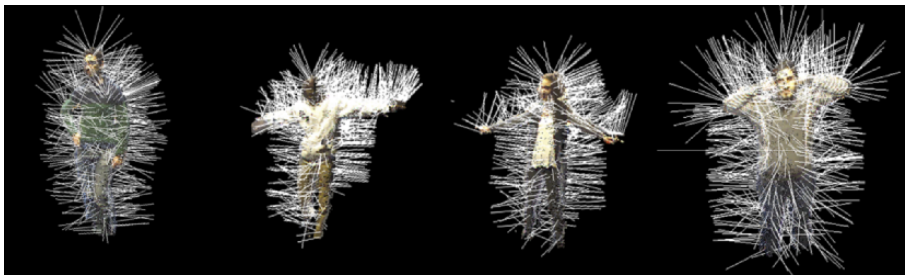


Figure 2.1: Exemple de nuage de points ou les normales à la surface ont été représentées sur une sélection aléatoire de points.

Le mouvement 3D, parfois appelé 4D, représente l'évolution temporelle des scènes spatiales 3D. Ces données peuvent être décrites par le calcul de ce que l'on appelle le flux de mouvement (flot optique) qui est composé de vecteurs de vitesse 3D. La Figure 2.2 montre un exemple de flux de mouvement dense (c'est-à-dire calculé pour chaque point) déterminé à partir de deux images consécutives par l'algorithme PD Flow [2]. Le flux de mouvement dense est riche en

informations, mais son estimation est également très exigeante en termes de calcul et entraîne une charge importante en termes de mémoire et de stockage. L'utilisation des données de flux de mouvement pour les tâches de classification n'est généralement pas simple et les chercheurs expérimentent différents schémas de quantification. Récemment, de nombreux algorithmes alternatifs ont été proposés afin de réduire cette charge.

Figure 2.2: À gauche, images d'intensité sur deux instants successifs. À droite, le flot de mouvement correspondant calculé par [2].

L'objectif de ce chapitre est d'examiner les descripteurs 3D appliqués au mouvement humain. Bien que de nombreuses méthodes récentes permettent, notamment par l'apprentissage profond, de réaliser la classification directement à partir des données et de déterminer la caractérisation des classes à partir d'une large base d'échantillons, nous avons choisi de présenter ici uniquement la façon de décrire le mouvement à partir des a-priori connus ou estimés du mouvement humain.
Nous avons divisé ces traitements en trois sous-catégories :

Reconnaissance d'actions	Classifie les actions du corps entier effectuées par un sujet humain.
Reconnaissance de gestes	Classifie les mouvements locaux.
Analyse de la marche	Issue de l'analyse du mouvement qui évalue la façon de marcher d'une personne.

Ces trois domaines d'application nous permettent de couvrir les spécificités des descripteurs de mouvements humains pour différents niveaux d'analyse : de plus général dans la reconnaissance d'actions à plus fin et particulier dans la reconnaissance de gestes et l'analyse de la marche. Ils font l'objet de recherches actives depuis de nombreuses années et ont bénéficié du développement des caméras 3D, qui permettent de supprimer facilement l'arrière-plan et de réduire l'ambiguïté des données 2D.

L'idée de ce chapitre est d'exprimer un contexte lié aux différents travaux que j'ai réalisés. Il retrace ainsi un état des lieux de l'existant lors des premières années après mon recrutement (en 2014). Les méthodologies que j'ai proposées se basent sur ces connaissances qui faisaient référence dans le domaine. Il est donc logique de partir de ces dernières pour expliquer mon cheminement. Mais bien évidemment la recherche avance sans cesse et certains choix que j'ai réalisés sont moins pertinents au vue des dernières avancées notamment technologiques. Je proposerai ainsi dans la suite de ce chapitre une interprétation personnelle sur les dernières évolutions ainsi que sur les dernières tendances.

2.1 Descripteur pour la reconnaissance d'actions

La description des mouvements est une partie essentielle de la reconnaissance des activités humaines. Les méthodes capables de discriminer la classe d'une action en cours sont basées sur l'analyse d'une séquence vidéo combinant un descripteur de mouvement avec un classificateur.

Les deux principaux groupes de méthodes se basent soit sur les articulations du squelette soit sur la carte de profondeur. Le squelette peut être extrait des données de profondeur à l'aide de la méthode proposée par Shotton et al [3] ou, avec une meilleure précision, par différents systèmes de capture de mouvement avec marqueurs. Considérons également les HMM (Hidden Markov Model) qui modélisent les données dans le temps. Notons que les HMM peuvent être utilisés avec les données du squelette et les caractéristiques de profondeur.

2.1.1 Méthodes basées sur les squelettes

Une approche très courante pour la reconnaissance des actions humaines consiste à suivre les articulations du squelette humain dont les états sont estimés à partir des cartes de profondeur [4, 5, 6]. Les angles et positions de ces articulations, et même la géométrie relative entre elles, sont ensuite utilisés pour directement modéliser l'activité humaine.

Exemples — Descripteurs basés sur les squelettes

- [7] → une chaîne de caractéristiques spatio-temporelles pour représenter les actions humaines à partir de la trajectoire de position des articulations.
- [8] → une représentation du squelette obtenue à partir de la géométrie issue des rotations et translations nécessaires pour faire correspondre la position et l'orientation de deux parties du corps.
- [9] → les positions des articulations dans le temps forment un histogramme spatial 3D (HOJ3D) qui est reprojeté en utilisant l'analyse discriminante linéaire et regroupé en k mots visuels de posture.

Les méthodes basées sur les articulations du squelette sont populaires, mais elles échouent si les articulations ont été initialement mal estimées. Ce peut être problématique, notamment dans le cas d'auto-occlusions prononcées. De plus, si une action très fine doit être reconnue, par exemple un geste ou un léger mouvement, les méthodes basées sur les articulations manquent d'informations précises concernant la forme et le mouvement. Pour cette raison, les attributs de bas niveau dans les images de profondeur sont souvent plus performants que les représentations de plus haut niveau [10].

2.1.2 Méthodes basées sur les cartes de profondeur

Bien que les méthodes basées sur la représentation de haut niveau soient très populaires, leur principal inconvénient est la difficulté à représenter les mouvements subtils, ce qui conduit à la recherche de descripteurs de mouvement basés sur des caractéristiques de bas niveau.
Des descripteurs simplifiés (non denses) basés sur le flux de mouvement peuvent également être favorisés [11, 12]. Dans ce cas, soit une estimation grossière du mouvement est utilisée, soit un flux de mouvement dense est calculé puis encodé dans une représentation plus compacte.
Lorsque des cartes de profondeur sont traitées, les normales constituent une source importante d'informations sur la forme d'un objet. Elles ont été exploitées avec succès dans de nombreux descripteurs de mouvement [13, 10].

Une notion de hiérarchie peut être ajoutée avec succès dans les descripteurs de mouvement 3D. Dans [14], les normales 3D sont calculées et projetées sur un ensemble appris de vecteurs de base compacts KPCA. Une méthode BoW est utilisée pour construire une structure hiérarchique sur les caractéristiques de bas niveau des patchs de la vidéo pour produire des vecteurs de caractéristiques.
Les descripteurs basés sur des caractéristiques spatio-temporelles ont récemment fait l'objet d'une grande attention [15, 16]. Ce groupe de méthodes localise généralement les caractéristiques spatiales et décrivent leur évolution temporelle.

Exemples — Descripteurs basés sur la carte de profondeur

- [17] → un modèle graphique extensible modélise explicitement la dynamique temporelle des actions et propose d'utiliser un ensemble de points 3D extraits d'une carte de profondeur pour modéliser les postures.
- [18] → la direction et la magnitude du mouvement de chaque partie du corps sont prises en compte par une grille 3D qui divise l'espace autour d'une personne en un certain nombre de cubes dont sont extraits la direction et l'intensité du flux.
- [12] → l'orientation 3D des vecteurs de flux autour de points d'intérêt est codée à partir d'un histogramme sphérique de la vitesse.
- [10] → un espace 4D (XYZ, t) est initialement quantifié afin d'obtenir une représentation par une grille en utilisant une extension 4D régulière d'un polygone 2D, à savoir un Polychoron de 600 cellules, puis la distribution de l'orientation des normales à la surface est calculée (HON4D).
- [16] → les points caractéristiques candidats sont échantillonnés de manière dense dans chaque image couleur et suivis à l'aide du flux optique pour former des trajectoires 3D.

2.1.3 Méthodes basées sur des caractéristiques multiples

Contrairement aux premières années qui ont suivi l'apparition de capteurs de profondeur non coûteux, les méthodes [19, 20] qui fusionnent les caractéristiques de couleur et de profondeur ont gagné en popularité. Avec les récentes avancées dans la reconnaissance de l'activité humaine, les chercheurs se sont également attaqués à la tâche difficile de la reconnaissance d'une action de groupe. Pour incorporer les informations spatio-temporelles, de couleur et de profondeur dans l'espace XYZt, [15] utilisent une cascade de trois filtres : un filtre de passage pour encoder les indices le long de la dimension de profondeur, un filtre gaussien pour encoder les indices dans l'espace XY et un filtre de Gabor pour encoder les informations temporelles.

Exemples — Descripteurs basés sur des caractéristiques multiples

- [21] → combinaison des données du squelette et des caractéristiques LST.
- [20] → un descripteur adaptatif MCOH (Multichannel Orientation Histogram) est appliqué à une région de support 4-D de chaque point d'intérêt puis les gradients d'image des patchs de profondeur sont calculés et quantifiés en utilisant une méthode basée sur les coordonnées sphériques.

TENDANCES

L'utilisation de réseaux profonds a profondément transformé l'approche du problème [22]. Partant de ce paradigme, les directions prioritaires de recherche se trouvent dans la compilation de données, l'apprentissage non supervisé ou la fusion de modalités. Les méthodes basées RNN/LSTM incorporent plus finement l'analyse de l'évolution temporelle comme [23] qui transforme les articulations du squelette en une représentation plus à même d'être exploitée par un LSTM. Mais ces méthodes sont souvent coûteuses à la fois en données et en traitements. De plus, l'évolution du geste sur l'intégralité de l'action n'est pas forcément discriminant. Les méthodes basées CNN sont plus légères et intègrent mieux la géométrie mais en minimisant l'influence de l'évolution temporelle. [24] utilise un réseau multi-stream pour conjointement prendre en compte les mesures de distance et angulaire. Les méthodes basées GCN apportent une flexibilité [25] mais souvent au prix d'une plus grande complexité.
Une ouverture vers des problèmes liant plusieurs entités comme pour l'étude des interactions entre personnes ou des activités de groupe [26] est également notable.
L'utilisation de la couleur seule est très courante [27] mais la profondeur reste un gage de robustesse. L'association de deux signaux est riche mais la fusion multi-modale à ses défauts. Dans [28], un réseau de transfert permet de forcer la correspondance entre les cartes RGB et de profondeur.

2.2 Descripteurs de mouvement pour la reconnaissance de geste

De nos jours, les applications informatiques nécessitent de nouveaux modes d'interaction, notamment dans le domaine en pleine expansion de la réalité virtuelle. C'est pourquoi l'interaction homme-machine, et en particulier la reconnaissance des gestes, est devenue un domaine de recherche très populaire.

Un geste peut être défini comme un mouvement physique des mains, des bras, du visage et du corps avec l'intention de transmettre une information ou un sens. Les descripteurs de mouvement dédiés à la reconnaissance de gestes sont similaires à ceux de la reconnaissance d'activité, à la différence que les descripteurs doivent être capables de capturer des mouvements plus subtils. Par exemple pour le suivi de la main, il y a plus de degrés de liberté et de sérieuses occlusions se produisent entre les doigts. Peu de méthodes de l'état de l'art exploitent les nuages de points 3D ni le mouvement 3D.
Comme pour les descripteurs de mouvement du corps entier, de nombreux types de caractéristiques visuelles ont été proposés pour les gestes. Les premiers travaux utilisaient des informations 2D et construisaient des descripteurs pour une silhouette 2D d'une main. En raison de l'ambiguïté des données 2D, la précision de ces méthodes n'était pas élevée. Les dernières méthodes de reconnaissance dynamique des gestes utilisent des informations de profondeur en 3D et leurs équivalents en 2D.
Les méthodes basées sur des modèles sont très populaires. Contrairement à la reconnaissance d'action, où les méthodes d'estimation du squelette sont bien développées, il n'existe pas vraiment d'algorithme d'estimation de la position des articulations de la main qui soit nettement supérieur aux autres (l'introduction de la caméra Leap-Motion a modifié ce constat). Les articulations du squelette de la main peuvent cependant être utilisées comme caractéristiques pour l'étape de classification. Une autre approche consiste à caractériser les modèles de mouve-

ment à partir d'indices de profondeur directs. Diverses caractéristiques géométriques peuvent être extraites d'une séquence de profondeur. Par exemple la forme d'une silhouette [29] ou l'occupation des cellules d'une grille [30] peuvent être utilisées comme caractéristiques. Il en résulte des approches qui sont moins dépendantes d'algorithmes de segmentation et de suivi conjoints.
Lorsque l'on parle de reconnaissance de gestes, il faut également mentionner le Dynamic Time Wrapping (DTW) et ses variantes. Ils sont utilisés pour aligner deux séquences sur la base de caractéristiques présélectionnées. En général, le DTW est appliqué à l'étape de la classification et n'affecte pas la conception du descripteur [31].

Exemples	Description pour la reconnaissance de geste
• [32] → une matrice de covariance composée des caractéristiques sélectionnées dans les images de profondeur d'une séquence vidéo exploite les représentations des interactions complexes entre les variations des caractéristiques 3D dans le domaine spatial et temporel.	
• [1] → les descripteurs de données de profondeur et de couleur ont été extraits séparément et leurs performances ont été comparées.	
• [33] → des réseaux de neurones sont appliqués de façon indépendante sur les données du squelette, les images de profondeur et les images couleur, puis leurs sorties sont fusionnées selon différents schémas.	

TENDANCES

La reconnaissance de geste tend à se spécialiser sur une application spécifique. Plutôt qu'un descripteur général, la spécialisation à la tache se manifeste par un apprentissage sur une base de données dédiée et une optimisation de la configuration du réseau choisi. Les avancées en génération de modèles ont également rendu très populaire la recherche de la prédiction du mouvement d'une personne [34]. Ici aussi l'apport de la 3D est significatif autant sur le point méthodologique qu'applicatif.

2.3 Descripteurs de mouvement pour l'analyse de la démarche

2.3.1 La démarche

La reconnaissance et l'analyse de la marche à partir de données 3D est devenue populaire depuis un quinzaine d'années. La démarche est une manière de marcher sur un substrat solide. Elle est périodique par nature et est donc généralement divisée en cycles, qui comportent à leur tour 8 événements clés [35] :

- Contact initial : la frappe du talon initie le cycle de marche et représente le point où le centre de gravité du corps est à sa position la plus basse.
- Réponse à la charge : le pied à plat est le moment où la surface plane du pied touche le sol.
- Midstance : se produit lorsque le pied controlatéral passe devant le pied d'appui.
- Stance terminale : le talon est décollé lorsque le talon perd le contact avec le sol et que la poussée est initiée par les muscles triceps.

- Avant l'élan : le décollement des orteils met fin à la phase d'élan lorsque le pied quitte le sol.
- Le swing initial : l'accélération commence dès que le pied quitte le sol et que le sujet active les muscles fléchisseurs de la hanche pour accélérer la jambe vers l'avant.
- L'élan moyen : il se produit lorsque le pied passe directement sous le corps, coïncidant avec l'élan moyen de l'autre pied.
- L'élan terminal : la décélération décrit l'action des muscles lorsqu'ils ralentissent la jambe et stabilisent le pied en préparation du prochain cycle de marche sur le talon.

L'observation de la démarche peut fournir des indices diagnostiques précoces pour un certain nombre de troubles du mouvement tels que la maladie de Parkinson, l'infirmité motrice cérébrale, les accidents vasculaires cérébraux, l'arthrite, les maladies pulmonaires obstructives chroniques et bien d'autres. Selon le domaine de recherche, différents paramètres de la démarche sont évalués. Les paramètres de la marche peuvent être divisés en paramètres cinématiques (comme l'angle de flexion du genou) et spatio-temporels (comme la vitesse). Une approche courante consiste à expérimenter différentes combinaisons de paramètres de la marche et à sélectionner la plus représentative pour une tâche donnée, comme dans le travail d'Agosti et al. [36], où les auteurs ont effectué une analyse spatio-temporelle et cinématique de la marche de patients atteints de démence fronto-temporale et de la maladie d'Alzheimer.

Les descripteurs pour la reconnaissance de la marche incluent généralement les paramètres biométriques, car la variabilité intra-personne n'est plus un problème comme dans le cas de la reconnaissance des actions. Les informations sur le mouvement font partie des informations utilisées pour décrire un modèle de la marche. C'est pourquoi les descripteurs de mouvement utilisés pour la reconnaissance de la marche en 3D sont généralement plus simples et plus compacts que ceux utilisés pour la reconnaissance des actions et des gestes. Bien que les caméras RGB-D soient des outils populaires dans les tâches d'évaluation de la marche, il est assez courant d'utiliser des projections 2D. De plus, une grande majorité des méthodes modernes de reconnaissance et d'analyse de la marche effectuent une transformation 3D-2D de la séquence de profondeur [37, 38] ou utilisent directement des capteurs 2D [39].

Exemples — Descripteurs pour l'analyse de la marche

- [40] → description par la carte d'énergie de la démarche c'est à dire la silhouette moyenne sur un cycle de marche.
- [37] → une optimisation par essaims de particules pour le suivi des mouvements par partie et une signature de la marche composée des distances entre les articulations projetées sur un plan 2D.
- [41] → modèle de marche sur des voxels 2,5D basé sur une combinaison de gaussiennes et de la courbure moyenne des données du nuage de points.
- [42] → suivi par filtrage particulaire des points correspondant aux différentes parties des jambes amélioré par un modèle de mouvement harmonique simple qui correspond à la manière de marcher de l'homme.

2.3.2 Descripteurs de la démarche basés sur la 3D

Comme pour la reconnaissance des actions et des gestes, les méthodes de reconnaissance de la démarche en 3D peuvent être classées en deux catégories : les méthodes basées sur

les articulations du squelette [14, 43] (à partir d'un modèle) et les méthodes basées sur les images de profondeur [42] (sans modèle). Les méthodes basées sur le squelette [43, 44] sont similaires aux méthodes analogues de reconnaissance d'actions : les descripteurs sont basés sur la position spatiale et temporelle des articulations du squelette humain ou un modèle de corps humain est utilisé. Les caractéristiques de la marche basées sur la profondeur utilisent des informations détaillées sur la forme et la variation de la profondeur d'un individu qui marche et ne nécessitent pas d'ajustement de modèle. Pour la reconnaissance et l'analyse de la marche, l'utilisation de HMM (et plus récemment de LSTM) entraînés sur différentes caractéristiques est une tendance populaire.

La conception d'un descripteur de démarche dépend fortement de l'application spécifique. Pour certaines applications, l'analyse basée sur les données du squelette est suffisante, tandis que d'autres nécessitent un suivi et une description plus précis des formes. L'analyse de la marche en 3D est le domaine le moins exploré parmi les trois passés en revue et qui bénéficierait d'une description plus poussée des indices de mouvement en 3D.

2.3.3 HMM et LSTM dans la recherche sur la démarche

Les modèles de Markov cachés (HMM) et les réseaux de mémoire à long et court terme (LSTM) [45] créent un modèle prenant en compte les paramètres temporels des données. Les HMM et les LSTM sont des outils d'apprentissage automatique très populaires qui représentent une séquence d'événements, comme une action [46, 47, 9] ou la marche [48]. Les HMM et les LSTM peuvent être utilisés pour la classification des données mais aussi pour leur codage.

Un HMM est défini comme un processus stochastique à double encastrement avec un processus sous-jacent qui est caché (non observable). Le processus caché ne peut être observé qu'à travers un autre ensemble de processus stochastiques qui produisent la séquence d'observations. Les HMM sont utilisés depuis longtemps pour la reconnaissance et l'analyse de la marche pour les données 2D [49, 48, 50] en raison de leurs propriétés statistiques et de leur capacité à refléter la nature temporelle de la marche. Les modèles HMM capturent la forme et la dynamique temporelle d'une personne qui marche et sont utilisés pour la description ou l'identification de la démarche, sur la base de différentes caractéristiques extraites de la silhouette humaine, telles que les distances entre le centre et les points extérieurs [50], la largeur d'un contour extérieur [48], le flux LBP [49], etc. Dans la recherche sur la marche en 3D, les HMM sont encore peu exploités, mais il existe quelques exemples performants [44, 51, 35].

Les LSTM ont progressivement remplacé les HMM dans de nombreuses applications. Les HMM et les LSTM peuvent tous deux capturer des caractéristiques transitoires en plus des caractéristiques structurelles extraites au préalable. Les HMM sont mieux adaptés lorsque le nombre de caractéristiques n'est pas excessif. C'est pourquoi ils sont souvent appliqués aux méthodes basées sur les squelettes [43]. Les réseaux neuronaux récurrents LSTM [45] sont pertinents pour apprendre les caractéristiques pour la reconnaissance ou la classification de la marche.

Les méthodes basées sur les HMM et les LSTM sont des moyens très prometteurs pour décrire les données de mouvement en 3D, car elles reflètent les états temporels des actions, des gestes et de la démarche.

TENDANCES

L'estimation du squelette est maintenant rapide et facile à obtenir mais aussi fiable. La plupart des travaux récents se focalisent donc sur la façon de tirer le mieux parti de ces données par différentes configurations de réseaux profonds comme [52] qui propose un réseau convolutionnel sur une structure en graphe amélioré par une définition spatio-temporelle de l'attention.
La méthodologie utilisée n'a pas subi de révolution importante. La reconnaissance d'une personne par sa démarche reste un problème ouvert [53] au vue de la faible variation dans le style de démarche de chaque individu. L'évolution de la démarche d'une personne (rééducation) outrepasse cette difficulté mais fait ceci dit face à une variation inter-classe très subtile.
La représentation 3D reste primordiale pour une analyse pertinente de la marche. L'acquisition de la profondeur est courante et peu coûteuse mais des méthodes par estimation monoculaire de la profondeur [54] sont possibles pour associer efficacement les méthodes basées 3D sur une acquisition standard.

2.4 Conclusion

L'apparition de capteurs de profondeur à faible coût a influencé de manière significative la recherche dans la reconnaissance et l'analyse des actions, des gestes et de la démarche. Le principal problème pour les applications basées sur le mouvement humain reste la plage de profondeur limitée imposée par la technologie utilisée. Ce problème est moins pertinent pour les applications de reconnaissance des gestes mais il limite l'utilisation des capteurs 3D dans l'analyse de la démarche et la recherche sur la reconnaissance des actions. La portée limitée en profondeur des capteurs 3D peut être surmontée par des caméras stéréo et plusieurs solutions intéressantes ont été proposées récemment dans ce domaine.
Les approches qui modélisent les statistiques spatiales et temporelles de manière holistique pour les données de nuages de points donnent des résultats moins prometteurs que les points caractéristiques LST et les méthodes basées sur la projection. Les méthodes basées sur des modèles ont toujours un grand potentiel mais ne sont pas nécessairement plus performantes que les méthodes basées sur des caractéristiques de bas/moyen niveau.
Le principal problème de nombreuses méthodes utilisant des caractéristiques de bas niveau reste le temps nécessaire pour extraire les caractéristiques des séquences de nuages de points. Les caractéristiques offrant les meilleures performances de reconnaissance sont souvent coûteuses en temps de calcul.
Dans le domaine de la reconnaissance d'actions, certaines méthodes récentes tendent à utiliser des informations multimodales pour obtenir une meilleure précision. Cela permet d'améliorer les résultats de la reconnaissance mais introduit d'autres difficultés, comme la nécessité d'une synchronisation et de calculs supplémentaires. Les méthodes basées sur l'estimation des articulations fournissent généralement des descripteurs compacts et significatifs et, avec les progrès de la reconnaissance des articulations des squelettes, elles ont un grand potentiel. Les méthodes d'ajustement de modèles restent populaires pour l'analyse de la démarche et la reconnaissance des gestes, mais pour la reconnaissance des actions, l'accent a été mis sur les méthodes sans modèle. Les dernières évolutions s'orientent, pour les trois thèmes évoqués, sur une spécialisation du résultat recherché. Plutôt qu'une course à la performance sur le thème général, de plus en plus de méthodes se focalisent sur une application précise, principalement depuis la popularisation de l'apprentissage profond. Cette spécialisation, outre la composition de bases

de données dédiées et l'augmentation artificielle de ces données, se traduit par des contraintes spécifiques et des acquisitions propres (multi-modalité, multiples points de vue, capteurs plus performants). L'intégration conjointe du mouvement et de la 3D est peu utilisée du fait de l'augmentation de la complexité du modèle engendrée ainsi que de la quantité plus élevée de données à fournir pour l'apprentissage. Il semble pourtant que, correctement combinée, elle soit particulièrement descriptive.

Chapitre 3

Analyse du comportement humain par le mouvement

L'analyse du comportement humain est une thématique très riche. Tout d'abord elle donne un cadre au thème plus large de la prise de décision. En effet elle définit une suite de contraintes et de critères qui caractérisent notre objectif. Ensuite elle apporte un cadre applicatif à de multiples concepts théoriques. Ainsi des algorithmes généraux sont adaptés pour correspondre au mieux à la classe des personnes.

Cette thématique m'est chère et était déjà au cœur de mes travaux avant mon recrutement. À cette époque, je me concentrais notamment sur le suivi de postures en vue de reconnaître un comportement d'achat dans un supermarché et sur l'alignement de vidéos de répétitions de théâtre avec le texte de la pièce. Depuis j'ai aussi organisé cinq éditions d'un workshop autour de cette thématique (HTBA : worshop of Human Tracking and Behavior Analysis) et une special issue du journal sensors (Human Activity Recognition Based on Image Sensors and Deep Learning). Je la considère ainsi comme le cœur de mon travail de recherche.

Le mouvement est la clé du décryptage du comportement. En effet, si la posture ou l'expression faciale produisent des indices ponctuels riches, c'est bien la variation de ces éléments dans le temps qui s'avère la plus révélateur.

Dans ce chapitre, je présente deux problématiques liées à cette lecture du mouvement pour la compréhension du ressenti et de l'attitude d'une personne.

Mouvements rapides et subtils

Par le biais de l'analyse des micro-expressions, je traite de l'expression minimaliste du mouvement. Le mouvement produit par une micro-expression faciale est très court (moins d'un quart de seconde) et de faible intensité. Bien que la nature des régions d'intérêt soit connue, cette courte extériorisation de l'émotion ressentie la rend très complexe à analyser.

Mouvements cycliques

La marche humaine est constituée d'une série d'étapes clés (de postures significatives) qui se répètent dans un ordre fixe. Cette régularité simplifie l'étude, permettant de s'attaquer à la recherche de déviations fines représentatives d'un défaut dans la démarche lors d'une phase de rééducation par exemple.

Chacune de ces deux problématiques, caractérisées dans le tableau 3.1, seront développées dans les deux sections suivantes.

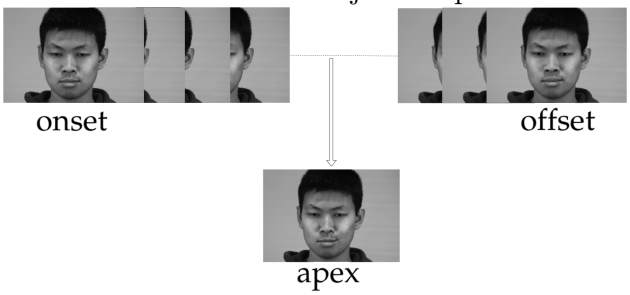
Mouvement rapide	Mouvement cyclique
Micro-expression	Marche humaine
Caméra rapide RGB (GO2400C)	Caméra RGB-D (Kinect)
Moins de 0,25 secondes	Quelques secondes
[ApSc2020] [ICIAP2019] [SITIS2019]	[AIM2019] [SITIS2018] [VISAPP2018]

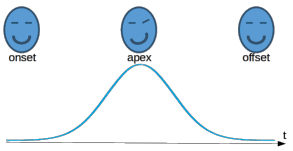
Table 3.1: Caractérisation des deux problématiques traitées concernant l'analyse du comportement humain.

3.1 Mouvements rapides et subtils : les Micro Expressions

Les Micro-Expressions

Initialement découvertes par Haggard et Isaacs [55], les Micro-Expressions (ME) sont un type d'expressions faciales involontaires extrêmement rapides et de très faible intensité. Leur durée est de l'ordre du quart de seconde, ce qui rend leur localisation et leur analyse assez compliquées. Les Micro-Expressions peuvent se produire dans deux situations : la suppression consciente et la répression inconsciente. La suppression consciente se produit lorsqu'une personne essaie intentionnellement de s'empêcher de montrer ses véritables émotions ou de les cacher. La répression inconsciente se produit lorsque le sujet lui-même ne se rend pas compte de ses véritables émotions Dans les deux cas, les micro-expressions trahissent les véritables émotions du sujet indépendamment de sa conscience de leur existence.





Instants déscriptifs de la ME :
onset→début
offset→fin
apex→intensité maximale

Deux tâches principales peuvent être recensées dans l'étude des expressions faciales : la reconnaissance et le spotting. La reconnaissance consiste à analyser le contenu de la séquence vidéo d'une expression et à estimer le type d'émotion qui lui est lié. Le spotting consiste quant à lui à détecter temporellement une expression dans une séquence. La reconnaissance se base sur une classification alors que le spotting a pour objectif la détection et la localisation des évènements. Les deux tâches sont intimement liées puisque le spotting fournit les expressions à classer par la reconnaissance. En fait, la sortie du spotting correspond à l'entrée de la reconnaissance. Bien qu'une utilisation pratique implique la combinaison des deux tâches, les études dans le domaine les séparent généralement en 2 étapes distinctes.

L'analyse des Micro-Expressions (ME) est un sujet d'actualité dans le domaine de la vision par ordinateur car elle constitue une passerelle importante pour saisir et comprendre

les émotions humaines quotidiennes. Il s'agit néanmoins d'un problème difficile, car la micro-expression est généralement transitoire (étant donné qu'elle dure moins de 250 ms) et subtile.

Les récents progrès du machine learning permettent d'adopter de nouvelles méthodes efficaces pour accomplir diverses tâches de la vision par ordinateur. En particulier, l'utilisation de techniques d'apprentissage profond sur de grands ensembles de données surpasse les approches classiques basées sur l'apprentissage classique avec des descripteurs. Même si les ensembles de données disponibles concernant la ME spontanée sont rares et beaucoup plus réduits, l'utilisation de réseaux neuronaux convolutionels (CNN) dans ce domaine donne des résultats de classification relativement satisfaisants. Cependant, ces réseaux sont exigeants en termes de consommation de mémoire et de ressources de calcul. Cela pose de grands défis lors du déploiement de solutions basées sur les CNN dans de nombreuses applications grand public, telles que la surveillance des conducteurs et la reconnaissance de l'émotion dans les classes virtuelles (e-learning), qui exigent une analyse précise, rapide et portable sur des systèmes embarqués.

Le travail que nous avons réalisé sur ce thème est varié allant de la définition de descripteurs de texture pour la classification à une étude de l'influence du contenu des bases de données disponibles sur les résultats obtenus. Dans ce manuscrit, je me suis concentré sur deux problèmes : l'optimisation du réseau profond pour une classification légère sans perte significative de précision pour la classification puis l'étude du spotting de ME et de son influence sur la classification. Les autres travaux réalisés sont disponibles dans [56].

3.1.1 Classification par optimisation de réseau profond

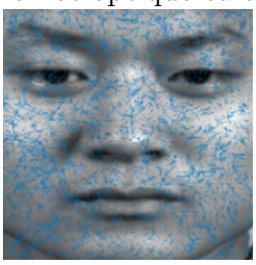
Les études appliquant l'apprentissage profond pour résoudre le problème de la classification des ME [57, 58, 59] ont généralement utilisé des CNN pré-entraînés tels que ResNet [60] et VGG [61] et ont appliqué l'apprentissage par transfert pour obtenir les caractéristiques des ME. Dans notre travail, nous avons d'abord choisi ResNet18 parce qu'il offrait le meilleur compromis entre la précision et la vitesse sur la classification difficile d'ImageNet et qu'il était reconnu pour ses performances en apprentissage par transfert. ResNet laisse explicitement les couches empilées s'adapter à une cartographie résiduelle. ResNet18 possède 20 couches convolutionnelles (CL) : 17 CL successives et 3 ramifications. Les liens résiduels après chaque paire d'unités convolutionnelles successives sont utilisés et la taille du noyau après chaque lien résiduel est doublée. Comme ResNet18 est conçu pour extraire les caractéristiques des images couleur RGB, il exige que les entrées soient codées sur 3 canaux.

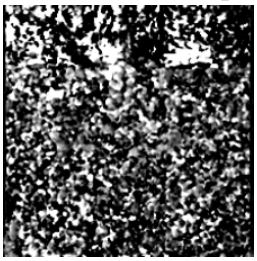
Pour obtenir le meilleur compromis entre la vitesse de traitement, le besoin de mémoire et la précision, nous lui avons apporté plusieurs modifications. Nous avons ensuite conçu nos propres architectures CNN pour exploiter au maximum les propriétés du flot optique extrait de ME afin de créer des structures spécifiques dédiées à la tâche de classification.

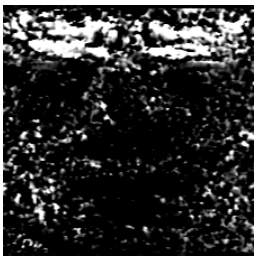
Étude sur la profondeur du réseau Les CNN sont créés pour des problèmes spécifiques et donc sur-calibrés lorsqu'ils sont utilisés dans d'autres contextes. ResNet18 a été conçu pour la reconnaissance d'objets de bout en bout : le jeu de données utilisé pour l'entraînement comporte des centaines de milliers d'images pour chaque classe et plus de mille classes au total. Mais une étude de reconnaissance des ME considère au maximum 5 classes, et les jeux de

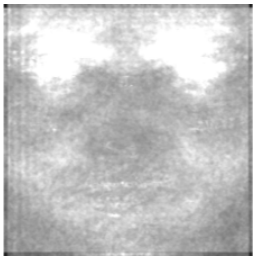
Le flot optique

Le flot optique est un moyen très efficace pour caractériser les ME en portant l'attention sur les mouvements des pixels lors d'un laps de temps précis.
À partir de l'hypothèse d'invariance de la luminosité, le mouvement de chaque pixel entre deux images sur une période de temps est estimé et représenté sous forme d'un vecteur indiquant la direction et l'intensité du mouvement.
La projection du vecteur sur l'axe horizontal correspond au champ V_x tandis que sa projection sur l'axe vertical est le champ V_y. La magnitude (M) est la norme du vecteur. Nous utilisons le flot optique calculé entre l'onset et l'apex.


vecteur $\{V_x,V_y\}$


V_x


V_y


magnitude M

données des ME spontanés sont rares et contiennent beaucoup moins d'échantillons. De plus le flux optique est une caractéristique de haut niveau contrairement à la couleur et nécessite donc des réseaux moins profonds. Nous avons ainsi empiriquement réduit l'architecture de ResNet18 en supprimant itérativement des couches résiduelles. Cela nous a permis d'évaluer l'influence de la profondeur du réseau sur ses capacités de classification dans notre contexte et donc d'estimer la configuration pertinente du réseau.

PROTOCOLE

- À chaque étape, la dernière couche résiduelle avec deux CL est supprimée et la précédente est connectée à la couche entièrement connectée. Seuls les réseaux avec un nombre impair de CL sont donc proposés.
- Les poids de tous les CNN sont pré-entraînés en utilisant ImageNet.
- Entrée : flot optique sur 3 canaux (V_x, V_y et M).
- Validation croisée LOSO.

Les précisions obtenues sont données dans le tableau 3.2. Nous pouvons observer que les meilleures performances ont été obtenues par la version avec 7 CL. Cependant la variation des scores en fonction du nombre de CL est limitée. En outre, au-delà de 7 CL, l'ajout de CL supplémentaires n'améliore pas la précision du modèle. Cela confirme que de multiples CL successives ne sont pas nécessaires pour obtenir une meilleure précision.
Le phénomène le plus intéressant se révèle être qu'avec une seule CL nous avons obtenu un score qui n'est pas très éloigné du score optimal alors que la taille du modèle est beaucoup plus réduite. Cela suggère qu'au lieu d'un apprentissage profond, une approche plus classique exploitant des réseaux neuronaux peu profonds présente un champ intéressant à explorer pour optimiser la portabilité et l'efficacité des calculs pour des systèmes embarqués. C'est la raison principale pour laquelle nous concentrons nos études sur des CNN compacts.

Étude sur la dimensionnalité des données d'entrée Les CNN prennent en entrée la carte de flux optique extraite entre les images correspondantes à l'onset et à l'apex. C'est entre ces deux moments que le mouvement est susceptible d'être le plus prononcé.

Nombre de CL	Nombre de paramètres	Accuracy (en %)
17	10 670 932	57,26
15	5 400 725	57,26
13	2 790 149	60,58
11	1 608 965	59,34
9	694 277	60,17
7	398 597	61,00
5	178 309	58,51
3	104 197	60,17
1	91 525	58,92

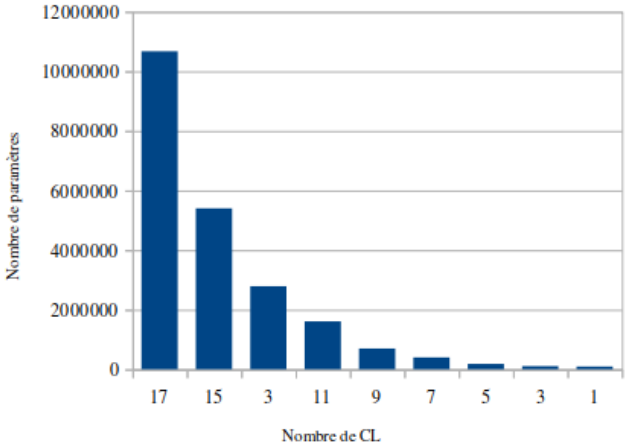


Table 3.2: Les performances varient en fonction du nombre de couches convolutionnelles (CL) et du nombre associé de paramètres apprenables.

La dimensionnalité des entrées détermine la complexité du réseau qui les utilise, puisque la réduction des canaux d'entrée dicte le nombre de filtres à utiliser dans toutes les couches suivantes du CNN. Pour correspondre aux trois canaux couleur dans le réseau pré-entraîné, le flux optique est généralement représenté selon 3 canaux. Lors de la classification des ME, les matrices résultantes V_x, V_y et M sont traditionnellement données comme entrées au CNN. Néanmoins, le troisième canal est intrinsèquement redondant puisque M est calculé à partir de V_x et V_y. En outre, nous supposons que même un champ de mouvement à un seul canal pourrait être suffisamment descriptif. Nous avons donc créé et évalué des réseaux prenant comme entrée une représentation du flot optique à deux canaux (V_x et V_y) et à un seul canal (M, V_x ou V_y).

PROTOCOLE

•Deux types de CNN ont été étudiés, l'un avec une entrée à 1 canal (V_x, V_y ou M) et l'autre utilisant la paire V_x-V_y à 2 canaux.
• Les réseaux proposés commencent par 3 CL liés à l'optimisation de la profondeur, suivis d'une normalisation par lots et un ReLU. Puis les réseaux se terminent par une couche de maxpooling et une couche entièrement connectée.
• Validation croisée LOSO.

En raison du fait que les CNN prennent généralement des entrées à 3 canaux et sont pré-formés en conséquence, l'adaptation de l'apprentissage par transfert à nos modèles aurait été une tâche non triviale. Au lieu de cela, nous avons créé des CNN personnalisés et les avons formés à partir de zéro.
Le tableau 3.3 montre les précisions de reconnaissance de différentes configurations en utilisant un petit nombre de couches CNN. Cela nous amène à penser que les caractéristiques les plus utiles pour la reconnaissance de ME pourraient être présentes dans les mouvements verticaux donnés par V_y. Cette hypothèse est logique en considérant les mouvements musculaires qui se produisent dans chaque expression faciale connue.
D'autre part, l'utilisation de la magnitude seule conduit à une précision similaire à celle de V_y et de la paire V_x-V_y avec un score de 59,34%. V_x a obtenu les plus mauvais résultats dans l'ensemble, avec un score maximal de 54,34%. Cette observation indique que les caractéristiques les plus importantes pour la classification de ME pourraient en effet être plus dominantes dans le mouvement vertical que dans le mouvement horizontal.

	1 CL	2 CL	3 CL	4 CL
V_x	52.24%	54.34%	53.92%	53.50%
V_y	58.09%	59.34%	**60.17%**	**60.17%**
V_x-V_y	58.51%	59.75%	**60.17%**	58.09%
M	58.09%	58.92%	59.34%	59.34%

Table 3.3: Performances obtenues sous différentes architectures CNN et représentations du flux optique.

Pour mieux visualiser la différence entre les caractéristiques de haut niveau présentes dans V_x, V_y et la magnitude M, nous en avons fait une moyenne sur tous les échantillons de la base de données CASMEII [62] en fonction de leur classe. Le résultat, visible sur la Figure 3.1, présente une quantité de bruit non négligeable pour V_x et des régions d'activité claires pour chaque classe avec M et V_y. Les régions d'activité sont alignées avec les muscles responsables de chaque expression faciale.

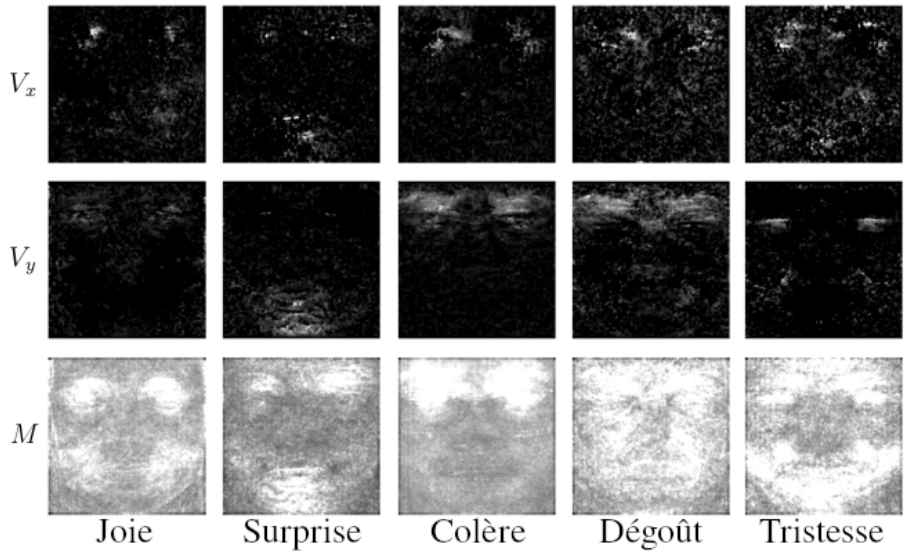


Figure 3.1: Flot optique moyen obtenu par classe de ME.

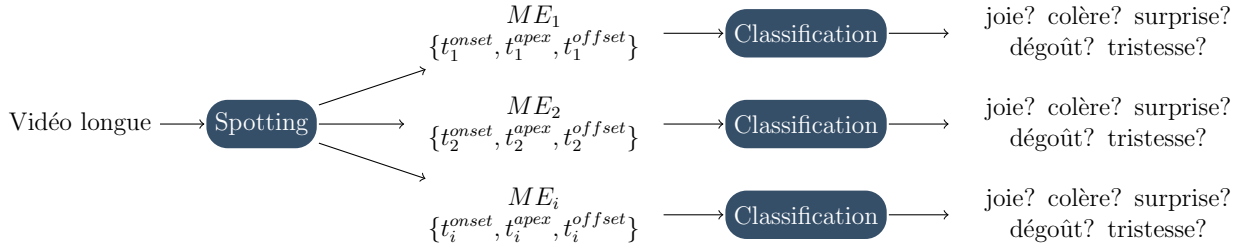
Une analyse détaillée des performances (similarité des caractéristiques extraites, comparaison avec l'état de l'art et évaluation de l'espace mémoire nécessaire) a été effectuée et est disponible dans l'article [ApSc2020].

3.1.2 Spotting

Le spotting est un véritable défi : détecter les mouvements musculaires subtils émanant des ME tout en les distinguant des autres mouvements musculaires tels que les tics nerveux, les clignements des yeux ou autres mouvements rapides. Étant donné leur faible occurrence d'apparition et la quantité réduite de mouvement pouvant apparaître, les ME sont excessivement difficiles à isoler correctement.

Le spotting

Le développement des méthodes basées sur la vision par ordinateur pour l'étude des ME se décompose, selon la communauté scientifique, en deux sous-disciplines : le spotting et la reconnaissance. Dans le spotting, rien n'est connu et il faut estimer où trouver l'onset et soit l'offset soit l'apex. Si ces deux étapes se suivent naturellement dans un processus complet, elles sont généralement étudiées séparément.



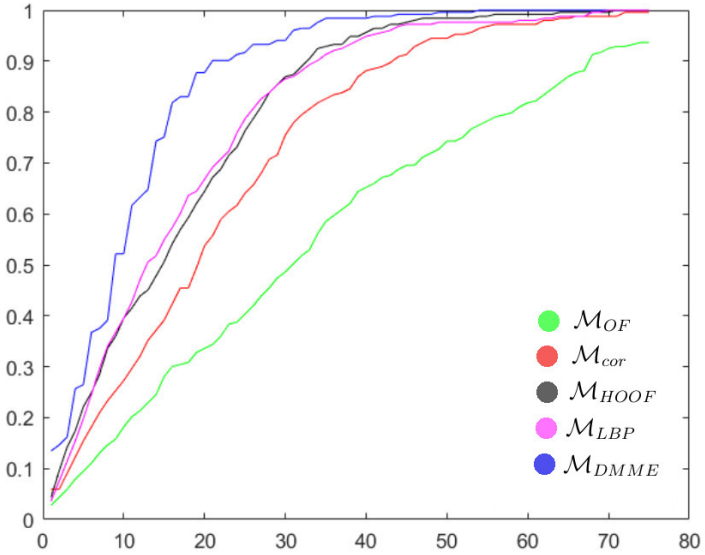
Travailler avec de longues vidéos, sans aucun a-priori sur le nombre de ME ni sur leurs positions, engendre un nombre d'erreurs trop important avec les algorithmes actuels pour pouvoir envisager de les utiliser pour concevoir un système complet apte à être déployé en conditions réelles. Pour notre étude exploratoire, nous avons décidé de contourner ces difficultés en proposant une simplification forte mais cohérente avec l'application recherchée.
De par leur nature même, les ME ont tendance à apparaître plus fréquemment dans des situations stressantes ou en relation directe avec un dialogue, un son ou une image. Les sentiments intrinsèques vis-à-vis d'un contexte sont très difficiles à estimer automatiquement mais ils donnent un a-priori fort sur la présence de ME. Devant une situation ou une ambiance spécifique, un utilisateur peut générer une ME qui traduit son émotion ressentie.
Rappelons que la ME est une représentation ponctuelle de l'émotion. Il ne faut pas la confondre avec un état émotionnel qui s'étend sur une certaine durée. Il est donc possible de supposer qu'un utilisateur puisse estimer quand une ME risque de se manifester.

Dans cette section, nous partons du postulat que nous connaissons le début (l'onset) de chaque ME. Donc, le nombre de ME est aussi connu et ceci réduit drastiquement le nombre de segments d'une séquence vidéo à étudier. Pour chaque ME, il reste maintenant à estimer les moments d'occurrence de l'apex et de l'offset. En pratique, les descripteurs utilisés pour la reconnaissance de ME se basent principalement sur la connaissance de l'apex mais très rarement sur celle de l'offset. Pour cette raison, je me concentrerai sur la seule détection de l'apex par la suite.

Approximation de la position de l'apex Pour estimer la position de l'apex, celle de l'onset étant connue, nous avons testé un certain nombre de méthodes classiques. Il s'agit de descripteurs bas niveau calculés sur chaque frame à partir de l'onset et sur une durée compatible à celle d'une ME.

PROTOCOLE

- 50 frames (soit 250ms) sont balayées à partir de l'onset et engendrent une courbe représentant l'évolution de la déformation du visage lors de la ME.
- La vraisemblance pour chaque frame en tant qu'apex est calculée par quatre critères (voir [56]). Le maximum obtenu correspond à notre estimation de l'apex.
- La tolérance correspond au nombre de frames d'écart accepté (pour être considéré comme une bonne détection) entre l'apex estimé et l'apex réel.



- $\mathcal{M}_{OF}$: distance entre la magnitude cumulée du flot optique pour les deux images.
- $\mathcal{M}_{cor}$: indice de corrélation entre les vecteurs LBP des deux images.
- $\mathcal{M}_{HOOF}$: distance χ^2 sur les vecteurs HOOF des deux images.
- $\mathcal{M}_{LBP}$: distance χ^2 sur les vecteurs LBP des deux images.
- $\mathcal{M}_{DMME}$: durée entre l'onset et l'apex supposée constante.

Figure 3.2: Comparaison entre les taux de bonne détection obtenus avec quatre méthodes basées descripteur puis en supposant une durée fixe (t_{ME}) entre l'onset et l'apex estimé.

La Figure 3.2 affiche les évolutions obtenues pour les quatre critères. Dans ce type de courbes, le point d'inflexion définit généralement le compromis le plus avantageux. Ici nous obtenons un point d'inflexion autour d'une tolérance de 30 frames. Avec notre framerate cela représente 150ms. Or la durée moyenne d'une ME est égale à 170ms sur la base de données CASME II & SAMM. En pratique cette tolérance n'est donc pas acceptable.
Prenons une tolérance de 10 frames qui est plus raisonnable. Le χ^2 du HOOF et du LBP donnent alors un taux de réussite inférieur à 40%. C'est une valeur très basse qui s'explique par la grande variation et l'extrême subtilité du mouvement des ME. D'ailleurs, l'évolution au cours du temps des vraisemblances obtenues avec les quatre méthodes est peu régulière.
À partir de ces constats, nous avons alors essayé d'inclure une donnée statistique en prenant en compte la durée moyenne d'une micro expression t_{ME}. Les mêmes tests que précédemment ont ensuite été effectués en prenant comme apex estimé la position $\widehat{t}_{apex} = t_{onset} + t_{ME}$.
La Figure 3.2 montre que cette nouvelle méthode donne des performances bien supérieures à celles obtenues avec des descripteurs. Pour une tolérance de 10 frames, elle obtient un taux de reconnaissance de 75%. Cela confirme le fait que les descripteurs sont trop imprécis et bien trop sensibles à diverses sources de bruit. De plus cette méthode est bien moins coûteuse en calcul.
Nous avons vu que l'estimation automatique de la position de l'apex est loin d'être précise. Une désignation statistique sans prise en compte des caractéristiques spécifiques à la séquence donne même les meilleurs résultats. Ces résultats semblent très insuffisants. Mais ce n'est pas le spotting en lui-même qui est le cœur de notre étude. Des écarts d'estimation ne sont pas vraiment problématiques en soi s'ils n'entraînent pas une dégradation significative de la classification qui est notre objectif principal. Dans la section suivante, je vais présenter une étude sur le protocole complet. L'évaluation portera alors sur l'association spotting-classification. En attendant la découverte d'une méthode efficace de spotting, il est important d'évaluer son impact sur la performance de reconnaissance.

Pipeline complet pour l'analyse de ME Dans cette partie, nous associons les étapes de spotting et de classification pour constituer un pipeline complet pour l'analyse de ME. Rappelons que la classification prend, en entrée d'un réseau CNN, le flot optique calculé entre

l'onset et l'apex.

PROTOCOLE

• Classification avec la configuration qui procure les meilleurs performances : le CNN qui considère V_y comme entrée avec 3 couches convolutionnelles.
• Spotting avec les cinq méthodes vues précédemment.
• Validation croisée LOSO.

La Figure 3.3 montre les résultats de détection/spotting obtenus à partir des différentes méthodes d'estimation de l'apex. Puisque la méthode $\mathcal{M}_{cor}$ a donné le plus mauvais résultat, elle n'est pas considérée dans la comparaison.

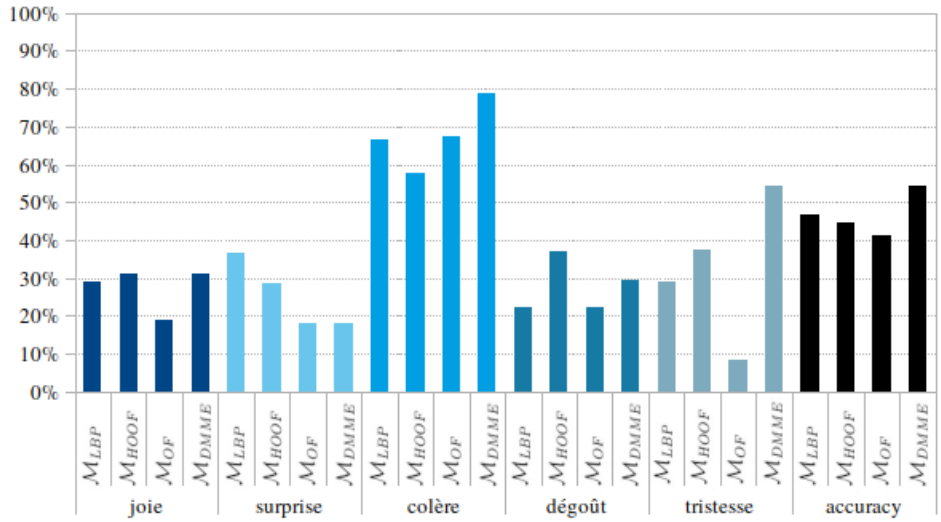


Figure 3.3: Score d'accuracy pour chaque type d'émotion à partir des 4 méthodes de pseudo-spotting présentées.

Les scores de classification confirment les observations du spotting : $\mathcal{M}_{DMME}$ donne les meilleurs résultats et $\mathcal{M}_{FO}$ obtient les moins bons. Cependant nous remarquons cette fois une distinction plus nette entre les $\mathcal{M}_{HOOF}$ et $\mathcal{M}_{LBP}$. La description basée sur la déformation de silhouette semble donc plus fructueuse.
L'écart entre $\mathcal{M}_{FO}$ et les autres méthodes est ici moins prononcé. Cela est probablement dû à la difficulté croissante d'amélioration du critère (une amélioration de 1% est plus significatif vers les hautes valeurs que vers les basses). Malgré cet aspect, l'amélioration apportée par $\mathcal{M}_{DMME}$ est confirmée. Cette donnée statistique (la durée moyenne d'une ME) est donc bien significative.
À titre de comparaison, le score atteint en utilisant un apex détecté manuellement est proche de 60.17%. Le score de 54.36% en utilisant $\mathcal{M}_{DMME}$ est inférieur mais reste proche des performances atteintes par l'être humain en ce qui concerne la détection et la reconnaissance de ME. Bien sûr, il reste des cas où l'apex réel est bien loin de celui estimé par une durée moyenne, mais ces cas particuliers sont relativement rares.

3.1.3 Conclusion

Le problème de la reconnaissance de ME est un problème encore ouvert. La capacité à estimer un ressenti ponctuel par ce biais est avérée. Cependant les taux de reconnaissance obtenus sont encore assez bas. De plus, ces résultats se basent sur la connaissance de l'apparition d'une ME et de sa localisation. Or nous avons vu qu'une localisation automatique (ou semi automatique) dégrade fortement les performances. Il faut aussi prendre en considération que le mauvais équilibre entre le nombre d'échantillons de chaque classe dans les bases de données disponibles (nous avons réalisé une étude pour quantifier l'influence de ce déséquilibre sur nos résultats dans une section du manuscrit de thèse de Reda Belaiche [56]) impacte ces résultats. Néanmoins les perspectives sont importantes et nous avons ciblé par notre étude des voies prometteuses pour des recherches futures.

LIEN RECHERCHE/ÉDUCATION

- Encadrement de la thèse de Reda Belaiche.
- Encadrement de projets étudiant autour de la thématique et sur les données récoltées :
  - Évaluation du spotting par Local Temporal Pattern [63].
  - Étude de la corrélation entre un gabarit (Magnitude de la Figure 3.1) et une image de flot optique par émotion.
  - Suivi de visage sur une caméra rapide.
- Étude sur l'influence du contenu des données de test et d'apprentissage dans l'évaluation des méthodes de classification.

3.2 Des mouvements cycliques : la marche

Je présente ici une étude sur la reconnaissance de défauts dans la démarche d'une personne. Le découpage temporel en périodes de marche, préliminaire à la classification, est assez facile à réaliser et très efficace. La décision, prise dans un but médical, se réalise ensuite sur l'ensemble des périodes pour produire un diagnostic sur la séquence complète.

Pour étudier la démarche d'une personne, nous plaçons une caméra RGB-D (Kinect) en face de la personne, selon l'axe de son déplacement.
Pour évaluer nos algorithmes, nous avons réalisé un jeu de données. Dans ce jeu de données, un certain nombre de personnes ont été filmées de face marchant selon trois façons différentes :

1. La personne marche normalement à la vitesse qui lui est la plus confortable.
2. Une semelle de rembourrage (voir Figure 3.4) a été rajoutée dans la chaussure droite de la personne. Ce procédé simule une personne qui boite et a aussi été utilisé dans [64].
3. Il a été demandé à la personne de ne pas plier le genou droit pendant la marche. Ce test simule les effets d'une prothèse sur la marche lors de la rééducation après une fracture [65].

Ces données sont disponibles en ligne : https://github.com/margokhokhlova/ et correspondent à la base que nous nommons MMGS.

Étapes de la marche

La marche est cyclique. Sur un cycle, 8 étapes primordiales (voir section 2.3.1) se succèdent : le contact initial (a), la réponse à la charge (b), la position intermédiaire (c), la position terminale (d), l'avant l'élan (e), l'élan initial (f), l'élan intermédiaire (g) et l'élan terminal (h).

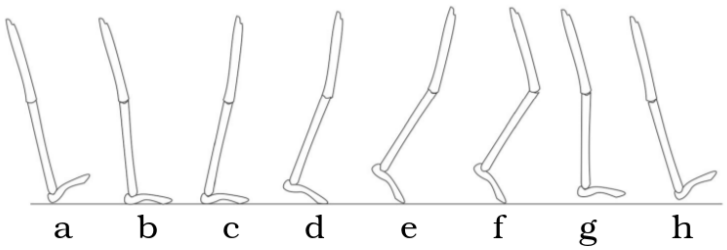


Figure 3.4: Semelle de rembourrage de 7cm de haut permettant de simuler une personne qui boite.

3.2.1 Fiabilité de l'estimation des articulations

Nous évaluons la fiabilité des angles de flexion du genou et de la hanche calculés à partir des données Kinect v.2 en les comparant à ceux obtenus par un dispositif Vicon. Les acquisitions des deux capteurs ont été segmentées manuellement en cycles.

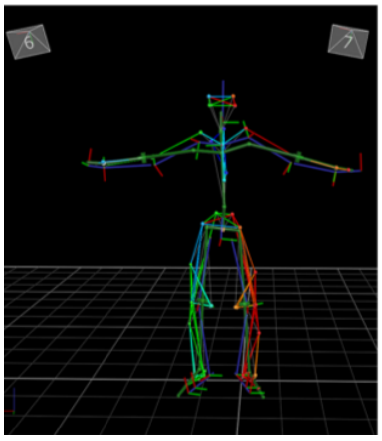
PROTOCOLE

- Segmentation des cycles en identifiant les pics de distance articulaires intra-pieds comme proposé dans [66].
- Calcul des coefficients de corrélation pour estimer la relation entre les données cinématiques Kinect et Vicon.
- Calcul du coefficient d'indice de corrélation pour la cohérence pour un seul évaluateur ICC(C,1), également connu sous le nom de fiabilité référencée par une norme.
- Calcul sur 32 cycles de marche acquis simultanément par les dispositifs Kinect et Vicon.

Cette étude démontre que l'évolution des angles de flexion de la hanche et du genou pendant la marche peut être obtenue de façon fiable à partir de l'estimation du squelette par la Kinect. Nous obtenons une fiabilité modérée pour la flexion du genou et une bonne fiabilité pour la flexion de la hanche avec de faibles intervalles de confiance. Les angles d'abduction ont été jugés non corrélés à l'exception des données de la hanche droite. Les résultats complets sont disponibles dans [67].

Vicon pour la vérité terrain

Les données de la marche ont été acquises simultanément avec une Kinect v.2 et le capteur Vicon en milieu clinique. Les articulations du squelette Kinect, les orientations et les états des articulations sont enregistrés. Les coordonnées des marqueurs et les angles de flexion du module Plug-in Gait Vicon sont enregistrés pour le système MOCAP basé sur les marqueurs. L'acquisition a été réalisée dans une plateforme du Centre Médical Universitaire (CHU) de Dijon, où la caméra Vicon est utilisée pour effectuer l'analyse du mouvement. Un seul chercheur expérimenté a placé des marqueurs rétro-réfléchissants sur chaque participant.



Squelette obtenu par le Vicon

Tous les marqueurs ont été placés directement sur la peau pour une plus grande précision. Le processus d'acquisition a été lancé simultanément par la Kinect et la Vicon. Les trajectoires des marqueurs de Vicon ont été enregistrées avec une fréquence d'échantillonnage de 120 Hz. La fréquence des données Kinect est d'environ 30 Hz. Nous avons ensuite sous-échantillonné les données Vicon pour avoir la même fréquence qu'avec la Kinect.

3.2.2 Covariance

Les angles de flexion sont des caractéristiques connues et largement utilisées pour l'analyse de la marche dans les cliniques et les hôpitaux. Nous combinons la compacité et la représentativité des matrices de covariance avec les caractéristiques très pertinentes des angles de flexion pour proposer une nouvelle méthode d'évaluation de la marche. Une matrice de covariance résume les relations entre les angles de flexion de la hanche et du genou pour une séquence de marche. Ainsi, nous cherchons à décrire la symétrie de la marche en fonction des angles de flexion des membres inférieurs. L'avantage des matrices de covariance est qu'elles fournissent une représentation des caractéristiques indépendante de la période du cycle.

Les équations suivantes caractérisent les matrices variance-covariance et de covariance avec P_a l'évolution temporelle des caractéristiques suivies sur une fenêtre de taille T frames.

$$cov = \frac{1}{T-1}(P_a - \mu)(P_a - \mu)^T \tag{3.1}$$

$$cov = \frac{1}{T-1} P_a \left(\frac{1}{T} I_T - 1_T \right) P_a^T \tag{3.2}$$

avec μ la moyenne de P_a et I_T (respectivement 1_T) la matrice identité (respectivement unité) de taille $T \times T$.

Dans notre étude nous définissons P_a de la façon suivante où $\theta_{a,c,t}$ est l'angle de flexion de l'articulation a (genou (g) ou hanche (h)) du côté s (droit (d) ou gauche (g)) à l'instant t.

$$P_a = \begin{bmatrix} \theta_{g,g,1} & \dots & \theta_{g,g,T} \\ \theta_{g,d,1} & \dots & \theta_{g,d,T} \\ \theta_{h,g,1} & \dots & \theta_{h,g,T} \\ \theta_{h,d,1} & \dots & \theta_{h,d,T} \end{bmatrix} \tag{3.3}$$

Nous associons ces données à l'équation 3.1 pour obtenir une matrice 4×4. Ensuite, la classification est réalisée par un classifieur k-NN. La distance de corrélation est utilisée et le

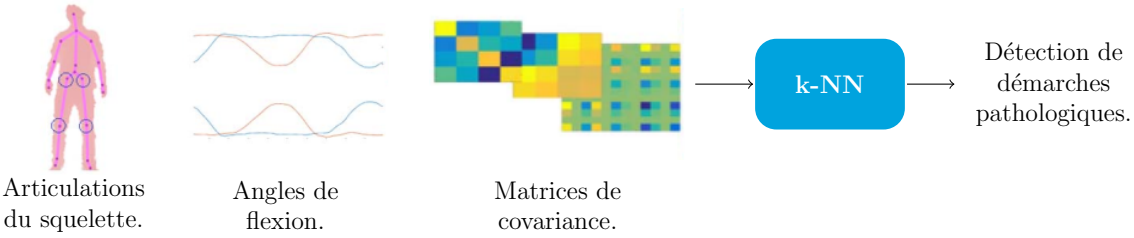


Figure 3.5: Procédé proposé pour la détection de démarches pathologiques à partir de matrices de covariance.

modèle de consensus est employé comme paramètre de l'algorithme k-NN. Nous avons également testé la distance entre matrices basée sur les valeurs propres comme [68] mais elle a été surpassée par la distance euclidienne et la distance de corrélation. Le procédé complet est illustré par la Figure 3.5.

PROTOCOLE

- Sur la base de données [64] avec annotation binaire : démarche normale ou pathologique.
- Taille de la fenêtre glissante $T = 90$ (4 cycles de marches).
- Répartition des données pathologiques réalisées 20 fois et résultats moyens gardés pour palier à la faible quantité de données.
- Comparaison avec une matrice standard de positions relatives des articulations [69] associée à l'équation 3.1 (matrice de covariance 24×24) puis à l'équation 3.2 (matrice de covariance 75×75).

Les résultats obtenus sont représentés dans le tableau 3.4. Les angles de flexion produisent des performances bien plus élevées que la position des articulations. La précision est proche de celle de l'état de l'art tout en proposant une approche bien plus légère.

Méthode	[66]	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
F-measure	0,87	0,86	0,82	0,69
précision	0,96	0,76	0,95	0,55
rappel	0,80	0,99	0,73	0,90
accuracy	0,83	0,84	0,79	0,13

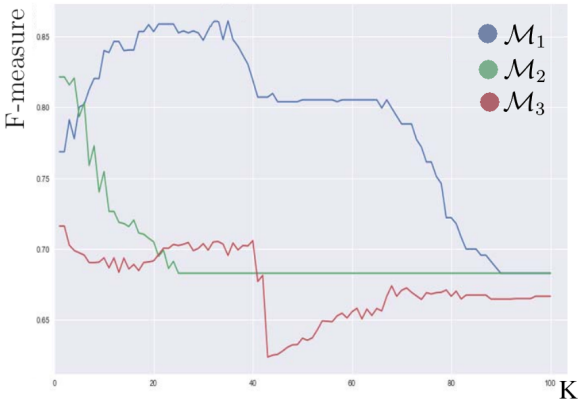


Table 3.4: Performances obtenues par notre méthode associant les angles de flexion et l'équation 3.1 ($\mathcal{M}_1$) et par l'utilisation des positions des articulations associée à l'équation 3.1 ($\mathcal{M}_2$) puis l'équation 3.2 ($\mathcal{M}_3$). Le tableau de gauche récapitule les valeurs obtenues pour la meilleure configuration.

Pour plus de détails sur cette partie, se référer à l'article [SITIS2018]. Le travail correspondant intègre également deux approches complémentaires :

- La modélisation de la démarche par une loi normale permettant de ne pas faire intervenir les données pathologiques (les plus difficile à obtenir) dans l'apprentissage.
- L'analyse de données croisées attestant que l'approche et les résultats obtenus ne sont pas limités à l'utilisation d'une certaine base de données.

3.2.3 LSTM

Les angles de flexion des membres inférieurs sont largement utilisés par les cliniciens pour évaluer la démarche d'une personne. Nous décrivons dans cette section une méthode d'apprentissage profond qui classe automatiquement la marche en fonction de la dynamique des membres inférieurs obtenue. Les **LSTM** sont un type particulier de réseaux de neurones récurrents, capables d'apprendre des dépendances à long terme. À l'instar des autres techniques basées sur le Machine Learning, un réseau LSTM doit être formé sur une quantité importante de données. L'architecture optimale et le choix des hyper-paramètres du réseau dépendent du problème spécifique.

Sur la base de certains tests initiaux, nous avons adopté un modèle LSTM bidirectionnel à 2 couches pour la tâche de classification de la marche. Nos tests ont validé que l'utilisation des états de cycle de marche passés et futurs est bénéfique pour le modèle de classification de la marche. L'architecture exacte du modèle utilisé dans notre travail est illustrée sur la Figure 3.6.

PROTOCOLE

- Sur le jeu de données MMGS que nous avons créé.
- Trois classes dont deux pathologies différentes.
- Optimiseur Adam.
- Comparaison entre trois vecteurs de données en entrée : les angles de flexion, les positions des articulations et les matrices cinématiques de covariance.
- Les paramétrages du réseau sont définis et justifiés dans l'article [AIM2019].

Nous observons un biais élevé entre les résultats d'entraînement et de validation, donc plus de données d'entraînement seront bénéfiques pour le modèle. De même nous notons une déviation de 3% sur nos performances selon la distribution des échantillons en apprentissage et validation. Les données sont insuffisantes car très complexes à obtenir, mais les résultats obtenus sont encourageants. L'utilisation en entrée des angles de flexion produit un score d'accuracy moyen de 82% contre 75% avec les matrices de covariance cinématique et 48% avec les positions des articulations du squelette.

Dans l'ensemble, les performances du système proposé sont comparables à celles rapportées par d'autres études dans le domaine [69, 66, 51]. De plus, nous utilisons plus de données que [69, 51] et prenons en compte trois classes et non une détection binaire comme [66]. Les angles de flexion de la hanche fonctionnent extrêmement bien pour reconnaître le problème de rigidité du genou. La pathologie simulée par la semelle rembourrée est plus difficile à différencier de la marche normale.

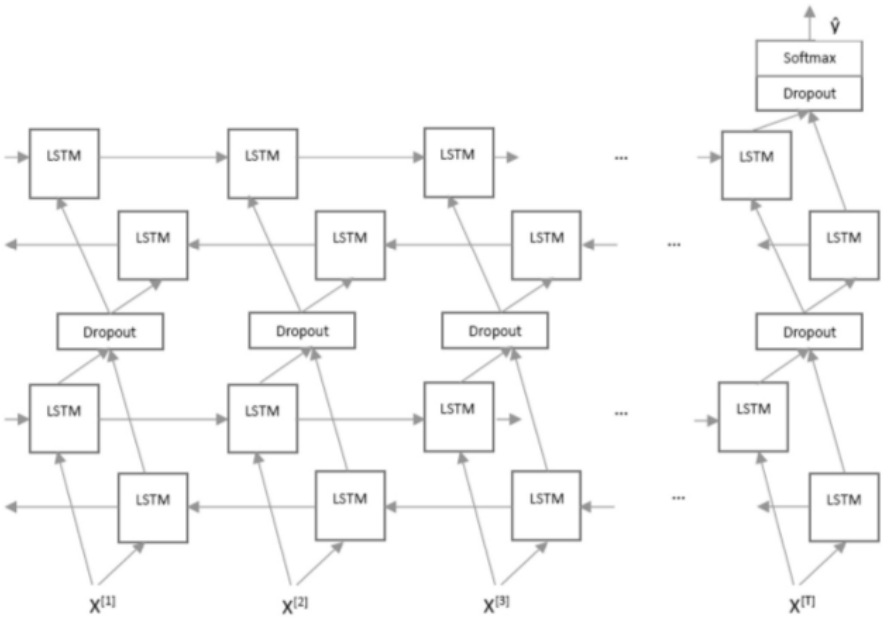


Figure 3.6: Structure LSTM bidirectionnelle adoptée : l'entrée correspond aux angles de flexion, la sortie correspond à la classe attribuée.

3.2.4 Conclusion

La variation de la démarche est très sensible. Nous avons démontré la capacité de nos méthodes à dissocier une démarche problématique (simulée pour correspondre à des effets reconnus des phases de rééducation). Néanmoins la possibilité de classer une démarche problématique en fonction du type de défaut semble inaccessible, en particulier en raison du peu de données disponibles et de la difficulté à les recueillir sans les simuler. La possibilité de produire une aide au diagnostic par une voie non intrusive (seulement par une capture vidéo) représente cependant une avancée cruciale.

LIEN RECHERCHE/ÉDUCATION

- Encadrement de la thèse de Margarita Khokhlova.
- Encadrements de projets étudiant autour de la thématique et sur les données récoltées :
  - Suivi du flot optique 3D à partir d'une Kinect.
  - Descripteur du mouvement humain pour l'étude de la marche par HMM.
  - Nettoyage de nuages de points 3D à partir de méthodes de débruitage 2D et évaluation selon des critères 3D.
- Cours de M2 sur le traitement numérique des vidéos (Suivi, HMM, LSTM ...).

Chapitre 4

Interprétation de scènes 3D

Dans l'interaction des personnes avec leur environnement, soit la personne est le sujet de l'étude, soit l'étude prend son point de vue. La prise de vue dans ce dernier cas est réalisée au niveau de la personne qui se déplace dans un environnement 3D. Il ne s'agit alors plus d'analyser un mouvement absolu de la personne (caméra fixe) mais un mouvement relatif par rapport à son déplacement.
La motivation est d'interpréter la scène pour fournir des descriptions et des indices pertinents à l'utilisateur équipé. La 3D est incontournable pour ne pas limiter l'interprétation et se rapprocher du ressenti réel d'interaction à un environnement.
La reconnaissance et la localisation de personnes sont des tâches essentielles de l'interprétation d'un environnement. Même en prenant le point de vue d'une personne, la classe des personnes reste le centre de l'attention dans un souci de communication et d'interaction sociale.

Dans ce chapitre, je traite cette problématique selon deux approches caractérisées dans la Table 4.1 :

Interaction avec une personne	Pour favoriser les interactions sociales, les personnes présentes dans la scène sont détectées et localisées. L'utilisateur doit être conscient de leur présence ainsi que de leur interaction avec l'environnement. La perception de la scène est donc enrichie par une informatique sémantique riche en applications pratiques.
Aide au déplacement	Le point de vue de l'acquisition est celui d'une personne. La personne se déplace et la vision de la scène varie donc constamment. L'objectif est de fournir les informations nécessaires pour que la personne se déplace sans danger. Il s'agit donc de détecter, d'évaluer et de localiser les dangers ainsi que les chemins possibles. Le changement de point de vue nécessite alors de séparer les variations provenant des mouvements de la caméra de celles relevant d'un mouvement dans la scène.

4.1 Interprétation de la présence d'une personne dans une scène 3D

La classe des personnes est l'une des plus étudiées du fait des nombreuses applications qui lui sont liées. Nous créons des machines pour nous aider et non pour qu'elles aient leurs objectifs propres. C'est d'autant plus vrai dans le cas de l'étude de l'interprétation de scènes. Que ce

Détection de personne	Aide au déplacement
Détecter une personne ou reconnaître une pose.	Représentation sonore d'une scène.
Une ou plusieurs caméras RGB-D.	Une caméra RGB-D.
Léger et robuste.	Temps-réel.
[MTAP2019] [ICASSP2020] [VISAPP2018]	[SITIS2022] [Frontiers2023]

Table 4.1: Caractérisation des deux problématiques traitées concernant l'interprétation centrée personne de scènes 3D.

soit pour un robot ou pour une personne, les interactions avec les humains sont essentielles : ce sont les plus présentes et celles contre qui une erreur est le plus préjudiciable.

4.1.1 Présence d'une personne

Avant de reconnaître un comportement, il convient préalablement de séparer les personnes des autres éléments de la scène. Les Histogrammes de Gradients Orientés (HOG) sont la référence pour les images en tant que descripteur de forme pour la reconnaissance des personnes. Nous avons introduit un descripteur transposant ce procédé sur les nuages de points complet (CPC - voir section 1.1). Ce descripteur extrait les statistiques d'orientation des normales sur une subdivision cylindrique de l'espace.
Sur un CPC les gradients n'ont pas de signification. C'est pourquoi nous les remplaçons par les vecteurs normaux à la surface estimés à partir d'un ajustement du plan par la méthode des moindres carrés. Le descripteur est calculé sur une portion du nuage de points en plusieurs étapes. Tout d'abord, le nuage de points est divisé en blocs à l'aide d'une subdivision en coordonnées cylindriques. Puis, dans chaque bloc, nous calculons un histogramme de l'orientation des normales. Enfin la concaténation des histogrammes de tous les blocs forme le descripteur final.

La subdivision de l'espace Le descripteur est calculé sur une sous-partie du nuage de points qui doit être classée comme contenant une personne ou pas. Ce sous-nuage de points est positionné à l'intérieur d'un volume cylindrique de dimension fixe, et toute partie située à l'extérieur du cylindre est éliminée. Les dimensions sont choisies en fonction de la taille moyenne d'une personne. Le nuage de points est ajusté à l'intérieur du cylindre en faisant passer l'axe principal du cylindre par le centre de gravité du nuage de points. L'espace 3D à l'intérieur du cylindre est divisé en sous-zones (blocs). Nous utilisons une subdivision cylindrique similaire à celle proposée par [70] pour ses travaux sur la reconnaissance de pose à partir de la reconstruction de voxels. Le bloc résultant est un secteur comme représenté sur la Figure 4.1. La coupe axiale à travers l'axe vertical rend le descripteur invariant à l'échelle. Mais le descripteur n'est pas invariant en rotation. C'est ce qui nous permet de déterminer l'orientation frontale du corps.

La quantification des normales 3D Chaque bloc résultant du processus de subdivision contient un certain nombre de points. Les normales sur ces points encodent les informations de courbure de cette portion de surface. Pour capturer ces informations, les normales doivent être transformées en mesures mathématiques qui forment la pierre angulaire du descripteur final. Cela peut être fait en construisant un histogramme 1D des orientations des normales. Mais comme chaque normale est un vecteur 3D, elle ne peut pas être associée directement à un histogramme 1D. Pour résoudre ce problème, nous nous sommes inspirés de la quantification

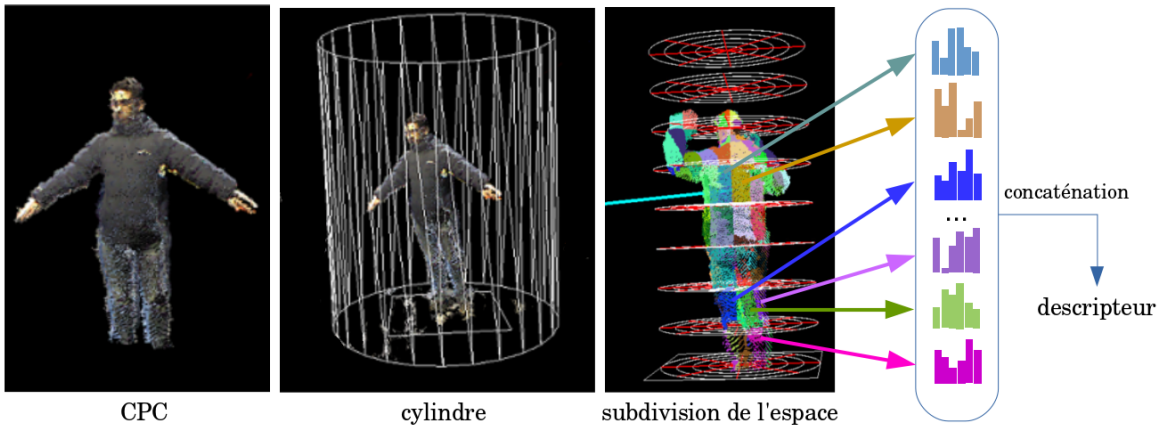


Figure 4.1: Réalisation du descripteur 3D : un cylindre centré sur le CPC est spatialement subdivisé en blocs, un histogramme d'orientation des normales est calculé pour chaque bloc, le descripteur final est la concaténation de ces histogrammes.

d'orientation générique proposée par Klaser [71] sur des vecteurs 2D. Dans ce cas, les vecteurs 2D produisent un histogramme n-bin à l'aide d'un polygone à n côtés. Chaque vecteur est positionné au centre du polygone. Le vecteur est ensuite affecté dans l'histogramme au côté vers lequel il pointe. La même idée est appliquée au vecteur normal 3D mais, dans ce cas, elle est réalisée au moyen d'un polyèdre régulier à n côtés parmi les cinq existants (Figure 4.2).

Figure 4.2: Les cinq polyèdres réguliers.

Chaque face représente un bin (une orientation). La normale associée à chaque pixel est projetée sur chaque face du polyèdre. La longueur de la projection (obtenue par produit scalaire et mise à 0 si négative) représente l'influence de ce pixel sur ce bin. Les longueurs ainsi calculées pour tous les pixels du bloc sont ajoutées afin d'obtenir un histogramme. Chaque histogramme est normalisé. Le descripteur final est la concaténation des histogrammes obtenus pour tous les blocs.

Estimation de l'orientation frontale Le descripteur proposé dépend de la rotation du nuage de points. Nous avons exploité ce critère pour estimer la direction frontale de la personne (Figure 4.3). Un bloc créé par le processus de subdivision du cylindre est associé à trois coordonnées qui correspondent aux trois coupes de subdivision appliquées (radial, azimut et axial). Lorsque le nuage de points est placé à l'intérieur du cylindre, puis pivote autour de l'axe principal, cela change la coordonnée d'azimut du bloc. Le score de classification est plus élevé pour une coupe azimutale qui commence en face de la personne. En testant un nuage de points pour chaque orientation, il est alors possible de déterminer l'orientation de la personne.

Évaluation La classification d'un CPC en une personne est réalisée par un classifieur SVM entraîné sur une base de données de 1000 exemples positifs (contenant une personne) et 1000

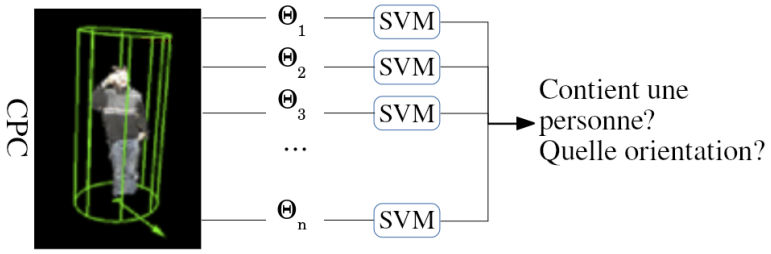


Figure 4.3: Un descripteur est calculé pour chaque orientation avant de passer par un classifieur SVM afin de donner une décision.

exemples négatifs (objets présents en intérieur de formes proches de celle d'une personne). L'orientation est correctement estimée sur 70% des cas. La classification obtenue est comparée avec celles obtenues avec l'algorithme des HOG sur une vue simple (SPC) et avec le meilleur résultat obtenu à partir de chacun des points de vue (Max-SPC). La Figure 4.4 montre que le CPC améliore significativement la classification.

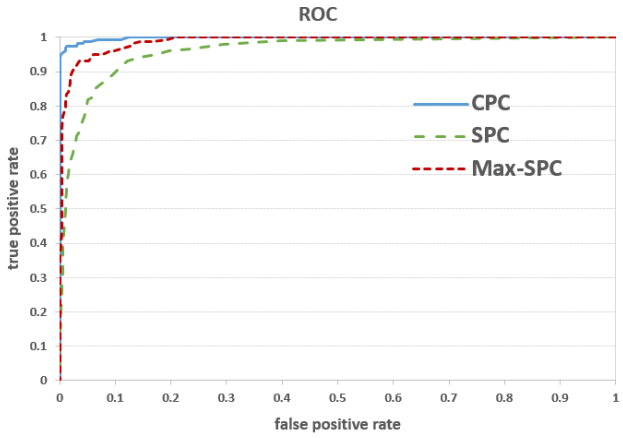
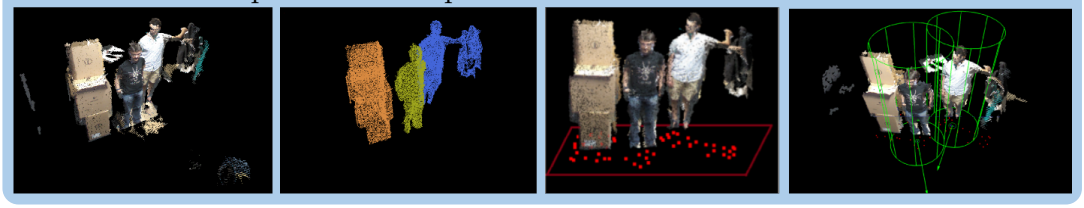


Figure 4.4: Courbes ROC obtenues par notre méthode avec un CPC ou avec la méthode [72] à partir d'un seul point de vue.

Pour plus de détails sur cette partie, se référer à l'article [MTAP2019]. Le travail correspondant intègre également une évaluation de la détection de personnes dans une scène 3D. Le cylindre balaye le nuage de points pour tester les candidats à représenter une personne.



4.1.2 Estimation de pose

Dans l'optique de reconnaître un comportement ou une action à partir d'une séquence vidéo (voir section 3.2), nous avons estimé la capacité d'un descripteur semblable à celui de la section précédente à reconnaître une posture. Le nuage de points contenant une personne est découpé de la même façon en sections, cercles et secteurs (12, 10 et 10 dans la meilleure configuration)

pour former des blocs. La personne est vue de face ce qui limite la variation de forme. Nous ne prenons donc plus en compte dans chaque bloc un histogramme d'orientation mais juste la proportion de points présents dans l'espace considéré (Figure 4.5). Tous les détails se trouvent dans l'article [VISAPP2018].

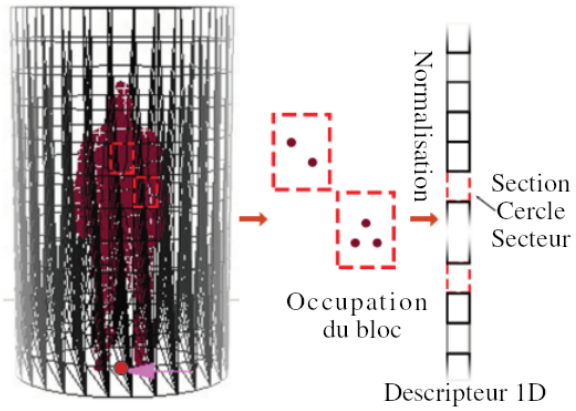
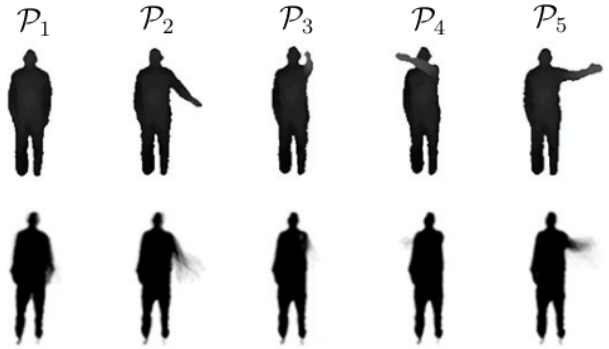


Figure 4.5: La structure du nuage de points est répartie à partir d'une grille cylindrique puis les taux d'occupation de chaque bloc sont concaténés pour former le vecteur de caractéristiques.

PROTOCOLE

- Classification par SVM One vs all.
- Triple validation croisée.
- À partir de la base de données MSR Action Dataset [17].
- Sur 5 poses clés de l'action "wave" (salut de la main) puis sur 18 poses extraites de l'ensemble des actions.

Nous nous sommes tout d'abord concentrés sur l'action "wave". Nous avons manuellement défini 5 poses clés confirmées par une classification par K-means (Table 4.2 à gauche). Nous obtenons alors, dans la configuration optimale, les résultats de la Table 4.2.



Posture	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_5$
Précision	0,94	0,81	1	1	1
Rappel	1	0,8	0,76	1	0,71
F-measure	0,97	0,82	0,86	1	0,83

Table 4.2: Cinq postures de l'action "wave" (en haut) et les clusters obtenus automatiquement par K-means (en bas) ainsi que les performances de classification sur ces 5 postures à partir de notre méthode.

Puis nous avons généralisé l'opération sur l'ensemble des actions à partir de 18 poses clés. Nous obtenons la matrice de confusion de la Figure 4.6 et une précision moyenne de 94%.

Les performances sont élevées et confirment que cette approche peut être intégrée efficacement à un processus de catégorisation d'action à partir d'une séquence vidéo.

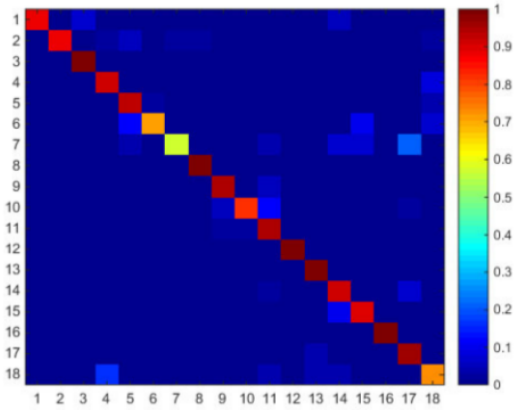


Figure 4.6: Cette matrice de confusion est obtenue en associant notre descripteur à un classifieur SVM sur les 18 postures clés de MSR Action Dataset [17]. Les postures sont toutes bien reconnues sauf la 7 (deux bras tendus à l'horizontale) parfois confondue avec la 17 (bras droit à l'horizontale).

4.1.3 Localisation d'une personne par un son

Dans cette section, nous proposons un système d'interface auditive homme-machine qui permet la localisation 3D d'une personne à partir d'une caméra RGB-D. Nous nous concentrons spécifiquement sur la localisation de personnes car c'est l'une des situations les plus fréquemment rencontrées lors de la marche urbaine et qu'elle est nécessaire pour de multiples applications de sociabilisation. Contrairement aux approches de l'état de l'art [73, 74], le traitement global est effectué en temps réel sur des plateformes de calcul standard ainsi que sur des unités de traitement dédiées à la conception de systèmes embarqués. Le système est mobile et compact avec une faible consommation d'énergie, et donc une plus grande autonomie.

Compte tenu des problèmes liés à une transmission à distance, nous intégrons tous les traitements vidéo et sonores dans le système embarqué.

Sonification d'une personne Notre méthode pour localiser une personne dans un environnement 3D et générer le son stéréophonique associé se décompose en plusieurs étapes :

① Tout d'abord, un réseau de détection CNN estime les positions des personnes à partir de la scène 2D. Nous avons choisi un modèle CNN basé sur l'architecture You Only Look Once (YOLO) car il présente le meilleur framerate [75] pour un réseau de détection qui produit une détection robuste à plusieurs échelles. Une boîte englobante autour de la partie visible de la personne détectée est disponible en sortie (Figure 4.7-a).

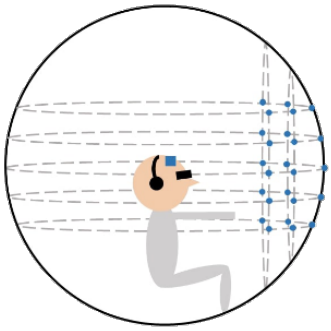
② Ensuite, la distance par rapport à l'utilisateur est estimée en utilisant la carte de profondeur (Figure 4.7-b).

③ Enfin, la position du centroïde est extraite (Figure 4.7-c) et sonifiée avec un signal stéréophonique généré en fonction de la localisation 3D des personnes. Sur la base de la spatialisation du son, l'utilisateur est alors en mesure de localiser la personne ciblée dans l'espace 3D.

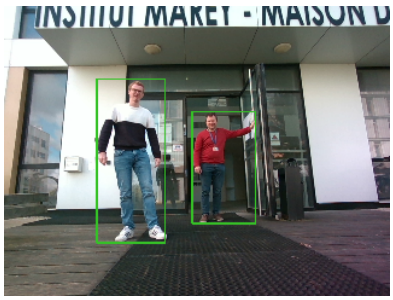
Implémentation Nous cherchons à minimiser le délai entre l'acquisition d'une image et la transmission du son à l'utilisateur. L'architecture de sonification LibreAudioView a été

Système de sonification

La méthode de sonification que nous utilisons est basée sur le système LibreAudioView [76] dans lequel chaque élément de l'image génère un signal audio 3D (dont une personne peut localiser l'origine par un point 3D de l'espace). Les coordonnées du pixel sont transposées en coordonnées sphériques sur une sphère de 2 mètres de rayon centrée sur la caméra. Nous avons ensuite utilisé l'ensemble des données HRIR de deux mètres enregistrées dans une chambre anéchoïque [77] pour spatialiser un son monophonique bref (33ms) de 440 Hz avec un fondu en cosinus de 5ms.

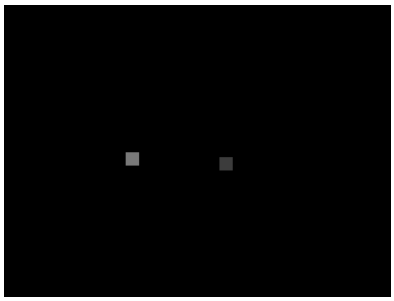


Une HRIR est une réponse impulsionnelle qui imite la déformation naturelle faite par le corps de l'auditeur (principalement la tête) sur le son pour en connaître l'origine. L'amplitude du son est modulée en fonction de la distance qui sépare la cible de l'utilisateur en utilisant une loi de carré inverse $A = 1/d^2$ où A est l'amplitude et d la distance entre l'utilisateur et la cible.



(a) Détection

(b) Carte de profondeur



(c) Pixels à sonifier

Figure 4.7: Vue d'ensemble de la méthode : d'abord un CNN estime les positions des personnes à partir de l'image couleur (a), ensuite la distance à l'utilisateur est extraite de la carte de profondeur (b) pour générer le pixel à sonifier (c).

optimisée dans un précédent article [78] : l'optimisation du logiciel de sonification a permis de réduire le temps de traitement requis de 86% par rapport à la version originale [76]. Après l'optimisation de l'étape de génération du son, la partie traitement d'image représente 95% du temps de traitement global sur une plateforme PC standard (CPU). Par conséquent, nous proposons des implémentations basées sur les GPU afin de réduire la vitesse de traitement. En effet, l'architecture spécifique multi-cœurs des GPU est particulièrement adaptée aux tâches régulières et au parallélisme intrinsèque de l'algorithme choisi. Enfin, nous proposons une seconde implémentation basée sur une cible GPU qui est dédiée aux conceptions de systèmes embarqués. L'objectif est de démontrer qu'une solution à faible latence peut être conçue autour d'une telle cible afin de proposer un système compact et embarqué. De plus, la consommation électrique a été ajustée pour augmenter l'autonomie énergétique du système tout en respectant les contraintes de performance de l'application. Différentes optimisations sont ensuite proposées, comme l'adaptation de la dynamique des données, pour diminuer la latence du système.

PROTOCOLE

- CNN Yolov5-small entraîné sur la base de données Microsoft Coco [79].
- Entraînement sur des images de résolution 640 × 640.
- Acquisition vidéo réalisée à l'aide d'une caméra stéréoscopique Intel RealSense D455 à 30 images par seconde.
- **Implémentation CPU** : basée sur un processeur Intel Core i7-6700HQ (4 cœurs - 8 threads, 2,60 GHz ; 16 Go de RAM).
- **Implémentation GPU** : basée sur un GPU Nvidia GTX 1070 (2048 cœurs CUDA, 6,738 Tflops, 8 Go de VRAM).

Les deux premières colonnes du Tableau 4.3 représentent la comparaison entre CPU et GPU ainsi que les performances du YOLOv5-small et son impact sur le fonctionnement global du dispositif de sonification. Le temps d'inférence du réseau YOLOv5-small sur un processeur d'ordinateur standard ne permet pas, au contraire d'une cible GPU, d'atteindre des performances en temps réel. De plus, le temps d'inférence du CNN sur GPU a été réduit en convertissant le modèle avec le SDK TensorRT (troisième colonne).

	CPU	GPU	
		LibTorch	TensorRT
Yolov5-small (ms)	135	16.3	7.5
Processus global (ms)	142	23.6	15.3

Table 4.3: Temps de traitement du réseau YOLOv5-small sur trois cibles différentes.

Compte tenu des problèmes liés à une transmission à distance, nous avons privilégié le développement d'un dispositif autonome. Une implémentation à base de GPU représente une solution appropriée pour accélérer le traitement et donc une solution pertinente pour développer in-fine un dispositif portable. Nous avons la possibilité de varier la plage dynamique des poids du CNN (16 bits et 18 bits) ainsi que de fonctionner dans deux modes de consommation (Max-Q de 7,5W (5,5V) et Max-P de 15W). Plus de détails sur les caractéristiques de la cible se trouvent dans [ICASSP2022].

Le Tableau 4.4 résume l'impact de ces variations sur le fonctionnement du CNN sur la cible embarquée. Une inférence du réseau YOLOv5-small est inférieure au framerate de la caméra. Les deux configurations fournissent une précision de 55,4% pour un recouvrement de 50%. La durée de fonctionnement du système à pleine puissance est estimée à 6 heures 55 minutes tandis que l'autonomie de la batterie à faible puissance est estimée à 12 heures et 11 minutes.

L'utilisation de ce mode basse puissance génère une latence plus importante mais permet d'obtenir des performances compatibles avec les contraintes de l'application. Ce choix de mode est donc pertinent compte tenu du gain significatif en termes de durée d'utilisation.

	FP32 15W	FP32 7.5W	FP16 15W	FP16 7.5W
Yolov5-small (ms)	55	71	35	47
Processus global (ms)	62	80	43	56

Table 4.4: Temps de traitement du réseau YOLOv5-small sur une cible embarquée selon plusieurs configurations.

Validation Nous avons mesuré les capacités offertes par un tel dispositif de substitution sensorielle auditive dans une tâche consistant à localiser des personnes se trouvant à proximité immédiate.

PROTOCOLE

- Participants assis sur une chaise tournante à 360° équipés de notre système.
- Personne cible située sur 8 positions comme présenté dans la Figure 4.8 (à gauche).
- Positions de la tête du participant et du torse de la cible estimées par un système HTC Vive.
- Le participant doit effectuer une rotation sur la chaise afin de trouver la cible et la placer devant lui en se basant uniquement sur les indications auditives fournies par le système de substitution (plus de détails dans [ICASSP2022]).
- Des tests où le participant peut regarder la scène sont ajoutés pour réaliser des comparaisons.

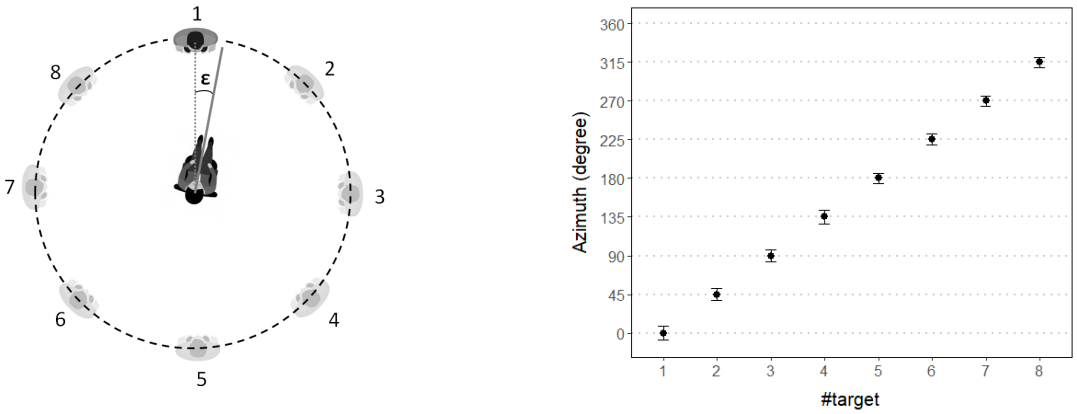


Figure 4.8: Expérience réalisée pour valider les capacités de notre système. Le participant est assis au centre de la zone d'expérimentation sur une chaise tournante. Une personne debout change aléatoirement sa position parmi 8 également espacées sur un cercle de deux mètres de rayon. Nous avons mesuré l'azimut réel de chaque cible (point noir à droite) avec la valeur moyenne associée de l'erreur angulaire de l'azimut estimé (barre verticale).

Les résultats présentés dans la Figure 4.8 (à droite) montrent l'erreur angulaire azimutale moyenne pour chaque position cible. L'erreur moyenne obtenue est de $6,72° \pm 5{,}82$. L'erreur moyenne obtenue lorsque la personne peut regarder est environ 2 fois plus petite ($2,85° \pm 1{,}99$). Malgré cette différence, ces résultats montrent que les participants peuvent localiser une personne avec une grande précision en utilisant notre dispositif de substitution sensorielle.

4.1.4 Conclusion

La détection des personnes dans une scène est un problème où beaucoup de solutions ont déjà été proposées. Nous avons apporté en robustesse vis-à-vis du point de vue de l'acquisition en proposant un descripteur basé sur des nuages de points complets, puis étudié la complexité et la consommation de la détection pour une application embarquée. La reconnaissance de pose est un problème un peu plus complexe qui dépend beaucoup des classes envisagées. Nous avons proposé une étude adaptée au suivi des poses d'une action spécifique.

LIEN RECHERCHE/ÉDUCATION

- Encadrement de la thèse de Kyis Essmaeel et Florian Scalvini.
- Encadrement d'un projet étudiant autour de la thématique et sur les données récoltées :
  - Suivi 3D de personnes à partir d'un nuage de points.
- Cours de M1 sur les descripteurs de forme dans les images.

4.2 Interprétation d'une scène 3D en vue d'y évoluer

La capacité à se déplacer dans un environnement est un élément fondamental pour une vie indépendante. Une personne utilise les informations spatiales acquises par ses modalités sensorielles pour s'orienter et se déplacer d'une position à une autre. Bien que toutes les modalités sensorielles soient nécessaires pour comprendre pleinement une scène, elles ne fournissent pas le même niveau d'information. La vision est le sens le plus important pour percevoir l'environnement spatial d'une personne. En effet, il est facile d'imaginer combien il serait difficile de naviguer en toute sécurité sans perception visuelle dans un environnement familier ou non.

Les méthodes classiques utilisées pour remédier à ces problèmes sont la canne blanche ou le chien dressé. Bien que ces méthodes soient largement utilisées et efficaces pour se déplacer en toute sécurité, elles ont des limites. La portée de la canne blanche est limitée aux objets proches, les chiens dressés nécessitent une formation longue et intensive et sont relativement coûteux. De nouvelles technologies pour améliorer les informations spatiales perçues par une personne mal-voyante, appelés technologies d'assistance, doivent être ergonomiques et présenter une faible latence pour détecter les objets en mouvement.

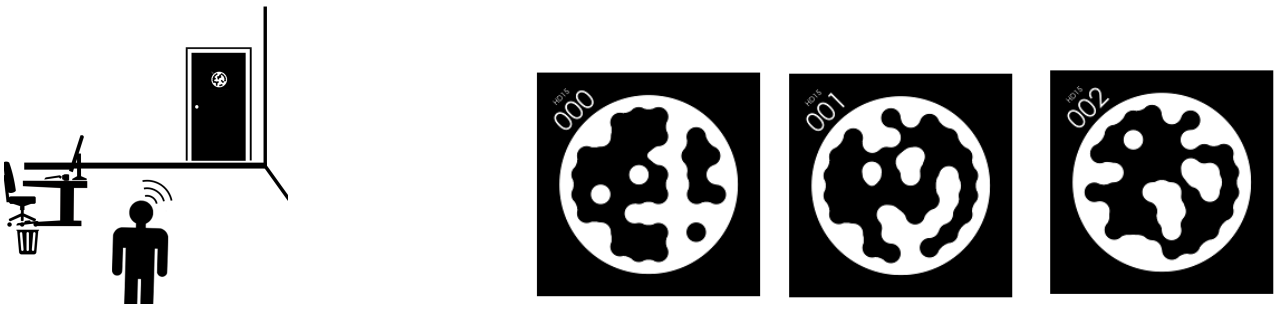


Figure 4.9: À gauche : vue schématique du système. À droite : exemples de marqueurs STag.

Nous proposons dans cette section un nouveau dispositif d'EOA (Electronic Orientation Assistance) en intérieur basé sur une approche de substitution visuelle-auditive. Il s'agit d'une aide à la navigation pour les personnes déficientes visuelles qui fournit des informations sur la trajectoire à prendre mais qui prend aussi en compte la présence ou l'absence d'obstacles qui gênent la progression de l'utilisateur. De plus, le système se doit d'être réactif avec un traitement rapide des données et une émission sonore courte pour permettre un déplacement fluide vers la destination souhaitée.
Nous nous restreignons dans un premier temps à la navigation dans un bâtiment. La méthode proposée est basée sur un maillage du bâtiment par des balises visuelles imprimables où

l'utilisateur navigue par des balises intermédiaires pour atteindre la destination souhaitée. La Figure 4.9 (à gauche) montre une vue schématique du fonctionnement du système dans lequel une personne reçoit un son indiquant la position relative du marqueur. La détection des obstacles est extraite de la carte de profondeur fournie par la caméra RVB-D.

4.2.1 La méthode de navigation

Notre système localise des marqueurs visuels placés dans le bâtiment. Les marqueurs sont positionnés à différents points d'intérêt tels que des portes. Le maillage du bâtiment avec les marqueurs permet une représentation graphique de l'environnement avec leurs connexions associées et permet ainsi l'utilisation d'un algorithme de recherche de chemin basé sur des graphes. Cependant, l'algorithme de recherche de chemin exige que les nœuds soient identifiables et, par conséquent, le symbole du marqueur visuel utilisé doit être unique. Le fonctionnement global de notre système de navigation est décrit ci-dessous :

① Un premier marqueur visuel est recherché pour définir le point de départ puis un algorithme de plus court chemin désigne la succession des balises à atteindre.

② La position du marqueur correspondant au prochain nœud à atteindre est sonifié pour que l'utilisateur s'y dirige.

③ Lorsqu'un marqueur est situé à une distance suffisamment proche de l'utilisateur, le système vérifie si le marqueur est référencé comme un point d'intérêt et alerte l'utilisateur si une action est requise (par exemple l'ouverture d'une porte). Ensuite, si le marqueur atteint n'est pas la destination finale, l'utilisateur est invité à scanner à nouveau l'environnement pour trouver le marqueur suivant.

Nous avons choisi le système de marqueurs fiduciaires STag [80] conçu pour être rapide, stable, robuste à la détection de marqueurs distants, à l'occlusion partielle et aux conditions d'angle de vue difficiles. Un exemple de trois STags est fourni dans la Figure 4.9 (à droite).

Notre représentation graphique non dirigée du maillage du bâtiment est constituée de nœuds qui sont liés les uns aux autres sans connaître la distance qui les sépare. Nous avons privilégié l'algorithme BFS afin de proposer à la personne malvoyante le chemin le plus court pour atteindre la destination.

Nous avons intégré dans notre méthode EOA des informations sur les obstacles proches de l'utilisateur afin d'améliorer la compréhension de la scène et d'éviter la collision. Les obstacles proches sont extraits du flux vidéo de profondeur par un simple seuillage.

Les données à sonifier sont une combinaison d'informations sur le chemin à suivre et les obstacles à éviter. Ces informations doivent être facilement compréhensibles et différenciables par l'utilisateur. Nous associons à chaque information un son monophonique unique et bref facilement identifiable de façon similaire à la section 4.1.3. Nous additionnons les sons résultants pour obtenir un son global. La Figure 4.10 résume notre traitement. Le son monophonique de l'obstacle a une fréquence plus basse que le son monophonique de la navigation.

En plus de ces sons spatialisés et en raison des multiples étapes pour atteindre la destination désirée, nous avons ajouté des informations verbales sur l'étape en cours, mais aussi des informations sur la nécessité de franchir une porte ou d'emprunter un ascenseur. Les informations sémantiques sont préenregistrés et sont jouées au début ou à la fin d'une étape.

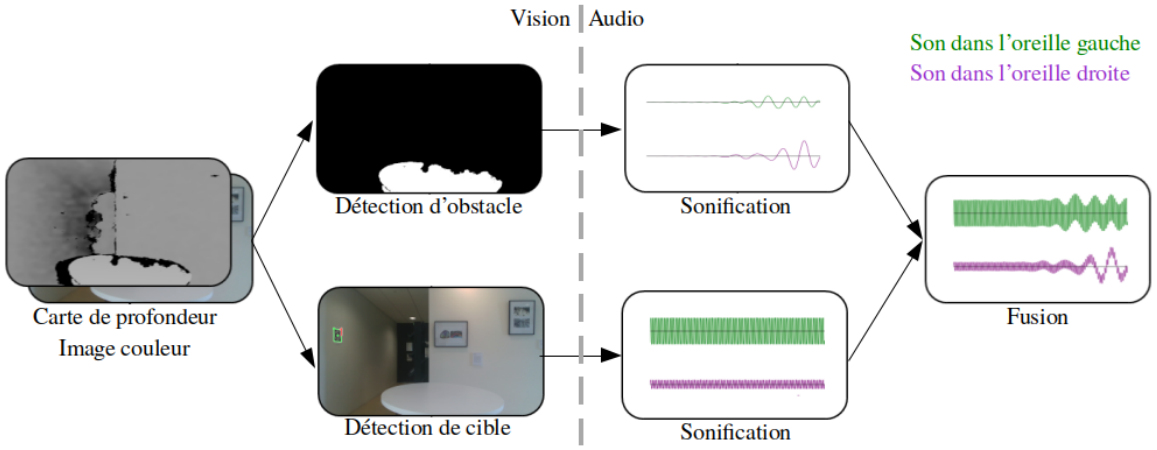


Figure 4.10: Fonctionnement du traitement vidéo et de la sonification pour la détection d'obstacles (en haut) et la détection de marqueurs (en bas). Le son stéréophonique est représenté par une couleur verte pour le canal gauche et une couleur violette pour le canal droit.

4.2.2 Expérimentation

Le système

Le pipeline de notre système est similaire à celui d'autres dispositifs de substitution sensorielle : une étape d'acquisition réalisée par une caméra, suivie du traitement des données visuelles par une unité de calcul, puis de la transcription et de la sortie en informations auditives par un dispositif audio.

PROTOCOLE

- Vidéos acquises par une caméra stéréoscopique Intel RealSense D435i (FOV profondeur de 87° × 58° et couleur de 69° × 42°).
- Images couleur et de profondeur réalignées de manière synchrone avec une résolution de 1280 × 720 pixels à 30 images par seconde.
- Module caméra placé au niveau des yeux sur des lunettes électroniques.
- Unité de traitement : ordinateur portable standard placé dans un sac à dos avec le système d'exploitation Ubuntu 20.04, un processeur Intel Core i7-6700HQ (4 Cœurs - 8 Threads avec une fréquence de 2,60 GHz), 8 Go de RAM, et une carte graphique mobile Nvidia GTX 1070.
- Transmission des informations auditives par écouteurs Bluetooth.
- Sonification par pixel sonore des objets proches effectuée à une résolution de 160 × 120 pixels.

La résolution en pixels sonores est limitée par le jeu de données HRIR utilisé pour spatialiser un son mais surtout par l'incapacité de l'oreille humaine à distinguer une petite variation angulaire dans l'émission sonore.

La limitation des délais est une nécessité. Dans cette optique nous utilisons des sons préchargés et avons divisé le traitement en 4 threads : acquisition des images, traitement des images, génération du son et chargement du son.

Le tableau 4.5 résume l'impact des différents processus (vidéo et audio) pendant une navigation entre deux marqueurs. Les mesures correspondent à une situation courante (voir [SITIS2022]) sur une implémentation CPU. Une optimisation sur GPU réduirait sensiblement

ces durées et permettrait une implémentation matérielle en temps réel sur une cible embarquée de faible puissance.

Table 4.5: Temps de traitement du système de navigation (en ms)

Traitement vidéo		**Sonification**		
STag	Obstacles	STag	Obstacles	Fusion
57	1.4	0.02	17.5	1.9

Protocole

Pour démontrer les capacités pratiques de notre système, nous avons demandé à des personnes ayant les yeux bandés et équipées de notre dispositif de substitution d'atteindre une destination inconnue à l'intérieur d'un bâtiment balisé.

L'espace expérimental contient plusieurs obstacles statiques (chaises, bureaux, tables ...) avec 6 marqueurs imprimés sur un papier A4 et placés à des positions pertinentes afin de créer un maillage du bâtiment.

La Figure 4.11 représente une vue de dessus du bâtiment reconstituée par plusieurs scans Lidar sur laquelle sont affichées des informations sur les positions des marqueurs visuels (flèches vertes et violettes), les positions de départ (point rouge) et de destination (point rose) du participant, et le chemin suivi (ligne blanche).

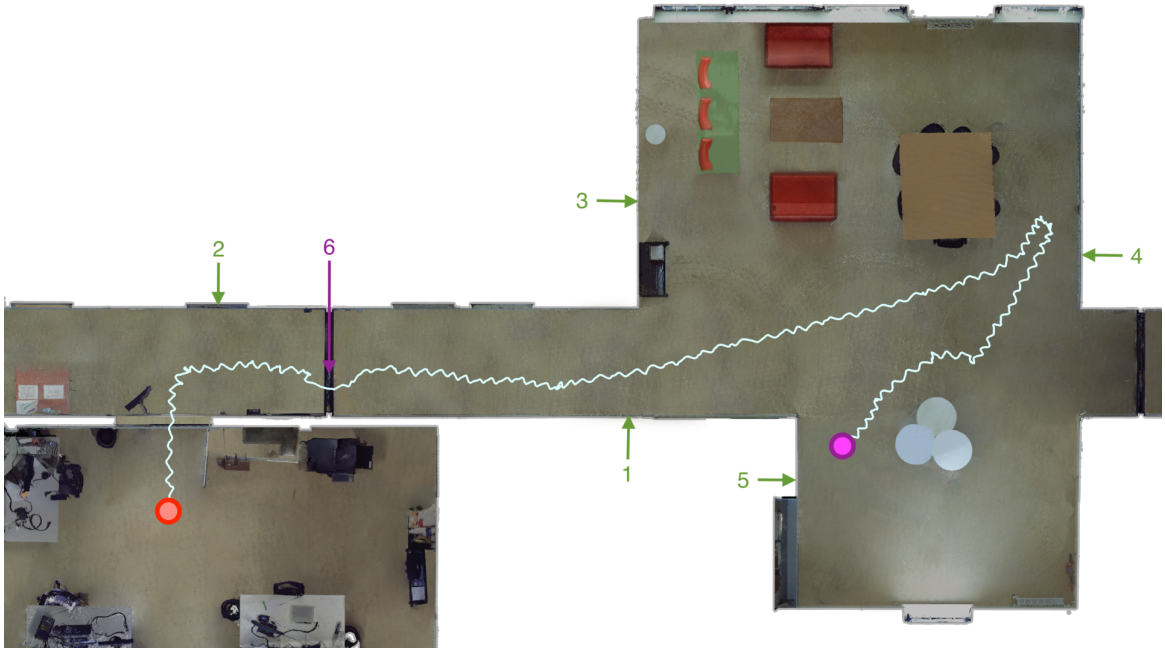


Figure 4.11: Vue de dessus des mouvements de l'utilisateur pour atteindre la destination souhaitée. Le point rouge et le point rose indiquent respectivement la position de départ et la destination. Les flèches représentent la position des marqueurs dans l'espace du bâtiment et la flèche violette indique un marqueur placé sur une porte. Le numéro associé à une flèche représente le numéro du marqueur.

Le chemin parcouru pour atteindre le marqueur final (la porte de l'ascenseur) a été enregistré afin d'analyser le comportement de l'utilisateur lorsqu'il rencontre un obstacle. La position

relative de la personne dans le bâtiment a été enregistrée en utilisant la centrale de mesure inertielle (IMU) d'un casque Oculus Quest 2 en parallèle avec notre système. Les oscillations de la trajectoire dans le chemin suivi sont dues au balancement d'un être humain lors de la marche mais aussi aux mouvements de la tête de l'utilisateur nécessaires pour scanner l'environnement et les éventuels obstacles. La distance maximale entre le marqueur actuel et la personne avant de passer au marqueur suivant est de 80 cm. Le chemin suivi par la personne aux yeux bandés représente les difficultés du parcours. En effet, la personne recule légèrement pour se repérer avant d'ouvrir la porte, puis détecte et évite les tables pour atteindre l'ascenseur. L'expérience peut être reproduite avec succès. Les résultats montrent qu'une personne utilisant notre dispositif de substitution sensorielle peut naviguer dans un environnement intérieur en évitant les obstacles statiques (y compris les murs) et les obstacles mobiles (c'est-à-dire toute personne lui passant devant) tout en passant par une porte pour atteindre la destination souhaitée.

4.2.3 Et maintenant?

Le système présenté dans la section précédente se limite à la navigation dans un bâtiment. Le projet continu et nous étendons le principe sur une navigation en extérieur. Cette méthode s'appuie sur l'information de positionnement spatiale de l'utilisateur fournie par le module GPS et une carte de balise spatiale extraite à partir de la carte coopérative et libre OpenStreetMap. Cette carte permet de connaître le chemin à suivre pour rejoindre une destination souhaitée en appliquant un algorithme de recherche de chemin similairement aux logiciels propriétaires comme Google Map, Bing Map ... et en filtrant uniquement les passages accessibles pour une navigation pédestre (Figure 4.12). Cependant, les dangers extérieurs comme la route, d'éventuels obstacles mobiles comme les vélos ou les trottinettes électriques, ou fixes comme les lampadaires et les arbres, requièrent une détection et une analyse du degré de danger.

Figure 4.12: À gauche, représentation d'une carte en 2D montrant la dangerosité des différentes routes d'une ville en fonction de la présence d'un trottoir ou du type de route. Les tons gris clairs indiquent les routes peu dangereuses et les noirs les plus dangereuses. À droite, itinéraire optimal en fonction de l'accessibilité de la navigation pour les aveugles.

Une phase de détection des obstacles et de prédiction de leur vélocité est donc intégrée au système. Ces détections et prédictions sont réalisées par un réseau de neurones convolutif combiné à un algorithme de suivi des déplacements. Le capteur de mouvement est associé à l'algorithme de suivi pour corriger l'erreur de prédiction du mouvement. Une personne aveugle a une faible connaissance d'où se situe la route et l'ensemble des trottoirs en ville n'intègre pas encore une bande d'assistance aux personnes aveugles. Dans le but de permettre à l'utilisateur de suivre l'axe du trottoir, une phase de différenciation du trottoir et de la route est réalisée via

un réseau de neurones de segmentation sémantique de la scène visuelle. Enfin, d'une manière similaire à la méthode de navigation en intérieur, un son spatialisé indique le chemin à suivre. Une hiérarchisation du degré de dangers des obstacles est réalisée afin de limiter le nombre d'information sonore simultanée et la confusion pour l'utilisateur associé à une saturation du canal auditif.

La méthode de navigation peut être améliorée par l'intégration d'un traitement de commande vocale pour interagir avec le système afin d'obtenir des informations supplémentaires sur la scène. Dans cette optique une collaboration avec le laboratoire TIL (Texte, Image, Langage - EA 4182) a été initiée et plusieurs étudiants travaillent sur le choix des mots ou phrases à sonifier.

4.2.4 Conclusion

Le design de notre EOA n'est pas ergonomique pour l'utilisateur, avec un poids et une consommation électrique excessifs dus à l'utilisation d'un ordinateur portable comme unité de calcul. Une optimisation de notre méthode sur un appareil à faible consommation avec une implémentation GPU est possible. Par ailleurs, le caractère intrusif du conduit auditif avec un casque pourrait interférer avec la compréhension de la scène auditive naturelle (bruit des voitures, paroles ...). Cependant, un remplacement par des écouteurs à conduction osseuse permet d'éviter l'obstruction auditive.
Mais tout ceci n'est que le premier pas. Or notre travail a démontré la faisabilité du processus et sa capacité à répondre au besoin. L'analyse de la scène est robuste et efficace. Le travail à fournir se situe maintenant au niveau de l'interprétation par une personne du signal sonore généré et sur l'optimisation du système en terme de puissance de calcul demandé et de consommation.

LIEN RECHERCHE/ÉDUCATION

- Encadrements des thèses de Florian Scalvini et Quentin Hérau.
- Encadrements de projets étudiant autour de la thématique et sur les données récoltées :
  - Suivi de la pose de la main vis-à-vis d'un objet.
  - Visualisation de la pose de la main dans un environnement virtuel interactif.
- Notre projet est associé à 4 projets d'étudiants du master T2M (Traduction multimédia) pour la recherche de termes pertinents pour la sonification sémantique.

Chapitre 5

Conclusions

5.1 Étude du mouvement 3D pour les personnes

Le mouvement est le reflet du comportement humain. Certes l'estimation de la pose donne des indices précieux. Mais c'est bien une évolution au cours du temps qui permet d'en interpréter la substance. Nous l'avons vu au niveau des micro-expressions où, contrairement aux macro-expressions, la position des points caractéristiques du visage ne suffit pas, mais aussi au niveau de la démarche, où c'est bien à travers l'enchaînement de poses caractéristiques qu'apparaissent les défauts. L'instantané peut suffire pour des expressions grossières et visibles, mais les éléments les plus subtiles, pourtant les plus révélateurs, ne peuvent être appréciés que sur une étude à travers le temps.
Le mouvement est aussi primordial sur l'interprétation d'un environnement. Notre attention est fixé sur le changement. Nous ne voyons pas notre nez bien qu'il soit en plein dans notre champ de vue car il se trouve toujours au même endroit par rapport à nos yeux. De même, dans notre analyse d'une scène, nous nous concentrons sur le changement et donc le mouvement. Il s'agit à la fois de réagir aux dangers mais aussi de filtrer l'immense quantité d'informations qui arrive constamment à nous. Nous nous en sommes rendu compte sur nos lunettes pour malvoyants, où une personne bouge la tête pour faire varier la scène de façon contrôlée et où le mouvement va sélectionner les dangers à sonifier.

La profondeur est aussi une donnée très riche qui complète et enrichit la première. Bien que l'interprétation sémantique se base principalement sur la couleur, la profondeur filtre les informations et aide à associer les éléments en interaction. Elle permet d'enrichir le signal pour une analyse de scène, notamment en levant des ubiquités, ce qui est très en phase avec l'idée d'une interaction, et principalement d'un déplacement, dans un environnement 3D. Mais la profondeur peut même devenir le signal principal, par exemple sur la classe des personnes ou l'enveloppe est plus discriminante que le motif ou la texture. Nous avons pu apprécier la supériorité d'un son spatialisé à partir d'une étude visuelle de la profondeur pour une interprétation de l'environnement.

Les caméras récentes permettent une acquisition simple et rapide de la profondeur. L'objectif étant une utilisation directe de l'utilisateur, le traitement doit être réalisé à la vitesse de l'acquisition, ou du moins à la fréquence du retour à fournir. Tout cela varie bien évidemment avec l'application : un système de sonification de l'environnement doit être temps réel car une latence serait très perturbante mais une information ponctuelle sur la présence d'une personne peut se permettre un léger décalage. Nous avons ainsi proposé des méthodes

légères : en configurant un CNN pour le rendre le plus simple possible tout en gardant des performances de classification des micro-expressions semblable, en étudiant les performances sur différentes cibles de notre système de lunettes pour mal-voyants.

Nos algorithmes accomplissent une assistance à la personne à partir d'appareillage courant. Les caméras utilisées sont simples et peu coûteuses et nous avons vérifié la faisabilité de leur utilisation en conditions réelles de façon non-intrusive. Si l'efficacité de l'étude des micro-expressions est encore limitée, nous avons analysé sous quelles conditions elles peuvent être utilisées en pratique. Que ce soit avec les micro-expressions ou la démarche, les variations sont extrêmement subtiles. Nos algorithmes ne sont pas des produits aboutis, mais nos résultats sont prometteurs vis à vis de l'extrême difficulté du problème, et encourageants puisque validant une méthodologie dans des conditions réalistes. Que ce soit avec des médecins (sur l'étude des démarches), des psychologues (sur l'étude des micro-expressions), ou des spécialistes de psychologie cognitive (sur l'interprétation des sons générés par nos lunettes), nous avons validé l'utilité de nos méthodes avec des spécialistes. Nos méthodes sont liées à nos applications non pas seulement par un contexte mais par une utilité vérifiée.

Nous avons de plus développé notre maîtrise de la description des personnes dans les images et les vidéos. Nous avons créé des descripteurs spécifiques basés sur la forme cylindrique d'un piéton. Nos résultats montrent que cette spécialisation est avantageuse en descriptivité et en efficacité.

Tous ces travaux ont mis en exergue de multiples verrous qui limitent l'utilisation de nos méthodologies mais ouvrent la voie à de nouvelles études.

5.2 Perspectives de l'étude du comportement humain

Ces dernières années, l'approche par apprentissage profond a bouleversé le monde de l'analyse d'image et principalement les problèmes de classification. Les approches basées descripteurs sont maintenant moins prisées et leur utilisation doit se justifier vis à vis des performances de l'apprentissage profond. Néanmoins cet apprentissage dépend des données sur lesquelles elles sont entraînées. La sélection et l'utilisation des bases de données est maintenant une problématique de premier choix. Les données doivent être très nombreuses, représentatives du problème et en proportion équitable. L'analyse vidéo rend le problème encore plus complexe. Les séquences vidéos sont plus volumineuses et plus contraignantes à acquérir que les images. Les bases de données de vidéos sont moins variées et de façon générale contiennent moins d'échantillons. Il a donc bien ici une limite car c'est bien la quantité et la variabilité de ces données qui justifient l'utilisation de l'apprentissage profond.
L'étude du comportement humain rend l'acquisition encore plus complexe puisque la génération du comportement doit être spontanée et est donc difficile à provoquer. Sur l'exemple des micro-expressions, nous avons vu que la taille limitée des bases de données disponibles et leur mauvais équilibre avait un impact important sur les résultats. Limiter l'impact des caractéristiques de ces bases (par exemple par l'usage de descripteurs en pré-traitement pour réduire la taille d'un échantillon ou bien par la réduction de la séquence en des instants clés - comme nous l'avons fait sur les micro-expressions en ne calculant le flot optique qu'entre deux instants) ou bien augmenter artificiellement leur taille (par exemple à partir de données de synthèse ou en déclinant les échantillons à partir de traitements pertinents) est ainsi un domaine d'étude nécessaire et prometteur.

D'un point de vue application, l'assistance à la personne propose de nombreux verrous très porteurs. Tout d'abord la nécessité de traitements légers et portables (la latence introduit un biais perceptif) impose une contrainte intéressante. Ensuite le mouvement étudié est spécifique. Contrairement aux classifications classiques répertoriant un nombre défini de classes (la reconnaissance d'action est souvent limitée à des actions basiques comme se lever ou agiter le bras), l'approche doit ici interpréter l'action effectuée, quelle qu'elle soit, pour produire une information adaptée afin, par exemple, d'anticiper un geste.

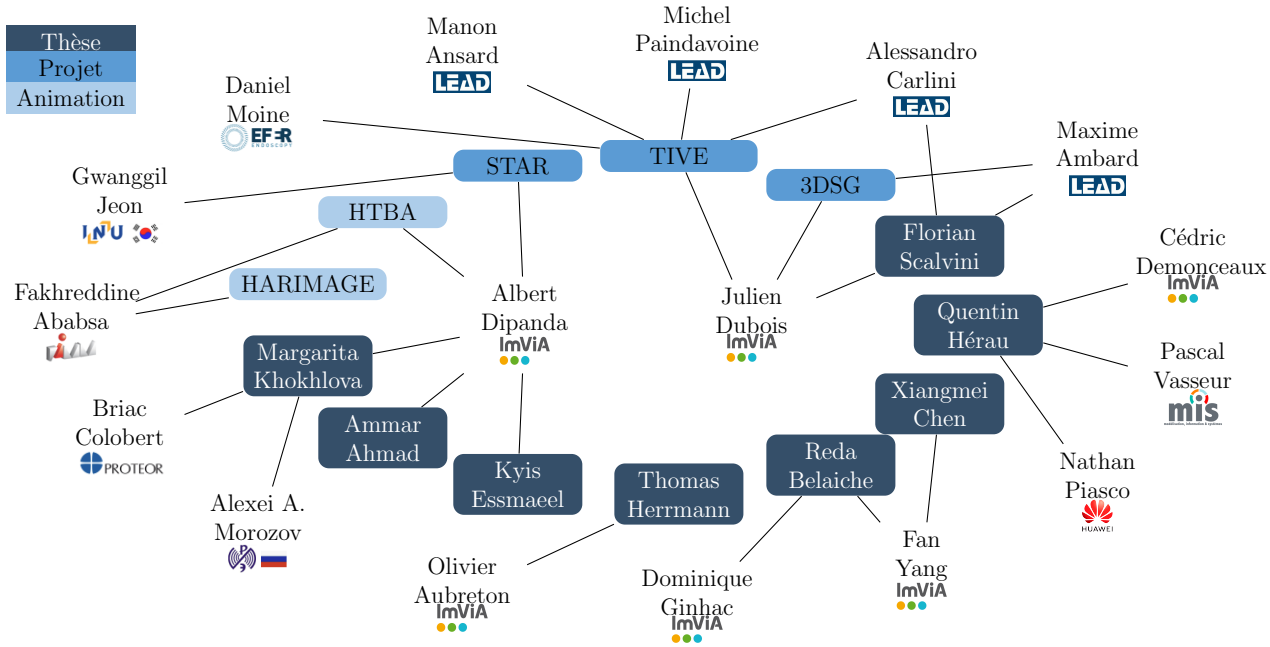


Figure 5.1: Mes collaborateurs privilégiés et les responsabilités que j'ai partagé avec eux.

5.3 Mon laboratoire, mon équipe et la communauté scientifique

J'ai intégré en 2014 le laboratoire Le2i dans le pôle 5 : Systèmes de vision et Méthodes d'imagerie. En 2018, le laboratoire s'est scindé et j'ai intégré le laboratoire ImViA à travers l'équipe CoReS : COmputer vision for REal time Systems. Étant d'origine plutôt concentré sur le traitements des images et des vidéos, j'ai évolué en intégrant les contraintes systèmes (temps-réel, embarquabilité et consommation de calcul) et l'utilisation de modes d'acquisition différents (3D, nuage de points complet, caméra thermique, caméra rapide...) pour coller au mieux aux thématiques de mon équipe. Au sein de CoReS, je m'inscris au niveau du thème sur les outils de mesure et de diagnostic basés sur la vision. Mes études, notamment sur la rééducation au niveau de la démarche et sur l'analyse des émotions, ont été motivées par la volonté d'avancer sur ce thème. Cela a également permis d'étendre les sujets d'étude du laboratoire comme par exemple sur l'étude des micro expressions qui n'avait pas encore été traitées dans le laboratoire.
Cependant, étant donné mon expérience et mes affinités sur des thèmes à la périphérie de ceux mis en avant par l'équipe, j'ai également œuvré à utiliser mes travaux comme une passerelle entre équipes et entre laboratoires (Figure 5.1). La thèse de Thomas Herrmann a permis de conforter les collaborations entre les deux sites principaux du laboratoire (Dijon et Le Creusot).

De même, les projets 3DSG et TIVE ont permis d'initier un rapprochement aujourd'hui important entre les laboratoires ImViA et LEAD. J'ai également interagi avec le tissu industriel régional par l'intermédiaire d'une bourse JCE avec la société PROTEOR basée à Dijon et spécialiste des prothèses.
D'un point de vue international, l'intégration de ma recherche est passée par la soumission de multiples projets avec la Corée du sud (PHC STAR), la Russie (PRC RFBBR), le Japon et l'Allemagne (ANR COBOT). De plus, les organisations depuis 2016 du workshop HTBA (congrès en Espagne, en Italie, en Inde et en France) et du special issue HarImage m'ont permis de partager expériences et points de vue avec la communauté scientifique dans mon domaine de recherche.

Chapitre 6

Projet

L'interprétation d'une scène se réalise à partir de l'ensemble des sens qui est à notre disposition. Lorsqu'une information est manquante, certaines actions sont plus compliquées à réaliser. Nous considérons deux exemples. Tout d'abord, une personne malvoyante désirant saisir un objet aura un mouvement hésitant si elle ne connaît pas précisément la direction vers laquelle tendre le bras ni la distance où se trouve l'objet. Ensuite, pour une personne qui manipule un objet virtuel (dans un contexte de réalité virtuelle), le manque de retour haptique rend la manipulation peu intuitive notamment pour saisir et relâcher l'objet.
Dans le projet que je propose, l'information manquante est substituée par un autre signal, par exemple un signal sonore. J'ai déjà étudié cette substitution au niveau de la navigation de personne : notre système capture la scène visuelle et génère un son qui permet à une personne non ou mal voyante de se déplacer dans un environnement complexe. Ce projet prolonge cette étude au niveau de la manipulation d'objet. Le signal de substitution doit informer l'utilisateur sur la nature de l'interaction entre sa main et l'objet. Dans un second temps, il sera enrichi pour anticiper le besoin ainsi que l'action de la personne, prévoir une trajectoire et évaluer si le mouvement correspond à une manipulation correcte.

6.1 Travaux préliminaires

Ce projet repose principalement sur deux de mes études préalables. La première consiste en la substitution de la vue par un signal sonore pour la navigation d'une personne mal-voyante; la seconde considère l'interaction entre une personne et un objet virtuel. Les retours et résultats obtenus sur ces deux études me permettent de porter de légitimes espoirs sur les perspectives ouvertes par leur association.

Dans le projet envergure 3DSD, nous avons substitué l'information visuelle par une information sonore. Une acquisition 3D de la scène est acquise et certains éléments spécifiques (pixels ou fenêtres de détection) sont transformés en sons. Le système complet est embarqué sur des lunettes et fonctionne en temps réel avec une autonomie acceptable. Le son permet à la personne d'évaluer la localisation mais aussi la distance d'un objet. Bien que moins riche que l'information disponible dans la scène visuelle, l'information sonore générée permet à la personne de se faire une représentation de la scène et de sa propre situation face à son environnement. Notre système permet tout à la fois d'interpréter la présence d'obstacles et de localiser un élément spécifique de la scène (une personne ou une cible à atteindre). Une personne peut ainsi se déplacer jusqu'à un lieu cible en évitant les obstacles (Figure 6.1).
Cette substitution est guidée par notre application qui a pour objectif de fournir une assistance aux personnes malvoyantes [ICASSP2022] [SITIS2022]. Notre système est adapté à l'analyse

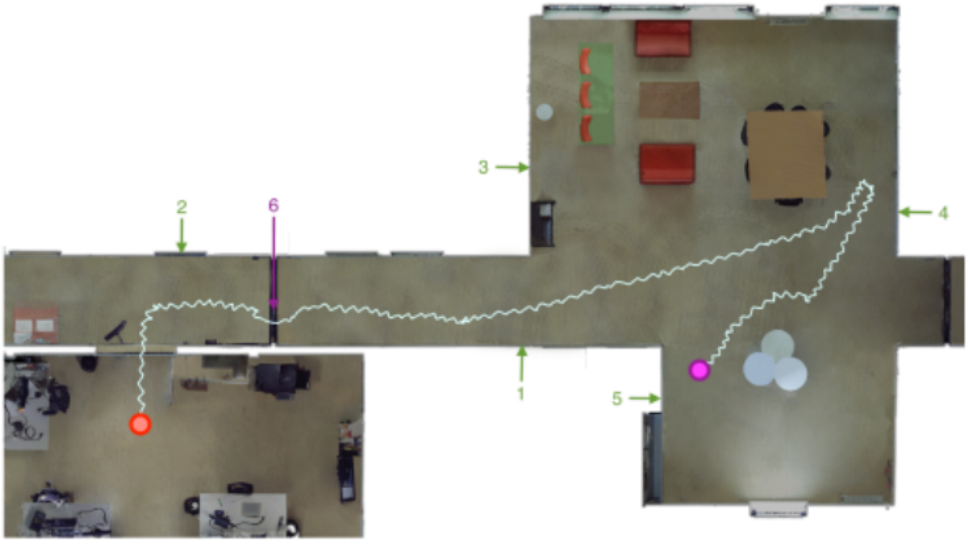


Figure 6.1: Notre système de sonification de la scène embarqué sur des lunettes permet à une personne privée de la vue de se déplacer efficacement vers un objectif précis tout en évitant les obstacles. Sur l'exemple, la personne (trajectoire en blanc) rejoint l'ascenseur (5) tout en évitant les tables et chaises et en ouvrant une porte (6).

de la scène (statique) et au déplacement d'une personne (dynamique). Néanmoins, il ne prend en compte ni le comportement ni l'action de la personne qui l'utilise. De plus, un a-priori sur l'opération effectuée peut affiner l'information contenue dans le signal de substitution pour une interprétation plus évoluée.

En résumé

- Substitution sonore pour une perception générale de l'environnement.
- Sans interaction avec l'environnement ni analyse de l'action.

Dans [IVC2019] et [SITIS2017] nous avons fait un état des lieux des méthodes de suivi des mains dans un environnement réel et virtuel. Le suivi de la main 3D basé sur la vision est un élément clé et populaire pour les études d'interaction dans un large éventail de domaines tels que la réalité virtuelle, la réalité augmentée et l'interaction homme-machine. Bien que ce domaine de recherche ait été bien étudié au cours des dernières décennies, la plupart des approches ont considéré la main humaine de manière isolée et non en action ou en interaction avec l'environnement. Cependant, de nombreuses applications actuelles nécessitent de plus en plus d'interactions main-objet. En outre, l'utilisation d'informations contextuelles sur l'objet dans la main (par exemple la forme, la texture et la pose) peut considérablement limiter le problème du suivi. Les contraintes contextuelles les plus étudiées concernent l'interaction avec des objets réels et non avec des objets virtuels.
Un utilisateur devrait être en mesure d'étendre les capacités de ses mains pour interagir avec des objets virtuels (en plus des objets réels) aussi naturellement que possible. Puisque les mains réelles de l'utilisateur sont remplacées par des mains virtuelles et intégrées dans un contexte d'interactivité virtuelle, la sensation d'immersion et de présence dans l'environnement virtuel peut être fortement perturbée. À cet égard, un phénomène qui se produit fréquemment dans le contexte de l'interaction virtuelle est l'estimation de poses des mains physiquement irréalistes et instables en raison de l'absence de retour de force.

Deux manifestations majeures de ce phénomène peuvent être distinguées (Figure 6.2) : l'interpénétration main-objet (un problème qui se produit pendant l'interaction virtuelle lorsque

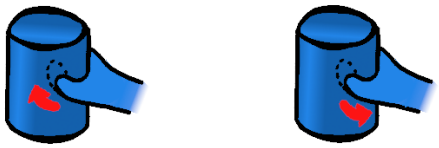


Figure 6.2: Comme il n'y a pas de retour haptique, la main virtuelle (issue du suivi de la main réelle) peut chevaucher le volume de l'objet virtuel (à gauche). La position estimée de la main sera donc modifiée en conséquence. Mais lorsque la personne ouvre de nouveau la main (à doite), cela revient à un retour à la position estimée et non à un relachement de l'objet.

les volumes de la main suivie et de l'objet virtuel se chevauchent) et la prise de l'objet (lors du relâchement après l'interpénétration main-objet, un mouvement des doigts permet de sortir du volume de l'objet virtuel mais ne correspond à aucun mouvement retranscrit dans le monde virtuel).
Le principal problème avec la manipulation d'un objet virtuel est l'absence de retour haptique. L'objet n'existant pas, l'utilisateur ne peut pas le toucher. Il est possible d'utiliser des gants haptiques qui fournissent un retour (une vibration correspondant à la pénétration ou au toucher de l'objet) mais il s'agit d'un équipement intrusif et coûteux inenvisageable pour de multiples applications. Prenons un exemple dans le domaine des musées du futur. Un visiteur s'intéresse à la manipulation d'un objet historique. Pour cela, il se place devant une borne et se voit (ou au moins voit ses mains) manipuler un double virtuel de cet objet. Il n'est pas envisageable de faire porter à chaque visiteur un équipement compliqué. Mais sans retour haptique, la personne vivra une expérience limitée et sa prise de l'objet sera grossière. Si ce retour est substitué par un autre signal, la manipulation peut devenir plus naturelle.

En résumé

- Correction du suivi de la main en fonction de la position de l'objet virtuel.
- Décalage de perception provoqué par l'absence de retour haptique.

6.2 Développement du projet

Le problème de l'interaction avec un objet virtuel est lié à celui des lunettes de sonification en cela qu'un sens fait défaut. Au lieu de la vue c'est le toucher qui manque. Si l'information perdue semble moins primordiale (l'action peut être réalisée uniquement avec un retour visuel) l'absence de résistance au contact d'un objet rend l'interaction plus artificielle. Le geste et le mouvement seront plus naturels et plus justes avec cette information, même simulée.
L'utilisation d'un son pour substituer cette information, de façon similaire aux lunettes du projet 3DSG, semble ainsi prometteuse. La faisabilité de l'utilisation d'un son 3D pour faciliter l'interaction avec un environnement 3D a été démontrée dans [81]. Selon [82], l'ajout d'un retour audio pour la saisie d'objet virtuel est plus confortable mais augmente le temps de réaction. [83] démontre qu'un retour multimodal (audio et tactile) améliore significativement les performances. Le système proposé dans [84] produit un guidage sonore pour permettre aux malvoyants d'agripper un objet (une bouteille dans l'article). Dans ces systèmes, le son généré est un signal indicateur de la saisie de l'objet.

Le son possède l'intérêt de pouvoir fournir une perception 3D de l'information. Par exemple le système [78], utilisé dans notre projet 3DSG, permet à partir de fonctions HRTF de générer un son que l'utilisateur associe à une position spatiale 3D. Cette représentation est très intéressante du point de vue de l'interaction avec un objet. Déterminer à la fois où se

diriger et à quelle distance se trouve un objet est une démarche intuitive pour réaliser le bon geste. De plus, la localisation 3D du son généré est naturellement et intuitivement interprétée, principalement au niveau de l'axe longitudinal grâce au décalage du son entre les deux oreilles. Il n'y a donc besoin de la part de l'utilisateur ni d'un apprentissage complexe ni d'un usage en continu du système.

Une variable d'ajustement digne d'étude est le changement de point de vue. Comme avec nos lunettes, la majorité des méthodes de substitution de l'environnement prennent comme origine le visage de l'utilisateur. C'est en effet selon ce point de vue qu'une personne perçoit naturellement son environnement. Mais dans l'optique d'une action spécifique, un autre point de vue aura ses avantages (Figure 6.3). Par exemple concernant la préemption d'un objet, la distance de la main à l'objet peut être plus informative que la distance du visage à l'objet, en particulier sur des distances réduites. Le référentiel pris en compte peut aussi tenir compte de l'orientation de la paume de la main. Le signal généré permettrait alors à la personne de plus finement disposer sa main.
Le capteur d'acquisition peut alors être positionné au point d'origine (la main) mais cela n'est pas pratique et trop dépendant de l'action. Un changement de repère à partir d'une acquisition fixe (la tête de l'usager ou bien son torse plus stable) est préférable. Il faut donc suivre et estimer précisément la position et l'orientation de la main pour correctement effectuer le changement de repère.
Les méthodes existantes de suivi de la main sont très efficaces et une caméra telle que la Leap-Motion permet une acquisition tout à fait adéquate. Il s'agit d'une caméra RGB-D qui permet un suivi efficace des mains par une vue du dessous avec une acquisition correcte de la profondeur à partir d'une distance de 10cm (ce qui correspondrait à la position d'une borne de manipulation dans un établissement). Pour une acquisition prise depuis la personne (comme sur des lunettes), les caméras de profondeur de type real-sense permettent une estimation suffisante compte tenu du mouvement et de l'éclairage. Comme nous l'avons expérimenté sur les personnes et sur les mains, la profondeur fournit une plus forte robustesse lorsqu'il s'agit de suivre ou de reconnaître une pose. Mais c'est surtout au niveau de la disposition des mains et de l'objet dans l'environnement 3D que l'information de profondeur prend de l'intérêt.

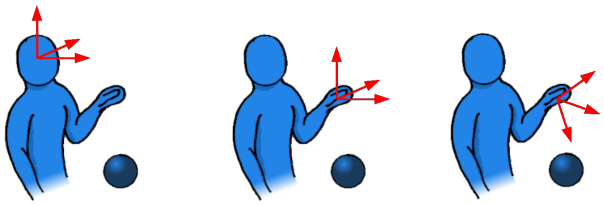


Figure 6.3: La relation à l'objet peut être vue selon différents points de vue : celui de l'utilisateur (à gauche) qui est le plus naturel, celui de la main (au centre) pour évaluer la distance de cette dernière à l'objet et celui de la paume (à droite) pour aussi gérer l'orientation.

Une première étape consistera à adapter le système existant à la manipulation d'un objet réel. De nouveau, c'est la vue qui sera substituée par un son. Mais cette fois l'étude sera centrée sur l'interaction proche d'un utilisateur avec un objet et à partir d'un capteur de profondeur plus adapté. Le champ d'étude sera plus restreint et spécialisé à une action mais demandera une plus grande précision. La validation de cette première étape débouchera sur une meilleure compréhension du problème principal.

La dernière phase du projet devrait concerner la spécialisation à une action. Si un a-priori sur l'action effectuée est intégré dans le processus, alors la solution sera moins générale mais devra produire une plus grande robustesse et un rendu plus fin. Il s'agit alors de reconnaître le mouvement de la personne et de suivre chaque étape de son exécution. Ceci permettra alors d'anticiper un problème voir de quantifier la pertinence d'un geste.

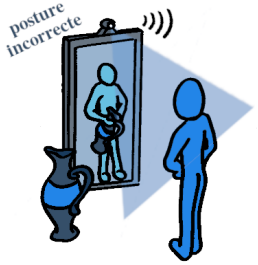
En résumé

- Utiliser le son pour remplacer le retour haptique afin de manipuler un objet virtuel.
- Utiliser le son pour remplacer la vue sur une action spécifique qui demande de la précision.
- Définir le signal en fonction de la pose et du mouvement de la personne.
- Déporter le point de vue au niveau de la main ou de la paume.
- Reconnaître une bonne manipulation de l'objet.

6.2.1 Applications pratiques :

Assistance aux personnes malvoyantes sur des actions de tous les jours : par exemple sélectionner et attraper correctement un aliment dans une assiette.

Musée du futur : une borne permettant à un visiteur de se visualiser en réalité augmentée manipulant un double virtuel d'un objet historique exposé. Cette borne pourrait permettre l'apprentissage et l'évaluation du bon maniement de l'objet : par exemple une prise correcte de couverts ou un mouvement avec une arme correspondant à une utilisation historique.



6.2.2 Premiers essais

Dans l'optique d'initier ce projet et de vérifier sa viabilité, j'ai proposé un projet dans le cadre du master IIA (Image et Intelligence Artificiel). Il s'agit d'une seconde année de master en alternance et ces projets (équivalent à 5 demi journées par semaine sur 4 mois) sont réalisés par les étudiants sans contrat avec une entreprise (notamment issus des recrutements RI). Le sujet proposé consistait à évaluer les capacités de la caméra LeapMotion dans un premier temps puis de lancer une première expérience pour estimer l'interprétation d'un signal sonore pour la saisie d'un objet.

Au niveau du capteur, la LeapMotion offre un champ de vue de $140 \times 120°$ efficace sur une distance de 10 à 80cm. Ce fonctionnement à courte distance est cohérent avec nos applications. L'étudiant a principalement testé l'algorithme de suivi implémenté dans le SDK fourni (Figure 6.4). La méthode est prévue pour un positionnement de la caméra sous les mains (comme si l'utilisateur pianotait juste au dessus d'elle - Figure 6.5 à gauche). Dans cette configuration le suivi est très robuste, quelque soit l'orientation ou l'inter-occlusion des mains. Cependant, dans l'idée d'une borne ou d'un système embarqué, cette configuration n'est pas conforme. L'étudiant a alors fait varier le point de vue. Nous avons alors noté que le suivi reste, en ajustant certaines variables, très robuste avec une vue de dessus (au niveau de la tête - Figure 6.5 au

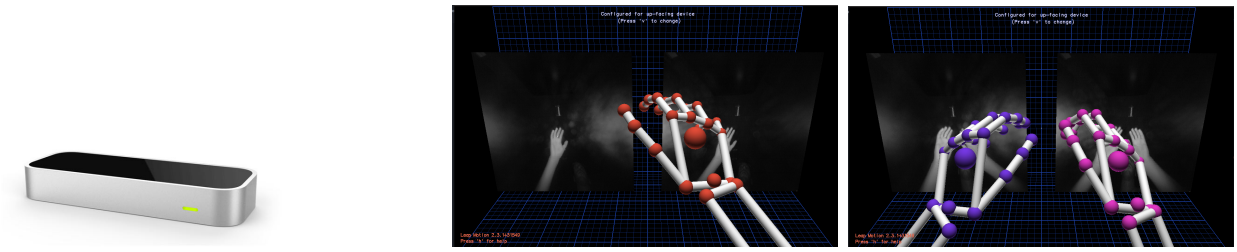


Figure 6.4: Le SDK de la caméra LeapMotion propose un algorithme de suivi de la pose des mains robuste et précis.

centre). Ces performances sont néanmoins très sensibles à la lumière; un bon fonctionnement nécessite donc une utilisation en intérieur (valable dans un musée ou un restaurant).

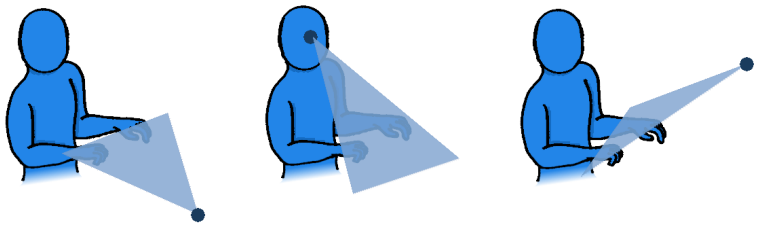


Figure 6.5: La caméra LeapMotion est conçue pour être positionnée sous les mains (à gauche). Le suivi fonctionne également avec une vue au niveau de la tête (au centre). Un ajustement doit être ajouté pour une acquisition depuis une borne extérieure (à droite).

Figure 6.6: La personne doit se saisir d'une sphère virtuelle à partir d'un son généré selon trois points de vue différents.

L'étudiant a ensuite préparé une expérience pour évaluer le système. L'utilisateur a les yeux bandés et est assis. L'algorithme génère une sphère virtuelle de position aléatoire. La personne reçoit un signal sonore et doit se saisir de l'objet virtuel. Le signal sonore est généré de trois façons différentes (Figure 6.6) à partir de la pose et de la localisation des mains :

① Seule la distance entre l'objet et la main est prise en compte. Plus la distance est réduite plus l'amplitude du son est élevée.

② La localisation de l'objet est aussi prise en compte. L'orientation est définie selon le point de vue de la tête. Le son 3D est obtenu de la même façon que sur le projet 3DSG. L'amplitude du son est toujours reliée à la distance de l'objet à la main.

③ La position de l'objet est prise du point de vue de la main avec l'orientation de la paume.

Le protocole est finalisé mais, par manque de temps, son application sur un ensemble de volontaires sera réalisé dans un autre projet. Sur nos premiers essais, l'utilisation du point de vue de la paume semble améliorer l'interaction mais cela devra être validé sur un panel plus important d'utilisateurs.

6.3 Connections avec la recherche et l'enseignement

6.3.1 Collaborations

Un projet s'enrichit par l'association de points de vues et la variétés des expertises. Je compte faire vivre ce projet en collaborations avec plusieurs personnes et entités.
Au sein de mon laboratoire, je compte déjà continuer de collaborer avec Julien Dubois, avec qui j'ai encadré la thèse de Florian Scalvini. La continuité du projet ainsi que les multiples perspectives qu'il a ouvertes nous donne pleinement envie de continuer sur ce thème porteur. Julien Dubois apporte une expertise sur les développements embarqués et l'implantation matériel qui complète mes compétences.
Je me suis rapproché de Carlos Manuel Mateo Agullo, un maître de conférences récemment recruté travaillant sur la saisie des objets par un bras robotique via un retour tactile. La saisie virtuelle et la saisie par un robot par retour haptique sont deux thématiques opposées concernant les données disponibles et à évaluer. Mais leurs traitements suivent la même logique et nos deux recherches profiteront d'un suivi en parallèle. De plus, Carlos Manuel Mateo Agullo appartient à l'équipe Vibot se trouvant sur le site creusotins du laboratoire. Un travail conjoint permettra de réaliser une passerelle pour renforcer les liens entre les deux entités.
Outre l'aspect traitement informatique, le projet proposé comporte une partie importante d'interprétation de l'information. Il est donc primordiale de l'associer avec une équipe spécialiste en psychologie cognitive. Le laboratoire LEAD est spécialiste du sujet. C'est avec cette équipe que s'est construite le projet 3DSG. Depuis trois ans les relations entre les deux laboratoires se sont considérablement confortées. J'ai personnellement été acteur de deux projets de collaboration entre les deux entités : le projet envergure région 3DSG et le projet rapid DGA TIVE. Je compte continuer à alimenter ce rapprochement notamment au travers de ce projet et de mes actions futures.
Un rapprochement avec le musée des Beaux-Arts de Dijon, comme initié sur la thèse d'Ammar Ahmad, serait judicieux, notamment pour définir le besoin et les contraintes.

En résumé

- Dans le laboratoire vers des compétences en implantation matériel et en robotique.
- Entre laboratoires en psychologie cognitive.
- Hors laboratoire vers les musées nationaux.

6.3.2 Liens avec l'enseignement

Le master M1 informatique de l'UFR Sciences et Techniques propose un parcours généraliste permettant de maîtriser les fondamentaux pour la conception et le développement des systèmes informatiques, notamment par la maîtrise des méthodes liées à l'image numérique. Je suis impliqué dans ce master en tant que responsable des relations internationales. Je suis aussi responsable d'un cours de traitement des images traitant particulièrement des opérations morphologiques et de la classification. Chaque année les élèves doivent réaliser un projet en groupe de 4 à 6 élèves sur plusieurs mois pour répondre à une problématique en traitement de l'image.

J'encadre chaque année plusieurs groupes de projets.

Ce master peut déboucher sur deux master 2 dont le master IIA (Image et Intelligence Artificielle). Dans ce parcours je suis responsable d'un module de traitement numérique des vidéos (compression, suivi, estimation du mouvement ...). Ce parcours est en alternance et les élèves sans entreprise réalise un projet long. J'encadre chaque année au moins un de ces projets.

Je réalise également un module de programmation dans la classe préparatoire intégrée de l'école d'ingénieur ESIREM. Je propose et j'encadre régulièrement des projets longs de 3 à 5 étudiants que j'associe traditionnellement avec des problématiques de vision.

Ces formations, surtout les deux premières, sont directement liées à mes thématiques. Mon expertise affine la formation et le retour des étudiants nourrit ma recherche. Ces projets sont aussi l'occasion d'intégrer des problèmes liés avec ma recherche ce qui me permet de tester des solutions et aux étudiants de traiter des problèmes concrets.

6.3.3 Investissement dans le futur du laboratoire

Depuis 2018 et sa création je fais partie du laboratoire ImViA. Je suis membre de l'équipe CoReS (COmputer vision for REal time Systems) et suis majoritairement intégré à l'axe "outils de mesures et de diagnostics basés sur la vision". Mes recherches s'insèrent naturellement dans les thématiques portées par le laboratoire. J'ai accepté de multiples responsabilités qui m'ont bien intégré au fonctionnement du laboratoire et que je compte continuer d'assumer.
En septembre 2023, une nouvelle direction prendra les rênes du laboratoire. Je me suis fortement impliqué dans le projet porté et ferai partie de la cellule de direction.

6.4 Projets soumis ou en cours

Si le projet proposé n'est pas directement la continuité des travaux déjà réalisés, il en découle par l'expérience acquise et les notions prises en compte. Je ne compte ceci dit pas abandonner les possibilités données par les perspectives directes des travaux que je viens de terminer.

6.4.1 Projets soumis

En compagnie de mes collègues, nous avons déposé cette année deux projets ANR.

Projet LAGUNA
La longévité, un acquis important de notre société, présente de nouveaux défis en termes de prise en charge d'une population âgée croissante. Les troubles neurocognitifs majeurs désignent les maladies du cerveau qui interfèrent avec la vie quotidienne, dont 70% pour la maladie d'Alzheimer. Avec l'augmentation des besoins en soins de santé, les coûts en Europe atteindront 250 Md€ en 2030. Pour conserver le niveau de service actuel fourni aux personnes fragiles, il manquera 10 millions d'aidants en Europe à cette date. Les robots peuvent améliorer la qualité de vie et le bien-être de ces patients et soutenir les aidants. Compte tenu de la complexité des soins aux malades, l'intégration de la technologie dans les soins constitue un défi de taille et une réelle opportunité. L'ambition du projet LAGUNA est d'améliorer les conditions de vie des patients atteints d'Alzheimer en prolongeant leur séjour à domicile, réduisant ainsi la pression sur les établissements de soin, la charge sur les soignants et les coûts liés à la maladie. Dans ce projet, un robot sera équipé de capteurs et exécutera des algorithmes d'intelligence artificielle (IA) pour mesurer des paramètres physiologiques sur

des patients au premier stage de la maladie d'Alzheimer. Parmi ces paramètres se trouvent le suivi oculaire et la variabilité du pouls acquise à distance par photopléthysmographie assistée par l'IA (détection vidéo de changements de flux sanguin micro vasculaires). Un second niveau d'IA permettra d'estimer la gravité de la maladie, grâce à une base de données acquise en milieu hospitalier. Le bénéfice potentiel (bien-être) des différents types de thérapies avec le robot sera exploré ainsi que l'acceptabilité des robots pour les malades d'Alzheimer et les soignants.
Le projet n'a pas passé le second tour mais sera resoumis l'année prochaine en prenant en compte les retours des reviewers.

En résumé

- Un robot pour suivre au quotidien des personnes atteintes de la maladie d'Alzeimer.
- Acquérir et traiter des paramètres physiologiques.
- Estimer l'évolution de la maladie.

Projet aiMotions

Comprendre les relations des consommateurs avec la nourriture est une question cruciale pour promouvoir des comportements plus sains. Afin de mieux comprendre le comportement des consommateurs, les mesures des émotions ont rapidement augmenté au cours de la dernière décennie, car les aliments et les émotions sont tous deux interconnectés. Dans le domaine des sciences de l'alimentation, la collecte des réactions émotionnelles repose principalement sur des mesures explicites directes, telles que les questionnaires d'auto-évaluation dans lesquels les consommateurs sont invités à verbaliser explicitement leurs sentiments. Cette approche peut être biaisée sur le plan cognitif car les réponses peuvent être influencées par des facteurs tels que la désirabilité sociale ou les conventions culturelles. En outre, sachant que les choix de la vie réelle sont souvent guidés par des mécanismes non conscients plutôt que conscients, il serait plus judicieux de recueillir les indices automatiques et incontrôlables des réponses émotionnelles des consommateurs. Ainsi, évaluer les sentiments des consommateurs de manière fiable, en se basant sur les mesures de leurs émotions, serait une réelle valeur ajoutée .
aiMotions propose un programme de recherche interdisciplinaire comprenant des enquêtes combinées avec des études expérimentales sur les comportements des consommateurs, et vise 3 objectifs scientifiques :

- Publier un grand ensemble de données ouvertes, extensibles et multimodales sur les réponses émotionnelles provoquées par les aliments et les boissons.

- Concevoir un modèle de bout en bout basé sur l'intelligence artificielle qui vise une analyse automatisée des émotions en exploitant simultanément les informations uniques et les relations complémentaires des différentes composantes disponibles dans les données multimodales.

- Apporter de nouvelles idées sur l'analyse des émotions liées à l'alimentation et plus généralement sur le comportement des consommateurs.

Le projet a été accepté.

En résumé

- Étude de la génération des émotions liée à la consommation de nourriture.
- Réalisation d'une base de données complète et équilibrée.
- Évaluation de l'émotion par un modèle multimodal.

6.4.2 Suites du projet 3DSG

La suite du projet 3DSG s'oriente vers le déplacement en extérieur. Sur le même principe que le système sonifiant la cible et les obstacles en intérieur, ce système sonifie un chemin à suivre et les éléments à éviter.
La cible était définie par des marqueurs visuels accrochés aux murs et porteurs d'information. Cette fois le chemin à suivre est obtenu par un repérage GPS associé à une carte sémantisée (rue, arbre, trottoir, feux tricolores, bâtiments ...). Le traitement vision permet aussi de segmenter les trottoirs. Le calcul est donc affiné pour que la trajectoire à suivre par la personne reste toujours dans de bonnes conditions de sécurité (rester sur le trottoir plutôt que se rapprocher de la route). De même, les algorithmes basés vision permettent d'enrichir la détection des obstacles par une catégorisation (danger, élément en mouvement, type d'obstacle ...).

Cette continuité du projet s'écarte de notre proposition dans le sens où elle s'intéresse à la motricité et au déplacement de la personne (donc sur un mouvement global de la personne sur de grandes distances plutôt que sur une analyse fine de petits mouvements). Cependant elle reste sur une même ligne puisqu'elle associe le mouvement à une substitution sensorielle. Les deux approches s'enrichiront l'une l'autre à être traitées séparément mais en parallèle.

Ce travail est en cours et un prototype réalisant le traitement est déjà fonctionnel. De plus, nous construisons une base de données pour la navigation de personnes dans un milieu urbain. En effet, la plupart des bases disponibles prennent en compte le point de vue d'une voiture.

En résumé

- Déplacement d'une personne non voyante dans un environnement extérieur.
- Étude des indices de circulation pour trouver une trajectoire adéquate.
- Substitution sonore à une information visuelle.

Part II

Dossier administratif

Chapitre 1

CV détaillé

Cyrille Migniot - Maître de Conférences (CNU 27)	
Date de naissance	9 août 1985
Établissement d'enseignement	UFR Sciences et Techniques Département IEM
Laboratoire	ImViA - EA 7535
Adresse	9 avenue Alain Savary, 21078 Dijon
Téléphone	03 80 39 36 92
Adresse mail	cyrille.migniot@u-bourgogne.fr
Site web	https://imvia.u-bourgogne.fr/equipe/cyrille-migniot

Défilé pour les 300 ans de l'université de Bourgogne

Principaux enseignements
- Traitement des images et des vidéos.
- Programmation.
- Systèmes et réseaux.

Thèmes de recherche
- Description et analyse du comportement humain.
- Description du mouvement.
- Interprétation d'un espace 3D.

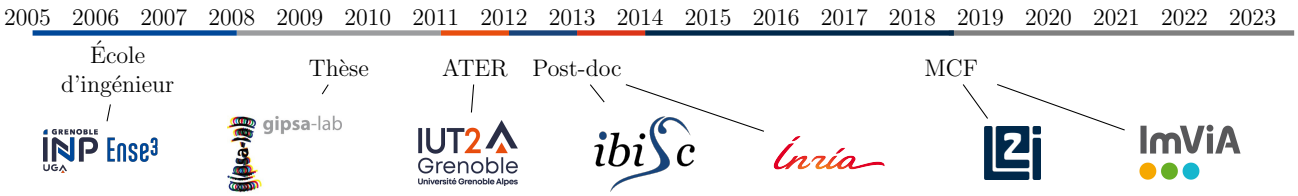
Publications
- 1 manuscrit de thèse
- 7 revues internationales
- 19 conférences internationales et workshops
- 5 conférences nationales

Encadrements
- 3 thèses soutenues, 2 arrêtées et 3 en cours
- 11 stages M2 ou école d'ingénieurs

Prime d'encadrement doctoral et de recherche depuis 2020 (avis B par le CNU 27).

1.1 Parcours

Depuis 2014	**Maître de conférences** section CNU 27. Université Bourgogne Franche-Comté. UFR Sciences et Techniques de Dijon. Laboratoire ImViA EA 7535.
2013-2014	**Post-Doctorat** INRIA Montbonnot. Équipe IMAGINE.
2012-2013	**Post-Doctorat** IBISC, Évry. Équipe IRA2.
2008-2012	**Doctorat** Université de Grenoble. *Segmentation de personnes dans les images et les vidéos.* Soutenue le 17 janvier 2012. Rapporteurs : S. Philipp-Foliguet, J.L. Dillenseger. Examinateurs : F. Brémond, B. Triggs (Président), J.-M. Chassery (dir. de thèse) et P. Bertolino (co-encadrant). Vacations à l'INP Grenoble et à l'IUT1 de Grenoble, ATER à l'IUT2 de Grenoble.
2008	**Diplôme d'Ingénieur** de l'ENSIEG de Grenoble (devenue ENSE3). **Master 2 Recherche** de l'université de Grenoble.

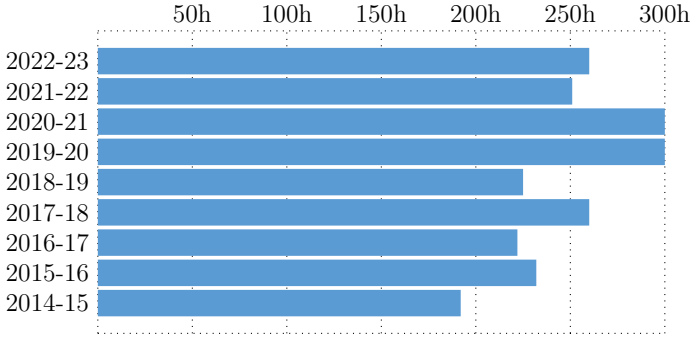


1.2 Activités d'enseignements

1.2.1 Description des modules enseignés

J'enseigne dans le département IEM de l'UFR Sciences et Techniques de Dijon. J'enseigne également un module pour la classe préparatoire de l'école d'ingénieur ESIREM de Dijon.

Année	2014-15	2015-16	2016-17
EqTD	192h	232h	222h
Année	2017-18	2018-19	2019-20
EqTD	260h	225h	300h
Année	2020-21	2021-22	2022-23
EqTD	300h	251h	260h



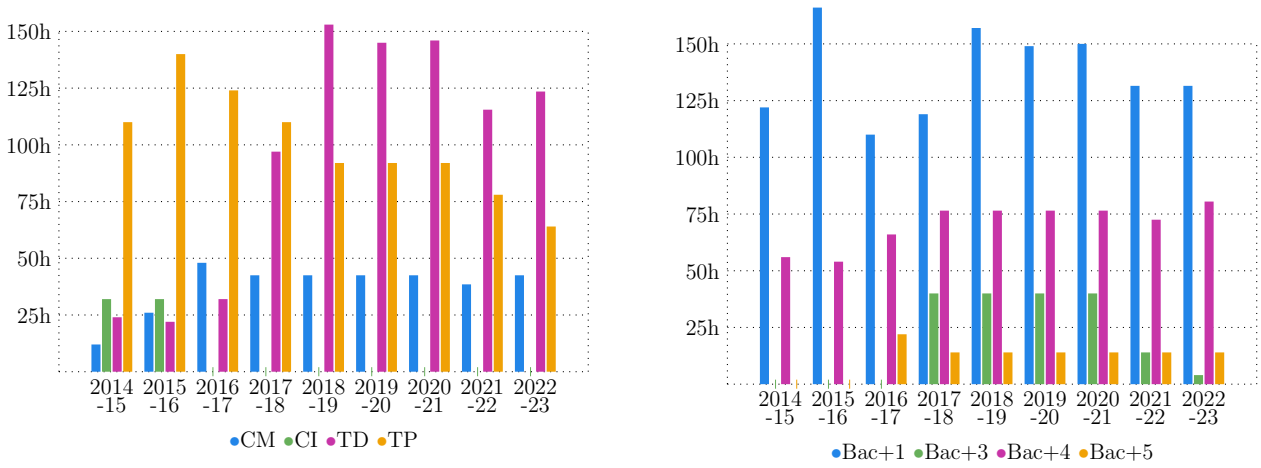


Figure 1.1: Évolution du service d'enseignement (à gauche) et des niveaux d'études de mes enseignements (à droite).

Je donne ci-dessous plus de détails sur les modules dont j'ai été responsable.

Traitement numérique des vidéos

Années : Depuis l'année universitaire 2016-17.
Public : Bac+5 - Master IIA - UFR Sciences et Techniques.
Volume horaire (pour 1 groupe d'étudiants) : 8h CM, 4h TD, 2h TP.
Objectif : Intégrer la temporalité (redondance et cohérence) aux traitements des images pour l'étude des séquences vidéos.
Contenu : Estimation et description du mouvement, compression, suivi, alignement.

Traitement des images

Années : Depuis l'année universitaire 2014-15.
Public : Bac+4 - M1 informatique - UFR Sciences et Techniques.
Volume horaire (pour 1 groupe d'étudiants) : 8,5h CM, 8h TD, 4h TP.
Objectif : Comprendre et manipuler des visions bas niveau et haut niveau du traitement numérique des images.
Contenu : Opérations morphologiques sur images binaires et en niveaux de gris, description de la texture et des contours, méthodes de classification.

Réseaux

Années : Depuis l'année universitaire 2014-15.
Public : Bac+4 - M1 informatique - UFR Sciences et Techniques.
Volume horaire (pour 1 groupe d'étudiants) : 8h CM, 8h TD, 8h TP.
Objectif : Comprendre les différents protocoles des couches réseau et liaison.
Contenu : Protocole IP, routage, détection d'erreurs, liaison à accès multiples.

Systèmes et réseaux

Années : De 2017 à 2023.
Public : Bac+3 - M1 informatique - UFR Sciences et Techniques.
Volume horaire (pour 1 groupe d'étudiants) : 4h CM, 6h TD, 6h TP.
Objectif : Comprendre et manipuler les principaux accès systèmes de Linux.
Contenu : Commandes Unix, processus, système de fichiers, droits d'accès.

Initiation à la programmation C

Années : Depuis l'année universitaire 2015-16.
Public : Bac+1 - classe préparatoire intégrée GEIPI - ESIREM.
Volume horaire (pour 1 groupe d'étudiants) : 14h CM, 17.5h TD, 16h TP.
Objectif : S'initier à la programmation et apprendre le language de programmation C.
Contenu : Création d'un programme, commandes de base, programmation modulaire, introduction vers la POO.

ScIn1B - Sciences de l'Informatique

Années : Depuis l'année universitaire 2018-19.
Public : Bac+1 - L1 AGIL - UFR Sciences et techniques.
Volume horaire (pour 1 groupe d'étudiants) : 24h TD, 24h TP.
Objectif : Découvrir les outils informatiques de l'université et s'initier à la programmation à travers le language Java.
Contenu : Découverte des outils de l'ENT, commandes de base de la programmation Java, création d'un programme.

Voici un récapitulatif des modules sur lesquels je suis intervenu :

Modules	Public	Période	Resp.	Type
Traitements numériques des vidéos	Bac+5 M2 IIA	2016-23	X	CM/TD/TP
Traitement des images	Bac+4 M1	2014-23	X	CM/TD/TP
Réseaux	Bac+4 M1	2014-23	X	CM/TD/TP
Système et réseaux	Bac+3 L3	2017-22	X	CM/TD/TP
Initiation à la programmation C	Bac+1 Prépa	2015-23	X	CM/TD/TP
Programmation Java	Bac+1 L1	2014-17		TD/TP
Programmation HTML	Bac+1 L1	2014-23		TD/TP
Initiation à la programmation Java	Bac+1 L1AGIL	2018-23	X	TD/TP

1.2.2 Responsabilités collectives au sein de la composante d'enseignement

- Responsable des relations internationales du Master 1 informatique d'UFR Sciences et Techniques depuis 2022 (étude des candidatures issues de campus France (218 en 2023), bourse Eiffel, suivi des étudiants en semestre à l'étranger ...).

- Responsable de la L1 AGIL à l'UFR Sciences et Techniques en 2023 sur les parcours Infomatique-Electronique et Physique-Chimie.

- Membre du jury Habilitation ISN en 2015, 2016, 2017, 2018 et 2019.
  Ce jury vise à évaluer le projet réalisé par des professeurs de lycée désirant devenir enseignant en informatique au lycée et ayant suivi la formation ISN au lycée Boivin à Chevigny Saint Sauveur.

- Membre du jury de BTS, spécialité Maintenance des systèmes option A systèmes de production, sur les sessions 2021, 2022 et 2023.

1.3 Activités de recherches

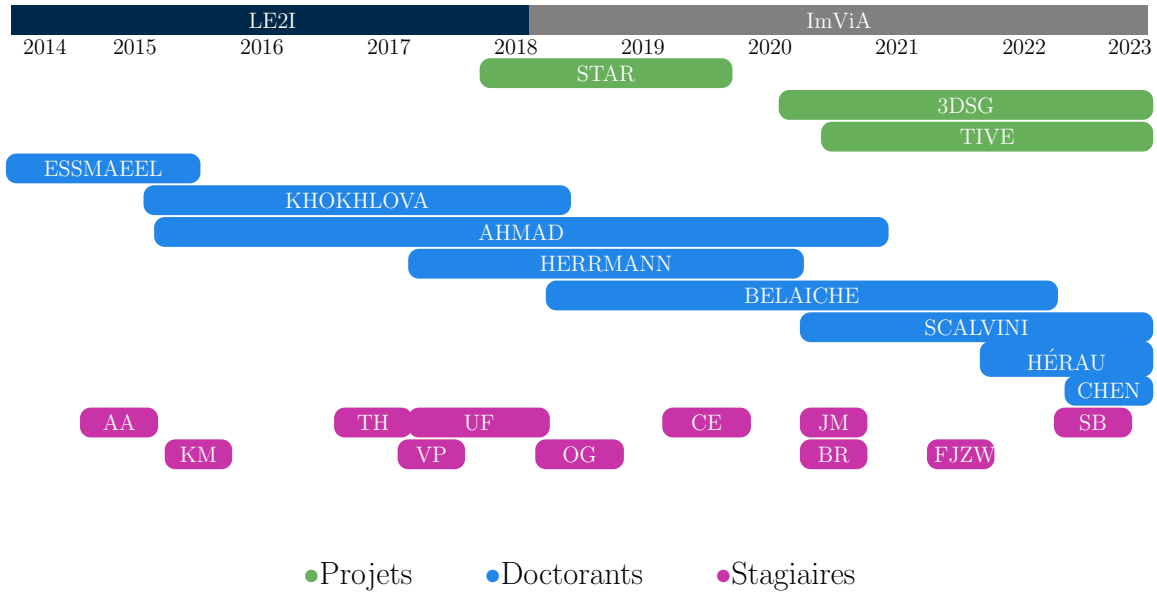


Figure 1.2: Répartition de mes responsabilités depuis mon recrutement.

1.3.1 Projets de recherche

J'ai été porteur ou associé à plusieurs projets de recherche.

TIVE
Rapid
DGA
2021-2024

Identification et reconnaissance assistées par IA de défauts sur des moteurs aéronautiques à partir d'un endoscope industriel.
Consortium : ImViA, LEAD, EFER et SAFRAN.
Responsabilité : responsable du WP (Work Package) 2 : comparaison temps réel de la zone observée à son modèle numérique.

Le projet a pour objectifs de développer la gamme EFER sur les trois axes et les fonctionnalités suivantes :

- amélioration de la représentation des zones inspectées complètes (construction d'images "gigapixels" résultant d'un balayage de zone par défilement de la sonde ou par défilement de la pièce, visualisation/mesure 3D haute résolution);
- assistance à l'inspection et au diagnostic local (corrélation avec des modèles numériques et identification de zones, reconnaissance temps réel et caractérisation de défauts assistée par IA, optimisation du geste opérateur);
- échange et assistance temps réel avec un expert déporté.

3DSG
Envergure
2020-2023

3D Sound Glasses : développement de lunettes électroniques servant de système sonore d'aide à la locomotion des déficients visuels.
Consortium : LEAD et ImViA.
Responsabilité : porteur du côté ImViA → budget 70k€.

Ce projet consiste à développer des lunettes électroniques servant de système sonore d'aide

à la locomotion des déficients visuels. Le caractère interdisciplinaire est constitué d'une part par un développement en Informatique-Electronique réalisé au laboratoire ImViA pour concevoir un système technologique opérationnel et d'autre part par une étude de Psychologie Expérimentale réalisée au LEAD pour valider l'adéquation du système avec les capacités cognitives des utilisateurs. La partie Informatique-Electronique réalise la conception d'algorithmes optimisés pour l'analyse de scènes vidéo en 3D et pour la synthèse sonore ainsi que leur implémentation sur systèmes mobiles et puces électroniques économes en énergie. La partie Psychologie Expérimentale cherche à optimiser l'information sonore fournie par le système pour favoriser son adéquation avec les capacités d'interprétation audio-spatiales de l'utilisateur tout en perturbant a minima son audition naturelle.

PHC STAR n°41603SA 2018-2020

Étude du débruitage de nuage de points pour une représentation plus précise d'une scène acquise par un système multi-kinects.
Consortium : ImViA et Université d'Incheon (Corée du Sud).
Responsabilité : co-chef de projet côté français → budget 24k€.

La carte de profondeur acquise par une caméra de type Kinect possède des trous qui réduisent la quantité d'informations fournie. De plus des points aberrants apparaissent et limitent la qualité des descripteurs. La représentation est inégale en fonction du point de vue (mal représenté si parallèle à l'axe de la prise de vue). L'objectif de ce projet est donc d'éliminer ces phénomènes à partir de post-traitements.

Le deux projets suivants ont été redéposés cette année après avoir été refusés l'année dernière. Le second a été accepté.

LAGUNA ANR 2021-22

Robot d'accompagnement pour les personnes âgées dépendantes atteintes de la maladie d'Alzheimer.
Consortium : KOMPAÏ, CEA, Université de Tours et ImViA.
Responsabilité : attaché aux WP 1, 2 et 4.

AIMOTIONAL ANR 2021-22

Système de vision multimodale alimenté par l'IA pour la reconnaissance automatique des émotions faciales.
Consortium : ImViA, IETR et CSGA.
Responsabilité : co-responsable du WP1.

1.3.2 Encadrements

Étudiants en thèse de doctorat

J'ai jusqu'à maintenant suivi et encadré huit doctorants. Trois ont déjà soutenu. Deux de mes doctorants ont arrêté leur thèse avant de la soutenir. Cette expérience, bien que douloureuse, a été instructive.

Deux de ces thèses ont été réalisées en collaboration avec une entreprise. En effet la thèse de M. Khokhlova est une thèse JCE avec la société PROTEOR située à Dijon. La thèse de Q. Hérau, elle, est un financement CIFRE avec la société Huawei située à Paris.

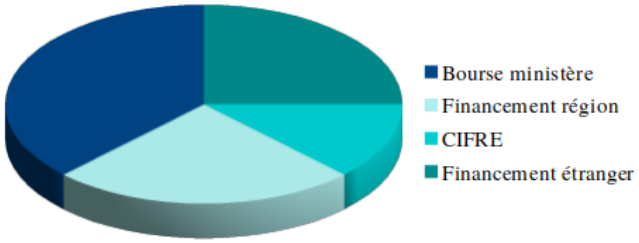


Figure 1.3: Répartition des financements de thèse.

Xiangmei CHEN **encadrement 50%**

Prédiction des performances de la communication acoustique sous-marine basée sur l'apprentissage automatique .
Thèse débutée en décembre 2022.
Encadrée avec F. Yang (ImViA).
Financement : CSC (China Scholarship Council).

Quentin HÉRAU **encadrement 50%**

Cartographie compressée et évolutive pour la navigation à long terme sur support embarqué à partir de données provenant de sources multiples.
Thèse débutée en avril 2022.
Encadrée avec C. Demonceaux (ImViA) et P. Vasseur (MIS).
Financement : CIFRE (Huawei).

Florian SCALVINI **encadrement 40%**

Développement de lunettes électroniques servant de système sonore d'aide à la locomotion des déficients visuels.
Thèse débutée en novembre 2020.
Encadrée avec J. Dubois (ImViA) et M. Ambard (LEAD).
Financement : projet envergure région - 3DSG.

Reda BELAICHE **encadrement 40%**

Étude des émotions par l'analyse temps-réel des micro-expressions basée sur la vision par ordinateur.
Thèse débutée en novembre 2018 et soutenue le 30 septembre 2022.
Encadrée avec F. Yang et D. Ginhac (ImViA).
Financement : bourse ministérielle.

Thomas HERRMANN **encadrement 50%**

Contrôle non destructif par thermographie active dynamique.
Thèse débutée en octobre 2017 et arrêtée en septembre 2020.
Encadrée avec O. Aubreton (ImViA)
Financement : allocation ministérielle.

Ammar AHMAD **encadrement 50%**

Manipulation virtuelle des représentations 3D d'objets culturels.
Thèse débutée en octobre 2015 et arrêtée en mai 2021.
Encadrée avec A. Dipanda (ImViA)
Financement : bourse du gouvernement Syrien.

Margarita KHOKHLOVA **encadrement 50%**

Analyse et suivi des déformations du corps humain à partir d'une acquisition multi-kinect.
Thèse débutée en septembre 2015 et soutenue le 19 novembre 2018.
Encadrée avec A. Dipanda (ImViA)
Financement : bourse région JCE.

Kyis ESSMAEEL **encadrement 50%**

Analyse du mouvement humain par stéréoscopie active : application à la détection de personne dans un système multi-kinects.
Thèse débutée en octobre 2012 et soutenue le 10 décembre 2015.
Encadrée avec A. Dipanda (ImViA)
Financement : bourse ministérielle.

Étudiants en stage/projet M2
Il s'agit principalement d'étudiants du master IIA. Je les ai tous encadrés à 100%.

Ulrick FERRET **octobre 2017 à septembre 2018**

Visualisation de la pose de la main dans un environnement virtuel interactif.

Thomas HERRMANN **mars à août 2017**

Réalisation d'un système hybride à partir d'un capteur infra-rouge et d'un capteur 3D (type kinect).

Karim MOULAY **novembre 2015 à mars 2016**

Descripteur du mouvement humain pour l'étude de la marche par HMM.

Ammar AHMAD **mars à août 2015**

Suivi 3D de personnes à partir d'un nuage de points.

Étudiant en stage de cursus ingénieur
Il s'agit d'un étudiant en dernière année à l'école d'ingénieur ESIREM. Je l'ai encadré à 100%.

Vivien PACEZNY **septembre à janvier 2018**

Suivi du flot optique 3D à partir d'une kinect.

Étudiants en projet long de master
Il s'agit principalement d'étudiants du master IIA. Je les ai tous encadrés à 100%.

Sami BENTEBBICHE **novembre 2022 à mars 2023**

Suivi de la pose de la main vis-à-vis d'un objet.

Faizath-Jedida ZOUMAROU-WALIS **novembre 2021 à mars 2022**

Motif local et temporel pour la détection de micro-expressions.

Baptiste REUNGOAT **novembre 2020 à mars 2021**

Reconnaissance de micro-expressions sur des ROI.

Jordan MERCIER **novembre 2020 à mars 2021**

Estimation de mouvement pour un dispositif d'aide aux personnes non-voyantes.

Ossama GHARBI **octobre à avril 2019**

Nettoyage de nuage de points 3D à partir de méthodes de débruitage 2D et évaluation selon des critères 3D.

Corentin EUVRARD **octobre à avril 2020**

Élaboration d'un critère quantitatif de débruitage pour la calibration.

Projets de groupe long (sur un semestre)

Cycle préparatoire GEIPI	Un projet en programmation par groupe de 3 encadré depuis 2021.
Master 1 UFR ST	De 3 à 4 projets en traitement de l'image par groupe de 4 à 6 encadrés depuis 2018.

1.3.3 Animation et responsabilités scientifiques

Organisation d'événements scientifiques

Workshop HTBA	En 2016, j'ai créé le worshop HTBA (worshop of Human Tracking and Behavior Analysis) qui se tient en périphérie de la conférence SITIS. Depuis j'ai organisé 5 éditions de ce workshop.
HARIMAGE Special Issue	**Human Activity Recognition Based on Image Sensors and Deep Learning** est un special issue de la revue Sensors (ISSN 1424-8220) appartenant à la section Sensing and Imaging. Editeurs : Fakhreddine Ababsa et Cyrille Migniot. Date limite de soumission : 21 mai 2021.
Session PEDR du CNU27	Organisation logistique de la session du 20 au 23 septembre 2021. Création et gestion du site web.

Activité éditoriale

Depuis mon recrutement je participe régulièrement à la relecture d'articles dans des conférences et revues internationales (109 articles reviewés) comme le montre la Figure 1.4.
J'ai aussi rapporté un projet ANR en avril 2023.

Reviewer pour des conférences

- VISAPP (International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications) depuis 2014, chairman en 2020.
- SITIS (International Conference on Signal Image Technology and Internet Based Systems) depuis 2015.
- ICISP (International Conference on Image and Signal Processing) depuis 2016.

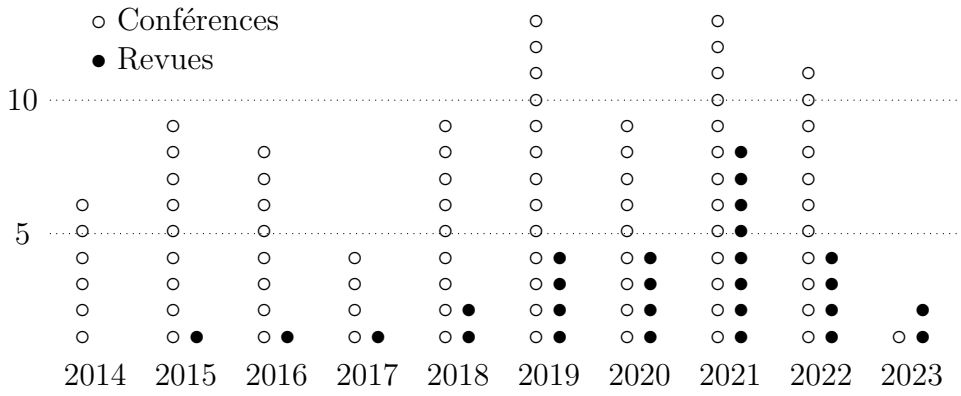


Figure 1.4: Répartition de mes reviews d'articles depuis mon recrutement.

- IROS, ICRA, ICLR ...

Reviewer pour des journaux

- IEEE : Transaction on Cybernetics, Transactions on Multimedia, Transactions in Emerging Topics in Computing, Transactions on Neural Network and Learning Systems.
- Springer : Journal of Multimedia Tools and Applications, The Visual Computer Journal, Journal of Reliable Intelligent Environments, Signal, Image and Video Processing.
- Elsevier : Pattern Recognition Letters, Journal of Systems Architecture.
- MDPI : Sensors, Applied Sciences, Future Internet, Algorithms, Multimodal Technologies and Interaction.
- SPIE : Journal of Applied Remote Sensing, Journal of Electronic Imaging.

1.3.4 Responsabilités pour le laboratoire

Depuis 2018 j'ai pris une double responsabilité au sein du laboratoire ImViA en plus d'une participation active à l'organisation des séminaires scientifiques en soutien de l'équipe animation.

Responsable du bâtiment I3M :
Ce poste m'apporte un rapport privilégié avec une grande partie des chercheurs étudiants et non permanents du laboratoire. Je m'occupe de leur installation, de leur accueil et les guide sur les différents protocoles administratifs et scientifiques qu'ils rencontrent.
Je suis aussi l'intermédiaire entre le laboratoire et le service patrimoine.

Membre de l'équipe communication :
Cette responsabilité s'exprime par de multiples tâches :

- administrateur principale du site web : https://imvia.u-bourgogne.fr/
- référent de la collection HAL imvia

- référent de la collection dat@IMVIA
- référent pour la protection des données personnelles
- administrateur des 16 listes de diffusion du laboratoire
- correspondant GdR-ISIS
- membre du réseau RESCOM
- membre du comité éditorial du journal interne de l'uB

J'ai également été membre de la commission de sélection de deux poste MCF de section 27 :

- Poste $n°0126$ en 2020 affecté à l'ESIREM (Dijon) ainsi qu'au laboratoire ImViA.
- Poste $n°0273$ en 2016 affecté à l'UFR Sciences et Techniques(Auxerre) ainsi qu'au laboratoire Le2i.

1.4 Activités au niveau de l'université

J'ai plusieurs activités au niveau de la communication et des sciences ouvertes pour l'université de Bourgogne.
Je suis en effet membre depuis 2022 du réseau RESCOM visant à discuter et élaborer la communication au niveau de l'université.
Je suis également membre depuis 2023 du comité éditorial du journal interne de l'université de Bourgogne. Ce comité, composé d'une dizaine de membres issus des différentes composantes, a pour vocation de décider du contenu du journal interne mis à disposition des personnels de l'université 3 fois l'an.
Je fais enfin partie des correspondants IST du réseau HAL-UB pour le référencement des publications, le dépôt de texte intégral et la curation des données des collections.

Chapitre 2

Liste des publications

2.1 Articles dans des revues internationales

[Frontiers 2023] Camille Bordeau, Florian Scalvini, Cyrille Migniot, Julien Dubois, Maxime Ambard - *Cross-modal correspondence enhances elevation localization in visual-to-auditory sensory substitution* - Frontiers, Volume 14, 2023

[ApSc 2020] Reda Belaiche, Yu Liu, Cyrille Migniot, Dominique Ginhac and Fan Yang - *Cost-effective CNNs for real-time Micro-Expression recognition* - Applied Sciences, MDPI, 2020

[IVC 2019] Ammar Ahmad, Cyrille Migniot, Albert Dipanda - *Hand Pose Estimation and Tracking in Real and Virtual Interaction : A Review* - Image and Vision Computing, Elsevier, Volume 89, Pages 35-49, 2019

[MTAP 2019] Kyis Essmaeel, Cyrille Migniot, Albert Dipanda, Luigi Gallo, Ernesto Damiani, Guiseppe De Pietro - *A New 3D Descriptor For Human Classification: Application For Human Detection in a Multi-Kinect System* - Multimedia Tools and Applications, Springer, Volume 78, Pages 22479–22508, 2019

[AIM 2019] Margarita Khokhlova, Cyrille Migniot, Alexei Morozov, Olga Sushkova, Albert Dipanda - *Normal and pathological gait classification LSTM model* - Artificial Intelligence in Medicine Volume 94, Pages 54-66 , 2019

[MTAP 2018] Margarita Khokhlova, Cyrille Migniot, Albert Dipanda - *Advances in Description of 3D Human Motion* - Multimedia Tools and Applications, Springer, Volume 77, Pages 31665–31691 2018

[JRTIP 2014] Cyrille Migniot, Fakhr-Eddine Ababsa - *Hybrid 3D-2D human tracking in a top view* - Journal of Real-Time Image Processing, Springer Verlag (Germany), Volume 11, Pages 769-784, 2014

2.2 Conférences internationales et workshops

[IROS 2023] Quentin Herau, Nathan Piasco, Moussab Bennehar, Luis Roldao, Dzmitry Tsishkou, Cyrille Migniot, Pascal Vasseur and Cédric Demonceaux - *MOISST: Multi-modal Optimization of Implicit Scene for SpatioTemporal calibration* - International Conference on Intelligent Robots and Systems, 2023 (accepté).

[SITIS 2022] Florian Scalvini, Camille Bordeau, Maxime Ambard, Cyrille Migniot, Stéphane Argon and Julien Dubois - *Visual-auditory substitution device for indoor navigation based on fast visual marker detection* - International Conference on Signal Image technology & Internet Based Systems, 2022.

[ICASSP 2022] Florian Scalvini, Camille Bordeau, Maxime Ambard, Cyrille Migniot and Julien Dubois - *Low-Latency Human-Compuetr Auditory Interface Based On Real-Time Vision Analysis* - International Conference on Acoustics, Speech, & Signal Processing, 2022.

[QIRT 2020] Thomas Herrmann, Cyrille Migniot and Olivier Aubreton - *Thermal camera calibration with cooled down chessboard* - Quantitative InfraRed Thermography Conference, 2020.

[ICIAP 2019] Reda Belaïche, Rita Meziati Sabour, Cyrille Migniot, Yannick Benezeth, Dominique Ginhac, Keisuke Nakamura, Randy Gomez and Fan Yang - *Emotional State Recognition with Micro-Expressions and Pulse Rate Variability* - International Conference on Image Analysis and Processing, 2019.

[SITIS 2019] Reda Belaïche, Cyrille Migniot, Dominique Ginhac and Fan Yang - *Time Unification on Local Binary Patterns Three Orthogonal Planes for Facial Expression Recognition* - International Conference on Signal Image technology & Internet Based Systems, 2019.

[QCAV 2019] Thomas Herrmann, Cyrille Migniot, Olivier Aubreton - *Cracks Detection on Glass Object based on Active Thermography Approach* - Quality Control for Artificial Vision, 2019.

[SITIS 2018] Margarita Khokhlova, Cyrille Migniot, Albert Dipanda - *Kinematic Covariance Based Abnormal Gait Detection* - International Conference on Signal Image technology & Internet Based Systems, 2018.

[VISAPP 2018] Margarita Khokhlova, Cyrille Migniot, Albert Dipanda - *3D Point Cloud Descriptor for Posture Recognition* - International Conference on Computer Vision Theory and Applications, 2018.

[SITIS 2017] Ammar Ahmad, Cyrille Migniot, Albert Dipanda - *Tracking Hands in Interaction with Objects: A Review* - International Conference on Signal Image technology & Internet Based Systems, 2017.

[SITIS 2016] Margarita Khokhlova, Cyrille Migniot, Albert Dipanda - *3D Visual-based Human Motion Descriptors: A Review* - International Conference on Signal Image technology & Internet Based Systems, 2016.

[VISAPP 2016] Kyis Essmaeel, Cyrille Migniot, Albert Dipanda - *3D Descriptor for an Oriented-Human Classification from Complete Point Cloud* - International Conference on Computer Vision Theory and Applications, 2016.

[CuHe 2015] Rémi Ronfard, Vineet Gandhi, Benoit Encelle, Pierre-Antoine Champin, Thomas Steiner, Nicolas Sauret, Cyrille Migniot - *Capturing and Indexing Rehearsals: The Design and Usage of a Digital Archive of Performing Arts* - Cultural Heritage, 2015.

[VISAPP 2014] Cyrille Migniot, Fakhr-Eddine Ababsa - *Part-based 3D multi-person tracking using depth cue in a top view* - International Conference on Computer Vision Theory and Applications, 2014.

[ISVC2013] Cyrille Migniot, Fakhr-Eddine Ababsa - *3D Human Tracking in a Top View Using Depth Information Recorded by the Xtion Pro-Live Camera* - International Symposium on Visual Computing, 2013.

[CAIP 2013] Cyrille Migniot, Fakhr-Eddine Ababsa - *3D Human Tracking from Depth Cue in a Buying Behavior Analysis Context* - International Conference on Computer Analysis of Images and Patterns, 2013.

[VISAPP 2013] Cyrille Migniot, Pascal Bertolino, Jean-Marc Chassery - *Iterative Human Segmentation from Detection Windows Using Contour Segment Analysis* - International Conference on Computer Vision Theory and Applications, 2013.

[ICIP 2011] Cyrille Migniot, Pascal Bertolino, Jean-Marc Chassery - *Automatic people segmentation with a template-driven graph cut* - International Conference on Image Processing, 2011.

[VISAPP 2010] Cyrille Migniot, Pascal Bertolino, Jean-Marc Chassery - *Contour segment analysis for human silhouette pre-segmentation* - International Conference on Computer Vision Theory and Applications, 2010.

2.3 Conférences nationales

[COMPAS 2023 1] Florian Scalvini, Camille Bordeau, Maxime Amabrd, Cyrille Migniot, Julien Dubois - *Système d'assistance à la mobilité en milieu urbain des personnes malvoyantes via une substitution de l'information visuelle par un signal auditif* - Compas : Conférence francophone d'informatique en Parallélisme, Architecture et Système, 2023

[COMPAS 2023 2] Sean Marotta, Alessandro Carlini, Vincent Brost, Cyrille Migniot, Manon Ansart, Michel Paindavoine, Julien Dubois - *Prototypage rapide pour l'inférence par réseaux de neurones convolutifs sur cibles matérielles hétérogènes* - Compas : Conférence francophone d'informatique en Parallélisme, Architecture et Système, 2023

[RENSIT 2018] Alexei A. Morozaov, Olga S. Sushkova, Margarita Khokhlova, Cyrille Migniot - *Development of agent logic programming means for multichannel intelligent video surveillance* - RENSIT: Radioelectronics. Nanosystems. Information technologies V. 10, No. 1. – p. 101-116, 2018

[ORASIS 2015] Kyis Essmaeel, Cyrille Migniot, Albert Dipanda - *Une nouvelle approche de classification de personnes à partir d'une plate-forme multi-kinect* - Journées francophones des jeunes chercheurs en vision par ordinateur, 2015.

[GRETSI 2011] Cyrille Migniot, Pascal Bertolino, Jean-Marc Chassery - *Segmentation automatique de personnes par coupe de graphe et gabarits* - colloque GRETSI, 2011.

Références

[1] E. Ohn-Bar and M. M. Trivedi, "A comparative study of color and depth features for hand gesture recognition in naturalistic driving settings," in *Intelligent Vehicles Symposium*, pp. 845–850, 2015.

[2] M. Jaimez, M. Souiai, J. Gonzalez-Jimenez, and D. Cremers, "A primal-dual framework for real-time dense rgb-d scene flow," *International Conference on Robotics and Automation*, pp. 98–104, 2015.

[3] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[4] M. Devanne, H. Wannous, S. Berretti, P. Pala, and M. Daoudi, "Objective classes for micro-facial expression recognition," *International Conference on Image Analysis and Processing*, no. 10, pp. 456–464, 2013.

[5] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Conference on Computer Vision and Pattern Recognition*, pp. 1290–1297, 2012.

[6] J. Zheng and Z. Jiang, "Learning view-invariant sparse representations for cross-view action recognition," in *International Conference on Computer Vision*, pp. 3176–3183, 2013.

[7] W. Ding, K. Liu, F. Cheng, and J. Zhang, "STFC: spatio-temporal feature chain for skeleton-based human action recognition," *Journal of Vision Communication Image Represention*, vol. 26, pp. 329–337, 2015.

[8] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Conference on Computer Vision and Pattern Recognition*, pp. 588–595, 2014.

[9] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 20–27, 2012.

[10] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *Conference on Computer Vision and Pattern Recognition*, pp. 716–723, 2013.

[11] S. Hadfield and R. Bowden, "Kinecting the dots: Particle based scene flow from depth sensors," in *International Conference on Computer Vision*, 2011.

[12] S. Hadfield, K. Lebeda, and R. Bowden, "Natural action recognition using invariant 3d motion encoding," *European Conference on Computer Vision*, pp. 758–771, 2014.

[13] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in *Conference on Computer Vision and Pattern Recognition*, pp. 804–811, 2014.

[14] P. Chattopadhyay, S. Sural, and J. Mukherjee, "Frontal gait recognition from incomplete sequences using rgb-d camera," *Transactions on Information Forensics and Security*, vol. 9, no. 11, pp. 1843–1856, 2014.

[15] H. Zhang, C. Reardon, C. Zhang, and L. E. Parker, "Adaptive human-centered representation for activity recognition of multiple individuals from 3d point cloud sequences," *International Conference on Robotics and Automation*, pp. 1991–1998, 2015.

[16] Y. Xiao, G. Zhao, J. Yuan, and D. Thalmann, "Activity recognition in unconstrained rgb-d video using 3d trajectories," in *Autonomous Virtual Humans and Social Robot for Telepresence*, 2014.

[17] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 9–14, 2010.

[18] M. Munaro, S. Michieletto, and E. Menegatti, "An evaluation of 3d motion flow and 3d pose estimation for human action recognition," *RGB-D: Advanced Reasoning with Depth Cameras*, 2013.

[19] Z. Gao, S. Li, Y. Zhu, C. Wang, and H. Zhang, "Collaborative sparse representation leaning model for rgbd action recognition," *Journal of Visual Communication and Image Representation*, vol. 48, pp. 442–452, 2017.

[20] H. Zhang and L. E. Parker, "Code4d: Color-depth local spatio-temporal features for human activity recognition from rgb-d videos," *Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 3, pp. 541–555, 2016.

[21] H. Zhang, P. Zhong, J. He, and C. Xia, "Combining depth-skeleton feature with sparse coding for action recognition," *Neurocomputing*, vol. 230, pp. 417–426, 2017.

[22] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "Rgb-d-based human motion recognition with deep learning: A survey," *Computer Vision and Image Understanding*, vol. 171, pp. 118–139, 2018.

[23] X. Jiang, K. Xu, and T. Sun, "Action recognition scheme based on skeleton representation with ds-lstm network," *Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 2129–2140, 2020.

[24] A. Banerjee, P. K. Singh, and R. Sarkar, "Fuzzy integral-based cnn classifier fusion for 3d skeleton action recognition," *Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 6, pp. 2206–2216, 2021.

[25] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Conference on Computer Vision and Pattern Recognition*, pp. 180–189, 2020.

[26] A. Sarkar, A. Banerjee, P. K. Singh, and R. Sarkar, "3d human action recognition: Through the eyes of researchers," *Expert Systems with Applications*, vol. 193, p. 116424, 2022.

[27] J. Lee and B. Ahn, "Real-time human action recognition with a low-cost rgb camera and mobile robot platform," *Sensors*, vol. 20, no. 10, 2020.

[28] X. Qin, Y. Ge, J. Feng, D. Yang, F. Chen, S. Huang, and L. Xu, "Dtmmn: Deep transfer multi-metric network for rgb-d action recognition," *Neurocomputing*, vol. 406, pp. 127–134, 2020.

[29] Z. Ren, J. Yuan, and Z. Zhang, "Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera," in *International Conference on Multimedia*, p. 1093–1096, 2011.

[30] P. Suryanarayan, A. Subramanian, and D. Mandalapu, "Dynamic hand pose recognition using depth data," in *International Conference on Pattern Recognition*, 2010.

[31] H. Cheng, Z. Dai, Z. Liu, and Y. Zhao, "An image-to-class dynamic time warping approach for both 3d static and trajectory hand gesture recognition," *Pattern Recognition*, vol. 55, pp. 137–147, 2016.

[32] P. Cirujeda and X. Binefa, "4dcov: A nested covariance descriptor of spatio-temporal features for gesture recognition in depth sequences," in *2014 2nd International Conference on 3D Vision*, vol. 1, pp. 657–664, 2014.

[33] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1583–1597, 2016.

[34] K. Lyu, H. Chen, Z. Liu, B. Zhang, and R. Wang, "3d human motion prediction: A survey," *Neurocomputing*, vol. 489, pp. 345–365, 2022.

[35] X. S. Papageorgiou, G. Chalvatzaki, C. S. Tzafestas, and P. Maragos, "Hidden markov modeling of human pathological gait using laser range finder for an assisted living intelligent robotic walker," in *International Conference on Intelligent Robots and Systems*, pp. 6342–6347, 2015.

[36] M. Belghali, N. Chastan, F. Cignetti, D. Davenne, and L. Decker, "Loss of gait control assessed by cognitive-motor dual-tasks: pros and cons in detecting people at risk of developing alzheimer's and parkinson's diseases," *Geroscience*, 2017.

[37] B. Kwolek, T. Krzeszowski, A. Michalczuk, and H. Josinski, "3d gait recognition using spatio-temporal motion descriptors," in *Intelligent Information and Database Systems*, pp. 595–604, 2014.

[38] J. Tang, J. Luo, T. Tjahjadi, and Y. Gao, "2.5d multi-view gait recognition based on point cloud registration," *Sensors*, vol. 14, no. 4, pp. 6124–6143, 2014.

[39] M. Alotaibi and A. Mahmood, "Automatic real time gait recognition based on spatiotemporal templates," in *Long Island Systems, Applications and Technology*, pp. 1–5, 2015.

[40] J. Han and B. Bhanu, "Individual recognition using gait energy image," *Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316–322, 2006.

[41] P. Tosranon, A. Sanpanich, C. Bunluechokchai, and C. Pintavirooj, "Gaussian curvature-based geometric invariance," in *International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, vol. 02, pp. 1124–1127, 2009.

[42] C. D. Lim, C.-Y. Cheng, C.-M. Wang, Y. Chao, and L.-C. Fu, "Depth image based gait tracking and analysis via robotic walker," in *International Conference on Robotics and Automation*, pp. 5916–5921, 2015.

[43] R. R. Drumond, B. A. D. Marques, C. N. Vasconcelos, and E. Clua, "Peek - an lstm recurrent network for motion classification from sparse data," *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pp. 215–222, 2018.

[44] J. Gu, X. Ding, S. Wang, and Y. Wu, "Action and gait recognition from recovered 3-d human joints," *Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 40, no. 4, pp. 1021–1033, 2010.

[45] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 11 1997.

[46] B. Mahasseni and S. Todorovic, "Regularizing long short term memory with 3d human-skeleton sequences for action recognition," in *Conference on Computer Vision and Pattern Recognition*, pp. 3054–3062, 2016.

[47] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. M. Campos, "Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pp. 252–259, 2012.

[48] A. Kale, A. Rajagopalan, N. Cuntoor, and V. Kruger, "Gait-based recognition of humans using continuous hmms," in *International Conference on Automatic Face Gesture Recognition*, pp. 336–341, 2002.

[49] M. Hu, Y. Wang, Z. Zhang, D. Zhang, and J. J. Little, "Incremental learning for video-based gait recognition with lbp flow," *IEEE Transactions on Cybernetics*, vol. 43, no. 1, pp. 77–89, 2013.

[50] A. Kolawole and A. Tavakkoli, "A novel gait recognition system based on hidden markov models," in *Advances in Visual Computing*, pp. 125–134, 2012.

[51] A. Paiement, L. Tao, M. Camplani, S. Hannuna, D. Damen, and M. Mirmehdi, "Online quality assessment of human motion from skeleton data," in *British Machine Vision Conference*, 2014.

[52] H. Tian, X. Ma, H. Wu, and Y. Li, "Skeleton-based abnormal gait recognition with spatio-temporal attention enhanced gait-structural graph convolutional networks," *Neurocomputing*, vol. 473, pp. 116–126, 2022.

[53] M. Kumar, N. Singh, R. Kumar, S. Goel, and K. Kumar, "Gait recognition based on vision systems: A systematic survey," *Journal of Visual Communication and Image Representation*, vol. 75, p. 103052, 2021.

[54] M. N. Favorskaya and V. V. Buryachenko, "Monocular depth-based visual tracker for gait recognition," *Procedia Computer Science*, vol. 207, pp. 205–214, 2022.

[55] E. A. Haggard and K. Isaacs, "Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy," in *Methods of Research in Psychotherapy*, 1966.

[56] R. Belaiche, *Analyse temps reel des micro-expressions par vision artificielle*. PhD thesis, Université of Burgundy, 2022.

[57] D. Patel, X. Hong, and G. Zhao, "Selective deep features for micro-expression recognition," in *International Conference on Pattern Recognition*, pp. 2258–2263, 2016.

[58] S. Wang, B. jun Li, Y. Liu, W.-J. Yan, X. Ou, X. Huang, F. Xu, and X. Fu, "Micro-expression recognition with small sample size by transferring long-term convolutional neural network," *Neurocomputing*, vol. 312, pp. 251–262, 2018.

[59] S.-T. Liong, Y. Gan, W.-C. Yau, Y.-C. Huang, and T. Ken, "Off-apexnet on micro-expression recognition system," *Signal Proc. Image Comm.*, 2019.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[61] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations,*, 2015.

[62] W. Yan, X. Li, S. Wang, G. Zhao, Y. Liu, Y. Chen, and X. Fu, "Casmeii: an improved spontaneous micro-expression database and the baseline evaluation," *PLOS One*, vol. 9, pp. 1–8, 2014.

[63] J. Li, C. Soladie, and R. Seguier, "Local temporal pattern and data augmentationfor micro-expression spotting," *Affective Computing, Institute of Electrical andElectronics Engineers,*, 2020.

[64] T.-N. Nguyen and J. Meunier, "Walking gait dataset : point clouds , skeletons and silhouettes technical report number 1379," 2018.

[65] A. Chaaraoui, J. Padilla-Lopez, and F. Lorez-Revuelta, "Abnormal gait detection with rgb-d devices using joint motion history features," *International Conference and Workshops on Automatic Face and Gesture Recognition*, vol. 7, pp. 1–6, 2015.

[66] T.-N. Nguyen, H.-H. Huynh, and J. Meunier, "Skeleton-based abnormal gait detection," *Sensors*, vol. 16, no. 11, 2016.

[67] M. Khokhlova, *Évaluation clinique de la démarche à partir de données 3D*. PhD thesis, Université of Burgundy, 2018.

[68] F. Ahmed, P. P. Paul, and M. L. Gavrila, "Dtw-based kernel and rank-level fusion for 3d gait recognition using kinect," in *The Visual Computer*, vol. 31, pp. 915–924, 2015.

[69] Q. Li, Y. Wang, A. Sharf, Y. Cao, C. Tu, B. Chen, and S. Yu, "Classification of gait anomalies from kinect," *The Visual Computer*, vol. 34, pp. 1–13, 02 2018.

[70] L. Gond, P. Sayd, T. Chateau, and M. Dhome, "A 3d shape descriptor for human pose recovery," in *Articulated Motion and Deformable Objects*, pp. 370–379, 2008.

[71] A. Klaser, M. Marszalek, and C. Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients," *British Machine Vision Conference*, pp. 275:1–10, 2008.

[72] M. Munaro, F. Basso, and E. Menegatti, "Tracking people within groups with rgb-d data," *International Conference on Intelligent Robots and Systems.*, pp. 2101–2107, 10 2012.

[73] A. Constantinescu, K. Müller, M. Haurilet, V. Petrausch, and R. Stiefelhagen, "Bring the Environment to Life: A Sonification Module for People with Visual Impairments to Improve Situation Awareness," in *International Conference on Multimodal Interaction*, pp. 50–59, ACM, 2020.

[74] G. Bologna, B. Deville, T. Pun, and M. Vinckenbosch, "Transforming 3D Coloured Pixels into Musical Instrument Notes for Vision Substitution Applications," *EURASIP Journal on Image and Video Processing*, vol. 2007, pp. 1–14, 2007.

[75] A.-A. Tulbure and E.-H. Dulf, "A review on modern defect detection models using DCNNs – Deep convolutional neural networks," *Journal of Advanced Research*, 2021.

[76] M. Ambard, Y. Benezeth, and P. Pfister, "Mobile Video-to-Audio Transducer and Motion Detection for Sensory Substitution," *Frontiers in information and communication technologies*, vol. 2, 2015.

[77] H. Wierstorf, M. Geier, and S. Spors, "A free database of head related impulse response measurements in the horizontal plane with multiple distances," *Journal of the audio engineering society*, 2011.

[78] M. Ambard, "Software Design for Low-Latency Visuo-Auditory Sensory Substitution on Mobile Devices," *Computer and Information Science*, vol. 10, no. 2, p. 1, 2017.

[79] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," *European Conference on Computer Vision*, pp. 740–755, 2015.

[80] B. Benligiray, C. Topal, and C. Akinlar, "Stag: A stable fiducial marker system," *Image and Vision Computing*, vol. 89, pp. 158–169, Sep 2019.

[81] V. Ortega-Gonzalez, S. Garbaya, and F. Merienne, "Using 3d sound for providing 3d interaction in virtual environment," *World Conference on Innovative Virtual Reality*, pp. 311–321, 2010.

[82] R. Canales and S. Jörg, "Performance is not everything: Audio feedback preferred over visual feedback for grasping task in virtual reality," *Motion, Interaction and Games*, 2020.

[83] N. Cooper, F. Milella, C. Pinto, I. Cant, M. White, and G. Meyer, "The effects of substitute multisensory feedback on task performance and the sense of presence in a virtual reality environment.," *PLOS ONE 13*, vol. 2, pp. 1–25, 2018.

[84] S. Gautam, K. Sivaraman, H. Muralidharan, and A. Baskar, "Vision system with audio feedback to assist visually impaired to grasp objects," *Procedia Computer Science*, vol. 58, pp. 387–394, 2015.