PROJET DE FIN D'ÉTUDES

RÉ-IDENTIFICATION DE PERSONNES À PARTIR D'UN RÉSEAU DE CAMÉRAS RGB-D

15 septembre 2014

Guilhem Marion
IMA 2014

Table des matières

1. Introduction, contexte et motivations

Ce stage de fin d'études d'une durée de six mois (du 17 Mars au 17 Septembre 2014) a été effectué dans le cadre de mon double diplôme d'ingénieur ENSEEIHT (option Informatique et Mathématiques Appliquées) auquel est adjoint un Master de Recherche en Informatique et Télécommunication, option Recherche d'information, Bases de Données et Multimédia (IT-RIBDM).

La totalité du stage s'est déroulée au laboratoire LAAS-CNRS (pour Laboratoire d'Analyse et d'Architecture des Systèmes) de Toulouse. Après avoir présenté le CNRS, le LAAS et l'équipe RAP dont j'ai fait partie en tant que stagiaire, on s'intéressera aux motivations du stage puis sa réalisation sera présentée dans les trois chapitres suivants.

1.1. Contexte du stage

Le CNRS (Centre national de Recherche Scientifique) est un organisme public français de recherche fondamentale (sous la tutelle du Ministère chargé de la Recherche). Son président est Alain Fuchs, qui est assisté de deux directeurs généraux délégués : Philippe Baptiste à la science et Xavier Inglebert aux ressources. Fort de près de 33 000 employés, le CNRS exerce son activité dans tous les champs de la connaissance, en s'appuyant sur plus de 1100 unités de recherche et de service. Principal organisme de recherche à caractère pluridisciplinaire en France, le CNRS mène des recherches dans l'ensemble des domaines scientifiques, technologiques et sociétaux. Il couvre la totalité de la palette des champs scientifiques, qu'il s'agisse des mathématiques, de la physique, des sciences et technologies de l'information et de la communication, de la physique nucléaire et des hautes énergies, des sciences de la planète et de l'Univers, de la chimie, des sciences du vivant, des sciences humaines et sociales, des sciences de l'environnement ou des sciences de l'ingénierie.

Le CNRS est présent dans toutes les disciplines majeures regroupées au sein de dix instituts dont trois sont nationaux :

- Institut des sciences biologiques (INSB)
- Institut de chimie (INC)
- Institut écologie et environnement (INEE)
- Institut des sciences humaines et sociales (INSHS)
- Institut des sciences de l'information et de leurs interactions (INS2I)
- Institut des sciences de l'ingénierie et des systèmes (INSIS)
- Institut national des sciences mathématiques et de leurs interactions (INSMI)
- Institut de physique (INP)
- Institut national de physique nucléaire et physique des particules (IN2P3)

— Institut national des sciences de l'univers (INSU)

Le CNRS cherche à favoriser la collaboration entre spécialistes de différentes disciplines, notamment dans le milieu universitaire. Cela permet d'ouvrir le champ de la recherche scientifique à de nouveaux horizons stratégiques vis-à-vis des besoins de l'économie et de la société.

1.1.1. La Délégation Midi-Pyrénées

Le CNRS est organisé en 19 délégations couvrant le territoire national. Elles assurent une gestion directe et locale des laboratoires et entretiennent les liens avec les partenaires universitaires régionaux, les autres établissements publics à caractère scientifique et technique (EPST) et les collectivités locales.

La Délégation Midi-Pyrénées couvre 8 départements : Ariège, Aveyron, Gers, Haute-Garonne, Hautes-Pyrénées, Lot, Tarn, Tarn et Garonne. Elle est dirigé par le délégué régional Patrick MOUNAUD.

1.1.2. Le Laboratoire d'Analyse et d'Architecture des Systèmes

Le Laboratoire d'Analyse et d'Architecture des Systèmes (LAAS-CNRS) est un laboratoire du CNRS rattaché à l'Institut des Sciences de l'Ingénierie et des Systèmes (INSIS) et à l'Institut des Sciences de l'Information et de leurs Interactions (INS2I). Le LAAS est actuellement dirigé par Jean ARLAT. Situé à Toulouse, il est associé par convention à cinq membres fondateurs de la COMUE (Communauté d'Universités et d'Établissements) « Université de Toulouse » qui comprend :

— l'Université Paul Sabatier (UPS),
— l'Institut National des Sciences Appliquées de Toulouse (INSA),
— l'Institut National Polytechnique de Toulouse (INP),
— l'Université du Mirail (UTM),
— l'Université Toulouse 1 Capitole (UT1).

Le LAAS mène par ailleurs des recherches en sciences et technologies de l'information, de la communication et des systèmes dans 8 thèmes :

— Informatique critique
— Réseaux et communications
— Robotique
— Décision et optimisation
— HF et optique : de l'EM aux systèmes
— Nano ingénierie et intégration
— MicroNanoBioTechnologies

— et Gestion de l'énergie

Le LAAS-CNRS traite de sujets en amont et en lien avec des centres d'intérêt du monde socio-économique en vue d'applications futures. Le LAAS-CNRS est Institut Carnot, label qui souligne la qualité et la pertinence de ces travaux relativement aux enjeux socio-économiques et est membre de l'Association des Instituts Carnot.

1.1.3. L'équipe Robotique, Action et Perception

Les travaux du groupe Robotique, Action et Perception (RAP) sont dans la lignée des recherches menées au LAAS-CNRS depuis de nombreuses années sur l'autonomie fonctionnelle et décisionnelle des systèmes robotiques devant intervenir dans des environnements dynamiques et/ou évolutifs. Au sein du pôle Robotique, le groupe RAP étudie essentiellement la couche fonctionnelle de ces systèmes robotiques : robots d'intérieur ou d'extérieur équipés généralement de plusieurs capteurs, ou robot humanoïde, équipé de capteurs visuels et sonores. Nous nous intéressons à l'intégration de fonctions évoluées dans un robot physique, mais aussi aux interactions entre Hommes, robots et environnement pour tirer parti d'une Intelligence Ambiante, et enfin aux applications des technologies de la robotiques à d'autres domaines.

Dans ce contexte, les travaux du groupe RAP portent sur quatre thématiques :

— Perception visio-auditive pour la reconnaissance d'activités humaines
— Perception pour l'exécution de tâches sur un robot
— Modélisation et commande de systèmes robotiques complexes
— Capteurs intégrés communicants

1.2. Motivations

Lorsqu'il faut qu'un système robotique (ou plus simplement informatique) prenne une décision relativement à son environnement, il est souvent nécessaire d'agir en fonction de et avec les humains présents. A cet effet, la vision par ordinateur offre des outils de détection et de suivi d'êtres humains, ainsi que d'estimation de leur pose (c'est-à-dire de la position et de l'orientation de certaines de leur parties corporelles).

Malgré qu'il existe des travaux sur la reconstruction de posture mono-caméra, le passage à la reconstruction de posture multi-caméra demande une stratégie de collaboration : il est nécessaire d'associer les données captées par chacune des caméras entre-elles. La littérature à ce sujet est assez dense comme on pourra le voir au chapitre 2.

Dans le contexte de mon stage j'ai assisté Jean-Thomas Masse, doctorant au LAAS dans l'équipe de Frédéric Lerasle, pour le développement de la démonstration de son travail de fin de thèse dont le fonctionnement est décrit dans [Masse et al., 2013]. Celui-ci, consiste en l'estimation robuste de la pose d'humains à partir des poses estimées par plusieurs capteurs RGB-D, et requiert l'association des différentes poses détectées entre-elles (il ne faut par exemple pas associer la pose d'un utilisateur avec

celle d'un autre utilisateur). Pour cela il est nécessaire de faire correspondre les détections de chacune des caméras entre elles.

Comme nous disposions des informations de couleur et de profondeur issues du ou des capteurs RGB-D, il nous a paru pertinent de tenter de les utiliser au maximum pour cette étape de ré-identification. Malheureusement, si les techniques de ré-identification à partir d'images RGB sont nombreuses il n'existe aujourd'hui que peu de méthodes se servant de l'information de profondeur, ce qui a motivé la recherche d'un processus qui combinerait le meilleur des deux mondes : la robustesse des méthodes basées RGB qui bénéficient d'années de recherche, ainsi que les informations supplémentaires de profondeur et de pose offertes par les capteurs RGB-D. De plus, l'utilisation d'informations de profondeur a un avantage certain lorsque les conditions d'illumination ne sont pas contrôlées : alors que l'apparence d'une personne telle que captée par une caméra RGB varie fortement en fonction des conditions d'illumination, la carte de profondeur n'est pas affectée par ce phénomène.

1.3. Début du stage

En guise d'introduction au projet pour lequel j'ai travaillé, la première partie du stage a consisté au port du programme développé par Jean-Thomas Masse dans le cadre de sa thèse depuis la plateforme ROS (très répandue, qui permet de faire fonctionner ces programmes sur un grand nombre de systèmes robotisés) vers la plateforme Genom3 développée par le LAAS.

Le LAAS disposant d'un appartement expérimental (dans le bâtiment ADREAM) au plafond duquel sont placées des caméras RGB-D qui permettent l'observation et la supervision des robots et des humains évoluant dans l'appartement. L'utilisation de Genom3 facilite l'intégration avec le logiciel de supervision installé (et donc l'observation des résultats du programme en situation et en temps réels et de manière agréable pour l'utilisateur) : c'est ce qui a motivé ce portage.

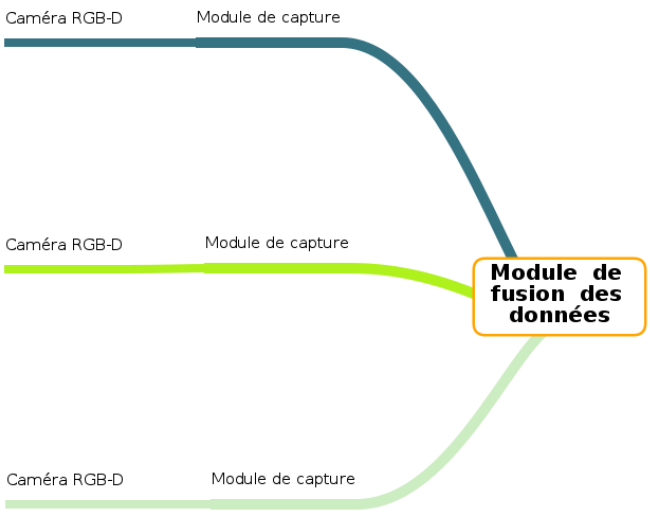

FIGURE 1 – Illustration du fonctionnement du programme original

Le programme d'origine contient deux parties : la première se charge de récupérer les informations d'une caméra RGB-D et estime la pose de l'utilisateur s'il se trouve devant la caméra. Ce module est

lancé un nombre de fois égal au nombre de caméras dont on dispose. Le second module reçoit les informations envoyées par les multiples instances du premier (image RGB + carte de profondeur + estimation éventuelle du squelette d'utilisateur et position de l'utilisateur sur l'image) et se charge d'estimer une pose plus probable. Après m'être fait présenter et expliquer le code par Jean-Thomas Masse, la décision a été prise de tenter de minimiser les modifications apportées au code existant, privilégiant une "décoration" des modules par du code Genom3 afin de pouvoir plus facilement mettre à jour mon code lorsque de nouvelles versions du programme ROS sortiront.

L'idée est donc d'encapsuler le code du second programme dans un module Genom3 qui saurait faire la traduction des types de données vers les types attendus par Genom3 et les autres programmes de démonstration du LAAS mis en place dans l'appartement de simulation (voir page 2). Cette démonstration, bien que très intéressante, n'est malheureusement valable que pour un utilisateur. En effet, en cas d'utilisation par de multiples personnes il serait nécessaire au préalable de séparer les poses détectées par chacune des caméras et de les associer en fonction de l'utilisateur dont elles proviennent, ce qui a servi de moteur à la suite de mon stage.

À la fin de cette première partie de stage, il a été possible de monter une petite démonstration dans l'appartement qui s'intégrait aux logiciels déjà installés (notamment, il a été possible de visualiser les poses estimées de l'utilisateur dans une simulation 3D de l'appartement, voir page 3).

Comme expliqué précédemment, afin de généraliser ce processus à plusieurs utilisateurs il est nécessaire d'associer les données capturées par les caméras RGB-D au bon utilisateur. C'est avec cet objectif que je me suis lancé dans la partie recherche du stage, en commençant notamment par un état de l'art des solutions existantes pour ré-identifier des personnes dans un réseau de caméras (en particulier de caméras RGB-D).

1.4. Plan du mémoire

Préalablement à la présentation de la partie recherche du stage, nous établirons un état de l'art des techniques de ré-identification de personnes par caméras. Nous détaillerons par la suite les méthodes retenues puis la manière dont elles ont été intégrées au projet global.

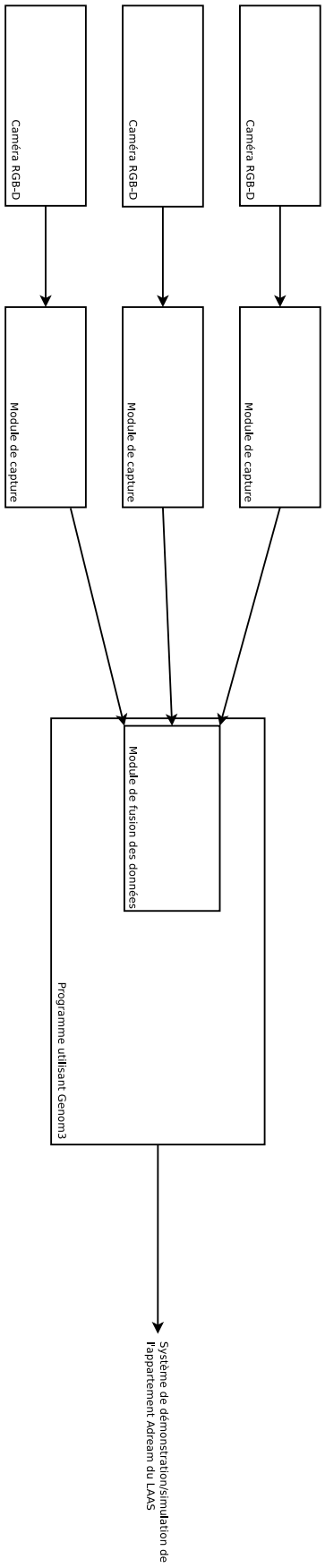

FIGURE 2 – Illustration du fonctionnement du module Genom3 développé

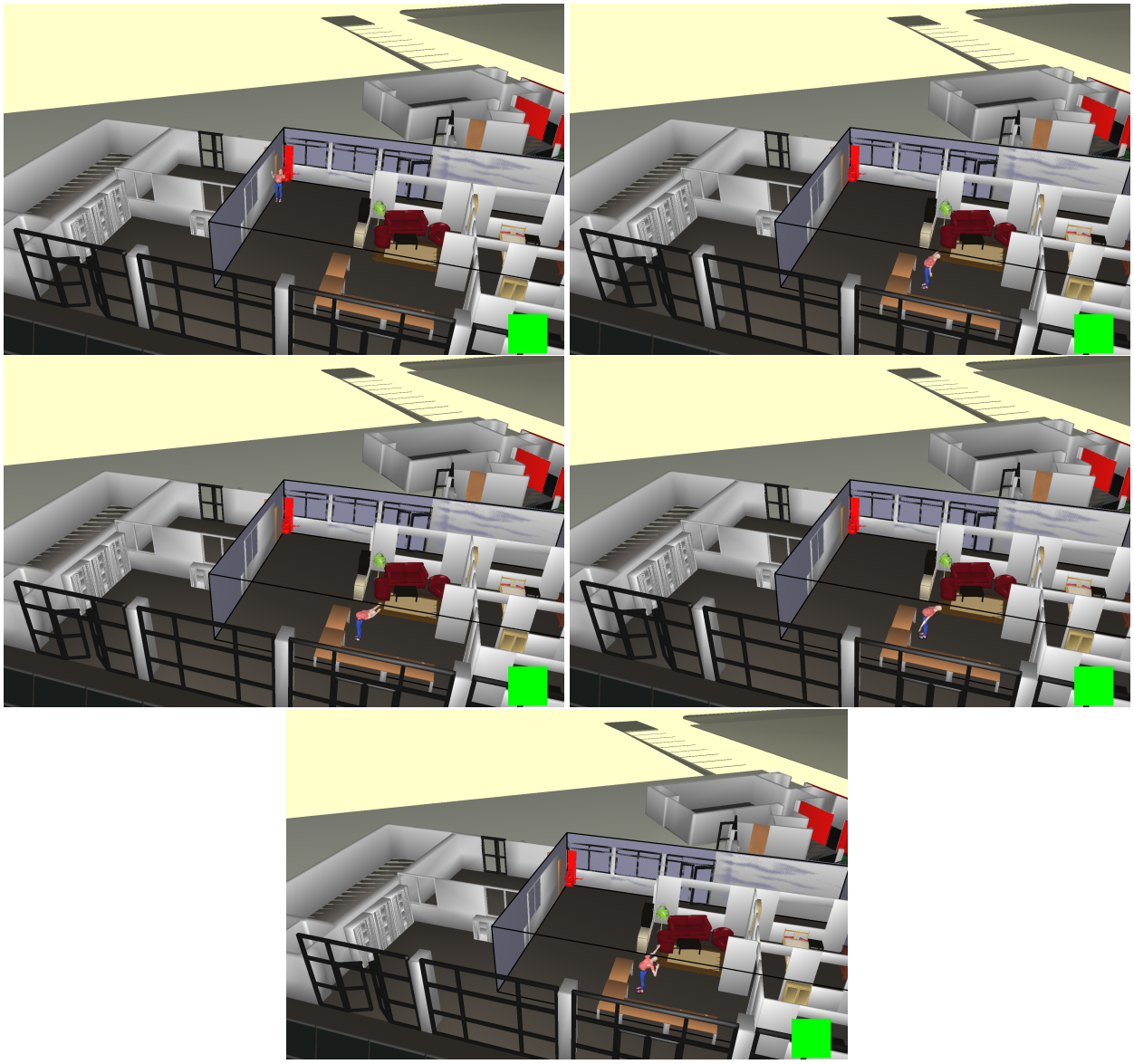

FIGURE 3 – Démonstration du module Genom3 développé : on voit bien la pose estimée de l'utilisateur qui se promène dans l'appartement. L'orientation du buste est erratique mais pas spécifique au module développé : le bug est d'origine externe.

2. Etat de l'art

Cet état de l'art porte sur les descripteurs pris en compte dans la ré-identification d'êtres humains dans un réseau de capteurs RGB-D (apparence avec une information de profondeur) à champs disjoints.

2.1. Les capteurs de profondeur

Avec l'annonce de la *Microsoft Kinect* (anciennement connue sous le nom *project Natal*) dans le courant 2008 et la sortie de divers SDK et pilotes associés les capteurs RGB-D (D pour *depth*, i.e. profondeur) commerciaux ont fait leur apparition sur un marché anciennement réservé à un public spécialisé, les technologies alors utilisées étant alors très coûteuses en grande partie à cause de la recherche d'une précision importante ainsi que du manque d'un marché de masse.

Il est alors devenu possible pour un budget restreint de travailler avec des capteurs de profondeur certes limités mais relativement adaptés à la recherche, notamment dans les domaines de la Robotique et de la Vision par Ordinateur.

Ces capteurs proposent une image très correcte (la résolution pouvant monter jusqu'à 1280x1024 pixels avec une information de couleur (8bits en RGB)) et surtout la possibilité de récupérer une carte de disparité (en 640x480), une carte segmentée des utilisateurs (en 320x240, jusqu'à six personnes) et une estimation des poses desdits utilisateurs (*via* un squelette composé de plusieurs articulations).

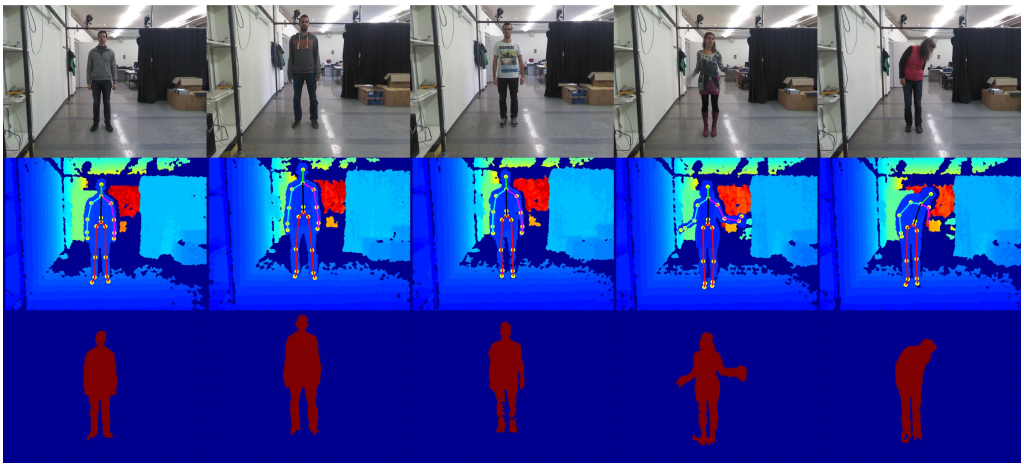

FIGURE 4 – Exemple d'informations fournies par les capteurs concernés : de haut en bas, une image RGB de la scène en haute résolution, le squelette évalué superposé à l'image de profondeur (dans laquelle le bleu signifie un point proche et le rouge un point éloigné) et une image de segmentation en basse résolution. [0]

Depuis, le marché s'est révélé florissant et est apparue la Xtion d'Asus, qui constitue d'ailleurs l'essentiel du parc du LAAS.

0. Ensemble d'images utilisé sans modification, aimablement fourni sous licence Creative Commons NC-BY-SA par Matteo Munaro (voir [Munaro et al., 2014])

2.2. Contexte et enjeux

Une interaction naturelle entre un environnement robotique et des personnes présuppose la capacité à reconnaître ces derniers. Dans le cadre du projet ADREAM par exemple, le LAAS possède un prototype d'appartement dans lequel des capteurs RGB-D sont placés au plafond et peuvent observer les différents individus et robots y évoluer. Cependant, il est très coûteux en temps comme en argent de s'assurer que les individus observés le restent en permanence car il faudrait garantir une couverture totale de l'endroit à observer, quel que soit le nombre de pièces dans l'appartement. Pouvoir associer deux enregistrements d'une même personne vue à des moments et des endroits différents devient une nécessité dans un réseau de caméras dont les champs seront à priori disjoints afin de limiter l'instrumentation en capteurs.

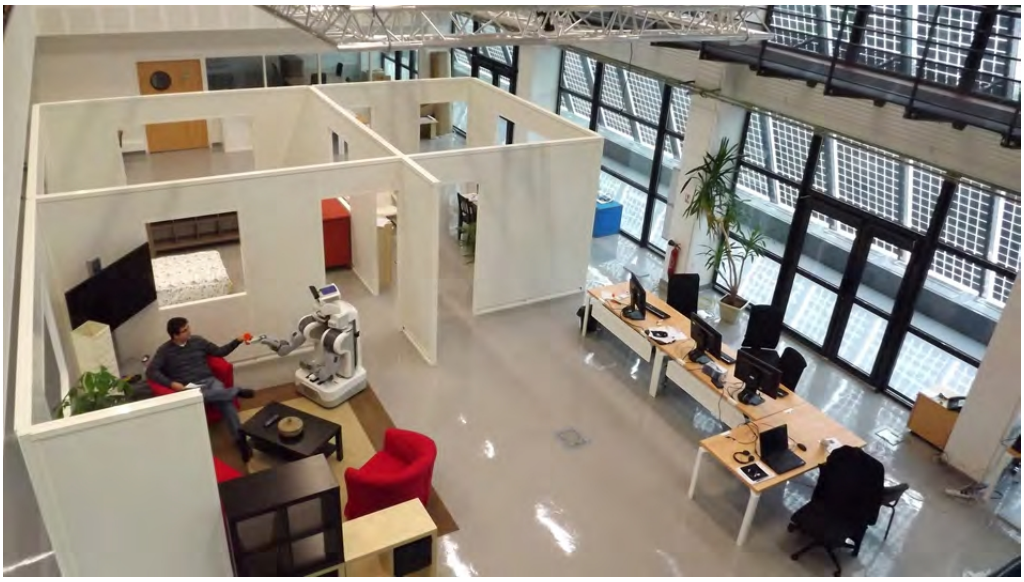

FIGURE 5 – L'appartement d'ADREAM. Dans sa version actuelle, des caméras RGB-D fixées au plafond permettent d'observer les pièces depuis des champs de vue disjoints[0]

L'objectif de cet état de l'art est donc d'évaluer les possibilités d'utilisation des avantages des capteurs RGB-D commerciaux pour une meilleure ré-identification des personnes vues par un réseau de caméras RGB-D à champs disjoints.

Il faut néamoins rester conscient des limitations des capteurs du commerce actuels, qui souffrent pour la plupart de l'utilisation de lumière structurée pour la génération de cartes de profondeur selon des stratégies similaires à celles évoquées par [Salvi et al., 2004], qui évaluent la surface des objets à partir des déformations des motifs projetés ce qui limite l'utilisation de tels capteurs en intérieur (à cause des interférences avec la lumière du soleil dans le domaine infrarouge) et entrave partiellement la génération simultanée de cartes de disparité avec plusieurs senseurs en raison de la superposition des motifs (un problème auquel [Berger et al., 2011] envisagent plusieurs réponses). Les champs de nos caméras étant disjoints nous ne sommes pas affectés par ce problème, il est cependant probable que cela s'améliore dans tous les cas avec l'arrivée de caméras temps de vol à faible coût (comme la *Kinect* fournie avec la *Xbox One*). De plus, la détection d'êtres humains ne peut se faire que dans une zone bien déterminée,

0. Extrait de `http://www.laas.fr/files/ADREAM/ADREAM-ResearchPlatform-11p.pdf`, ©CNRS-LAAS

entre 0.7 et 5 à 6 mètres de la caméra.

2.2.1. Généralités sur la ré-identification

Le processus de ré-identification de personnes consiste en un apprentissage supervisé des descripteurs de ces personnes. Il est la plupart du temps implémenté selon un pipeline assez générique : en effet, l'image est d'abord prétraitée afin notamment d'éliminer la distorsion de l'image, de la rendre moins sensible aux variations d'illumination... et isolée par une segmentation préalable de la personne, souvent grâce à un algorithme de détection/suivi (pour le problème en main il s'agira souvent d'une boîte englobante de l'utilisateur concerné) ce qui permet de réduire fortement la complexité du problème. Il est également possible d'inférer la pose des utilisateurs des données capturées.

Certains algorithmes (dits *part-based models*) sépareront alors les données en plusieurs sous-parties, pour prendre en compte la complexité de l'apparence de la personne considérée (par exemple la différence de couleur entre le bas et le haut du corps) dans le but d'obtenir une meilleure robustesse aux occultations et changements d'illumination. Dans tous les cas, il devient nécessaire de trouver un moyen de décrire la classe (groupement des représentations d'une même personne) de la détection : toutes les détections d'une même classe doivent avoir une signature proche les unes des autres, et si possible éloignée de celles d'une classe différente. Le moyen choisi pour la description s'appelle descripteur (*feature* en anglais).

La dernière étape est un problème de classification puisque l'on souhaite associer à notre signature une des signatures apprises à partir de l'ensemble d'apprentissage. Un challenge supplémentaire sera de détecter si la détection appartient à une classe encore non-apprise (c'est à dire que la personne a ré-identifier est en fait une personne à identifier) afin de le rajouter à la base de données. Il est important pour cela de connaître plusieurs particularités des classes en jeu : les variances extra- et intra-classes sont-elles grandes ? Cette dernière dépend-elle de la classe ?

Il y a plusieurs défis à relever ici : non seulement il s'agit de choisir les caractéristiques des personnes détectées qui seront les plus discriminantes et permettront une ré-identification efficace et rapide (les contraintes CPU sont importantes dans notre application), mais il faut aussi choisir le classifieur (méthode de classement des descripteurs) et la règle de décision associée. Nous nous focaliserons ici sur la catégorisation via des descripteurs pertinents dans notre application.

Dans la littérature scientifique actuelle ressortent plusieurs approches :

— La plus ancienne est probablement l'**approche basée sur l'apparence**, dont les mécanismes sont étudiés depuis les débuts de la Vision par Ordinateur. Dans cet état de l'art, nous séparerons les méthodes s'intéressant à des **caractéristiques biométriques** de l'humain (visage, iris...) et celles s'intéressant à son apparence de manière plus générale.

— Lorsque l'analyse de démarche (*gait analysis* en anglais) s'est imposée comme une des méthodes utilisées par le cerveau humain pour distinguer plusieurs personnes (ce qui a entre autres été mis en évidence par [Stevenage et al., 1999]) s'est ouvert un nouveau champ de recherche pour la ré-identification humaine. Ces méthodes seront observées dans la partie **approches basées sur la dynamique**.

— Déjà étudiée en deux dimensions, l'arrivée de la carte de profondeur a semble-t-il grandement amélioré l'analyse des paramètres géométriques propres au corps de l'être humain : taille, rapports entre les os... Ceci sera exploré dans la partie **approches basées sur la morphologie**.

2.3. Approches biométriques

Parmi les traits caractéristiques d'un être humain, le visage ressort immédiatement comme l'un des plus importants. Hors circonstances exceptionnelles, le visage varie peu au cours du temps comparé aux parties du corps humain recouvertes par les vêtements ce qui donne aux signatures extraites une stabilité plus grande à long terme.

En termes de reconstruction faciale, plusieurs approches sont possibles :

— Reconnaissance sur la base d'images fixes bidimensionnelles : [Baüml et al., 2010] présente par exemple une implémentation basée sur la transformée en cosinus discrète des sous-blocs de 8x8 pixels d' images réorientées représentant les visages à reconnaître. Les *features* sont alors calculées en normalisant les coefficients de la DCT et en n'en gardant qu'un sous-ensemble déterminé. Cela requiert en outre un suivi robuste des visages, la gestion de l'orientation éventuellement non-frontale et des occultations ([Baüml et al., 2010] considèrent à ce sujet qu'après une occultation trop importante une nouveau visage doit être détecté).

— Reconnaissance sur la base d'un scan 3D du visage : [Blanz and Vetter, 2003] proposent par exemple le *fitting* d'un modèle à géométrie transformable de visage sur le visage détecté en deux dimensions, et la comparaison des paramètres de la transformation avec ceux de l'apprentissage. La reconnaissance 3D a un avantage certain sur son homologue bidimensionnel (à condition que l'information soit de qualité) : elle permet notamment de corriger la pose par rotation dans l'espace en plus d'apporter des informations supplémentaires de structure et d'être plus robuste aux variations d'illumination.

L'approche tridimensionnelle peut, d'après [Abate et al., 2007], grandement améliorer les performances en reconaissance lorsqu'elle est combinée à l'approche bidimensionnelle, mais se heurte à ses propres difficultés comme l'acquisition précise de la géométrie du visage et l'occultation (par des cheveux, les lunettes...). De façon plus générale, l'acquisition d'informations suffisamment précises pour entamer une reconnaissance faciale restreint son utilisation à des environnements contraints. Ces méthodes ne sont donc pas adaptés à notre application.

2.4. Approches basées sur la dynamique

Il est alors envisageable de recourir à des concepts de biométrie douce (*soft biometrics*) tels que la reconnaissance de démarche ou plus largement la dynamique du corps humain. En effet, [Nixon and Carter, 2004] notent que la reconnaissance de démarche est une méthode efficace et non-intrusive, utilisable à grande distance et faible résolution.

En 2002 par exemple, suite à la publication de plusieurs rapports touchant à l'analyse de la démarche humaine pour l'utiliser comme donnée biométrique [Yam et al., 2002] propose après avoir constaté qu'il est possible de relier les paramètres de marche et de course chez un être humain d'en extraire un modèle

caractéristique de la personne à identifier par l'analyse du mouvement périodique de la cuisse et de la jambe, le tout avec des capteurs 2D classiques.

[Bouchrika and Nixon, 2008] montrent, plus tard et avec des caméras RGB, que la reconnaissance de démarche bidimensionnelle est relativement robuste lorsque l'on considère l'utilisation de vêtements de différentes longueurs, le changement de chaussures ou le port de bagages.

Du côté de l'analyse tridimensionnelle, [Urtasun and Fua, 2004] proposent d'extraire les vitesses angulaires paramétrisant les degrés de liberté d'un modèle cinématique du squelette humain lors de la marche. L'idée est d'effectuer ensuite une analyse en composantes principales et de conserver les premiers coefficients pour une classification optimale. Cependant, il est difficile de conclure sur les résultats car l'ensemble d'apprentissage est relativement réduit (quatre sujets). Enfin, il faut noter que cette identification ne se fait pas en temps réel mais à partir des données moyennes d'une séquence vidéo, pour être résistante à l'occultation. L'approche choisie est toujours 2D. [Zhao et al., 2006] ont en conséquence proposé de ne conserver que les informations de dynamique sur les membres inférieurs (considérant que la contribution des membres supérieurs à la démarche n'est pas suffisamment importante) et de compléter ces informations par les longueurs des os du squelette. A noter que cette fois les informations de dynamique portaient sur des distances (inter-chevilles, inter-rotules...), cependant l'ensemble d'apprentissage est encore une fois très réduit.

S'appuyant sur les possibilités d'extraction de squelette offertes par les capteurs RGB-D, [Kumar and Babu, 2012] proposent alors un approche axée sur la comparaison des distances et facteurs d'échelles impactant la position des articulations du corps relativement au centre des hanches, à l'aide d'un capteur de profondeur. Il est noté que les trajectoires obtenues sont très bruitées, la solution retenue étant l'application d'un filtre passe-bas sur ces dernières. Si les séquences de test excluent des scénarios contraignants (comme l'occultation par-exemple), les données sélectionnées sont déjà un peu plus larges (vingt sujets, chacun dans dix situations différentes).

[Sivapalan et al., 2011] proposent en 2011 une approche originale, s'appuyant sur les *Gait Energy Images* présentées dans [Han and Bhanu, 2006] qui consistent à représenter en superposition sur une seule image toutes les silhouettes capturées lors d'un cycle de marche, et les étendant en 3D grâce à l'aide d'une caméra Kinect. Si les résultats sont très encourageants ; les auteurs notent qu'il s'agit d'une méthode qui dépend fortement de la pose de la caméra.

[Preis et al., 2012] ont recherché des traits utiles pour identifier une personne avec les informations extraites d'une *Kinect*, mais utilisent au final peu d'informations dynamiques (uniquement la vitesse de déplacement et la longueur d'un pas) ce qui amène les auteurs à ne considérer que des informations de longueur de membres. Les conclusions sont donc moins intéréssantes lorsque l'on s'intéresse à la dynamique qu'à l'aspect statique pour lequel les auteurs estiment qu'il constitue une information de qualité pour la ré-identification (encore une fois, sur un ensemble très limité de huit personnes).

L'étude de la dynamique du corps humain dans le cadre de la réidentification apparaît donc comme un critère prometteur. Les avantages de cette méthode est qu'elle est robuste aux changements vestimentaires, et n'est pas affectée directement par les changements d'illumination tout en donnant des résultats très concluants ([Satta, 2013, paragraphe 4]). Les approches basées sur la squelettisation nécessitent néamoins une connaissance précise de la personne ce qui est difficile en environnement non contraint, tandis que les approches comme la Gait Energy Image ont besoin d'un alignement parfait des silhouettes à comparer ainsi que d'une segmentation très précise.

Pour ce qui est de l'utilisation des informations renvoyées par une caméra RGB-D, il est difficile de conclure de manière fiable sur les résultats obtenus tant les ensembles utilisés pour les expériences sont petits et restreints à des situations la plupart du temps idéales.

2.5. Approches basées sur la morphologie

Comme l'utilisation de caméras RGB-D permet aujourd'hui d'estimer de manière fiable la pose d'un être humain (voir notamment [Shotton et al., 2013]) il est devenu intéressant de tenter de ré-identifier des humains à partir de leurs caractéristiques anthropométriques (comme on a pu le voir dans la partie précédente avec [Preis et al., 2012]).

Dans [Barbosa et al., 2012], de telles informations sont extraites directement avec le *Kinect SDK*, auxquelles sont rajoutées des mesures de distances géodésiques sur un maillage reconstruit de l'abdomen. Les auteurs déterminent alors les données les plus discriminantes et gardent les dix retenues comme les meilleures. Les essais effectués sur un ensemble de 79 individus semblent prometteurs même s'il peut se révéler assez difficile d'extraire de telles informations dans des poses quelconques. De plus, la reconstruction de maillage et le calcul de distances géodésiques peut se révéler très consommateur de ressources et de temps. Une approche plus simple, basée simplement sur des longueurs et la hauteur calculée de la personne est présentée par [Araujo et al., 2013], qui mettent d'ailleurs en exergue la nécessité de filtrer les informations afin de retirer les données aberrantes (la méthode retenue par les auteurs étant de simplement retirer toute image contenant des données trop éloignées de la moyenne).

Si les approches se basant sur les caractéristiques anthropométriques semblent prometteuses, la plupart se basent uniquement sur des informations bidimensionnelles (comme [Madden and Piccardi, 2005] qui utilisent l'information de hauteur pour associer des personnes détectées par deux caméras à champs disjoints) et la fiabilité des squelettes fournis par les capteurs RGB-D pour la ré-identification reste encore floue.

2.6. Approches basées sur l'apparence vestimentaire

Les approches les plus courantes semblent être les méthodes se basant sur l'apparence des personnes. Celles-ci sont plutôt sensibles aux variations de pose et de point de vue, à l'occultation partielle et aux changements d'illumination et présupposent souvent que la personne à identifier ne change pas de vêtements entre deux identifications.

[Farenzena et al., 2010] subdivise le corps suivi selon ses propriétés de symétrie : sont donc séparés le torse des deux jambes par antisymétrie et la gauche de la droite par symétrie. Le descripteur retenu se compose de trois parties qu'il est possible de pondérer différemment voire de retirer : l'histogramme en HSV, puis les MSCR (*maximally stable colour regions*) et les RHSP (*recurrent high-structured patches*) qui correspondent à l'information de texture, ces derniers étant obtenus par échantillonage puis classification de points proches des axes de symétrie afin de trouver les régions/*patches* les plus importants. Un gros intérêt de cette composition est qu'il est possible de retirer au choix des descripteurs de n'importe quel type (selon le but recherché : par exemple, le calcul des descripteurs touchant à la texture est très lourd, et apporte peu lorsque la qualité de l'image est insuffisante ou que les vêtements sont unis), voire d'en ajouter d'autres. Il est aussi possible de jouer en fonction de la situation sur la pondération des différents descripteurs (ce à quoi [Liu et al., 2012] se sont intéressés).

Cela permet une plus grande robustesse aux variations d'illumination, de pose et de vue que des algorithmes ne prenant pas en compte autant d'informations. Est par ailleurs posée la question du nombre d'images utilisées pour la reconnaissance et l'apprentissage, les auteurs concluant que l'utilisation de plusieurs images est un énorme plus pour les résultats de reconnaissance tout en soulignant que leurs descripteurs (appelés SDALF pour *Symmetry-Driven Accumulation of Local Features*) obtiennent des résultats excellents même en n'utilisant qu'une image pour la ré-identification.

FIGURE 6 – Différentes étapes de création des SDALF : chacune des lignes représente une instance de la même personne. La seconde colonne montre le découpage par symétrie de l'image segmentée. La troisième colonne montre l'importance des couleurs vis-à-vis de l'histogramme (les pixels clairs indiquent une couleur plus importante). Viennent ensuite les MSCR, puis les RHSP [0]

[Satta et al., 2013] se situent dans un scénario similaire au nôtre (ré-identification d'humains dans un réseau de capteurs RGB-D à champs disjoints) mais n'utilisent l'information de profondeur que pour le suivi et la segmentation, et retombent sur la segmentation par symétrie de [Farenzena et al.,

0. Extrait de `http://www.lorisbazzani.info/code-datasets/sdalf-descriptor/`, ©Loris Bazzani

2010] lorsque cette information n'est pas disponible. Discutant des principaux travaux effectués (dont les SDALF), les auteurs constatent qu'il s'agit le plus souvent de méthodes demandant beaucoup de temps de calcul et de mémoire et proposent de réduire la complexité du problème, indépendamment des descripteurs retenus, en définissant une personne par sa dissimilarité par-rapport à des modèles ce qui permet d'éliminer des informations redondantes.

La littérature est par-ailleurs abondante sur le sujet, puisque la question était ouverte bien avant l'arrivée des capteurs RGB-D. [Bak et al., 2010] ont par exemple montré que les caractéristiques pseudo-Haar, très utilisées dans la détection d'objets et de visages et qui présentent l'avantage d'être extrêmement rapides à calculer mais qui ne retiennent chacune que très peu d'information, étaient utilisables dans ce domaine. Comme on l'a vu précédemment, l'utilisation de caméras RGB-D servirait surtout aux étapes de suivi et de segmentation et bien moins à l'étape de calcul de descripteurs dans ces conditions.

2.7. Conclusion

Le tableau ci-contre présente un résumé des différentes approches vues dans cet état de l'art.

Les informations fournies par les caméras RGB-D dont nous disposons dans notre cas ne se restreignent donc pas simplement à servir au suivi de personnes et à la segmentation de l'image : l'approche morphologique montre par exemple que les articulations estimées peuvent directement servir de descripteurs pour la ré-identification. Plusieurs idées peuvent ensuite venir compléter ces approches pour leur apporter soit une plus grande rapidité d'exécution soit de meilleures performances en ré-identification (et potentiellement les deux en même temps) : comme l'ont fait [Farenzena et al., 2010] il est par exemple possible de combiner plusieurs types de descripteurs (les nouveaux descripteurs étant alors appelés **hétérogènes**) mais également, comme [Meden et al., 2012] l'ont abordé, d'apporter une connaissance de la topologie du réseau afin de peaufiner la justesse de la détection et d'éliminer certains candidats. [Javed et al., 2008] proposent d'ailleurs un modèle spatio-temporel qu'il est possible d'apprendre et montrent qu'il s'agit d'une méthode efficace de ré-identification, même dans des situations contraignantes.

Nous retenons donc la possibilité de combiner des descripteurs de différents types afin de former des descripteurs hétérogènes. Durant la suite de ce stage, nous souhaitons évaluer l'impact qu'aurait la combinaison de descripteurs morphologiques comme ceux présentés par Barbosa et al. et de descripteurs SDALF, tout en utilisant les cartes de segmentation et de profondeur pour un suivi et une segmentation rapides et précis.

Type d'approche	Avantages	Inconvénients
Biométrique	Caractéristiques extrêmement discriminantes	Dépend de l'orientation - Vulnérable au changement de pose/illumination - Demande de l'information de qualité (haute résolution…), peu pertinent à longue distance et en environnement non contraint : inadapté en vidéosurveillance - Coût CPU élevé
Analyse de démarche	Assez peu impacté par des facteurs covariants (port de sacs, vêtements courts) - Possibilité de lier marche/course - Utilisable avec peu d'informations, faible coût CPU	Demande un mouvement - Très impacté par les vêtements longs, la marche pieds nus ou en nu-pieds - Descripteurs difficiles à extraire en environnement non contrôlé
Morphologique	Très prometteur - Utilise directement les informations du capteur - Charge CPU légère hors informations de distances géodésiques	Demande une sélection des descripteurs pertinents - Dépend fortement de la qualité des squelettes détectés - Encore peu étudié
Apparence	Littérature très extensive - Fort pouvoir discriminant entre les cibles	Sensible aux variations de pose/illumination (peut être contourné) - Sensible aux variations fréquentes d'apparence (changement de vêtements…) - Certains descripteurs (notamment ceux qui concernent la texture dans SDALF) sont très lourds à calculer

3. Descripteurs retenus

3.1. Approche retenue

À la lumière de l'état de l'art présenté à la section 2, il a été choisi de tenter de combiner l'approche basée sur l'apparence et l'approche basée sur la morphologie. Le descripteur proposé par [Barbosa et al., 2012] semble être intéressant pour la partie morphologie : en effet, si le descripteur retenu est peu discriminant comparé à certains autres, il a le mérite d'être extrêmement rapide à calculer, les comparaisons de descripteurs étant très simples (une comparaison de distances entre vecteurs) par ailleurs.

De plus, comme déjà indiqué dans [Barbosa et al., 2012] ce descripteur peut être facilement utilisé en complément d'un autre plus discriminant et se basant sur un type d'information différent : si le premier a du mal à classifier la détection, le second peut prendre le relais en apportant une nouvelle connaissance de la détection (par-exemple si l'apparence de la détection n'est pas assez discriminante, sa morphologie le sera peut-être). Cela laisse tout de même un choix d'implémentation : en effet, il est possible de combiner les descripteurs et de leur affecter à chacun un poids (ce qui se fait déjà dans [Barbosa et al., 2012], dans lequel le descripteur final est déjà une combinaison de descripteurs) auquel cas il est nécessaire de choisir les poids optimaux (ce qui peut se révéler complexe lorsque le nombre de poids augmente). Une autre solution pourrait être d'essayer de classifier la détection avec une première méthode, et de se servir de ce second descripteur comme d'une solution de secours si la classification semble imparfaite. Cela évite d'avoir à calculer tous les descripteurs à chaque détection si un seul suffit, mais demande de bien maîtriser les conditions auxquelles il va être fait appel à la solution de secours, et dans quelle mesure cela améliore ou détériore la classification.

Il est à noter qu'il aurait été possible de choisir une autre méthode basée sur l'information de profondeur. Cependant, il s'agit la plupart du temps d'ajouter à l'information de couleur une dimension spatiale (comme par exemple dans [Baltieri et al., 2011] où les auteurs mappent les pixels de la détection à un modèle ressemblant à un "sarcophage" du corps humain, ce qui permet de mieux classifier les apparences (en évitant de comparer la couleur de la tête d'une détection avec la couleur du torse d'une autre par exemple). Ayant retenu SDALF comme descripteur, l'information de localisation des pixels est déjà prise en compte lorsque l'imagette détectée est séparée selon les axes de symétrie/antisymétrie : le descripteur basé sur la morphologie apparaît alors comme une approche plus originale et moins redondante.

Nous avons donc retenu le descripteur SDALF amputé des informations de texture (RHSP, en raison de la lourdeur de son calcul comme on le verra en section 3.2), qui fait actuellement partie des méthodes les plus performantes de ré-identification en RGB (en particulier dans de mauvaises conditions : variations de pose, d'illumination, très faible résolution des détections...) combiné avec le descripteur basé sur la morphologie proposé par [Barbosa et al., 2012] sauf l'information géodesique. En effet, celle-ci requiert une vue frontale de l'utilisateur pour être pertinente, est plus lourde à calculer que les autres, et est

très impactée par le bruit (la conception des capteurs actuels fait que les cartes de profondeur sont de très basse résolution et très bruitées. De plus, une grande partie des valeurs de la carte est obtenue par interpolation des valeurs environnantes car la carte comporte à l'origine de nombreux "trous").

Ce second descripteur apporte par ailleurs deux avantages : avec l'amélioration des capteurs de profondeur (la Kinect 2 de Microsoft par exemple) les mesures prises en compte deviennent bien plus précise qu'avec la première génération de capteurs. Le descripteur basé sur la morphologie en devient d'autant plus pertinent. De plus, comme il se base sur une information infrarouge, le classifieur fonctionnera dans l'obscurité à l'inverse d'un simple classifieur RGB. L'idée est alors d'explorer un domaine peu répandu dans la littérature scientifique.

3.2. Evaluations et discussions associées

L'objectif avoué de ce stage est d'obtenir une librairie capable d'effectuer de la ré-identification de personnes en temps réel (voire même plus rapidement, étant donné qu'il ne s'agit que de la première chaîne d'un *pipeline* plus étendu). A cet effet, il pourrait paraître pertinent d'identifier quels sont, parmi les combinaisons de descripteurs retenues, les descripteurs les plus pertinents et les descripteurs les plus consommateurs de temps CPU. C'est pourquoi quelques séries de tests ont été effectuées sur des ensembles de données publics (notamment le *dataset* ETHZ1 [1]).

Les bases publiques sont avantageuses, car elles permettent de comparer différentes méthodes de ré-identification entre-elles. Elles correspondent d'ailleurs souvent à un cas d'utilisation précis (marche dans un couloir, dans la rue, supermarché...) qui posent chacun des défis différents, comme des variations de pose/illumination plus fortes, de nombreuses occultations ou des variations d'apparence faibles (si l'on disposait d'une base de détections prises à la sortie d'une salle d'expériences de chimie par exemple, il est très probable que la plupart des utilisateurs soient habillés de blouses blanches).

En revanche, si l'on souhaite évaluer une méthode de ré-identification dans un contexte très particulier (pour nous par-exemple il s'agit de regarder différentes personnes depuis un réseau de caméras RGB-D à champs joints ou disjoints, au plafond, dans un appartement avec peu d'utilisateurs simultanés) il est peu probable qu'il existe une base publique : il faudra alors se constituer une base propre.

3.2.1. Protocole d'évaluation

Les évaluations sont faites en calculant les différents descripteurs (et leurs combinaisons) sur des sets de données et en comparant leurs performances en reconnaissance. Pour cela plusieurs critères peuvent être importants, mais généralement la littérature retient l'aire normalisée ($0 \leq nAUC \leq 100$) sous la courbe CMC (Cumulative Matching Curve) qui représente la probabilité pour une détection d'être correctement ré-identifiée parmi les X premiers résultats, X étant lisible sur l'axe des abscisses. La courbe SRR représente elle la probabilité de ré-identifier correctement une détection parmi une base de X individus.

1. ETHZ (`http://www.umiacs.umd.edu/~schwartz/datasets.html`) provided by : A. Ess, B-Leibe, and L. V. Gool. Depth and appearance for mobile scene analysis. In IEEE International Conference on Computer Vision, 2007.

Il est également possible, dans la mesure où seules les classes les plus probables seront utilisées par la suite du pipeline, de considérer les probabilités de classification aux premiers rangs (les points d'abscisses 0, 1... sur le CMC) comme un étalon pour l'évaluation des méthodes de classification.

Dans un premier temps, il a fallu réduire le temps de calcul du descripteur SDALF dont le coût en temps CPU est assez haut.

Les premières évaluations seront effectuées sur la base ETHZ-1 qui met en scène 83 personnes dans des situations de marche, dans des espaces publics (notamment dans la rue ou à l'intérieur de stations de métro). Les images, en format PNG (Portable Network Graphics) sont très petites et de tailles variables (rarement au-dessus de 50*150 pixels) et le nombre d'images par personne peut varier grandement en-fonction du sujet (226 images sont étiquetées comme représentant la personne "p024", tandis que 7 seulement représentent "p068"). En voici des exemples :

FIGURE 7 – Exemples d'images trouvées dans le jeu de données ETHZ-1 : les deux premières images représentent l'individu étiqueté "p008", la troisième représente "p033" et la dernière "p055"

Cette base n'ayant pas d'information de profondeur, nous avons ensuite utilisé le jeu de données RGBD-ID, dont le cas d'utilisation est un peu plus proche de celui recherché : il s'agit de personnes vue de relativement près par une caméra RGB-D dans des salles de cours, ou des couloirs. Celui-ci présente en fait quatre bases distinctes : "backwards" contient des détections de personnes dos à la caméra, "collaborative" contient des personnes vues de face, mais ayant les bras écartés (afin d'obtenir une meilleure estimation de la pose par la caméra), "walking 1" et "walking 2" présentent des personnes marchant face à la caméra, les bras le long du corps (marche normale). Les visages sont floutés, et pour chaque personne on dispose :

- des images RGB prises par la caméra (en format JPEG)
- d'une estimation de l'équation paramétrique du sol dans le repère caméra (dans un fichier texte)
- des masques de segmentation de l'utilisateur dans l'image (ie. les pixels représentant le sujet relativement à ceux représentant l'arrière-plan) (en format JPEG)
- de reconstructions (en format .ply) 3D de ce qui est vu de l'utilisateur (cela veut dire qu'il s'agit de la surface visible par la caméra)
- du squelette estimé par la caméra pour chaque image de l'utilisateur, dans un fichier texte.

Ces informations (à l'exception du fichier .ply) sont celles que l'on peut directement obtenir avec la librairie OpenNI de contrôle de caméras RGB-D. Pour chaque individu (les 79 personnes ont chacune participé à chacun des "sous-bases" présentées précédemment, changeant éventuellement leurs vêtements entre-temps) on dispose de cinq images (donc cinq masques, cinq squelettes... etc). Voici les

informations que l'on peut y voir par-exemple :

FIGURE 8 – Exemples d'images trouvées dans le jeu de données RGBD-ID : les deux premières images représentent l'individu étiqueté "10" (la première est l'image RGB, la seconde le masque associé), les autres correspondent à "66"

3.2.2. SDALF complet

La première étape est d'évaluer les performances de SDALF [2] sur un ensemble de données complet, afin d'avoir un étalon. Sur l'ensemble appelé ETHZ1, les temps de calcul sont les suivants :

- Subdivision en 3 parties et calcul des poids de l'histogramme : Négligeable
- Calcul des MSCR [3] : Négligeable
- Histogramme pondéré HSV : Négligeable
- Calcul RHSP [4] : 14845 secondes

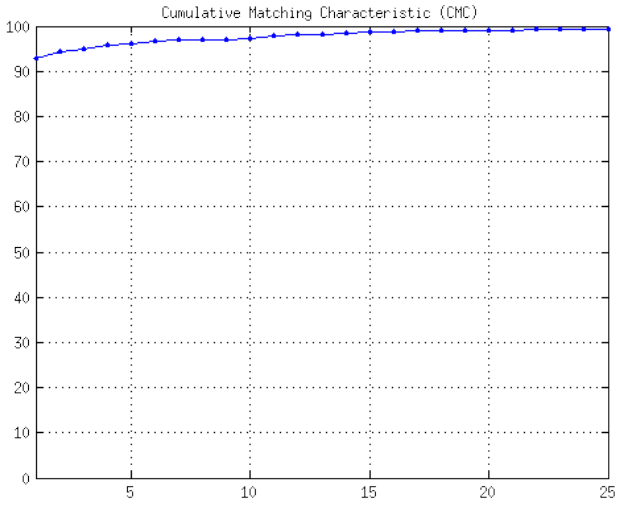

FIGURE 9 – Courbe CMC de la classification du set de données ETHZ1 avec tous les composants de SDALF. nAUC = 99.085

2. Voir le glossaire au chapitre 6.
3. Voir le glossaire au chapitre 6.
4. Voir le glossaire au chapitre 6.

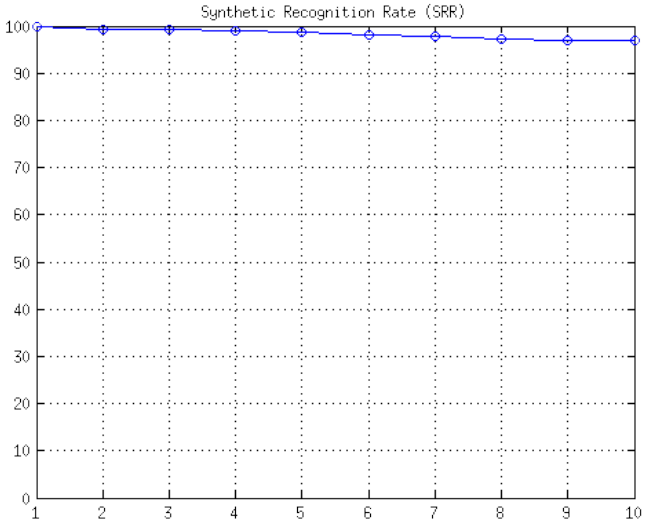

FIGURE 10 – Courbe SRR de la classification du set de données ETHZ1 avec tous les composants de SDALF

Il ressort immédiatement que le calcul des RHSP (qui représentent l'information de texture du descripteur) est extrêmement long en comparaison des autres composants de SDALF. Afin de gagner en performance il a alors été envisagé de le retirer si son apport aux résultats de la classification est suffisamment faible.

3.2.3. SDALF sans l'information de texture

La première étape est d'évaluer les performances de SDALF sur un ensemble de données complet, afin d'avoir un étalon. Sur l'ensemble appelé ETHZ1, les temps de calcul sont les suivants :

— Subdivision en 3 parties et calcul des poids de l'histogramme : 165 secondes
— Calcul des MSCR : 51 secondes
— Histogramme pondéré HSV : 24 secondes

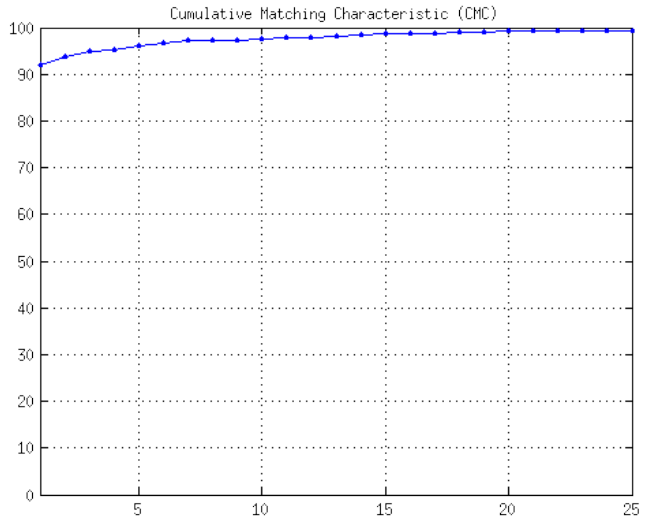

FIGURE 11 – Courbe CMC de la classification du set de données ETHZ1 avec SDALF mais sans l'information de texture.
nAUC = 99.056467

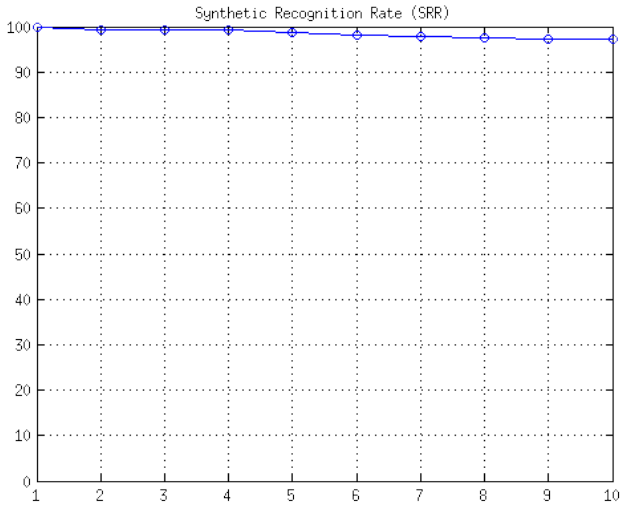

FIGURE 12 – Courbe SRR de la classification du set de données ETHZ1 avec SDALF mais sans l'information de texture

Comme on le voit, l'information de texture, au moins dans le cas du set de données ETHZ1, ne contribue pas énormément à la réussite de la classification. Au rang 1, on obtenait 92.89% de reconnaissance avec la prise en compte des informations de texture. Une fois omise, on obtient tout de même 91,69% de classifications correctes soit une diminution relative de 1,3% des classifications correctes pour un temps de calcul des ordres de grandeur plus faible.

3.2.4. Descripteur basé sur la morphologie

Puisque l'on souhaite évaluer les performances de ce descripteur, il faut utiliser un ensemble de données RGB-D. A cet effet, j'ai utilisé le set de données RGBDID [5] qui présente une situation proche de notre cas d'utilisation (les images sont prises à courte distance dans une petite pièce, la grosse différence étant que le capteur est ici relativement bas par-rapport aux sujets). J'ai ensuite codé le descripteur en Matlab moi-même.

Pour cette combinaison de descripteurs, les poids associés à chaque descripteur jouent sur la qualité finale de la ré-identification. Les auteurs du rapport original ont choisi des valeurs pour chacun de ces poids mais n'indiquent pas la méthode employée pour leur détermination. Par une exploration (non exhaustive) de l'espace des poids, j'ai déterminé mon propre jeu de poids mais il faut noter que l'on pourrait obtenir des poids plus optimaux avec d'autres méthodes d'optimisation numérique.

5. [Barbosa et al., 2012], accessible à l'adresse `http://www.iit.it/en/datasets-and-code/datasets/rgbdid.html`

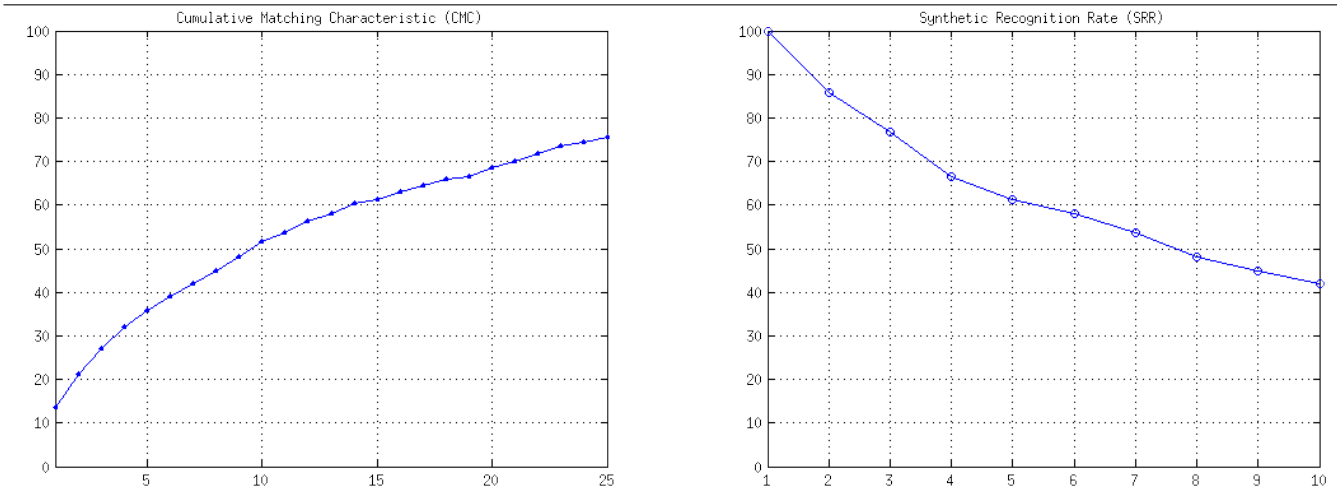

FIGURE 13 – Courbes CMC et SRR de la classification du set de données RGBDID avec le descripteur basé sur la morphologie

Comme attendu les performances en reconnaissance sont insuffisantes seules, mais déjà largement au-dessus d'un classificateur aléatoire. J'ai donc tenté d'observer les résultats de la combinaison de ce descripteur avec le descripteur SDALF.

3.2.5. Combinaison "naïve" des deux descripteurs

J'ai donc réalisé une combinaison des deux descripteurs que l'on appellera "naïve" car elle est basée sur une simple addition de distances (comme l'était SDALF, d'ailleurs) plutôt que sur une méthode de classification plus complexe. Cela a nécessité de modifier le code de SDALF afin de l'adapter au jeu de données RGBD-ID et l'injection de mon code calculant le descripteur basé sur la morphologie. Les résultats sont les suivants :

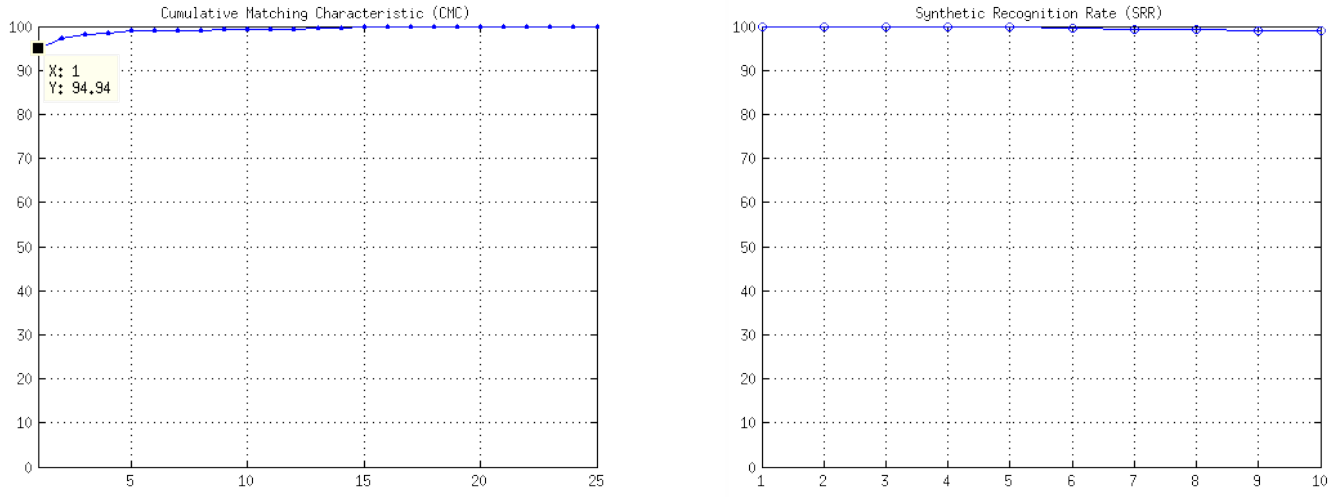

FIGURE 14 – Courbes CMC et SRR de la classification du set de données RGBD-ID avec la combinaison des deux descripteurs.
nAUC = 99.796507

Que l'on peut comparer avec les résultats de SDALF seul sur ce jeu de données :

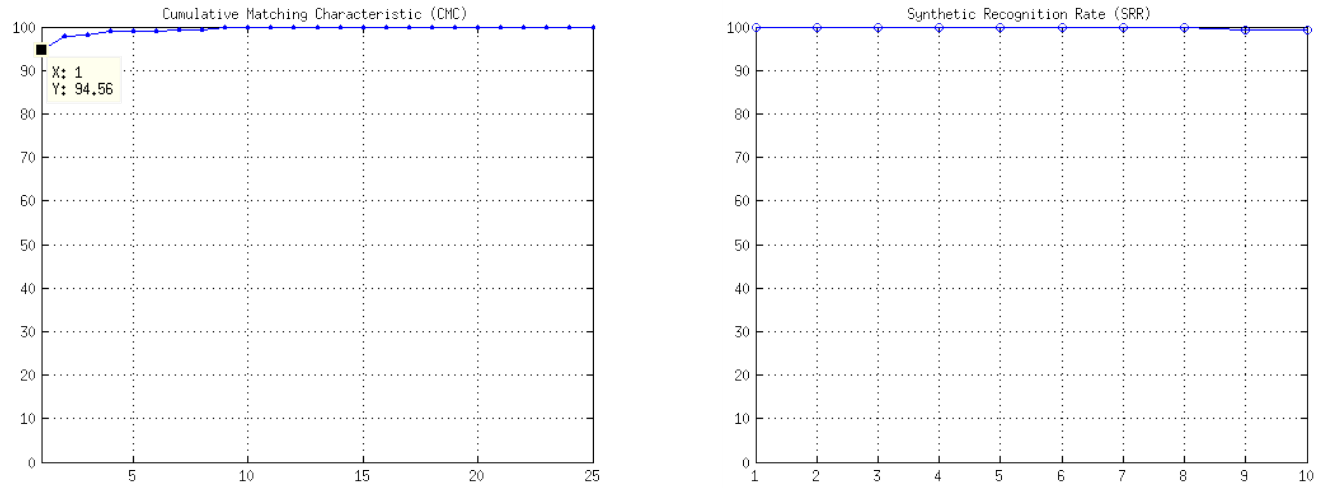

FIGURE 15 – Courbes CMC et SRR de la classification du set de données RGBD-ID avec SDALF seul.
nAUC = 99.791700

Les deux descripteurs obtiennent sensiblement les mêmes résultats, ce qui est probablement dû au taux de reconnaissance déjà élevé de SDALF sur ce jeu de données. On a cependant ajouté l'information sur la morphologie de l'utilisateur qui peut faire la différence dans l'obscurité et dans les situations où deux individus de morphologies différentes ont une apparence similaire (mêmes vêtements), ce qui sera vérifié en créant un jeu de données propre au LAAS.

4. Intégration, démonstration

Une fois le descripteur choisi, il a été choisi d'implémenter le processus de ré-identification dans une librairie C++, afin de pouvoir facilement réutiliser les fonctions offertes. Le descripteur basé sur la morphologie étant relativement simple à mettre en place, le descripteur basé sur l'apparence a été développé en premier, en émulant l'implémentation Matlab du descripteur SDALF mise à disposition par ses auteurs sur le Web [6].

4.1. Implémentation du descripteur basé sur l'apparence

Voici l'algorithme général de cette partie :

Algorithme 1 Calcul du descripteur basé sur l'apparence

Découper l'image en trois parties, en utilisant les axes de symétrie et d'antisymétrie de l'image
En déduire une carte de l'importance des pixels pour la ré-identification
Calculer le descripteur basé sur l'histogramme pondéré de l'image
Calculer le descripteur MSCR
Joindre le descripteur MSCR et les histogrammes pondérés

4.1.0.1. Découpage de l'image, calcul des poids des pixels.

L'axe d'antisymétrie tête/torse est obtenu en maximisant la différence entre le nombre de pixels dans le masque au-dessus de l'axe et en dessous, dans une fenêtre d'un quart de la hauteur de l'image d'origine

FIGURE 16 – Exemple de détection de l'axe tête/torse sur une image du set de données ETHZ1 (à gauche la détection du code de SDALF original, à droite l'axe horizontal représente celle réalisée par mon code). L'axe détecté est plus bas que le cou car la fenêtre choisie empêche de commencer la recherche plus haut.

L'axe d'antisymétrie torse/jambes est lui obtenu en maximisant la différence de couleur (représentée par une distance euclidienne dans l'espace des couleurs en HSV) tout en minimisant la différence en nombre de pixels dans le masque au-dessus et en-dessous de l'axe. On minimise donc une combinaison linéaire de ces distances (de manière expérimentale les deux ont le même poids).

6. http ://www.lorisbazzani.info/code-datasets/sdalf-descriptor/

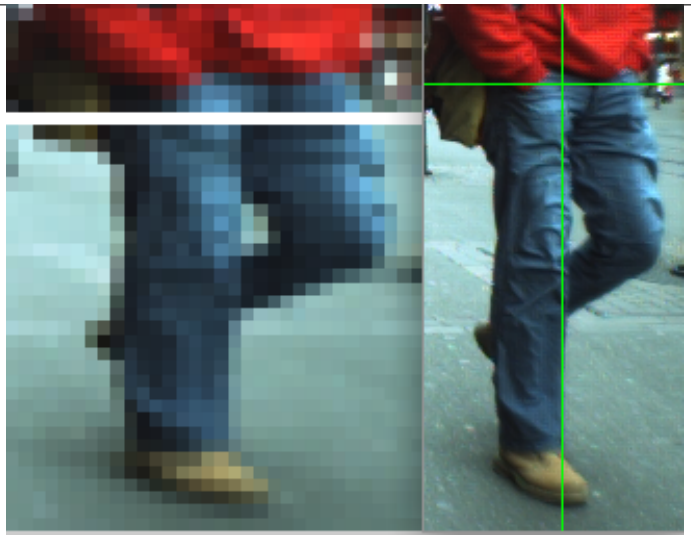

FIGURE 17 – Exemple de détection de l'axe torse/jambes sur la même image. Le code original de SDALF (à gauche) comportait une erreur que j'ai corrigé afin de rendre l'axe calculé plus fiable (voir à droite le résultat de mon code).

Les deux axes de symétrie restants (jambe gauche/jambe droite et torse gauche/torse droit) sont obtenus en minimisant simultanément les deux différences (mais en cherchant des axes verticaux cette fois). On obient donc le résultat final suivant :

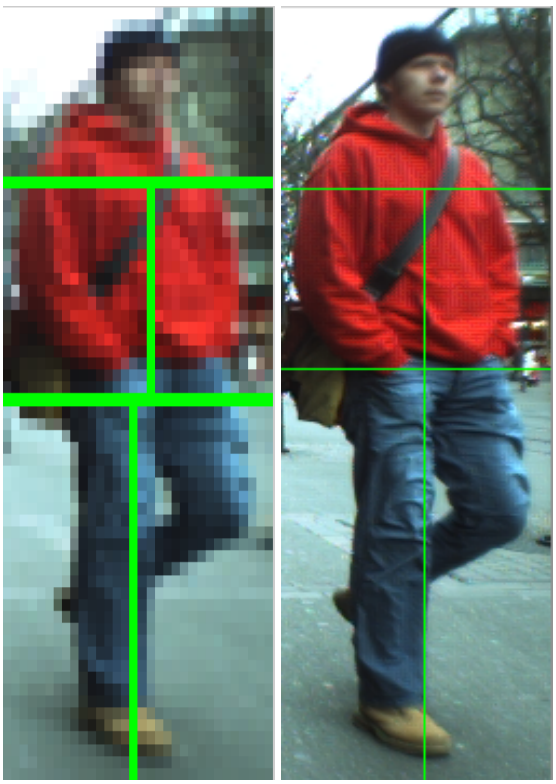

FIGURE 18 – Résultat de la segmentation de l'image (avec, pour comparaison, à gauche le résultat de la version originale sous Matlab et à droite le résultat de l'implémentation CPP). L'axe Torse/jambes est manifestement mieux détecté suite à la correction faite au code original.

4.1.0.2. Calcul de la carte des poids pour l'histogramme pondéré. Les poids sont ensuite calculés séparément sur la région du torse et les jambes, en fonction de la distance à l'axe de symétrie correspon-

dant. Les poids sont calculés à partir d'un noyau gaussien centré sur l'ordonnée de l'axe de symétrie, et de variance $\frac{Largeur\ de\ l'image}{4}$. Pour un point situé sur le torse, de coordonnées (x,y)

$$Poids(x,y) = \mathcal{N}(\mu,\sigma)(y) = \frac{\exp\frac{(y-\mu)^2}{2\sigma^2}}{(\sqrt{2\pi}*\sigma)}$$

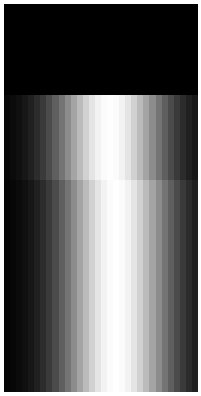

FIGURE 19 – Un exemple de carte des poids générée, mettant en évidence la séparation entre la tête, le torse et les jambes ainsi que le décalage horizontal de chaque partie dû au fait que les axes de symétrie du torse et des jambes ne sont à la même abscisse (l'image de base n'est pas la même que précédemment afin de mieux mettre en évidence la séparation entre le torse et les jambes).

4.1.0.3. Calcul de l'histogramme pondéré. L'implémentation fournie en Matlab de SDALF effectue avant le calcul une égalisation d'histogramme sur le canal V (pour pallier aux variations d'illumination). Afin d'éviter de prendre en considération la couleur du sol et de l'arrière plan (qui peuvent changer entre deux détections) les pixels de l'image ne sont inclus que s'ils appartiennent au masque de l'utilisateur (dit masque de segmentation).

Les histogrammes sont ensuite concaténés dans un seul vecteur, en commençant par l'histogramme complet (c'est à dire l'histogramme sur le canal H, puis S, puis V) de la tête, puis celui du torse et enfin celui des jambes. J'ai réalisé cela en utilisant les structures de données de la librairie C++ OpenCV, cependant la fonction de calcul d'histogramme fournie par OpenCV ne permet pas d'obtenir un histogramme pondéré, ce qui m'a poussé à la réécrire.

4.1.0.4. Calcul de la composante MSCR. Pour finir la composante basée apparence de notre descripteur, il reste à calculer la composante MSCR (c'est à dire l'information sur les régions où la couleur est relativement constante). La version Matlab de SDALF utilise en externe la librairie C fournie par Per-Erik Forssén [7] légèrement modifiée pour prendre en compte le masque de segmentation.

J'ai donc réutilisé la même version que j'ai moi-même adapté pour lui permettre de fonctionner avec ma version d'OpenCV (OpenCV 2, qui n'utilise pas les mêmes structures de données qu'OpenCV 1) et que j'ai directement inclus dans ma librairie après en avoir retiré les parties qui dépendaient de Matlab. Le résultat final n'est pas exactement le même que le résultat obtenu dans Matlab (probablement à cause des différences de représentation entre OpenCV et Matlab), mais en est assez proche :

7. Aimablement fourni sous licence GPL à l'adresse `http://www.cs.ubc.ca/~perfo/mscr/`

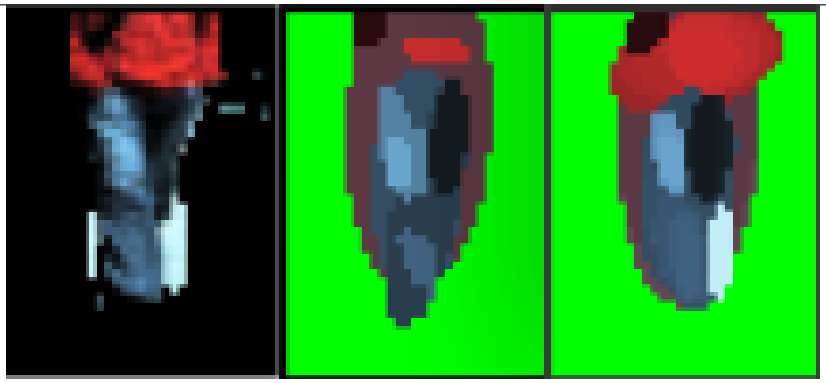

FIGURE 20 – A gauche, image d'origine (sur laquelle a été appliqué le masque de segmentation). Au centre, les régions de couleurs telles que détectées par la version Matlab de SDALF. A droite, celles détectées par ma version C++. La couleur verte indique une transparence.

Le vecteur retourné par la fonction de détection des MSCR et le vecteur des histogrammes pondérés ne sont pas comparés simultanément entre deux détections : pour chacune de ses composantes, SDALF calcule une distance au moment de la ré-identification. Ces distances sont ensuite combinées linéairement (avec des poids déterminés expérimentalement, [Farenzena et al., 2010] indiquant que ceux par défauts ont été calculés afin de maximiser l'aire sous la courbe CMC dans le cas du dataset VIPeR [8]).

4.2. Implémentation du descripteur basé morphologie.

Ce descripteur est bien plus direct à calculer. Les distances entre articulations retenues parmi celles présentées par [Barbosa et al., 2012] étant les suivantes :

- d_1 : distance entre la tête et le sol
- d_2 : distance entre le cou et l'épaule gauche
- d_3 : distance entre le cou et l'épaule droite
- d_4 : rapport entre la taille du torse et la taille des jambes
- d_5 : estimation de la taille de la personne tel que calculée par [Barbosa et al., 2012] (hauteur des jambes + taille du torse + taille du cou)
- d_6 : distance entre le cou et le sol
- d_7 : distance entre le milieu du torse et l'épaule droite

La paramétrisation du sol est fournie par la caméra RGB-D, le sol étant approximé par un plan, on dispose alors d'un point $P = \begin{bmatrix} x_p \\ y_p \\ z_p \end{bmatrix}$ de ce plan et d'un vecteur $\vec{N} = \begin{pmatrix} x_n \\ y_n \\ z_n \end{pmatrix}$ qui lui est normal. On peut alors calculer l'équation paramétrique du plan D de la manière suivante :

$$\forall P_x \begin{bmatrix} x \\ y \\ z \end{bmatrix} \in (D) : ax + by + cz + d = 0 \tag{1}$$

8. Disponible à l'adresse : `http://vision.soe.ucsc.edu/?q=node/178`

Avec :

$$a = x_n \tag{2}$$

$$b = y_n \tag{3}$$

$$c = z_n \tag{4}$$

$$d = -(x_n * x_p + y_n * y_p + z_n * z_p) \tag{5}$$

Les distances au sol sont alors calculées de la manière suivante :

$$dist(\begin{bmatrix} x \\ y \\ z \end{bmatrix}, (D)) = \frac{|ax + by + cz + d|}{\sqrt{a^2 + b^2 + c^2}} \tag{6}$$

Afin de ne pas donner plus d'importance à une de ces distances qu'à l'autre, elles sont individuellement normalisées en utilisant une loi normale de moyenne et de variance égales à celles calculées pendant la phase d'apprentissage.

Les résultats de l'implémentation de ce descripteur ont déjà été vus dans le chapitre précédent. Au moment de l'écriture de ce rapport, le descripteur est toujours en train d'être intégré au code (mis à jour) développé en première partie de stage afin de créer un programme de démonstration, que nous souhaiterions pouvoir faire fonctionner avant le début de ma thèse.

5. Conclusion générale et apport personnel

Rappel des objectifs du stage

L'objectif du stage était de développer un système de ré-identification de personnes et de l'intégrer aux travaux de l'équipe pour l'estimation de la pose de personnes par un réseau de caméras RGB-D.

Le premier objectif du stage était le portage du logiciel existant depuis la plate-forme ROS vers la plate-forme Genom3 durant le premier mois de stage. **Cet objectif a été atteint** et le logiciel a avec succès été testé en situation réelle.

Le second objectif était l'établissement d'un état de l'art des techniques de ré-identification de personnes utilisant les informations extraites par les caméras RGB-D. **Cet objectif a été atteint** et l'état de l'art en question a été intégré à ce rapport de stage. Cette partie a été réalisée entre la fin du mois d'avril et la fin mai.

Le troisième objectif était l'établissement et le prototypage d'une technique de ré-identification de personnes utilisant efficacement les informations RGB-D. **Cet objectif a été atteint**, l'évaluation présentée au chapitre chapitre 3.2 en présente par ailleurs les résultats. Cela s'est fait lors des mois de juin et début juillet.

A ce point, il m'a fallu avoir l'opportunité de récupérer et stocker les informations envoyées par le capteur RGB-D afin, par exemple, de créer des jeux de données. Les outils que j'ai trouvé m'ayant posé des problèmes, j'ai fini par en développer un moi-même. Cela ne fait cependant pas partie des objectifs du stage à proprement parler.

L'objectif suivant était la création d'une librairie C++ permettant de ré-identifier des personnes avec la technique développée. **Cet objectif a été en grande partie atteint** : le descripteur a été développé dans sa totalité mais la version finale ne comporte pas la comparaison des signatures MSCR, afin d'obtenir une ré-identification plus rapide (et pour des contraintes de temps). Cela a été réalisé entre le mois de juillet et fin août.

Enfin, l'objectif final du stage est la réalisation d'une démonstration combinant cette technique de ré-identification de personnes avec le logiciel porté en première partie de stage (qui devra être mis à jour). **Cet objectif est en cours de réalisation**, avec une échéance prévue pour mi-septembre.

Perspectives

A l'avenir, la librairie développée pourrait être utile aux travaux de vision par ordinateur du LAAS, notamment dans le cadre de ma thèse dont le début est prévu au premier octobre dans la même équipe. Des travaux prévus avec un ingénieur de recherche visant à la conception d'un *framework* unifié pour la vision par ordinateur pourraient la voir y être intégrée.

Il y a également au LAAS des personnes travaillant sur la détection automatique de chute par des caméras chez les personnes âgées. Ces chutes pouvant subvenir la nuit, il est possible que la librairie soit utilisée, ou au moins considérée, à l'avenir dans ce projet.

La démonstration prévue doit servir à Jean-Thomas Masse pour la soutenance de sa thèse prévue fin 2014.

Outils utilisés, difficultés rencontrées

Lors du premier mois de stage j'ai dû me familiariser avec les plateformes Robot Operating System [9] (ROS) de Willow Garage et Genom3 [10] (du LAAS), ainsi qu'avec OpenNI [11], librairie permettant d'interagir avec les capteurs RGB-D du LAAS. La programmation s'est faite avec le langage C++. L'état de l'art a été l'occasion pour moi d'affûter mes connaissances en LaTeX tandis que le prototypage du descripteur a été fait sous Matlab. L'écriture de la librairie C++ et la création de la démonstration m'ont fait travailler avec OpenCV [12], la librairie MSCR [13] et la librairie standard du C++. De plus, la mise en place de mon environnement de travail m'a fait comprendre plus en profondeur les subtilités de la compilation C++ et des systèmes GNU/Linux.

Au long de ce stage, j'ai rencontré plusieurs obstacles. L'adaptation à un nouvel environnement de travail a tout d'abord été difficile : habitué que j'étais à maîtriser ma station de travail, je n'étais pas administrateur ici. De plus, l'endroit et la manière de se procurer les différentes ressources (caméras,... etc) n'était pas toujours très explicite, d'autant plus que le personnel du LAAS se déplace beaucoup au sein des bâtiments. Il a donc fallu que je sois plus souple lorsque je prévoyais une séquence de choses à faire dans la journée, car cette manière de faire s'est vite révélée trop rigide en cas d'impossibilité de finir un travail avant de passer au suivant.

Du point de vue informatique, plusieurs problèmes rarement rencontrés à l'école d'ingénieurs se sont présentés. L'incompatibilité des structures de données entre différentes librairies a par-exemple pris une place importante dans le travail d'ingéniérie, d'autant plus qu'il m'était nécessaire de concevoir sagement mon code pour obéir à la contrainte temps-réel. L'utilisation de librairies non-conventionnelles, pas forcément créées avec la ré-utilisabilité en tête (notamment MSCR et sa version modifiée utilisée dans SDALF) font qu'il est parfois difficile d'obtenir un logiciel "propre" : on aboutit facilement à une suite de petites rustines destinées à faire fonctionner des morceaux de code qui ne font pas toujours ce à quoi on s'attend (et bien souvent non documentés...). C'est notamment à cette occasion que l'avantage de faire partie d'une équipe expérimentée s'est fait sentir : avoir les conseils de personnes compétentes s'est révélé salvateur.

Conclusion personnelle

Ce stage de recherche au LAAS-CNRS de Toulouse a enfin été pour moi l'occasion de beaucoup mieux cerner les aspects de l'informatique qui m'intéressent : si l'enseignement a toujours été ma motivation principale, grâce à la combinaisons d'activités de recherche et d'ingéniérie j'ai compris que l'aspect recherche m'attire beaucoup plus que la conception et la programmation. Cela ne ressortait pas forcément de mes expériences précédentes. Mon autonomie dans la prise de décisions s'est également

9. http://www.ros.org/
10. http://www.openrobots.org/wiki//
11. http://structure.io/openni
12. http://opencv.org/
13. http://www.cs.ubc.ca/~perfo/mscr/

améliorée, et j'ai pu avoir un premier aperçu (un second en réalité, mon stage précédent s'étant déroulé au laboratoire IRIT de Toulouse) du domaine de la recherche publique en France.

Ces apports à mon expérience professionnelle me serviront sans aucun doute tout au long du doctorat que je préparerai au LAAS, dans la même équipe et sous la direction de Frédéric Lerasle mon maître de stage, à partir du premier octobre 2014.

6. Glossaire

Genom3

Plate-forme pour la robotique développée par le LAAS-CNRS

Kinect

Modèle de caméra RGB-D commercialisé par Microsoft

MSCR (Maximally-Stable Colour Regions)

Partie centrée sur les zones de couleurs stables d'une image (l'image est approximée par un ensemble d'ellipses d'une couleur chacune) de SDALF.

RGB-D

Se dit d'une caméra pouvant capturer des informations de couleur (RGB pour Red-Green-Blue) et des informations de profondeur (D pour depth)

RHSP (Recurrent, Highly-Structured Patches)

Un des descripteurs dont SDALF est une combinaison. Contient des informations de texture d'une image.

ROS (Robot Operating System)

Plate-forme pour la robotique développée par Willow Garage

SDALF (Symmetry-Driven Accumulation of Local Features)

Descipteur basé sur l'apparence présenté par [Farenzena et al., 2010].

Xtion

Modèle de caméra RGB-D commercialisé par Asus

Paramètres libres

Dans le descripteur final proposé, il y a plusieurs paramètres libres. Voici ceux sur-lesquels on peut jouer :

La fenêtre de recherche lorsque l'on cherche les axes de symétrie et d'antisymétrie dans SDALF peut être raccourcie si l'on souhaite faire une recherche plus près des bords de l'image. En revanche on prend en compte moins d'information, donc la recherche est moins précise.

On peut jouer sur les différents ratios d'importance entre disparité de couleur et disparité en nombre de pixels , toujours lors de la recherche des axes de symétrie et d'antisymétrie afin d'obtenir des axes plus précis en fonction des conditions de capture (si la segmentation est parfaite, il faut lui donner plus d'importance, par exemple)

Le paramètre σ dans l'équation qui calcule les poids des pixels pour le calcul d'histogramme pondéré de SDALF est important : il permet de rejeter de l'information non importante et de donner beaucoup d'importance à des pixels pertinents. Il est tout à fait possible de jouer sur ce paramètre.

Les paramètres de calcul des MSCR que j'ai choisi sont ceux du code Matlab de SDALF. Peut-être qu'une petite optimisation de ces paramètres entraînerait de meilleures performances en ré-identification.

Les poids de fusion de (ie. l'importance accordée à) l'histogramme pondéré, les MSCR et l'information de morphologie pourraient être analysés plus en profondeur. La version Matlab de SDALF contient des poids optimisés pour la base VIPeR, mais il est fort probable qu'un cas d'utilisation différent comme le notre nécessite des poids différents.

Enfin, les poids accordés à chaque composante du descripteur basé sur la morphologie pourraient être revus, notamment en utilisant des méthodes d'optimisation numérique.

Références

[Abate et al., 2007] Abate, A. F., Nappi, M., Riccio, D., and Sabatino, G. (2007). 2d and 3d face recognition : A survey. *Pattern Recognition Letters*, 28(14) :1885 – 1906. Image : Information and Control.

[Araujo et al., 2013] Araujo, R. M., Graña, G., and Andersson, V. (2013). Towards skeleton biometric identification using the microsoft kinect sensor. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 21–26. ACM.

[Bak et al., 2010] Bak, S., Corvee, E., Brémond, F., and Thonnat, M. (2010). Person re-identification using haar-based and dcd-based signature. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pages 1–8. IEEE.

[Baltieri et al., 2011] Baltieri, D., Vezzani, R., and Cucchiara, R. (2011). Sarc3d : A new 3d body model for people tracking and re-identification. In Maino, G. and Foresti, G., editors, *Image Analysis and Processing – ICIAP 2011*, volume 6978 of *Lecture Notes in Computer Science*, pages 197–206. Springer Berlin Heidelberg.

[Barbosa et al., 2012] Barbosa, I. B., Cristani, M., Del Bue, A., Bazzani, L., and Murino, V. (2012). Re-identification with rgb-d sensors. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 433–442. Springer.

[Baüml et al., 2010] Baüml, M., Bernardin, K., Fischer, M., Ekenel, H., and Stiefelhagen, R. (2010). Multi-pose face recognition for person retrieval in camera networks. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pages 441–447.

[Berger et al., 2011] Berger, K., Ruhl, K., Schroeder, Y., Bruemmer, C., Scholz, A., and Magnor, M. A. (2011). Markerless motion capture using multiple color-depth sensors. In *VMV*, pages 317–324.

[Blanz and Vetter, 2003] Blanz, V. and Vetter, T. (2003). Face recognition based on fitting a 3d morphable model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(9) :1063–1074.

[Bouchrika and Nixon, 2008] Bouchrika, I. and Nixon, M. S. (2008). Exploratory factor analysis of gait recognition. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE.

[Farenzena et al., 2010] Farenzena, M., Bazzani, L., Perina, A., Murino, V., and Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2010)*, San Francisco, CA, USA. IEEE Computer Society.

[Han and Bhanu, 2006] Han, J. and Bhanu, B. (2006). Individual recognition using gait energy image. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(2) :316–322.

[Javed et al., 2008] Javed, O., Shafique, K., Rasheed, Z., and Shah, M. (2008). Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109(2) :146–162.

[Kumar and Babu, 2012] Kumar, M. S. N. and Babu, R. V. (2012). Human gait recognition using depth camera : A covariance based approach. In *Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing*, ICVGIP '12, pages 20 :1–20 :6, New York, NY, USA. ACM.

[Liu et al., 2012] Liu, C., Gong, S., Loy, C. C., and Lin, X. (2012). Person re-identification : what features are important ? In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 391–401. Springer.

[Madden and Piccardi, 2005] Madden, C. and Piccardi, M. (2005). Height measurement as a session-based biometric for people matching across disjoint camera views. In *Image and Vision Computing New Zealand*, pages 282–286.

[Masse et al., 2013] Masse, J.-T., Lerasle, F., Devy, M., Monin, A., Lefebvre, O., and Mas, S. (2013). Human motion capture using data fusion of multiple skeleton data. In Blanc-Talon, J., Kasinski, A., Philips, W., Popescu, D., and Scheunders, P., editors, *Advanced Concepts for Intelligent Vision Systems*, volume 8192 of *Lecture Notes in Computer Science*, pages 126–137. Springer International Publishing.

[Meden et al., 2012] Meden, B., Sayd, P., Lerasle, F., et al. (2012). Suivi par ré-identification dans un réseau de caméras à champs disjoints. In *Actes de la conférence RFIA 2012*.

[Munaro et al., 2014] Munaro, M., Fossati, A., Basso, A., Menegatti, E., and Van Gool, L. (2014). One-shot person re-identification with a consumer depth camera. In Gong, S., Cristani, M., Yan, S., and Loy, C. C., editors, *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 161–181. Springer London.

[Nixon and Carter, 2004] Nixon, M. S. and Carter, J. N. (2004). Advances in automatic gait recognition. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 139–144. IEEE.

[Preis et al., 2012] Preis, J., Kessel, M., Werner, M., and Linnhoff-Popien, C. (2012). Gait recognition with kinect. In *1st International Workshop on Kinect in Pervasive Computing*.

[Salvi et al., 2004] Salvi, J., Pages, J., and Batlle, J. (2004). Pattern codification strategies in structured light systems. *Pattern Recognition*, 37(4) :827–849.

[Satta, 2013] Satta, R. (2013). Appearance descriptors for person re-identification : a comprehensive review. *arXiv preprint arXiv :1307.5748*.

[Satta et al., 2013] Satta, R., Pala, F., Fumera, G., and Roli, F. (2013). Real-time appearance-based person re-identification over multiple kinect cameras. In *Proceedings of the 8th International Conference on Computer Vision Theory and Applications (VISAPP), Barcelona, Spain*, pages 21–24.

[Shotton et al., 2013] Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., and Moore, R. (2013). Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1) :116–124.

[Sivapalan et al., 2011] Sivapalan, S., Chen, D., Denman, S., Sridharan, S., and Fookes, C. (2011). Gait energy volumes and frontal gait recognition using depth images. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–6. IEEE.

[Stevenage et al., 1999] Stevenage, S. V., Nixon, M. S., and Vince, K. (1999). Visual analysis of gait as a cue to identity. *Applied Cognitive Psychology*, 13(6) :513–526.

[Urtasun and Fua, 2004] Urtasun, R. and Fua, P. (2004). 3d tracking for gait characterization and recognition. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 17–22.

[Yam et al., 2002] Yam, C., Nixon, M. S., and Carter, J. N. (2002). On the relationship of human walking and running : automatic person identification by gait. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 1, pages 287–290. IEEE.

[Zhao et al., 2006] Zhao, G., Liu, G., Li, H., and Pietikainen, M. (2006). 3d gait recognition using multiple cameras. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 529–534. IEEE.