Université de Montréal

Vidéosurveillance pour le contrôle de la prise de médicaments

par

Myriam Valin

Institut de génie biomédical

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences appliquées (M.Sc.A.)
en génie biomédical

Août, 2006

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

Vidéosurveillance pour le contrôle de la prise de médicaments

présenté par

Myriam Valin

a été évalué par un jury composé des personnes suivantes :

Michel Bertrand
président-rapporteur

Jean Meunier
directeur de recherche

Guillaume-Alexandre Bilodeau
membre du jury

Résumé

Dans le contexte de l'augmentation de la proportion d'aînés dans la population occidentale et des efforts fournis dans les services de maintien à domicile, un système de vision par ordinateur a été développé pour procéder au contrôle de la prise de médicaments. Le système détecte automatiquement la prise de médicaments en utilisant une seule webcaméra (*webcam*) de faible coût. La détection et le suivi de la personne au cours de la séquence vidéo sont effectués à l'aide de techniques basées sur la couleur. Quant à la reconnaissance de la prise de médicaments, elle est effectuée en utilisant un modèle de scénario à trois niveaux qui constitue la partie importante de la contribution de ce travail. Dans des conditions contrôlées, le système a une efficacité de détection de plus de 95 % tout en ayant un taux de fausses détections autour de 5 %. Compte tenu des résultats expérimentaux, la faisabilité d'un tel système est démontrée.

Mots clés – Vidéosurveillance, prise de médicaments, détection, suivi, reconnaissance d'activité, maisons intelligentes.

Abstract

In the context of the growing proportion of seniors in the Western World population and the efforts provided in home care services, we have developed a computer vision system for monitoring medication intake. The system detects automatically medication intake using a single low-cost webcam. Person detection and tracking over the video sequence is done using color-based techniques while the recognition of the medication intake activity is performed using our main contribution, a three-level scenario model. In controlled conditions, the system has a detection accuracy of over 95% with a false detection rate around 5%. Considering the experimental results, the feasibility of such a system is demonstrated.

Keywords – Video surveillance, medication intake, tracking, activity recognition, smart homes.

Table des matières

Liste des tableaux

Table des figures

Liste des sigles et abréviations

EM Expectation/Maximization

HMM Hidden Markov Model (modèle de Markov caché)

HSV Hue Saturation Valeur

MS Multi-state (scénario états-multiples)

OpenCV Open Source Computer Vision

RGB Red Green Blue (rouge vert bleu)

ROC Receiver Operating Characteristic (courbe ROC)

À tous ceux qui ont stimulé mon
intérêt pour le génie biomédical

Remerciements

Je souhaite remercier le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) pour son soutien financier. Je souhaite également remercier Alain St-Arnaud pour son aide et l'enthousiasme qu'il a démontré face au projet. Finalement, je voudrais remercier tous ceux et celles qui m'ont appuyée, tant moralement que techniquement, tout au long de ma maîtrise.

Avant-propos

Quand on parle des diverses applications médicales de l'informatique, on fait souvent référence au traitement d'images médicales obtenues à l'aide d'appareils d'imagerie. Que ce soit avec des images radiologiques, échographiques ou autres, l'objectif est de détecter une pathologie chez le patient. La vidéosurveillance intervient à un tout autre niveau de la santé du patient. Appliquée dans le contexte du maintien à domicile, elle permet de réduire les risques pour la santé du patient et d'obtenir une intervention rapide en cas de problème. Il s'agit davantage d'offrir un soutien au patient dans son milieu que de déterminer son état physique. Ce complément deviendra sûrement de plus en plus utile et nécessaire avec le vieillissement de la population auquel on assiste.

Chapitre 1

Introduction

Les applications en vidéosurveillance sont aussi nombreuses que diversifiées. Par exemple, dans le domaine biomédical, la vidéosurveillance peut être utilisée pour la détection de chutes, l'analyse d'habitudes alimentaires, la détection de situations à risque, comme une cuisinière allumée, ou encore pour le contrôle de la prise de médicaments. Dans ce mémoire, il sera question de cette dernière application.

Problématique

Le Canada assiste actuellement à un vieillissement important de sa population. Une étude de l'*Agence de santé publique du Canada* [1] estime qu'en 2016, la proportion de Canadiens de 65 ans et plus dans la population sera de 16 % et passera à plus de 22 % en 2041. Quand on sait que le nombre de maladies chroniques ou aiguës tend à augmenter avec l'âge, les répercussions sur la santé inquiètent.

L'augmentation des problèmes de santé entraîne une augmentation de la médication et plusieurs études démontrent que les aînés sont un groupe à risque élevé en ce qui a trait au mauvais usage des médicaments. En effet, de 18 % à 50 % des médicaments pris par les personnes âgées ne le sont pas de façon appropriée. L'impact est important puisqu'entre 19 % et 28 % des hospitalisations de patients de plus de 50 ans sont dues à des problèmes de médicaments, dont 40 % à la non-conformité aux indications [1].

Les personnes âgées ne sont toutefois pas les seules à être à risque élevé, les personnes ayant des troubles cognitifs le sont aussi. En 2003, on estimait que près

de 18 millions de personnes étaient atteintes de démence dans le monde et que ce nombre allait grimper à 34 millions d'ici 2025 [3]. De ce nombre, environ 55 % sont atteints de la maladie d'Alzheimer.

Les répercussions se font également sentir au niveau monétaire. En effet, des études ont estimé que les coûts engendrés par l'usage non conforme de médicaments sont, au Canada, de l'ordre de plusieurs milliards de dollars, soit un coût comparable à celui des maladies coronariennes. La prise inappropriée de médicaments se trouve parmi les facteurs contribuant à ces coûts. Le contrôle de la prise de médicaments semble donc inévitable.

Une étude effectuée par *Statistique Canada* [2] mentionne qu'en 1991, 28 % de la population de 65 ans et plus vivait seule. De plus, la Société Alzheimer du Canada rapporte que 50 % des gens atteints d'une démence vivent à domicile ; 29 % d'entre eux vivent seuls (34 800 personnes au Canada) et 2,4 % n'ont aucun aidant [4]. Un système de vidéosurveillance autonome permettrait d'effectuer un contrôle de la médication dans les cas où la personne ne peut se fier à quelqu'un d'autre. En plus de réduire les risques pour la santé de la personne, ce système permettrait de rassurer les membres de la famille.

Dispositifs d'assistance existants

De nombreux dispositifs d'assistance pour la prise de médicaments ont été développés, plusieurs fournissant un rappel verbal à l'utilisateur. Toutefois, la dépendance envers ces dispositifs est difficile à accepter et mène souvent à de la colère ou à l'abandon. Un autre problème avec ces dispositifs est que le système fournit des messages de sollicitation, qu'ils soient nécessaires ou non, et requièrent souvent une interaction avec la personne, causant ainsi des frustrations chez l'utilisateur. Un avantage de l'utilisation d'un système de vidéosurveillance est qu'il n'interfère pas dans les habitudes de la personne et, utilisé à des fins d'assistance, n'intervient qu'en cas d'oubli.

Caractéristiques du système

Le système créé fait partie d'un projet plus vaste de vidéosurveillance pour le maintien à domicile des personnes âgées. L'objectif est de déterminer la faisabilité d'un système de détection de prise de médicaments dans des conditions définies. Celui-ci a été conçu pour effectuer la vérification de la prise de médicaments dans le contexte d'une personne vivant seule et dont les médicaments se trouvent dans des bouteilles.

Afin de procéder à cette vérification, le système doit détecter si la personne prend un médicament. Pour ce faire, le processus est divisé en deux parties :

- Traitement bas niveau : détection et suivi de régions d'intérêt pour chaque image de la séquence vidéo.
- Traitement haut niveau : reconnaissance de l'activité de prise de médicaments basée sur les caractéristiques de ces régions d'intérêt.

Dans l'optique de rendre le système le plus accessible possible, une seule webcaméra est utilisée. Le traitement doit donc pouvoir s'effectuer avec des images numériques de faible qualité.

Organisation du mémoire

Ce mémoire est principalement divisé en quatre parties. La première partie consiste en une description globale du système de vidéosurveillance et de ce qui a été fait dans le domaine, et est constituée des chapitres 2 et 3. Les deux parties suivantes traitent des deux niveaux de traitement, soit l'un portant sur la détection et de suivi, et l'autre sur la reconnaissance d'activité. Celles-ci forment les chapitres 4 et 5. Finalement, la quatrième partie consiste en une présentation et en une discussion des résultats obtenus.

Plus précisément, une brève revue de l'état de l'art est présentée au chapitre 2. Un article portant directement sur un système de contrôle de prise de médicaments est analysé, de même que quelques articles traitant de diverses parties du traitement. Au chapitre 3, une présentation de la structure globale du système, de même que de ses caractéristiques, est faite.

Au chapitre 4, les régions d'intérêt sont définies et les divers algorithmes utilisés

pour la détection et le suivi sont présentés en détail.

Le processus de reconnaissance d'activité est ensuite décrit en profondeur au chapitre 5. Ce chapitre constitue la partie la plus importante de la contribution de ce mémoire.

Finalement, les résultats sont présentés au chapitre 6, auquel fait suite une discussion quant aux limites du système et aux travaux futurs au chapitre 7. Une brève conclusion est faite au chapitre 8.

Chapitre 2

Revue de l'état de l'art

Si le domaine de la vidéosurveillance a suscité beaucoup d'intérêt, celui-ci a peu été orienté vers le contrôle de la prise de médicaments. Toutefois, un article publié en 2005 traite directement de ce sujet. Cet article est donc analysé dans un premier temps. Dans un deuxième temps, il sera question de quelques articles portant sur divers aspects de la détection et du suivi de même que sur la reconnaissance d'activités humaines.

2.1 Surveillance de la prise de médicaments

Dans l'article de Batz [6], les auteurs présentent un système de vision par ordinateur permettant de surveiller le comportement de prise de médicaments.

Description de la méthode

Dans leur approche, les auteurs procèdent d'abord à la segmentation de régions de peau dans l'image. Puis, en se basant sur les boîtes englobantes minimales et les centroïdes de ces régions au cours du temps, ils détectent la présence d'occlusions main/visage ou main/main. La segmentation des mains et du visage se fait ensuite en analysant la différence entre la forme du visage dans l'image précédente et dans l'image courante pour les occlusions main/visage, et en analysant le squelette de la région en occlusion pour les occlusions main/main. Une fois les régions segmentées, le visage est identifié d'après des règles quant à sa forme et sa taille. Les yeux sont

alors localisés selon les intensités lumineuses et les gradients, et la bouche selon la couleur pour les lèvres. Le suivi des mains se fait finalement en plaçant des cercles aux deux endroits les plus probables de contenir des mains. Les auteurs se basent alors sur les dimensions des régions de couleur peau et sur leur forme.

La détection des bouteilles de médicaments se fait en cherchant des objets de formes rectangulaires et de proportion d'environ 2 : 1. Une reconnaissance d'objets est ensuite effectuée, sur les régions possibles, à l'aide d'une librairie de bouteilles définie. Le suivi se fait en mettant à jour le modèle de la bouteille et en effectuant des corrélations entre les régions possibles et le modèle soumis à des transformations géométriques, soit des rotations et des translations.

Finalement, les actions d'ouvrir et de fermer une bouteille de médicaments sont détectées en analysant l'orientation des lignes des doigts dans les régions circulaires des mains. L'activité d'avaler le médicament est détectée si la région formée par une main englobe une partie de la région de la bouche. La prise de médicaments complète est détectée si la séquence formée par les actions « ouverture de la bouteille », « mouvement main sur bouche » et « fermeture de la bouteille » se produit.

Analyse de la méthode

Un problème avec cette méthode est que pour fonctionner correctement, la segmentation initiale de l'image en régions de peau doit être assez précise, ce qui n'est pas nécessairement atteignable avec des images de faible qualité comme celles acquises avec les webcaméras utilisées. De plus, pour déterminer l'orientation des doigts nécessaire à la reconnaissance de l'ouverture et de la fermeture de la bouteille, le suivi des mains doit aussi être très précis et les doigts toujours visibles, ce qui n'est pas toujours possible. En effet, l'algorithme de segmentation des mains semble être sensible aux occlusions majeures des mains. Finalement, le fait que les auteurs considèrent qu'il y a prise de médicaments simplement si une séquence d'actions, vérifiées à chaque image, se produit, indépendamment de la durée de chacune de ces actions ou du temps entre celles-ci, implique que le taux de fausses détections pourrait devenir important dans des systèmes réels de vidéosurveillance à domicile.

Étant donné le manque de précision atteignable dans la partie de détection et

de suivi et la complexité de l'activité de prise de médicament, un algorithme plus complexe semble nécessaire.

2.2 Détection et suivi

Étant donné que l'accent n'a pas été mis sur la partie de détection et de suivi, seuls quelques aspects seront traités. Il sera principalement question ici de détection de visage et de mains, et de suivi.

2.2.1 Détection de couleur peau

Afin de procéder à la détection de visage et de mains, plusieurs méthodes effectuent d'abord la segmentation de régions de couleur peau. Dans [7], le visage est divisé en 5 régions (voir figure 2.1) et la classification des pixels se fait selon leur ressemblance avec des vecteurs couleur dans l'espace RGB $(\bar{r}_i, \bar{g}_i, \bar{b}_i)$, avec $i \in \{A, B, C, D, E\}$, représentant la couleur moyenne de chacune des régions. Les auteurs supposent alors que les positions relatives de ces régions demeurent fixes pour retrouver la position du visage.

Plusieurs auteurs semblent toutefois préférer l'espace de couleur YC_bC_r pour la détection de peau. Cet espace de couleur sépare la luminance (Y) de la chrominance (C_bC_r), ce qui permet une meilleure représentation de la couleur peau humaine. En effet, tel que mentionné dans [14], les composantes C_bC_r des pixels de couleur peau de personnes de toutes les ethnies se situent dans la même région du plan $C_b - C_r$. Les auteurs de l'article classifient les pixels comme étant de couleur peau si la distance de Mahalanobis entre leur couleur dans l'espace 2D $C_b - C_r$ et le modèle, créé à partir de pixels d'entraînement de couleur peau dans ce même espace, est sous un certain seuil.

Un modèle de couleur peau dans le même espace est aussi utilisé dans [8]. Toutefois, plutôt que de créer un modèle en extrayant la moyenne et la matrice de covariance d'un ensemble d'entraînement, les auteurs proposent d'utiliser des histogrammes. Un histogramme en deux dimensions, soit pour les composantes C_b et C_r, est créé pour la couleur peau et un autre pour les couleurs autres (non-peau).

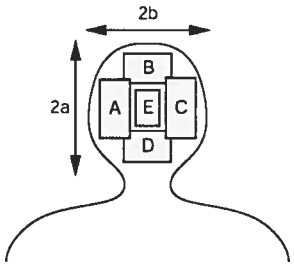


FIG. 2.1 – Zones caractérisant la couleur du visage.

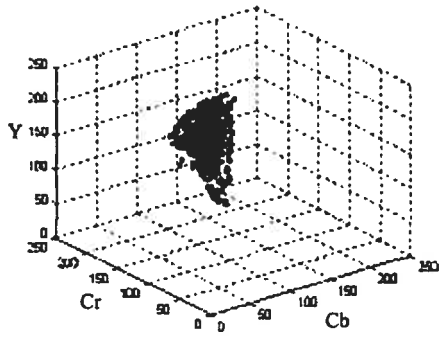


FIG. 2.2 – Représentation de la couleur peau dans l'espace YC_bC_r.

Ceux-ci sont ensuite convertis en distributions de probabilités $\hat{P}(C_bC_r|peau)$ et $\hat{P}(C_bC_r|\neg peau)$ utilisées pour déterminer la distribution de probabilité $\hat{P}(peau|C_bC_r)$ servant à la classification.

Dans [9], le modèle de couleur proposé fait aussi intervenir la composante de luminance dans la classification. Les auteurs font ressortir le fait que l'utilisation des simples composantes de chrominance ne permet de représenter la couleur peau que sur une petite plage d'intensités lumineuses, l'étendue de la projection de la couleur peau dans le plan $C_b - C_r$ semblant plus grande pour des valeurs de Y moyennes que pour de petites ou de grandes valeurs de Y (voir figure 2.2). Les frontières de décisions devraient alors être différentes selon la luminance. La couleur peau est donc représentée par un modèle formé de trois sous-modèles, un pour chaque plage de luminance. Dans l'étape de classification, chaque pixel est comparé aux trois sous-modèles et, si sa distance de Mahalanobis avec l'un de ceux-ci est inférieure à un seuil, le pixel est classifié comme étant de couleur peau.

Un autre espace couleur utilisé pour la détection de couleur peau est HSV. Le système de couleur HSV permet de séparer la teinte (H) de la saturation (S), soit la concentration de la couleur, de l'intensité lumineuse (V). Dans [17], la classification des pixels se fait selon un histogramme de couleur à une dimension n'utilisant que la composante H. En effet, cette composante s'avère être la même pour toutes les ethnies, tout comme c'est le cas pour les composantes C_bC_r de l'espace YC_bC_r. Afin d'éviter des erreurs dues à la faible luminosité, les pixels avec une faible valeur de V sont ignorés.

Finalement, dans [13], l'auteur utilise un autre espace couleur que YC_bC_r ou HSV, mais toujours séparant la chrominance de la luminance, pour créer un histogramme modèle. Contrairement à ce qui est fait dans [14] et [8], la composante de luminosité n'est pas ignorée. L'algorithme présenté dans l'article étant utilisé dans la partie de détection et de suivi du système présenté dans ce mémoire, il est décrit en détail à la section 4.1. L'espace de couleur utilisé s'est d'ailleurs montré, dans les conditions expérimentales du projet, plus efficaces que YC_bC_r et HSV.

Les couleurs d'une image sont affectées par les changements d'intensités lumineuses, mais elles le sont aussi par le type d'éclairage dans la pièce. En effet, les couleurs ne sont pas identiques selon que la source lumineuse est naturelle (comme le soleil) ou artificielle (comme une lampe halogène par exemple). Dans les cas où la lumière ambiante n'est pas blanche, une étape de balance des blancs permettrait de corriger le problème. Pour ce faire, le système devrait avoir de l'information *a priori* quant à la position d'un élément blanc dans l'image, soit par exemple la position du réfrigérateur.

2.2.2 Détection de visage

Afin de détecter le visage et les mains, la plupart des algorithmes extraient d'abord des régions de couleur peau (objets) à l'aide d'algorithmes de détection de couleur peau et de composantes connexes (*connected component*). Les régions peu susceptibles de contenir une partie du corps sont ensuite éliminées selon diverses règles, notamment quant à leur taille. Les objets restants sont alors analysés afin d'identifier la tête et les mains.

Dans [14], la méthode utilisée pour distinguer le visage des mains est basée sur le fait que le visage est un objet rigide alors que les mains sont des objets non rigides. Considérant que la forme du visage peut être approximée par une ellipse, les auteurs utilisent trois tests pour déterminer si un objet correspond à un visage. Une fois que l'ellipse représentant le mieux chaque objet est déterminée, ils vérifient d'abord si l'orientation de celle-ci est comprise entre -40^o et 40^o par rapport à la verticale. Ensuite, le rapport de forme (*aspect ratio*) est calculé à l'aide des moments d'inertie minimum et maximum et doit se trouver entre 1,4 et 1,8. Finalement, une mesure de similarité, soit la différence entre la région de l'objet à l'intérieur et à l'extérieur de l'ellipse divisée par l'aire de l'ellipse, est effectuée et doit être supérieure ou égale à 0,8. Si les trois conditions sont respectées, l'objet est identifié comme étant un visage.

Des critères de forme sont aussi utilisés dans [9]. En effet, pour chaque objet, la circularité C et le rapport de forme R sont calculés et seuls les objets dont les valeurs obtenues sont supérieures à des seuils θ_C et θ_R définis sont conservés. Les formes des objets restants sont alors raffinées pour leur donner une forme elliptique. Ensuite, une boîte englobante rectangulaire est extraite pour chaque objet, de même qu'une image d'intensité correspondant à cette région. Cette dernière est alors redimensionnée et placée sous forme de vecteur. Ce vecteur permettra d'obtenir, à partir d'un ensemble de propriétés de visage, un vecteur caractéristique w. Les propriétés de visage sont obtenues d'après une analyse en composantes principales (ACP). Un objet est alors classifié comme étant un visage si $\frac{p(w|visage)}{p(w|\neg visage)} > \tau_1$ et $p(w|visage) > \tau_2$ où τ_1 et τ_2 sont déterminés expérimentalement. Les densités de probabilité sont déterminées par la somme de gaussiennes sur chacune des composantes de w et pondérées par des facteurs. Les paramètres des gaussiennes de même que les facteurs impliqués sont obtenus en appliquant l'algorithme EM sur un ensemble d'entraînement. De par l'entraînement nécessaire et le nombre de paramètres à déterminer, cette méthode semble plus complexe à implémenter que celle utilisée dans [14].

Les articles présentés considèrent ensuite que les objets ne correspondant pas à des visages sont des mains. D'autres méthodes de détection de visage ont également été développées, mais ne seront pas traitées dans ce mémoire.

2.2.3 Suivi d'objets

Dans [8], les auteurs ont recours aux filtres de Kalman discrets pour le suivi du visage. Les filtres de Kalman discrets permettent d'estimer de façon efficace, et récursive, l'état d'un système linéaire, soit dans ce cas-ci la position du centre de l'objet ou sa taille, à partir d'informations *a priori* et de mesures réelles (informations *a posteriori*). Les détails des filtres de Kalman discrets sont présentés dans [15].

Les filtres de Kalman ont aussi été utilisés dans [16], mais pour le suivi de mains. Plutôt qu'extraire des régions de peau dans toute l'image et sélectionner celle se trouvant le plus près de la position prédite, l'estimation de la nouvelle position sert de point de départ pour la recherche des régions de couleur peau, limitant ainsi les calculs nécessaires. On suppose que la main possède une certaine surface minimale, ce qui permet de définir la région de recherche.

D'autres méthodes d'estimation pour le suivi existent également. Les auteurs de [7] mentionnent, entre autres, l'approche d'hypothèses multiples en arbre et l'approche par maximum de vraisemblance. Les filtres à particules peuvent aussi s'avérer une bonne solution, particulièrement dans les cas de déplacements rapides ou très aléatoires.

2.3 Reconnaissance d'activité humaine

2.3.1 Modèle déclaratif

Dans [18], les auteurs présentent une approche à deux niveaux : un niveau « événement », représentant les changements significatifs dans l'état de la scène, et un niveau « scénario », dont l'objectif est de reconnaître des situations à long terme modélisées par des combinaisons d'événements. La notion de *fait* est alors introduite. Un fait correspond à un objet défini par une série d'attributs, soit un nom, un type, une date, etc. Ceux-ci sont vérifiés à chaque image de la séquence vidéo. Deux types de faits sont définis : concrets et abstraits. Les faits concrets correspondent à des personnes, à de l'équipement ou encore à des environnements, alors que les faits abstraits correspondent à des états, à des événements ou à des scénarios. Chaque fait

abstrait est modélisé par un ensemble d'attributs correspondant aux variables (faits concrets) impliquées, aux conditions et au résultat produit (création du fait abstrait). Un scénario est donc reconnu si l'ensemble des conditions de son modèle est respecté. Pour ce faire, l'ensemble des faits produits au cours du temps est considéré.

Dans cet algorithme, les modèles impliquent généralement peu de variables et les conditions définies sont également peu nombreuses, simples et fixes. L'algorithme semble donc bien se prêter à des scénarios relativement simples, mais être plus difficilement utilisable à mesure que ceux-ci se complexifient.

2.3.2 Réseaux de Pétri

Dans [19], la reconnaissance se fait à l'aide de réseaux de Pétri. Un réseau de Pétri est un modèle abstrait de flux d'information dans un système, qui permet de représenter des événements de façon séquentielle, simultanée et synchronisée. Celui-ci est formé d'événements, de transitions et de jetons. La transition vers un événement ne peut être faite que lorsque tous les jetons le précédant sont présents et que celui-ci est détecté. Une fois la transition faite, un jeton est placé à la suite de l'événement. Un scénario est donc reconnu lorsqu'un jeton est placé à la suite du dernier événement de la séquence. Un événement pouvant dépendre de plusieurs jetons, du traitement en parallèle et du pipelinage peuvent être effectués, ce qui permet aux réseaux de Pétri d'être très efficaces. Toutefois, aucune contrainte de durée n'est introduite pour effectuer les transitions.

2.3.3 HMM

Un HMM est utilisé dans [20] pour analyser le comportement des personnes âgées pendant leurs repas. Les auteurs déterminent le début et la fin du repas pour en déterminer la fréquence et la durée. Le modèle créé possède trois états *:* « *mouvement main s'approche de la bouche* », « *mouvement main s'éloigne de la bouche* » et « *mouvement non relié à l'activité de manger* », soit, par exemple, le déplacement de la main entre différents éléments de vaisselle. Un état « *pas important* » (*don't care*) est aussi considéré pour les cas où aucun mouvement n'est détecté. Cet algorithme

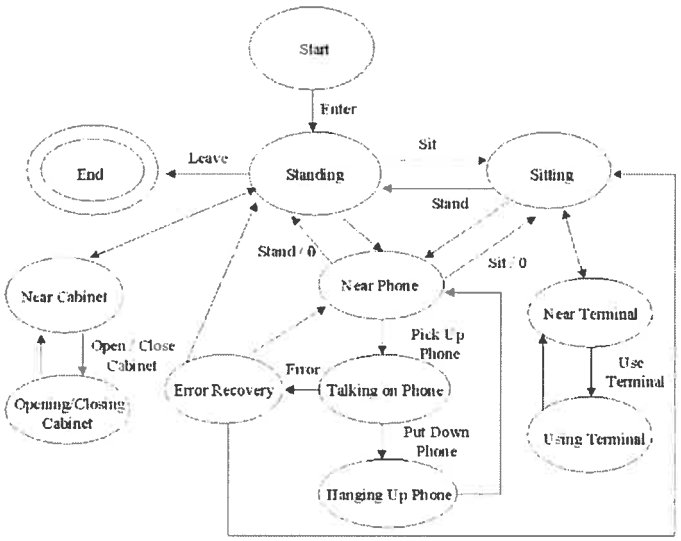


FIG. 2.3 – Schéma de la machine à états utilisée dans [21].

semble davantage intéressant lorsqu'il y a une forte composante d'apprentissage, ce qui n'est pas le cas ici car l'activité de prise de médicaments est bien définie.

2.3.4 Machine à états

Dans [21], les auteurs modélisent des actions propres à des environnements de travail par une machine à états, constituée d'états et de transitions entre les états. Les transitions se font lorsqu'une action est détectée. Chaque action est liée à une condition de reconnaissance. Par exemple, l'action « *prend un objet* » est liée à la condition « *la position initiale de l'objet a changé après que la personne et l'objet se soient éloignés* ». L'approche utilisée dépend donc beaucoup du contexte. En effet, le système doit avoir de l'information quant aux zones d'entrée et de sortie de la pièce et quant aux positions des objets d'intérêt et à la façon dont ils sont utilisés. Un schéma des relations entre les états pour un environnement comprenant un téléphone, une armoire et un ordinateur est présenté à la figure 2.3.

La boîte « Error Recovery» est utilisée afin d'éviter de rester pris dans l'état « Talking on Phone » en cas d'erreur pendant cet état. De telles boîtes sont utilisées pour les cas où un état dépend d'une seule action pour procéder à une transition vers un autre état.

Cette méthode semble intéressante dans les cas où l'on veut modéliser plusieurs actions se produisant dans un même environnement. La reconnaissance de l'action correspondant à chaque transition nécessite toutefois que chacune d'entre elles soit relativement simple.

2.3.5 Approche bayesienne

Finalement, Hongueng utilise une approche bayesienne dans le cadre d'un système automatique de vidéosurveillance [11]. L'auteur calcule la probabilité d'occurrence d'une activité en utilisant des réseaux bayesiens à plusieurs niveaux du traitement. La méthode développée et présentée dans ce mémoire est basée sur les travaux de l'auteur. En effet, celle-ci se prête bien à la reconnaissance d'activités relativement complexes qui n'impliquent pas nécessairement de nombreuses actions. Cette méthode est expliquée plus en détail au chapitre 5.

Chapitre 3

Structure globale du système

Avant de se lancer dans la description détaillée des algorithmes, il convient de décrire le contexte dans lequel la prise de médicaments est détectée, de même que les paramètres techniques entourant le système, notamment au niveau de la caméra et des bouteilles de médicaments.

3.1 Contexte

Les façons de prendre les médicaments sont multiples. Ils peuvent être placés dans des dosettes ou dans des bouteilles et être pris dans la salle de bains, la chambre à coucher, la cuisine, etc. Un système permettant de détecter la prise de médicaments, indépendamment du contexte, est très complexe. Afin de simplifier le problème, on considère que la personne prend ses médicaments toujours au même endroit, soit à la table de la cuisine, face à la caméra et que les médicaments se trouvent dans différentes bouteilles. La validité de ce contexte pour des applications réelles a été confirmée par le neuropsychologue Alain St-Arnaud, travaillant avec les personnes âgées et affilié au projet. Selon Dr St-Arnaud, si les habitudes de prise de médicaments varient beaucoup d'une personne à l'autre, elles varient très peu chez la même personne, permettant ainsi de considérer un contexte d'application particulier.

La table de cuisine a été préférée aux autres endroits, car cela permettait d'avoir une vue des bouteilles de médicaments et de la personne tout au long de la prise de médicaments. En effet, si la détection s'était faite pour des bouteilles situées dans

l'armoire de la pharmacie ou encore dans une armoire de la cuisine, il aurait été possible que la caméra ne voit pas les bouteilles, celles-ci étant d'abord cachées par l'armoire puis par les mains de la personne.

Le système considère également qu'une seule personne est présente dans la pièce et qu'elle n'est pas trop penchée lors de la prise de médicaments, de sorte que le visage demeure visible du point de vue de la caméra. La contrainte quant au nombre de personnes dans la pièce est davantage liée aux algorithmes de détection et de suivi, et pourrait être éliminée si d'autres algorithmes étaient utilisés. Toutefois, celle-ci est valide dans le contexte dans lequel le système est utilisé chez des personnes vivant seules.

3.2 Considérations techniques

Le contexte entourant l'activité de prise de médicaments étant donné, il convient de traiter des aspects plus techniques entourant celle-ci.

3.2.1 Bouteilles de médicaments

Tout d'abord, tel que soulevé au chapitre 2, les images acquises avec les webcaméras utilisées sont de faible qualité, ne permettant pas de reconnaître des bouteilles blanches selon les inscriptions qui s'y trouvent. De plus, les bouteilles pouvant se cacher en partie les unes les autres, celles-ci doivent être modifiées. La solution choisie est l'application de bandes de couleur. Là encore, M. St-Arnaud confirme la faisabilité d'une telle option. Afin de ne pas causer de confusion, les bandes de couleur sont simplement collées sur la bouteille et non sur le couvercle, évitant d'obtenir deux régions pour la même bouteille lorsque celle-ci est ouverte.

3.2.2 Caméra

Maintenant que l'environnement de prise de médicaments est défini, on doit déterminer l'emplacement de la caméra. Plusieurs positionnements ont été envisagés. Tout d'abord, de haut, de façon à avoir une vue plongeante de la personne, tel que montré à la figure 3.1a. Cette option a rapidement été éliminée puisque le visage

n'est pas visible et que les mains, de même que les bouteilles de médicaments lorsque manipulées, se trouvent souvent en occlusion avec la tête. Ce problème est encore davantage marqué pour une personne dont le dos est courbé et donc davantage repliée sur elle-même. De plus, les bandes de couleur recouvrant les bouteilles ne sont souvent pas visibles.

Une autre possibilité aurait été de placer la caméra de côté à la personne. Cette option a toutefois dû être rejetée puisqu'il aurait été difficile de procéder au suivi des mains. En effet, ces dernières, généralement placées de part et d'autre du corps, paraissent souvent incorrectement en contact du point de vue de la caméra, comme illustré à la figure 3.1b.

L'option de placer la caméra directement en face de la personne a ensuite été envisagée. Cette solution a comme avantage de permettre de bien voir tant les bandes de couleur sur les bouteilles de médicaments que le visage de la personne, en plus de limiter les faux contacts entre les mains. Toutefois, il pourrait arriver que du point de vue de la caméra, une main se trouvant devant une autre ou se trouvant devant une bouteille de médicaments soit perçue comme étant en contact. Cette situation est présentée à la figure 3.1c, où les mains paraissent en contact avec les trois bouteilles de médicaments.

Finalement, ce problème a été limité en plaçant la caméra devant la personne, mais légèrement au-dessus de sa tête, comme à la figure 3.1d, diminuant ainsi les occlusions. D'autres solutions possibles seront également discutées plus loin. Toutefois, il est à noter que la reconnaissance d'activité n'est pas nécessairement faussée par la présence d'occlusions.

En ce qui a trait à la distance entre la caméra et la personne, les résultats semblent meilleurs lorsque la caméra se trouve relativement près de la personne. Ainsi, la taille des objets d'intérêt est plus grande, ce qui facilite leur détection et leur suivi.

Pour ce qui est des paramètres de la caméra, la fréquence d'acquisition des images a été fixée à 15 images par seconde, et la taille des images a été fixée à 320x240 pixels, soit la taille maximale car, à cette fréquence, la caméra ne duplique pas trop souvent les images (permet d'éviter les retards ou *lags*).

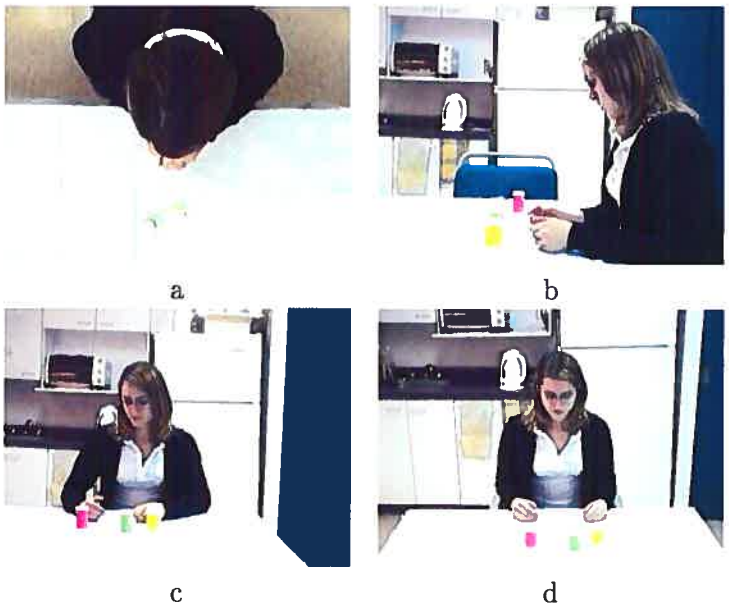


FIG. 3.1 – Vues obtenues selon une position de la caméra a) de haut, b) de côté, c) directement de face et d) de face et surélevée.

3.2.3 Traitement temps réel

Afin de pouvoir procéder à la détection de prise de médicaments en temps réel, le processus de détection et de suivi ne s'effectue que lorsque la personne se trouve complètement dans l'image, soit au moment où elle est assise à la table. Ainsi, lorsque la personne n'est pas en situation de prise de médicaments, le système ne travaille pas. Ceci est un élément important puisque, même si le système procède à la détection et au suivi en temps réel, plus de temps peut être nécessaire pour la reconnaissance d'activité. Ainsi, le système pourra utiliser le temps où rien ne se passe après la prise de médicaments pour terminer le processus de reconnaissance et ainsi ne pas causer de débordement d'informations à traiter (*overflow*).

Chapitre 4

Détection et suivi

Comme mentionné dans l'introduction, la première partie du processus de détection de prise de médicaments est un traitement bas niveau consistant à détecter et à suivre des régions d'intérêt pour chaque image de la séquence vidéo. La reconnaissance est ensuite basée sur l'évolution de ces régions que l'on appellera objets mobiles.

Dans le système, trois types d'objets mobiles sont utilisés : la tête de la personne, ses mains et les bouteilles de médicaments. La figure 4.1 présente ces objets. Dans cette image, trois bouteilles de médicaments sont impliquées.

Afin de procéder à la détection et au suivi des objets mobiles, trois algorithmes sont utilisés. Ce chapitre présente ces algorithmes. Il est toutefois important de garder en mémoire que cette partie ne constitue pas l'essentiel de la contribution du travail et donc que la recherche d'algorithmes s'est faite dans l'optique d'obtenir une solution assez bonne pour permettre de procéder à la reconnaissance, sans toutefois

FIG. 4.1 – Objets mobiles détectés et suivis.

être nécessairement robuste et rapide. En effet, on supposera que les prises de vidéos se font dans un environnement contrôlé avec des conditions relativement fixes. Pour chacun des algorithmes, les hypothèses émises sont mentionnées.

4.1 Tête

L'algorithme de détection et de suivi de tête utilisé est celui créé par Stan - Birchfield et est présenté dans [13]. L'algorithme a été choisi, car les résultats obtenus semblaient être bons et facilement reproduisibles, et parce qu'il peut être utilisé en temps réel. Son implémentation étant disponible sur le site internet du chercheur, celle-ci a directement été utilisée. L'utilisation de la librairie d'images *OpenCV* [5] développée par *Intel* a toutefois dû être intégrée, celle utilisée par l'auteur n'étant pas disponible. Les détails de l'algorithme sont présentés ici.

4.1.1 Principe

Dans l'algorithme, on suppose que la forme de la tête peut ête modélisée par une ellipse en deux dimensions dont le rapport hauteur/largeur α est fixe, mais dont la taille peut varier d'une image à l'autre. Le modèle est présenté à la figure 4.2.

Le principe est de déterminer, pour chaque image, l'ellipse qui représente le mieux la tête, en effectuant une recherche locale autour d'une position initiale. Pour déterminer la meilleure ellipse, l'algorithme se base sur l'intensité du gradient sur son périmètre et sur la vraisemblance de couleur peau à l'intérieur de celle-ci.

En utilisant la notation $\mathbf{s} = (x, y, \sigma)$ pour représenter une ellipse de centre (x, y) et de longueur de demi-petit axe σ, la meilleure ellipse $\mathbf{s}^*$ est

$$\mathbf{s}^* = arg \max_{\mathbf{s}_i \in S} \{\phi_g(\mathbf{s}_i) + \phi_c(\mathbf{s}_i)\}$$

où $\phi_g(\mathbf{s}_i)$ et $\phi_c(\mathbf{s}_i)$ correspondent respectivement à des scores de correspondance pour l'intensité du gradient et la vraisemblance de couleur peau et S correspond à l'ensemble des ellipses possibles dans une région définie autour d'une position

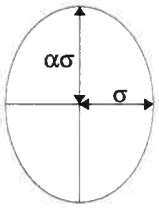


FIG. 4.2 – Ellipse servant à modéliser la forme de la tête.

initiale. L'étendue de cette région, fixée expérimentalement, dépend d'hypothèses quant à la vitesse possible de la tête.

4.1.2 Gradient

L'intensité du gradient d'un pixel correspond au taux de changement de l'intensité dans une petite région avoisinante. La carte de gradients de l'image est obtenue en appliquant un filtre gaussien puis un filtre de Sobel à l'image convertie en niveaux de gris. La figure 4.3 présente un exemple de carte d'intensités de gradients.

Plutôt que de simplement calculer la somme des intensités des gradients sur le périmètre de l'ellipse, le score de gradient calculée de sorte que la direction du gradient soit perpendiculaire au périmètre. Celui-ci est alors défini selon l'équation

$$\phi_g(\mathbf{s}) = \frac{1}{N_\sigma} \sum_{i=1}^{N_\sigma} |\mathbf{n}_\sigma(i) \cdot \mathbf{g}_\mathbf{s}(i)|, \tag{4.1}$$

où $\mathbf{g}_\mathbf{s}(i)$ est l'intensité du gradient au pixel i du périmètre de l'ellipse $\mathbf{s}$, N_σ est le nombre de pixels sur le périmètre d'une ellipse de demi-petit axe σ et $\mathbf{n}_\sigma(i)$ est le vecteur normal à l'ellipse au pixel i.

Pour pouvoir comparer le score de gradient à celui de la couleur, celui-ci est normalisé selon l'équation

$$\bar{\phi}_g(\mathbf{s}) = \frac{\phi_g(\mathbf{s}) - min_{\mathbf{s}_i \in S} \phi_g(\mathbf{s}_i)}{max_{\mathbf{s}_i \in S} \phi_g(\mathbf{s}_i) - min_{\mathbf{s}_i \in S} \phi_g(\mathbf{s}_i)}. \tag{4.2}$$

Plutôt que d'utiliser l'image en niveaux de gris, la carte de gradients aurait pu être remplacée par l'image de contour de la silhouette de la personne calculée à

FIG. 4.3 – Exemple de carte d'intensité de gradients.

partir de l'image d'avant-plan (*foreground*). On entend ici par image d'avant-plan l'image binaire correspondant à la région en mouvement, par opposition à l'arrière-plan (*background*) qui lui est fixe. Cette approche n'a pas été retenue, car elle exige une mise à jour de l'arrière-plan, ce qui est une opération relativement complexe et qui augmenterait le temps de calcul.

4.1.3 Vraisemblance de couleur peau

Contrairement à ce qui a été fait par plusieurs chercheurs, le modèle de couleur peau utilisé ici n'est pas représenté par une moyenne et une matrice de covariance [14], mais par un histogramme. L'avantage avec l'utilisation d'un histogramme est que ce modèle non paramétrique est très simple comparativement à l'utilisation d'une approche utilisant des gaussiennes. Ainsi, on pourra obtenir un score élevé sur l'ensemble de la région de l'ellipse, soit sur toute la surface de la tête, et non pas seulement sur la région du visage. L'ellipse ayant le plus haut score quant au gradient pourra aussi avoir un score élevé quant à la vraisemblance de couleur même si la tête est inclinée.

La carte de vraisemblance de couleur peau est obtenue en calculant l'intersection entre l'histogramme du modèle de couleur peau M et l'histogramme de couleur de l'image à l'intérieur de l'ellipse I. Ainsi, pour chaque ellipse s, on calcule le score de couleur selon l'équation

$$\phi_c(\mathbf{s}) = \frac{\sum_{i=1}^{N} min(I_{\mathrm{s}}(i), M(i))}{\sum_{i=1}^{N} I_{\mathrm{s}}(i)}, \tag{4.3}$$

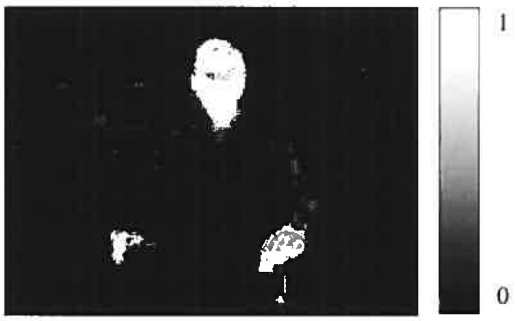


FIG. 4.4 – Exemple de carte de vraisemblance de couleur peau normalisée.

où $I_s(i)$ et $M(i)$ sont le nombre de pixels dans la i^e case des histogrammes, de taille N, de l'image à l'intérieur de l'ellipse et du modèle. Un exemple de carte de vraisemblance de couleur peau est présenté à la figure 4.4.

L'espace couleur utilisé est un espace combinant les composantes R, G et B. Cet espace couleur, que l'auteur appelle color123, est défini comme suit :

color1 = max(0, min(255, (b-g) * 10 + 128))

color2 = max(0, min(255, (g-r) * 10 + 128))

color3 = max(0, min(255, (b+g+r) / 3)).

Tout comme l'espace couleur YC_bC_r, celui-ci contient deux composantes de chrominance, soit color1 et color2, et une composante de luminosité, soit color3. De plus, contrairement à ce qui a été fait dans [8, 14, 17], la composante de luminosité n'est pas ignorée, ce qui permet de différencier des cheveux bruns très foncés (ou noirs) d'un arrière-plan blanc par exemple. Dans le contexte expérimental, les résultats obtenus avec ce nouvel espace couleur ont été meilleurs qu'avec l'espace couleur YC_bC_r, mais il serait intéressant de vérifier lequel est le plus robuste dans des conditions changeantes.

Dans l'implémentation de l'algorithme, il est possible de faire en sorte que l'histogramme s'adapte à mesure que la tête se déplace, ce qui permet de suivre la tête même si celle-ci est complètement tournée. Toutefois, cette option n'est pas nécessaire ici puisque, selon les hypothèses émises à la section 3.1 quant au contexte, la personne est toujours de face, ou légèrement inclinée, lorsqu'elle est dans le champ de vision de la caméra. De plus, comme il en sera question plus loin, la carte de

vraisemblance de couleur est utilisée dans une autre partie pour la détection des mains.

Tout comme pour le score du gradient, le score de couleur est normalisé selon l'équation

$$\bar{\phi}_c(s) = \frac{\phi_c(s) - min_{s_i \in S}\phi_c(s_i)}{max_{s_i \in S}\phi_c(s_i) - min_{s_i \in S}\phi_c(s_i)}. \quad (4.4)$$

4.1.4 Initialisation du système

Comme il a été vu au chapitre 3, l'étape de détection et de suivi ne s'effectue que lorsque la personne se trouve complètement dans l'image, soit lorsque la région d'avant-plan est totalement incluse dans l'image (ne touche pas aux bords). Ceci est nécessaire afin de s'assurer que la tête et les mains se trouvent dans l'image avant de commencer la détection et le suivi. Il convient de préciser que l'on considère que le cadrage de l'image est fait de sorte que la personne ne touche pas aux bords lorsqu'elle prend ses médicaments.

L'image d'avant-plan (binaire) est calculée selon la distance entre l'image courante et l'image de référence, soit dans ce cas-ci une image prise au début de la séquence, alors que la personne n'est pas encore visible. La distance d pour un pixel de l'image est définie comme étant

$$d = min(255, 10 * \sqrt{0.4 * (Y - Y_{ref})^2 + (Cb - Cb_{ref})^2 + (Cr - Cr_{ref})^2},$$

où Y et Y_{ref} correspondent respectivement aux composantes Y de la couleur du pixel dans l'image courante et dans l'image de référence. Si cette distance est supérieure ou égale à 255 (blanc), le pixel appartient à l'avant-plan. Le facteur 0.4 permet de diminuer l'influence de la luminosité, sans toutefois la négliger complètement. Dans le système, l'avant-plan n'ayant pas à être très précis, l'image de référence n'est pas mise à jour.

Une fois la personne complètement dans l'image, l'initialisation de la position

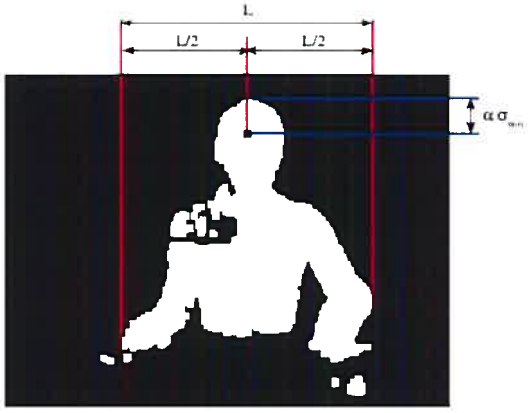


FIG. 4.5 – Exemple d'image d'avant-plan et initialisation de la position de la tête.

de l'ellipse s'effectue. La grandeur du demi-petit axe σ est alors fixée à la valeur moyenne, soit $\frac{\sigma_{max}+\sigma_{min}}{2}$ (voir plus bas). Tel qu'illustré à la figure 4.5, la position horizontale du centre de l'ellipse est fixée à la position horizontale du centre de la région d'avant-plan la plus haute dans l'image. La position verticale du centre est fixée, quant à elle, à une distance correspondant à la longueur du demi-grand axe moyen du haut de la région. Ceci n'est toutefois valide que dans l'hypothèse où la tête est la région du corps la plus haute dans l'image.

4.1.5 Détermination des paramètres

Pour que le système puisse calculer le score de couleur, l'histogramme formant le modèle de couleur peau doit être créé. Pour ce faire, un ensemble d'images comprenant un visage dont la taille respecte les dimensions fixées doit être présenté au système. Pour chaque visage, les paramètres de l'ellipse représentant le visage (x_i, y_i, σ_i) doivent être donnés.

Les longueurs minimale et maximale du demi-petit axe (σ_{min} et σ_{max}), le rapport hauteur/largeur (α) et les plages de déplacements horizontal et vertical sont fixés, mais pourraient être calculés à partir des données fournies au système et de la fréquence d'acquisition.

4.2 Mains

Au chapitre 2, il a été question de méthodes de suivi permettant de prédire la prochaine position d'un objet. Toutefois, les mains étant souvent en situation d'occlusion, ces algorithmes ne semblent pas bien se prêter à leur suivi. D'autre part, l'utilisation de modèle fixe pour représenter les mains, tel que le cercle utilisé dans [6], ne semble pas pouvoir s'appliquer puisque la forme de la main peut être très variable, spécialement lorsque celle-ci est en partie cachée par un objet comme un verre d'eau.

La détection et le suivi des mains sont donc effectués en déterminant les régions de couleur peau les plus susceptibles d'être des mains, puis en appliquant des règles définies selon le nombre de régions de peau obtenues et les positions précédentes des mains.

4.2.1 Hypothèses

Afin de limiter le nombre et la complexité des règles, plusieurs hypothèses sont émises. Tout d'abord, comme l'algorithme est basé sur une recherche de régions de couleur peau, les objets de cette couleur sont enlevés de l'environnement. Ceci permettra de réduire le nombre de régions susceptibles de contenir des parties du corps. Il serait toutefois possible d'éliminer cette contrainte en ne considérant que les régions de couleur peau appartenant à l'avant-plan.

Ensuite, on suppose que les mains sont visibles et ne se trouvent pas en occlusion lors de l'entrée de la personne dans l'image. Les positions de celles-ci ne sont donc pas initialisées tant qu'elles ne sont pas visibles et bien séparées des autres régions de couleur peau, puisqu'il est moins grave de ne pas détecter de mains que de les positionner aux mauvais endroits. En effet, le fait que les mains ne soient pas détectées au début de la séquence n'a généralement pas d'impact puisque, à ce moment, l'activité de prise de médicaments n'a pas débuté. Il est donc préférable de s'assurer que les mains soient détectées aux bons endroits au moment de l'activité d'intérêt.

De plus, on suppose que la personne porte un chandail à manches longues dont

le col se trouve relativement près du cou. Cette restriction a été imposée afin de simplifier le problème. En effet, il est difficile d'identifier les extrémités d'une forme dont le contour est mal défini et ensuite de déterminer laquelle correspond à une main. La difficulté est d'autant plus grande quand la région du bras est séparée en plusieurs régions, soit en raison d'erreurs au niveau de la carte de vraisemblance de couleur peau, soit par la présence d'une montre ou d'un autre objet cachant une partie du bras. Les occlusions entre mains et avec le visage ajoutent également à la difficulté.

Finalement, on pose que la main gauche se trouve toujours à gauche de la main droite et vice-versa, ce qui est vrai dans la grande majorité des cas (ici, la droite et la gauche sont définies selon le point de vue de la caméra et non selon celui de la personne). Cette hypothèse est nécessaire en raison des mouvements d'associations et de dissociations entre les mains. Il est toutefois à noter que dans les cas où cette règle n'est pas vraie, les inversions des positions réelles des mains n'affectent pas directement la reconnaissance.

4.2.2 Extraction de régions de mains possibles

La segmentation de la couleur peau étant déjà faite dans l'étape de détection et de suivi de la tête, la carte de vraisemblance de couleur alors calculée est celle utilisée. Les régions de peau sont extraites d'après un algorithme de composantes connexes disponible dans la librairie *OpenCV* et les boîtes englobantes de chacune des régions sont définies. Les boîtes se chevauchant sont alors jointes et les trop petites sont éliminées. Comme l'objectif est de déterminer les régions représentant possiblement une main, les boîtes chevauchant l'ellipse de la tête sont aussi éliminées, celles-ci pouvant correspondre à une région du visage ou du cou. Finalement, les boîtes se situant trop loin des positions précédentes des mains sont rejetées. À la fin de ce processus, il reste, dans la majorité des cas, deux boîtes englobantes, soit une pour chaque main.

4.2.3 Suivi

Dans le cas où deux boîtes englobantes sont obtenues, le positionnement des mains est généralement facile. Il ne suffit que de positionner la main à l'endroit de la boîte englobante la plus près. Le suivi se complique toutefois lorsque le nombre de boîtes obtenues est différent. Les règles deviennent alors plus complexes.

4.2.3.1 Moins de deux régions extraites

Si une seule boîte englobante est obtenue, on vérifie d'abord si elle peut correspondre à une occlusion entre les mains, c'est-à-dire que sa taille est plus grande que la taille d'une seule main et qu'elle se situe à une distance raisonnable des positions précédentes de chacune des mains, soit une distance inférieure au déplacement maximal permis pour chacune d'entre elles. Dans un tel cas, on suppose que la main gauche correspond à la moitié gauche de la région et la main droite à la moitié droite.

Si aucune occlusion n'est détectée entre les mains, la boîte est identifiée comme étant la main se situant le plus près dans l'image précédente et une occlusion possible avec le visage est vérifiée. Si une occlusion est détectée, la main est définie comme le coin inférieur de la boîte englobant la tête, du côté de la main concernée. Si aucune occlusion n'est détectée, on considère que la main est perdue et on la positionne au même endroit que dans l'image précédente. Le déplacement maximal permis de la main est alors ajusté pour permettre un plus grand déplacement à l'image suivante. Ainsi, si la main est perdue pour la première fois, le déplacement permis sera deux fois plus grand que le déplacement maximal initial, trois fois plus grand si la main est perdue de nouveau, et ainsi de suite.

Des règles similaires sont appliquées dans le cas où aucune boîte englobante n'est extraite.

4.2.3.2 Plus de deux régions extraites

Dans le cas où plus de deux boîtes englobantes sont obtenues, les mains sont positionnées aux endroits des boîtes englobantes les plus près de leur position précédente si aucune occlusion entre mains n'est détectée et si les deux mains ne sont pas jumelées à la même boîte englobante. Si ce dernier cas se produit, la boîte en-

globante est identifiée comme étant la main dont la seconde boîte englobante est la plus loin et l'autre main est positionnée à l'endroit de la seconde boîte englobante la plus près.

Finalement, si la main droite est positionnée à la gauche de la main gauche, ou vice-versa, les mains sont inversées.

4.3 Bouteilles de médicaments

Tout comme pour le suivi des mains, divers algorithmes de suivi auraient pu être utilisés pour le suivi de bouteilles de médicaments. L'algorithme de suivi *camshift* de la librairie *OpenCV* a d'ailleurs été testé. Cependant, l'algorithme s'est avéré être inutilisable dans ce contexte en raison du fait que la bouteille pouvait ne plus être visible lorsque prise dans les mains. L'algorithme perd alors la position de la bouteille et, plus souvent qu'autrement, n'arrive pas à la retrouver même lorsque la bouteille redevient visible.

4.3.1 Description de l'algorithme

Un autre algorithme, basé sur le même principe que pour la détection et le suivi des mains, est utilisé. Pour chaque bouteille, une carte de vraisemblance de couleur est calculée et des régions sont extraites par analyse de composantes connexes. Les régions trop petites sont ensuite éliminées par processus d'érosion sur la carte de vraisemblance. Au moment de l'initialisation de la position des différentes bouteilles, l'algorithme suppose que les bouteilles ont une forme rectangulaire de taille limitée et qu'elles sont regroupées. Les positions suivantes sont ensuite assignées aux endroits des régions respectant les contraintes de couleur et de taille se trouvant les plus près, si toutefois la distance n'excède pas le déplacement maximal défini.

Si aucune région de couleur n'est extraite, les occlusions possibles avec les mains sont vérifiées. Une occlusion est détectée si, à l'image précédente, la bouteille se trouve à proximité d'une main, soit à une distance inférieure au déplacement maximal permis pour la bouteille. Si une occlusion avec une main est détectée, la bouteille est positionnée au centre du côté intérieur de la main, se trouvant ainsi au centre

a b c

FIG. 4.6 – Exemples de cartes de vraisemblance de couleur pour des bouteilles a) rose, b) jaune et c) verte.

des mains si une occlusion entre les deux mains est détectée. Si aucune occlusion n'est détectée, la position de la bouteille demeure inchangée.

4.3.2 Modèle de couleur

La méthode utilisée pour créer le modèle de couleur des bouteilles est celle présentée dans [14], mais en utilisant l'espace couleur YC_bC_r complet plutôt que seulement les composantes C_b et C_r. La vraisemblance de couleur est définie comme étant fonction de la distance de Mahalanobis d avec le modèle définie par

$$d^2 = (\mathbf{x} - \mu_b)^T \sum_b^{-1} (\mathbf{x} - \mu_b), \tag{4.5}$$

où $\mathbf{x}$ est le vecteur couleur en trois dimensions du pixel et μ_b et $\sum_b$ sont respectivement le vecteur moyen et la matrice de covariance de la distribution du modèle.

La vraisemblance qu'un pixel soit de la couleur de la bouteille est définie à partir d'un seuil choisi expérimentalement. Un exemple de carte de vraisemblance pour une bouteille est présenté à la figure 4.6.

4.3.3 Détermination des paramètres

Pour créer les modèles de couleur des bouteilles et en déterminer la distribution, un ensemble d'images comportant des régions de chacune des couleurs doit être présenté au système. Pour chaque région, les caractéristiques de la boîte englobante doivent être données, soit la position du coin supérieur gauche (x, y) et les dimensions

(*largeur*, *hauteur*). Les bouteilles elle-mêmes peuvent être utilisées ou encore des cartons aux couleurs de celles-ci. Afin de faciliter la classification, des couleurs vives et ne se retrouvant généralement pas dans un environnement de cuisine habituel sont choisies.

Chapitre 5

Reconnaissance d'activité humaine

La reconnaissance d'activité humaine est basée sur le concept de scénarios. Un scénario correspond à une activité impliquant des objets mobiles se déroulant sur une certaine période de temps. L'algorithme de reconnaissance d'activité développé est formé de trois niveaux de scénarios : état-simple, état-multiple et complexe. Les concepts de scénarios états-simples et états-multiples ont été développés dans [12].

La figure 5.1 illustre la relation entre les différents niveaux de scénarios. Les scénarios états-simples sont calculés pour chaque image de la séquence vidéo alors que les autres scénarios sont vérifiés sur de plus longues périodes de temps. Un schéma détaillé de l'ensemble des scénarios impliqués est présenté à la fin du chapitre, à la figure 5.7.

5.1 Scénarios états-simples

Un scénario état-simple est défini par un ensemble de caractéristiques d'objets mobiles. Par exemple, le scénario *« une main touche la tête »* dépend des distances entre les mains et la tête. Quoiqu'évalués pour chaque image de la séquence, les scénarios peuvent dépendre de plusieurs images. Par exemple, le scénario *« une main s'approche de la tête »* évalué au temps t dépend des positions des mains et de la tête aux instants t et $t-1$. Pour chaque scénario, une probabilité d'occurrence $0/1$ est déterminée de façon heuristique. Les scénarios états-simples du système, sont définis comme suit :

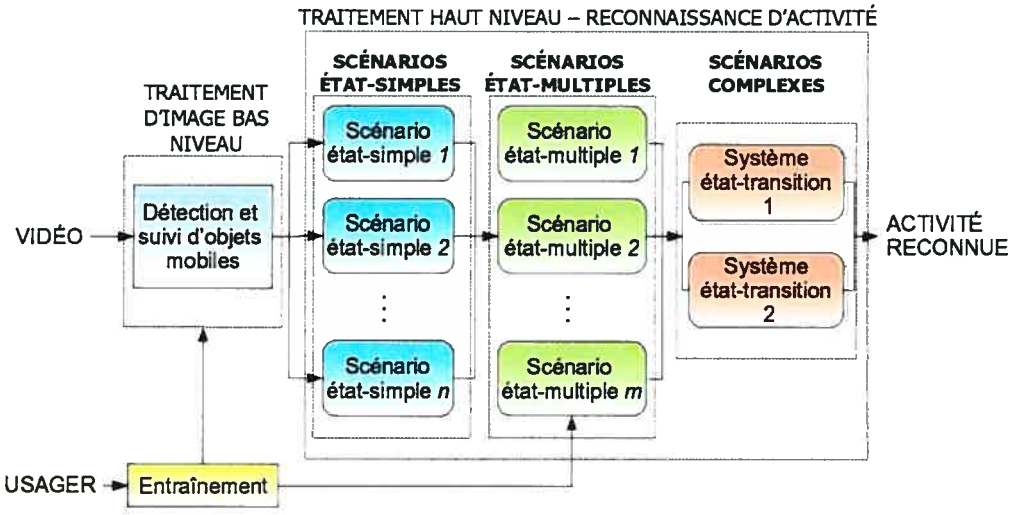


FIG. 5.1 – Structure du modèle de scénario à trois niveaux.

- S_1 : Une seule main manipule la bouteille de médicaments,
- S_2 : Deux mains manipulent la bouteille de médicaments,
- S_3 : Une main touche à la tête,
- S_4 : Une main s'approche de la tête,
- S_5 : Une main s'éloigne de la tête.

Pour chacun de ces scénarios, les objets mobiles impliqués sont conservés. La description détaillée des scénarios, illustrés à la figure 5.2, est la suivante :

- S_1 : La distance séparant les régions d'une main (rectangle vert) et d'une bouteille (rectangle bleu) est inférieure à 2 pixels.
- S_2 : La distance séparant les régions des deux mains (rectangles vert et rose) et d'une bouteille (rectangle bleu) est inférieure à 2 pixels OU la distance séparant les régions d'une main et d'une bouteille est inférieure à 2 pixels et la distance séparant les régions des deux mains est inférieure à 8 pixels (non illustré).
- S_3 : La distance entre la région d'une main (rectangle rose) et la région rectangulaire englobant la tête (ellipse rouge) est inférieure à 8 pixels.
- S_4 : La distance entre le centre de la région d'une main (rectangle rose) et celui de la région rectangulaire englobant la tête (ellipse rouge) à l'instant $t-1$ est supérieure d'au moins 5 pixels à la distance à l'instant t.

FIG. 5.2 – Exemples de scénarios états-simples.

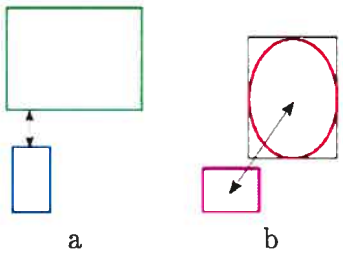


FIG. 5.3 – Exemples de distances séparant a) deux régions rectangulaires et b) le centre de deux régions.

- S_5 : La distance entre le centre de la région d'une main (rectangle rose) et celui de la région rectangulaire englobant la tête (ellipse rouge) à l'instant $t-1$ est inférieur d'au moins 5 pixels à la distance à l'instant t.

Ici, la distance séparant deux rectangles est illustrée à la figure 5.3a. Celle-ci est nulle si les deux régions se chevauchent. La distance entre le centre de deux régions est illustrée à la figure 5.3b.

Les distances, en pixels, mentionnées ont été déterminées expérimentalement, et dépendent de la précision des algorithmes de détection et de suivi ainsi que de la position de la caméra. Toutefois, elles pourraient aussi être déterminées en fonction

de la taille des objets impliqués.

Dans [12], les auteurs proposent de déterminer les distributions de probabilités S_i selon des règles bayesiennes et des probabilités conditionnelles des caractéristiques d'objets mobiles apprises à partir d'ensembles d'entraînement. Toutefois, ceci ne s'est pas avéré nécessaire pour cette application, les scénarios étant relativement simples.

5.2 Scénarios états-multiples

Un scénario état-multiple correspond à une séquence de scénarios états-simples et est évalué sur une longue période de temps. Un scénario état-multiple (*MS* pour *Multi-State*) est reconnu si l'ensemble de ses scénarios états-simples sont reconnus consécutivement. L'idée est de déterminer la probabilité qu'une séquence de N scénarios états-simples se soit produite étant donné une séquence d'observations O. La probabilité maximale que, à l'instant t, le *MS* se soit produit peut être exprimée par

$$P(MS^*|O) = \max_{\forall(t_1,t_2,\ldots,t_N)} P(S_{1_{(t_1,t_2-1)}} S_{2_{(t_2,t_3-1)}} \ldots S_{N_{(t_N,t)}}|O), \tag{5.1}$$

où $S_{i_{(t_i,t_{i+1}-1)}}$ signifie que le scénario état-simple i se produit de l'image t_i à l'image $t_{i+1} - 1$. Les temps t_1 et t correspondent respectivement à la première et à la dernière image du scénario état-multiple dans la séquence vidéo. Les observations O correspondent aux caractéristiques des objets mobiles (positions, tailles, etc.).

Soit S_i étant $S_{i_{(t_i,t_{i+1}-1)}}$ (similairement O_i étant $O_{(t_i,t_{i+1}-1)}$) et S_1^N étant $S_1 S_2 \ldots S_N$, la probabilité $P(S_1^N|O)$ peut être exprimée par[1]

$$P(S_1^N|O) = \prod_{1\leq i\leq N} a_{i,i-1} P(S_i|O_i) \tag{5.2}$$

où $a_{i,i-1}$ est la probabilité (normalisée) *a priori* de passer de l'état $i-1$ à l'état i au temps t_i, soit $\frac{P(S_i|S_{i-1})}{P(S_i)}$. On suppose ici que non seulement cette probabilité est constante à travers le temps, mais qu'elle est la même pour chacun des états et

[1] La preuve de l'équation 5.2 est donnée dans [12].

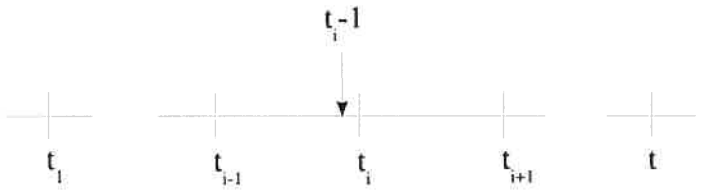


FIG. 5.4 – Relation entre les temps de transition.

qu'elle est égale à 1. On obtient alors

$$P(S_1^N|O) = \prod_{1\leq i\leq N} P(S_{i_{(t_i,t_{i+1}-1)}}|O_{(t_i,t_{i+1}-1)}). \tag{5.3}$$

L'algorithme consiste alors à trouver les temps de transition t_i d'un état au suivant de façon à maximiser la probabilité que la séquence soit complétée. La figure 5.4 illustre la relation entre les différents paramètres temporels.

5.2.1 Calcul des temps de transition

Soit $R_i(t)$, la probabilité que le *MS* occupe l'état i au temps t avec les temps de transition entre les états les plus vraisemblables, on a

$$R_i(t) = \max_{\forall(t_1,t_2,\ldots,t_i)} \prod_{1\leq j\leq i} P(S_{j_{(t_j,t_{j+1}-1)}}|O_{(t_j,t_{j+1}-1)}). \tag{5.4}$$

La solution de l'équation 5.1 est donc équivalente à celle de $R_N(t)$. En exprimant le produit sous forme récursive, on obtient

$$R_i(t) = \max_{t_{i-1}\leq t_i\leq t} P(S_{i_{(t_i,t)}}|O_{(t_i,t)})R_{i-1}(t_i-1). \tag{5.5}$$

Les temps de transition les plus vraisemblables t_i^* peuvent être obtenus selon

$$t_i^* = \underset{t_{i-1}\leq t_i\leq t}{argmax} P(S_{i_{(t_i,t)}}|O_{(t_i,t)})R_{i-1}(t_i-1) \tag{5.6}$$

Ainsi, en commençant avec $R_0(t) = 1$, l'équation 5.5 est calculée de façon récursive jusqu'à ce que l'état N soit atteint.

Si aucune restriction n'est émise sur le temps maximal entre t_i et t_{i-1}, la récursion s'effectuera, pour chaque $R_i(t)$, jusqu'au début de la séquence vidéo. Ceci aura pour effet d'augmenter considérablement le temps de calcul et ce, sans améliorer les résultats. En effet, comme il en sera question plus loin, les *MS* dont la longueur varie trop par rapport à leur durée moyenne sont pénalisés (jusqu'à permettre un rejet dans certains cas). La récursion permise pour chaque $R_i(t)$ est donc limitée en fonction de la longueur moyenne du scénario état-simple qu'il représente selon une équation décrite à la section 5.2.3.

5.2.2 Calcul de $P(S_{i_{(t_i,t)}}|O_{(t_i,t)})$

La probabilité $P(S_{i_{(t_i,t)}}|O_{(t_i,t)})$ peut être obtenue en calculant la moyenne temporelle (ou valeur attendue ou *expected value*) des $P(S_{i_{t_m}}|O_{t_m})$ pour $t_i \leq t_m \leq t$. Pour trouver le t_i qui maximise $R_i(t)$, toutes les valeurs de t_i devraient être considérées. Soit t'_i une valeur possible de t_i, on réduit le nombre de t'_i potentiels en ne conservant que ceux avec lesquels la probabilité cumulée que le scénario i se produise pendant l'intervalle $[t'_i, t]$ soit supérieure à celle de ne pas se produire.

Soit $S_i^+(t'_i, t)$ étant la densité de probabilité cumulée que le scénario i se produise sur l'intervalle $[t'_i, t]$ et $S_i^-(t'_i, t)$ la densité de probabilité cumulée que le scénario i ne se produise pas sur le même intervalle, on définit

$$\begin{aligned} S_i^+(t'_i, t) &= \textstyle\sum_{t'_i \leq j \leq t} P(S_{i_j}), \\ S_i^-(t'_i, t) &= \textstyle\sum_{t'_i \leq j \leq t} (1 - P(S_{i_j})). \end{aligned} \tag{5.7}$$

Rappelons ici que les probabilités $P(S_{i_j})$ ont des valeurs binaires 0/1.

La valeur attendue de reconnaissance $E[S_i]_{(t'_i,t)}$ est alors définie comme suit :

$$\begin{aligned} &\text{Si}(S_i^+(t'_i, t) > S_i^-(t'_i, t)), \\ &\text{alors } P(S_{i_{(t'_i,t)}}|O_{(t'_i,t)}) = \tfrac{S_i^+(t'_i,t)}{t - t'_i} \\ &\text{sinon } P(S_{i_{(t'_i,t)}}|O_{(t'_i,t)}) = 0 \text{ et } t'_i \text{ n'est pas considéré.} \end{aligned} \tag{5.8}$$

Il est intéressant de noter le fait qu'obtenir une forte probabilité pour un scénario n'a

pas d'effet sur les autres scénarios. Deux scénarios états-multiples différents peuvent donc être détectés simultanément ou se chevaucher.

5.2.3 Application de l'algorithme

La première étape pour détecter les *MS* consiste à déterminer les moments auxquels ils risquent de se terminer. Pour ce faire, on cherche des séquences où le dernier état (S_N) du *MS* est présent pour un certain nombre d'images consécutives. On procède alors à la reconnaissance du *MS* à partir de la fin de cette séquence.

Trois scénarios états-multiples sont définis pour la reconnaissance de prise de médicaments :

- MS_1 : La personne ouvre la bouteille de médicaments et en extrait les pilules,
- MS_2 : La personne avale les pilules,
- MS_3 : La personne referme la bouteille de médicaments.

Les séquences de scénarios états-simples requises pour chacun sont présentées à la figure 5.7. Chaque *MS* a une probabilité d'occurrence entre 0 et 1 et un numéro d'image de début et de fin.

Une fois les *MS* détectés, un facteur de *vraisemblance* est utilisé pour pénaliser les scénarios qui sont trop longs ou trop courts, en termes de nombre d'images, comparativement à leur durée ℓ moyenne en diminuant leur probabilité. Ce facteur est défini par la fonction de transfert d'un filtre de Butterworth d'ordre 2 et est exprimé par

$$f(\ell) = \frac{1}{1+\left(\frac{\ell-\bar{\ell}}{w_c}\right)^{2n}} \tag{5.9}$$

avec $n = 2$ (ordre 2) et où $\bar{\ell}$ est la durée moyenne du scénario, variant d'un *MS* à l'autre et déterminée à partir de données d'entraînement. w_c est défini comme étant $20\sqrt{\bar{\ell}}$. Comme on peut le voir à la figure 5.5, la fonction ne pénalise pas la reconnaissance pour de petites variations de durées autour de la moyenne. Les valeurs des durées moyennes utilisées dans le système varient en fonction de la personne et de la vitesse d'acquisition des images.

Pour ce qui est de la récursion permise pour le calcul de $R_i(t)$ dans l'équation

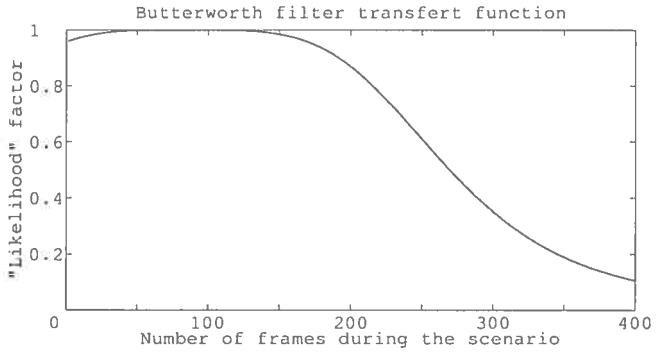


FIG. 5.5 – Fonction de transfert d'un filtre de Butterworth pour $\bar{\ell} = 85$, $w_c = 184$ et $n = 2$.

5.5, dont il a été question précédemment, elle représente la longueur à laquelle le facteur de vraisemblance atteint 40 % dans l'équation 5.9 où $\bar{\ell}$ représente la durée moyenne du scénario état-simple correspondant à l'état i. Là encore, ce paramètre est obtenu à partir de données d'entraînement.

Un autre point important dans l'application de l'algorithme est que le calcul des $P(S_{i_j})$ dans l'équation 5.7 dépend des objets mobiles impliqués. Par exemple, dans le contexte du scénario état-multiple de la manipulation de la bouteille verte, si au temps j la probabilité du scénario état-simple *« une main touche la bouteille verte »* $P(S_{1_{j_{verte}}})$ est 0, la probabilité $P(S_{1_j})$ sera 0, même si celle du scénario état-simple *« une main touche la bouteille jaune »* $P(S_{1_{j_{jaune}}})$ est 1.

5.3 Scénarios complexes

Les scénarios états-multiples peuvent représenter des activités simples, mais ne peuvent être utilisés pour des situations plus complexes. Par exemple, l'activité de prise de médicaments peut être effectuée de différentes façons : la personne peut prendre les pilules dans la bouteille de médicaments, avaler les pilules puis refermer la bouteille ou encore prendre les pilules dans la bouteille de médicaments, refermer la bouteille et finalement avaler les pilules. De plus, des actions non directement reliées peuvent se produire pendant la séquence, comme boire de l'eau ou encore déposer les pilules sur la table. Finalement, si plus d'un médicament est pris, les actions simples

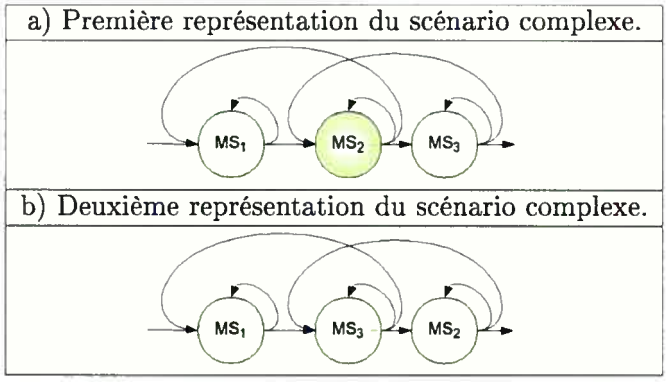


FIG. 5.6 – Systèmes état-transition représentant le scénario complexe formé des séquences de scénarios états-multiples a) $\{MS_1,MS_2,MS_3\}$ et b)$\{MS_1,MS_3,MS_2\}$.

formant la séquence complète peuvent ne pas se produire de façon consécutive. On ne peut donc pas procéder à la reconnaissance de la prise de médicaments en cherchant une séquence fixe.

Le problème est résolu en séparant l'activité complexe en séquence de scénarios états-multiples (états) et en utilisant un système *état-transition* pour la reconnaître. Ainsi, une fois les scénarios états-multiples détectés, ceux-si sont placés en ordre chronologique puis introduits dans le système.

Les deux représentations du scénario complexe sont présentées à la figure 5.6. Les cercles représentent les scénarios états-multiples et les flèches représentent les transitions. Des boucles autour d'un même état sont possibles étant donné que plusieurs médicaments peuvent être pris. Les pilules n'étant pas détectées spécifiquement et plus d'une pouvant être avalée à la fois, le même événement MS_2 peut être utilisé pour reconnaître la prise de médicaments pour toutes les pilules.

Les transitions sont utilisées pour pénaliser la reconnaissance si le nombre d'images augmente trop entre deux états. La pénalisation se faisant uniquement entre les états, la reconnaissance de la prise du médicament 1 n'échoue pas si la personne prend les pilules dans une autre bouteille de médicaments avant d'avaler les pilules de la bouteille 1. Le facteur de *vraisemblance* utilisé ici est également défini par la fonction

de transfert d'un filtre de Butterworth de second ordre et est exprimé par

$$f(\ell) = \frac{1}{1+\left(\frac{\ell}{w_c}\right)^{2n}} \tag{5.10}$$

avec encore $n = 2$ et $w_c = 20\sqrt{\bar{\ell}}$. Toutefois, plutôt que de dépendre des longueurs des états, celui-ci dépend des longueurs des transitions, aussi apprises à partir de données d'entraînement. Le paramètre w_c, tout comme pour l'ordre du filtre, est fixe.

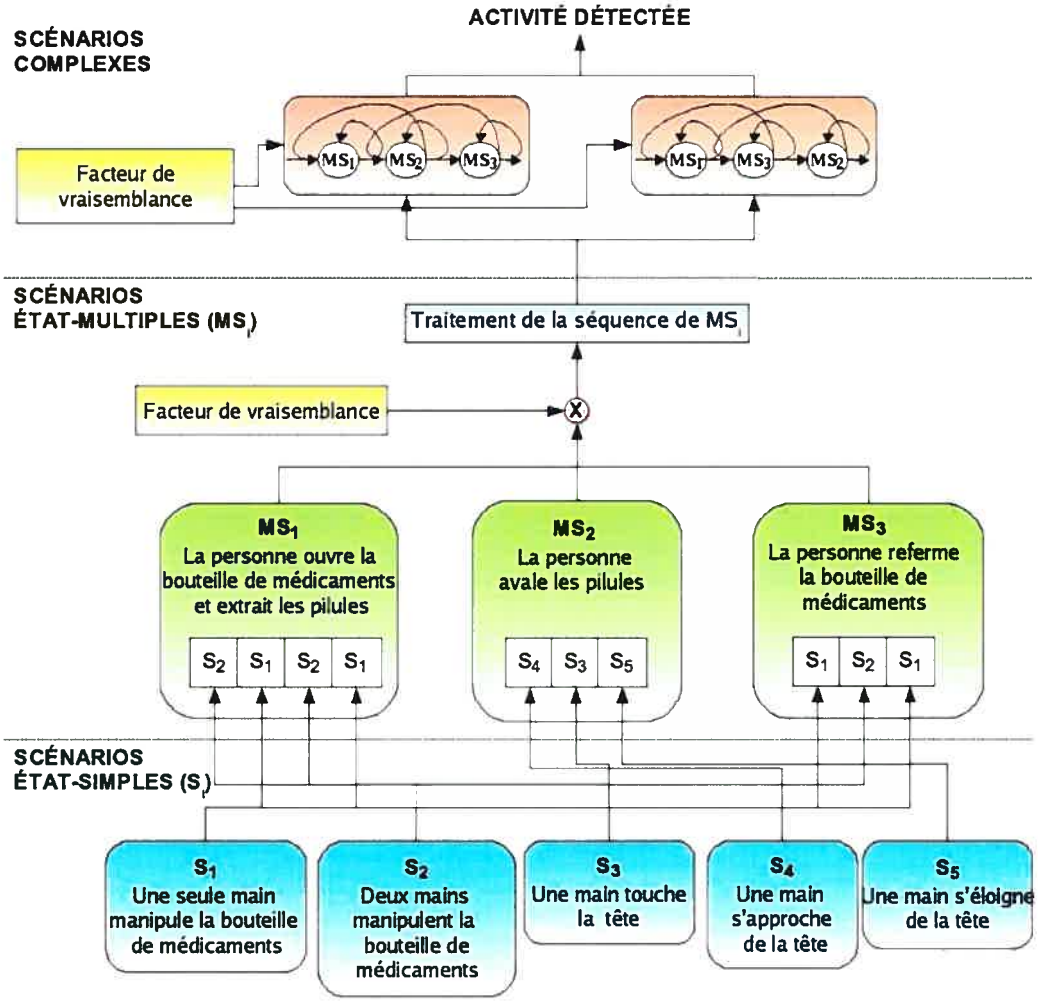


FIG. 5.7 – Schéma détaillé du processus de reconnaissance d'activité.

Chapitre 6

Résultats

Les modèles de couleurs, pour la peau et les bouteilles, utilisés pour la détection et le suivi sont créés lors d'une première phase de détermination des paramètres, à partir d'exemples identifiés manuellement. Le système utilise une webcaméra Quickcam Personal Pro pour ordinateurs portables de Logitech (voir figure 6.1) ayant une fréquence d'acquisition de 30 images par seconde et une résolution de 640x480 pixels. Toutefois, en pratique, la résolution maximale de l'image afin d'éviter des retards est de 320x240 pixels. Étant donné qu'une fréquence d'acquisition de 15 images par seconde est suffisante pour effectuer la reconnaissance d'activité, seulement une image sur deux est considérée, réduisant ainsi les calculs nécessaires.

En ce qui a trait à l'entraînement nécessaire à la phase de reconnaissance, le système détermine les valeurs des paramètres de durées à partir de séquences de prise de médicaments où les scénarios état-simples et état-multiples sont identifiés manuellement.

FIG. 6.1 – Webcaméra Quickcam Personal Pro pour ordinateurs portables de Logitech.

FIG. 6.2 – Exemple de suivi dans une séquence de prise de médicaments.

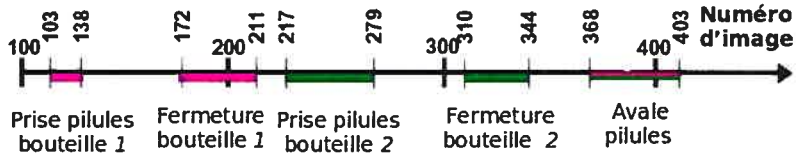
FIG. 6.3 – Exemple de leurre.

6.1 Détection de prise de médicaments

Les efforts étant dirigés vers la partie de reconnaissance, les performances de la partie de détection et de suivi n'ont pas été évaluées en profondeur. Notons simplement que les résultats étaient assez satisfaisants pour permettre la reconnaissance d'activité. Un exemple de suivi dans une séquence vidéo présentant une prise de médicaments est présenté à la figure 6.2. Malgré le fait que l'ellipse ne soit pas toujours centrée sur le visage, l'algorithme fonctionne tout de même correctement en raison des tolérances de distances reliées aux contacts entre les objets.

La détection de prise de médicaments a été testée avec 48 séquences provenant de 33 vidéos prises avec trois personnes différentes. De ces séquences, 31 présentent une prise de médicaments réelle alors que les 17 autres présentent d'autres activités (leurres), comme manger ou manipuler les bouteilles de médicaments. Un exemple de leurre est présenté à la figure 6.3. Dans cet exemple, on peut remarquer qu'une main paraît en contact avec la bouteille de médicaments jaune.

a) Séquence de scénarios état-multiples.



b) Séquence de scénarios complexes.

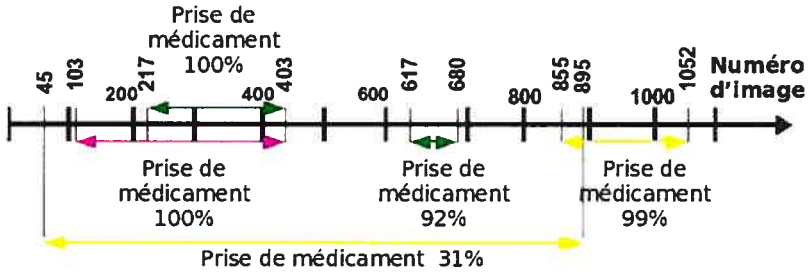


FIG. 6.4 – Résultats typiques : a) séquence de scénarios états-multiples formant le scénario complexe $\{MS_{1_1}, MS_{3_1}, MS_{1_2}, MS_{3_2}, MS_2\}$ et b) séquence de scénarios complexes impliquant plusieurs bouteilles avec les probabilités correspondantes.

Activité	Détectée	Non détectée	Efficacité
Prise de médicaments	30	1	97%
Autres (leurres)	1	16	94%

TAB. 6.1 – Résultats de reconnaissance pour 48 séquences vidéo.

Des résultats typiques sont présentés à la figure 6.4. La figure 6.4a présente une séquence de scénarios état-multiples obtenue à partir d'une vidéo avec prise de médicaments. Pour chacune des bouteilles, la prise de médicaments a été détectée au bon endroit et avec une probabilité de 99,6 %. Le scénario complexe de la séquence est défini par sa seconde représentation, présentée à la figure 5.6b.

La figure 6.4b montre une séquence de scénarios complexes impliquant trois bouteilles. On peut voir que la détection de la prise du médicament 3 chevauche les autres scénarios. Ceci est dû au fait que la bouteille 3 est manipulée au début de la séquence. La faible probabilité obtenue montre l'efficacité de la pénalisation qui permet de rejeter la fausse détection.

Considérant que le scénario est reconnu si sa probabilité est supérieure à 75 %, les résultats pour l'ensemble des séquences vidéos sont présentés au tableau 6.1. En

FIG. 6.5 – Exemple de situation causant une erreur de détection.

fait, la plupart des scénarios reconnus avaient une probabilité supérieure à 90 %, et les fausses détections une probabilité inférieure à 40 %. Le choix du seuil à 75 %, quoiqu'arbitraire, n'est donc pas critique pour les résultats.

On peut remarquer qu'une prise de médicaments n'a pas été détectée. Le problème est dû à un faux contact perçu du point de vue 2D de la caméra. En effet, lorsque la bouteille est déposée sur la table après avoir été utilisée, la main gauche de la personne se trouve derrière la main droite manipulant la bouteille et le scénario S_2 est reconnu à la place du scénario S_1. La figure 6.5 illustre la situation. Ainsi, seulement les deux premiers scénarios état-simples du scénario état-multiple MS_3 sont reconnus et non le troisième (voir figure 5.7). Dans ce cas-ci, le problème pourrait être résolu en utilisant une forme autre que le rectangle pour identifier les mains.

Pour ce qui est de la fausse détection, il est important de mentionner que pour la majorité des séquences avec des leurres, on demandait aux participants de manipuler les bouteilles et de toucher leur visage de façon à provoquer des erreurs de détection. Ces séquences ne sont donc pas très représentatives des activités normales s'effectuant à la table de cuisine, celles-ci n'impliquant généralement pas les bouteilles de médicaments. Toutefois, ces séquences ont contribué à déterminer les limites du système. Le fait que seulement une tentative ait réussi est donc un signe positif.

Un point également important à mentionner est que dans l'une des séquences, la prise de médicaments a été détectée, mais au mauvais endroit. La prise réelle de médicaments a été initialement bien détectée, mais une seconde prise, chevauchant

FIG. 6.6 – Exemple d'erreur de suivi de bouteille.

la première, a été détectée en erreur avec une probabilité supérieure. Dans ces cas-là, seule la prise avec la plus grande probabilité est conservée. La fausse détection est due au fait que la personne a déposé la bouteille de médicaments rose derrière la jaune et que, n'étant pas détectée par le système, celui-ci a continué de la considérer en occlusion avec la main, tel qu'illustré à la figure 6.6. Comme un autre médicament (vert) a été pris immédiatement après, la seconde prise de médicaments a également été attribuée au médicament de la bouteille rose. Le problème pourrait probablement être minimisé en utilisant un algorithme de détection d'occlusions plus sophistiqué.

6.2 Temps d'exécution

Même si l'acquisition des images se fait à une fréquence de 15 images par seconde, le système ne peut actuellement pas procéder au traitement en temps réel, aucune optimisation n'ayant encore été faite. Les temps de traitement de différentes parties du processus de détection et de suivi sont présentés au tableau 6.2.

L'initialisation du système ne s'effectue qu'une fois, soit lors du lancement du programme. Le système crée alors les modèles pour la détection et le suivi des objets mobiles et lit le fichier contenant les paramètres nécessaires à la partie de reconnaissance. Comme cette opération ne se réexécute pas, elle ne doit donc pas être prise en compte dans l'évaluation de la vitesse du système.

Pour ce qui est de la reconnaissance de prise de médicaments, elle ne doit pas nécessairement se faire en temps réel ; celle-ci pouvant être effectuée lorsque la per-

Partie	Durée (par image)
Initialisation du système (1 fois)	8 000 ms
Détection de présence de la personne	30 ms
Extraction de l'avant-plan	150 ms
Suivi de la tête	150 ms
Suivi des mains et des bouteilles	7 ms
Total détection et suivi	350 ms

TAB. 6.2 – Temps de traitement typiques de différentes parties du processus de détection et de suivi.

sonne a quitté le champ de vision de la caméra alors que le suivi est arrêté et que beaucoup moins de calculs sont effectués. La fréquence de traitement est donc limitée par la partie de détection et de suivi qui, actuellement, s'effectue à environ 3 images par seconde. Toutefois, cette partie de traitement, qui n'a pas été approfondie, n'est pas optimisée. La composante temporelle devrait être prise en compte lors de l'amélioration de cet aspect du système.

Quant au temps de calcul lors de la reconnaissance d'activité, celui-ci varie énormément en fonction de la séquence vidéo. Quatre exemples de temps de traitement pour cette partie sont présentés au tableau 6.3. Dans la première séquence, il n'y a pas de prise de médicaments. La situation de leurre est celle présentée à la figure 6.3. Plus la séquence contient d'interactions avec les bouteilles, plus le nombre de scénarios état-multiples est élevé et plus le temps de calcul est grand. Ainsi, même si la personne est présente pendant une grande période de temps dans le champ de vision de la caméra, c'est davantage le temps passé à manipuler les bouteilles de médicaments, soit le moment de la prise des médicaments, qui influence le temps de traitement que la durée de la présence.

Considérant un temps de traitement moyen est d'environ 30 secondes par période de prise de médicaments (voir tableau 6.3 pour les temps lors de la prise de 3 médicaments), si la personne procède à 3 prises de médicaments par jour, il suffit qu'elle ne se trouve pas dans le champ de vision de la caméra pendant environ 90 secondes pour que le traitement puisse s'effectuer totalement sans causer de débordements de mémoire. Il est important de noter que l'on considère ici que la partie de détection et de suivi s'effectue en temps réel.

Partie	Durée
Aucune prise de médicaments, séquence de 309 images (total)	*44 ms*
Extraction séquence scénarios état-simples	35 ms
Détection scénarios état-multiples	9 ms
Détection scénarios complexes	-
Prise de 2 médicaments, séquence de 695 images (total)	*9,54 s*
Extraction séquence scénarios état-simples	59 ms
Détection scénarios état-multiples	9 476 ms
Détection scénarios complexes	1 ms
Prise de 3 médicaments, séquence de 700 images (total)	*24,93 s*
Extraction séquence scénarios état-simples	55 ms
Détection scénarios état-multiples	24 874 ms
Détection scénarios complexes	4 ms
Prise de 3 médicaments, séquence de 920 images (total)	*29,01 s*
Extraction séquence scénarios état-simples	105 ms
Détection scénarios état-multiples	28 988 ms
Détection scénarios complexes	4 ms

TAB. 6.3 – Exemples de temps de traitement de différentes parties du processus de reconnaissance d'activité.

Chapitre 7

Discussion

Les résultats du système ayant été présentés, il convient de discuter de ses limitations dans des environnements moins contrôlés ainsi que des travaux futurs. Toutefois, il sera d'abord question des paramètres du système.

7.1 Paramètres fixés

Dans le système, certains paramètres sont fixés et d'autres sont déterminés à partir de données expérimentales. Dans cette section, il sera principalement question des paramètres fixés et des façons de les déterminer à partir de données réelles.

7.1.1 Détection et suivi

Dans cette partie du système, et plus particulièrement pour la partie de détection, les paramètres des modèles servant à créer les cartes de vraisemblance de couleur, que ce soit pour la peau ou pour les bouteilles, sont déterminés à partir d'images, prises dans les conditions d'utilisation, fournies au système. En ce qui a trait au suivi, les paramètres, tels que les dimensions de la tête et le déplacement maximal des mains entre deux images consécutives, ont été fixés de façon expérimentale. Ces paramètres dépendent principalement de l'éloignement de la caméra par rapport à la scène et très peu de la personne. En effet, quoique les dimensions des personnes et des bouteilles de médicaments puissent varier, à distance fixe de la caméra, celles-ci demeurent généralement dans les marges fixées. Toutefois, des changements plus

importants peuvent survenir si on éloigne ou rapproche considérablement la caméra. Une modification automatique de ces paramètres serait nécessaire dans un tel cas. Toutefois, comme la caméra doit demeurer relativement près de la scène afin de permettre une bonne classification des pixels de couleur tout en demeurant assez loin pour avoir une vue d'ensemble de la personne et des bouteilles, la position de la caméra demeure relativement fixe. Plutôt que de modifier les paramètres de suivi, il s'agit de bien positionner la caméra dans l'environnement.

7.1.2 Reconnaissance d'activité

La majorité des paramètres fixés dans cette partie sont reliés aux définitions des scénarios états-simples. En effet, ils servent, entre autres, à déterminer la distance maximale entre deux objets pour qu'il y ait contact ou encore la distance parcourue par une main entre deux images consécutives pour considérer un déplacement. Ces paramètres dépendent non seulement de la position de la caméra et des dimensions des objets mobiles, mais aussi de la précision de la détection et du suivi. Tout comme pour la partie de suivi, les paramètres demeurent valides si la caméra est bien positionnée. Les valeurs des paramètres pourraient être ajustées en fonction des améliorations faites à cette première partie. Par exemple, la distance entre deux objets pour considérer un contact, soit la marge de tolérance, pourrait diminuer avec une détection plus précise.

Les autres paramètres fixés sont davantage fonction du niveau de confiance souhaité de la reconnaissance d'activité. En effet, ceux-ci sont soit reliés à la décroissance des probabilités selon l'éloignement des durées par rapport à la moyenne (paramètres des filtres de Butterworth) ou encore aux probabilités minimales pour conserver les scénarios (états-multiples et complexes). Ainsi, avec une décroissance plus rapide des courbes des fonctions de Butterworth et un seuil de probabilité minimale élevé, le système sera plus restrictif et le taux de fausses détections (faux positif) tendra à diminuer, mais le taux de bonnes détections (vrai positif) aussi. Les paramètres utilisés semblent offrir un bon compromis, mais des tests plus approfondis (avec courbe ROC) permettraient peut-être de déterminer des valeurs plus optimales. Une telle étude n'a pas effectuée dans le cadre de ce projet, car une banque de séquences vidéo

beaucoup de plus grande auraient été nécessaire.

Dans la partie de détection des scénarios états-multiples, les facteurs $a_{i,i-1}$ de l'équation 5.2, plutôt qu'être fixés à 1 pour chaque scénario, pourraient aussi être déterminés à partir de données expérimentales et varier non seulement en fonction des scénarios, mais également en fonction du temps. Ainsi, par exemple, le facteur $a_{2,1}$ du scénario MS_2 *« la personne avale les pilules »*, soit pour le passage du scénario *« une main s'approche de la tête »* au scénario *« une main touche à la tête »*, pourrait être plus ou moins grand selon que le scénario MS_1 *« la personne ouvre la bouteille de médicaments et en extrait les pilules »* s'est produit un peu avant. Toutefois, les résultats obtenus étant déjà bons, cet élément n'a pas été testé.

7.2 Limitations

Plusieurs facteurs peuvent causer des erreurs de détection et de suivi ou fausser la reconnaissance d'activité.

7.2.1 Détection et suivi

Premièrement, la faible qualité des images obtenues à partir de la webcaméra a pour effet d'augmenter le niveau de bruit et de rendre la classification des pixels de couleur plus difficile. De plus, le logiciel d'acquisition des images créant des vidéos en format *avi* compressés, il y a perte d'information au niveau des couleurs, ce qui peut aussi fausser la classification. Par exemple, une ombre sur une nappe de couleur crème peut avoir exactement la même couleur que la peau dans les images obtenues, ce qui ne serait pas nécessairement le cas avec une caméra plus sophistiquée, dont l'étendue de la gamme de couleurs disponibles serait plus large et le bruit plus faible.

Deuxièmement, comme on peut le remarquer à la figure 6.2, le modèle elliptique utilisé n'est peut-être pas celui le plus approprié pour la détection et le suivi de la tête dans ce type de séquences vidéo. En effet, celui-ci se prête davantage à des images où la personne est directement de face, comme à la figure 4.1. Dans certains cas, des contacts main-tête pourraient ne pas être détectés en raison de ce genre d'erreur de positionnement de la tête et ainsi affecter la reconnaissance d'activité.

Cependant, ce cas ne s'est présenté que rarement en pratique.

Troisièmement, le fait que les bouteilles de médicaments soient souvent cachées par les mains rend leur suivi plus difficile. En effet, comme on peut le voir sur l'image 206 de la figure 6.2, la bouteille de médicaments rose, alors en occlusion totale, est assignée à la mauvaise main. Toutefois, ce genre d'erreur de suivi n'a pas d'impact sur la reconnaissance d'activité.

Quatrièmement, la présence d'objets similaires aux bouteilles de médicaments, soit quant à leur couleur et leur taille, peut également créer de la confusion au niveau du suivi de celles-ci. Par exemple, une bouteille de médicaments pourrait être confondue avec une salière, ou un carton de jus, de même couleur posée sur la table ou encore avec une tasse de même couleur qui, située plus loin dans la pièce, paraît de même grandeur du point de vue de la caméra. Ce problème pourrait être limité en utilisant des bouteilles un peu plus grosses et comprenant un motif particulier et uniforme (invariant selon l'orientation de la bouteille). L'utilisation de bouteilles plus grosses permettrait aussi de réduire les occlusions totales de celles-ci et donc d'en faciliter le suivi. Cette solution, qui, tout comme pour les bandes de couleurs, est facilement envisageable selon le neuropsychologue Alain St-Arnaud, n'a toutefois pas été testée.

Cinquièmement, la détection peut être affectée par la présence d'objets causant des occlusions entre la caméra et la personne, et ainsi provoquer des erreurs de suivi. Par exemple, la présence d'une boîte de céréales directement devant les bouteilles de médicaments peut rendre celles-ci indétectables et donc leur suivi impossible. Le fait de surélever la caméra limite ce genre d'occlusions, mais ne les élimine pas totalement. Une autre solution serait d'utiliser plusieurs caméras, mais en plus d'être plus coûteuse, celle-ci est assez complexe à implémenter, notamment en ce qui a trait à la synchronisation des caméras.

Finalement, la classification de pixels de couleur peau peut également être affectée par de grandes variations d'intensités lumineuses ou par la distance entre les objets et la caméra. Quant au suivi, il peut aussi être influencé par les mouvements de la personne, par exemple si la personne place une main sous la table.

7.2.2 Reconnaissance

Si la reconnaissance est indirectement affectée par les erreurs de détection et de suivi, certains autres facteurs ont un impact plus direct sur celle-ci. En premier lieu, une erreur peut survenir si, par exemple, dans le processus de prise de médicaments, la personne touche son front plutôt qu'avaler ses pilules ; la bouche n'étant pas spécifiquement détectée.

En second lieu, la détection peut échouer si la personne quitte la table entre la prise des pilules dans la bouteille et l'ingestion de celles-ci, par exemple pour aller se chercher un verre d'eau. Cette situation pourrait être tenue en compte dans les délais permis entre les différents états (scénarios états-multiples) du scénario complexe, mais cela risquerait de provoquer des fausses détections dans bien d'autres cas. De plus, il serait très difficile de déterminer si, à son retour à la table, la personne avale bien ses pilules ou si elle les a déposées ailleurs et ne fait que boire l'eau (séquence complète, mais sans prise de médicaments).

Enfin, le fait que le scénario MS_3 *« la personne referme la bouteille de médicaments »*, du niveau de la séquence de scénarios états-simples, soit un sous-ensemble du scénario MS_1 *« la personne ouvre la bouteille de médicaments et en extrait les pilules »* ajoute à la difficulté de reconnaissance du scénario complexe. En effet, pour chaque scénario MS_1, deux scénarios MS_3 peuvent être détectés.

Il convient de mentionner ici que dans cette application, on considère que la personne accepte de prendre ses médicaments et que la non-conformité est due à des problèmes de mémoire plutôt qu'à la non-acceptation du traitement. De plus, il est à noter que dans un contexte d'utilisation réelle du système, l'intervention d'une personne ressource affiliée au système de la santé demeurera nécessaire. Ainsi, dans la mesure ou les erreurs de détections ne sont pas trop fréquentes, celles-ci ne sont pas trop problématiques.

7.3 Travaux futurs

Comme la reconnaissance d'activité dépend beaucoup de la détection et du suivi des objets mobiles, toute amélioration des algorithmes impliqués dans cette partie

serait bénéfique. Tout d'abord, du travail pourrait être fait pour trouver un modèle de couleur plus adapté au type d'images utilisées et pour permettre une meilleure identification des régions détectées. Ensuite, un positionnement en trois dimensions, par exemple en utilisant des techniques de photogrammétrie monocaméra ou de stéréoscopie, pourrait permettre de réduire le nombre de faux contacts entre les objets. Enfin, la segmentation des régions du visage et l'obtention d'une distance main-bouche, plutôt que main-tête, permettraient de diminuer les fausses détections de prise de médicaments. En ce qui a trait aux algorithmes présentement utilisés, il serait nécessaire de procéder à la normalisation des distances exprimées en pixels, par exemple celles définissant les tolérances pour les contacts, en fonction de la taille des objets dans la scène. De plus, une analyse exhaustive des effets des différents paramètres utilisés, définis empiriquement, permettrait de mieux déterminer leur impact sur la reconnaissance et de les optimiser.

De façon générale, la reconnaissance de la prise de médicaments semblant bien fonctionner dans les cas où la détection et le suivi des objets mobiles sont corrects, les travaux futurs devraient davantage se concentrer, dans un premier temps, sur cette partie. Il serait toutefois intéressant de développer une méthode permettant un apprentissage autonome des scénarios états-multiples et des représentations du scénario complexe. Le système pourrait, à partir d'un ensemble de scénarios états-simples définis et d'un ensemble d'apprentissage constitué de séquences vidéo présentant une prise de médicaments, déterminer les séquences de scénarios états-simples qui formeraient les scénarios états-multiples et les différentes représentations du scénario complexe. Ainsi, d'autres scénarios états-simples pourraient être utilisés, par exemple boire de l'eau, selon les habitudes de la personne. Le système serait alors plus personnalisé. Il suffit de déterminer les différents scénarios états-simples possibles et de s'assurer de pouvoir procéder à la détection et au suivi des objets mobiles impliqués.

Enfin, l'algorithme de reconnaissance d'activité pourrait être appliqué à d'autres situations afin de vérifier tant son efficacité que la facilité à l'adapter à d'autres scénarios, définis par des séquences d'actions précises. Par exemple, il pourrait être utilisé dans un système de détection de situations à risque, comme le rond d'une

cuisinière resté allumé. Un tel système pourrait vérifier que si la personne allume un rond de la cuisinière, elle place une casserole sur celui-ci, puis éteint le rond une fois la cuisson terminée. Un autre exemple d'application serait un système d'assistance, comme pour le lavage des mains (assiste la personne dans les diverses étapes en cas d'oubli) [22]. Finalement, l'algorithme pourrait être utilisé dans des distributeurs automatiques de médicaments, nécessitant une assistance externe (pharmacien, ...) qui vérifie que le patient prend véritablement ses médicaments [23].

Chapitre 8

Conclusion

Le système de détection de prise de médicaments présenté dans ce mémoire fait partie d'un projet plus vaste de vidéosurveillance de personnes âgées pour le maintien à domicile. Quoique le système effectue les deux parties du traitement, soit une première de détection et de suivi et une seconde de reconnaissance d'activité, l'essence de la contribution de ce travail se situe dans la seconde partie. Compte tenu des résultats obtenus, les algorithmes utilisés se sont montrés efficaces et prometteurs, et la faisabilité d'un système de détection automatique de prise de médicaments est démontrée.

Toutefois, en raison des sources possibles d'erreurs présentées dans la discussion, des adaptations seront nécessaires afin d'adapter le système à des environnements réels.

Les algorithmes ont été testés pour la reconnaissance de prise de médicaments, mais pourraient aussi être appliqués à d'autres activités de la vie quotidienne à domicile. Appliqué à un domaine autre que le biomédical, ce système pourrait aussi servir à des fins de sécurité, afin de détecter des séquences d'événements à risques dans les endroits publics, soit par exemple un bagage abandonné.

Bibliographie

[1] Lindsay, Colin. Un portrait des aînés au Canada. *Statistique Canada*, Ottawa, octobre 1999.

[2] Division du vieillissement et des aînés de Santé Canada. Comité des hauts Fonctionnaires (aînés), L'usage des médicaments et de l'alcool chez les aînés : une approche concertée sur l'usage des médicaments par les aînés. Ottawa, juin 1996.

[3] A. Mihailidis et G. Fernie. Context-aware assistive devices for older adults with dementia. *Gerontechnology*, vol.2, pp. 173–189, 2002.

[4] Groupe de travail de l'étude sur la santé et le vieillissement au canada : Méthode de soins pour les personnes atteintes de démence au Canada. *Journal canadien sur le vieillissement*, vol.13, no 4, pp. 470–487, 1994.

[5] Intel. Open Source Computer Vision Library. [en ligne]. [juin 2006] Disponible sur internet : < http ://www.intel.com/technology/computing/opencv/>.

[6] D. Batz, M. Batz, N. da Vitoria Lobo et M. Shah. A computer vision system for monitoring medication intake Dans *IEEE Proc. of the 2nd Canadian Conference on Computer and Robot Vision*, pp. 362–369, 2005.

[7] P. Fieguth et D. Terzopoulos. Color-based Tracking of Heads and Other Mobile Objects at Video Frame Rates. Dans *IEEE Proc. Computer Vision and Pattern Recognition*, pp. 21–27, 1997.

[8] S. Spors et R. Rabenstein. A Real-time Face Tracker for Color Video. Dans *IEEE Int. Conf. on Acoustics, Speech & Signal Processing*, Utah, États-Unis, 2001.

[9] S.L. Phung, A. Bouzerdoum et D. Chai. A Novel Skin Color Model in YCbCr Color Space and its Application to Human Face Detection. Dans *IEEE International Conference on Image Processing*, pp. 289–292 , 2002.

[10] C. Nugent, D. Finlay, R. Davies, C. Paggetti, E. Tamburini et N. Black. Can Technology Improve Compliance to Medication ? Dans *Proc. of the 3rd International Conference On Smart homes and health Telematic*, Sherbrooke, Canada, pp. 65–72, 2005.

[11] S. Hongeng, F. Brémond et R. Nevatia. Representation and Optimal Recognition of Human Activities. Dans *IEEE Proc. Computer Vision and Pattern Recognition*, vol.1, pp. 818–825, 2000.

[12] S. Hongeng et al. Video-based event recognition : activity representation and probabilistic recognition methods. Dans *Computer Vision and Image Understanding*, vol.96, no 2, pp. 129–162, novembre 2004.

[13] S. Birchfield. Elliptical Head Tracking Using Intensity Gradients and Color Histograms. Dans *IEEE Proc. Computer Vision and Pattern Recognition*, pp. 232–237, 1998.

[14] N. Habili, C. Lim et A. Moini. Hand and Face Segmentation Using motion and Color Cues in Digital Image Sequences. Dans *IEEE International Conference on Multimedia and Expo*, pp. 377–380, 2001.

[15] G. Welch and G. Bishop. An introduction to the kalman filter. Technical Report 95-041, University of North Carolina at Chapel Hill, Department of Computer Science, 2002. [en ligne]. [décembre 2005] Disponible sur internet : < http ://www.cs.unc.edu/~welch/media/pdf/kalman_intro.pdf>.

[16] N. Liu et Brian C. Lovell. MMX-Accelerated Real-Time Hand Tracking System. Dans *Proc. of Image and Vision Computing New Zealand 2001*, Dunedin, Nouvelle-Zélande, pp. 381–385, 26-28, 2001.

[17] G. Bradski. Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal*, 2nd Quarter, 1998.

[18] N. Rota et M. Thonnat, Activity Recognition from Video Sequences using Declarative Models. Dans *14th. European Conference on Artificial Intelligence 2000,* Berlin, Allemagne, pp. 673–680, 2000.

[19] N. Ghanem, D. DeMenthon, D. Doermann et L. Davis. Representation and Recognition of Events in Surveillance Video Using Petri Nets. Dans *IEEE Proc.*

Computer Society Conference on Computer Vision ans Pattern Recognition Workshops 2004, Washington DC, États-Unis, vol. 7, p. 112, 2004.

[20] J. Gao, A. G. Hauptmann, A. Bharucha et H. D. Wactlar. Dining Activity Analysis Using a Hidden Markov Model. Dans *Proc. of the 17th International Conference on Pattern Recognition*, Cambridge, Royaume-Uni, vol.2, pp.915–918, 2004.

[21] D. Ayers et Mubarak Shah. Monitoring Human Behavior from Video Taken in an Office Environment. *Image and Vision Computing*, vol.19, no 12, pp. 833–846, octobre 2001.

[22] A. Mihailidis, J. C. Barbenel, G. Fernie. The efficacy of an The efficacy of an intelligent cognitive orthosis to facilitate handwashing by persons with moderate-to-severe dementia. Neuropsychological Rehabilitation, vol.14, no 1-2, pp. 135-171, 2004.

[23] S. G. Patil et T. J. Gale. Preliminary Design of Remotely Used and Monitored Medication Dispenser. Dans *Proc. of the 28th IEEE EMBS Annual International Conference*, New York, États-Unis, pp. 3616– 3619, 2006.

[24] Office québécois de la langue française. *Le grand dictionnaire terminologique* [en ligne]. [mai 2006] Disponible sur internet : <http ://www.granddictionnaire.com/btml/fra/r_motclef/index1024_1.asp>

[25] M. Valin, J. Meunier, A. St-Arnaud, J. Rousseau. Video Surveillance of Medication Intake. Dans *IEEE Proc. of Engineering in Medicine and Biology Society*, New York, États-Unis, pp. 6396–6399, 2006.

Annexe A

Article publié à EMBC

L'article présenté ici [25] a été publié lors de la 28^{e} conférence internationale *Engineering in Medicine and Biology Society* de IEEE qui s'est tenue à New York du 29 août au 3 septembre 2006.

Video Surveillance of Medication Intake

Myriam Valin, Jean Meunier, Alain St-Arnaud and Jacqueline Rousseau

Abstract— **In the context of the growing proportion of seniors in the western world population and the efforts provided in home care services, we have developed a computer vision system for monitoring medication intake. The system detects automatically medication intake using a single low-cost webcam. Person detection and tracking over the video sequence is done using color-based techniques while the recognition of the medication intake activity is performed using our main contribution, a three-level scenario model. Experimental results in controlled conditions are shown and we discuss improvements to our system.**

I. Introduction

In Canada, more than 30% of all people aged 65 and over live alone [1], with the associated risks for their health. According to the *Public Health Agency of Canada*, 18% to 50% of all medication used by seniors is used inappropriately and between 19% and 28% of hospital admissions for patients over 50 years of age occur as a result of medication problems [2]. Moreover, as mentioned in [3], approximately 125 000 people with treatable ailments die each year in the USA because they do not take their medication properly. A video surveillance system for monitoring seniors medication intakes could help reducing the risks as well as family member concerns.

The system presented here is part of a vaster project of video surveillance for home care services for seniors. It has been developed for monitoring medication intake in the context of a person living alone and whose medications are in bottles. In order to be as accessible as possible, a single low-cost webcam is used. The system has therefore to deal with low-quality digital images.

Recently, efforts have been made on developing computer vision systems for monitoring medication intake [4]. One problem with this approach is that the skin segmentation needs to be very precise to detect effectively occlusions, hand positions and orientations and face regions, which cannot be reached with low-quality images. Moreover, the authors consider that the medication intake activity occurs if a sequence of actions, verified every frame, is observed, disregarding the action durations and time laps between actions. Thus, the number of false alarms might become very high in real home situations. Because of the expected lack of tracking precision and the complexity of the activity to recognize, we developed a more complex algorithm to improve medication intake activity recognition, which forms our main contribution.

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

M. Valin and J. Meunier are with the Biomedical Engineering Institute and the Department of Computer Science and Operations Research, Université de Montréal, Montreal, Canada (e-mail: {[redacted]

A. St-Arnaud and J. Rousseau are respectively with the Centre de santé et de services sociaux Lucille-Teasdale and the Research Center of the Institut Universitaire de Gériatrie de Montréal, Montreal, Canada (e-mail: [redacted]

II. Overview

In order to perform medication intake monitoring, the system must first detect if the person is taking his medication. The proposed method is inspired by that in [5]. In our method, processing is divided in two parts:

- Low level processing: moving objects (regions) are detected and tracked at every frame (section III).
- High level processing: activity recognition is performed based on moving object characteristics, using our three-level scenario model (section IV).

When medication intake is detected, the used bottle and the time of detection are recorded. Results are presented in section V and system limitations and future work are discussed in section VI.

III. Detection and Tracking

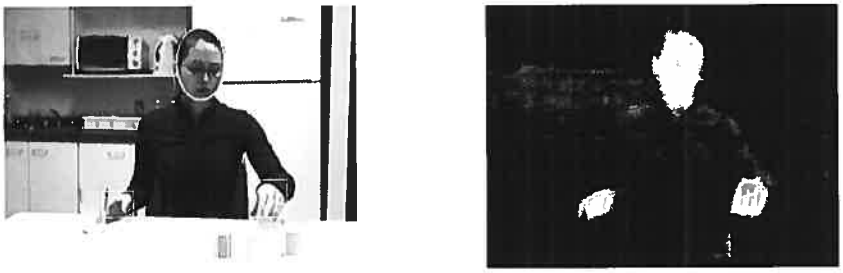
In our approach, three types of mobile objects are tracked: the person's head and hands and the medication bottles. Fig. 1a shows these objects. In this experimental setting, three medication bottles are present.

The head tracking algorithm is described in [6]. The head is modeled as an ellipse whose size can vary from one frame to the other. For each image, a local search determines the best fitting ellipse, based on the gradient magnitude around its perimeter and the likelihood of skin color inside it. The gradient magnitude of a pixel corresponds to the rate-of-change of intensities in the gray-scale image over a small local neighborhood. The color histogram used to compute skin color likelihood includes the person's hair color since the head might be turned or leaned.

Hands are positioned based on regions with high skin color likelihood. These regions are extracted from the skin color likelihood map (Fig. 1b) created during the head tracking process. Possible occlusions between the head and the hands or between both hands are also dealt with based on previous positions and a few assumptions.

Since we want to detect which medication is being taken and labels cannot be identified automatically in low resolution video, color bands are affixed on the bottles to better differentiate them.

The detection of medication bottles is done using color models as described in [7]. Pixels classification is performed based on their Mahalanobis distance to the color models,



a b

Fig. 1. a) Mobile objects detection and tracking and b) corresponding skin color likelihood map.

which are created from training samples. Possible bottle regions (objects) are extracted using connected component analysis. Bottle positioning is done considering the size of these objects and bottle previous positions, since bottles can be completely occluded by the hands.

IV. Activity Recognition

Activity recognition is based on the concept of scenarios, which correspond to long-term activities of mobile objects. There are three levels of scenarios: single-state scenarios, multi-state scenarios and complex scenarios. Fig. 2 shows how these scenario levels are related. The concepts of single-state and multi-state scenarios have been developed in [5] and are briefly described here.

A. Single-state Scenarios

Single-state scenarios are defined by a set of mobile object properties and verified at every frame. For example, the scenario *"the person touches his head"* depends on the distances between the hands and the head. For each scenario, an all-or-none probability of occurrence is determined heuristically using logical functions. The single-state scenarios in our system are defined as follows:

- S_1: Exactly one hand manipulates a bottle,
- S_2: Both hands manipulate a bottle,
- S_3: One hand touches the head,
- S_4: One hand approaches the head,
- S_5: One hand moves away from the head,

For each scenario, the implicated mobile objects are kept.

In [5], the authors present another way of computing probabilities by using Bayesian classifiers and learned conditional probabilities of the mobile object properties, but this did not prove to be necessary for this application.

B. Multi-state Scenarios

A multi-state scenario corresponds to a sequence of single-state scenarios and is verified over a longer sequence of frames. A multi-state scenario (*MS*) is recognized if all its single-state scenarios are recognized consecutively.

The main idea is to compute the maximum probability that, at time t, the sequence of N single-state scenarios occurs given a sequence of observations O. This can be expressed as follows:

$$P(MS^*|O) = \max_{(t_1,t_2,\ldots,t)} P(S_{1_{(t_1,t_2-1)}} S_{2_{(t_2,t_3-1)}} \ldots S_{N_{(t_N,t)}}|O), \quad (1)$$

where $S_{i_{(t_i,t_{i+1}-1)}}$ means single-state scenario i occurs between frames t_i and $t_{i+1}-1$. t_1 and t are respectively the first and last frames of the multi-state scenario in the video sequence. The observations O correspond to the mobile object properties. The algorithm then consists of finding the best transition times t_i from one state to the next so that the whole sequence is completed with the maximum probability.

In the algorithm, $P(S_{i_{(t_i,t_{i+1}-1)}})$ is approximated by the expected recognition value of $S_{i_{(t_i,t_{i+1}-1)}}$ which is the mean of $P(S_{i_t})$ for $t_i \leq t \leq t_{i+1}-1$ and whose value is between 0 and 1.

The system developed for medication intake is divided in three multi-state scenarios:

- MS_1: The person opens a medication bottle and takes the pill(s),
- MS_2: The person swallows the pill(s),
- MS_3: The person closes the medication bottle,

each scenario having a probability of occurrence and starting and ending frames.

C. Complex Scenarios

Multi-state scenarios can represent simple activities, but do not work with more complex situations. For example, the activity of medication intake can be done in many different ways: the person can take the pills out of the medication bottle, swallow the pills and then close the medication bottle or take the pills out of the medication bottle, close the medication bottle and then swallow the pills. Moreover, actions not related might happen during the sequence, such as drinking water or putting the pills on the table. Finally, if more than one medication are being taken, the simple states which form the entire sequence might not occur consecutively. Thus, activity recognition cannot be done by searching for a fixed sequence.

We solve this problem by splitting the complex scenario into a sequence of multi-state scenarios (states) and then use a *"state-transition"* system to recognize the complex activity.

Once the multi-state scenarios are detected, before pushing the states into the state-transition system, a *"likelihood"*

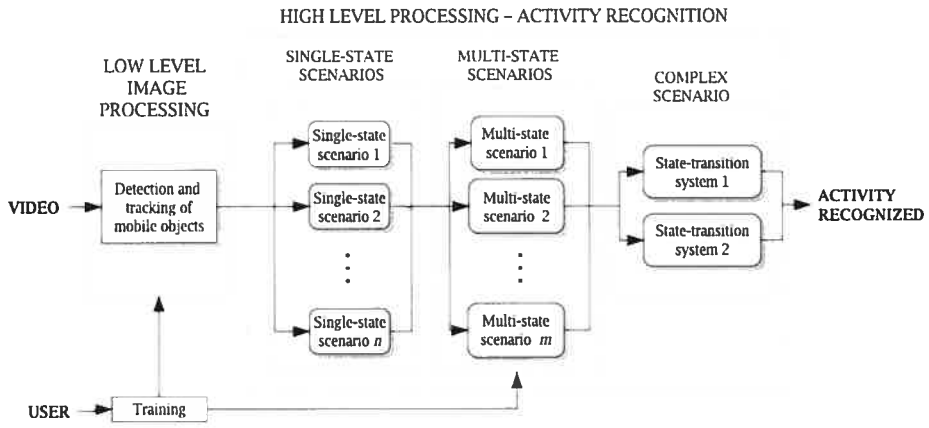


Fig. 2. Overview of the system.

factor is used to penalize the scenarios that are too long (or too short) compared to their mean durations, or lengths, by decreasing their probability. This factor is defined by a Butterworth filter transfer function and is expressed as follows:

$$f(\ell) = \frac{1}{1 + \left(\frac{\ell - \bar{\ell}}{w_c}\right)^{2n}} \tag{2}$$

where $\bar{\ell}$ is the mean scenario length, which varies from one multi-state scenario to the other and is manually defined. w_c and n (order of the filter) are fixed. As shown in Fig. 3, this function does not penalize recognition for small variations of duration around the mean. The resulting multi-state scenarios are then sorted according to their starting frame.

The two possible representations of the complex scenario used are shown in Fig. 4. The circles represent the multi-state scenarios and the arrows represent the transitions. Loops over the same state are possible because more than one medication can be taken. Since pills are not tracked specifically and more than one might be swallowed at the same time, the same event MS_2 might be used to recognize the medication intake for all pills.

Transitions are used to penalize the recognition if the number of frame increases too much between two states. Penalization is done only between states so the recognition of the intake of medication 1 does not fail if the person takes pills out of another medication bottle before swallowing the pills from bottle 1. The *"likelihood"* factor used here is also defined by another Butterworth filter transfer function. However, instead of depending on the state lengths, it depends on the transition lengths. The parameters have been set by hand after some experimentation, but could easily be learned automatically.

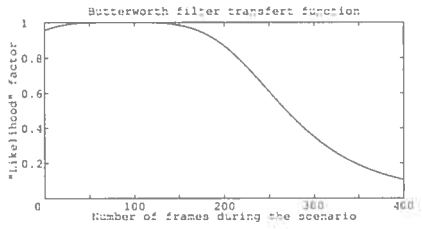
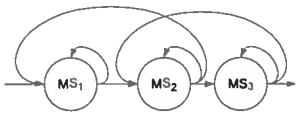


Fig. 3. Butterworth filter transfer function for $\bar{\ell} = 85$, $w_c = 184$ and $n = 2$ used for MS_1.

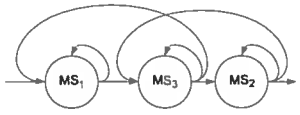
V. Results

The skin and bottle color models used for tracking have been computed in a first training step from hand labeled examples. The video acquisition is done at 15 frames per second with an image resolution of 320x240 pixels. The positions of the head, hands and bottles are automatically initialized as the person completely enters the camera field of view. We tailored the initialization procedure to the experimental conditions. Since the efforts were directed toward the recognition part, we do not extend on the performances of the tracking system. We just report here that it is good enough to support high-level activity recognition.

The detection of medication intake has been tested with 41 sequences from 26 videos taken with three different persons. 31 sequences show real medication intake and the 10 others show other activities (lures), like eating or manipulating the bottles.



a) First representation of complex scenario.



b) Second representation of complex scenario.

Fig. 4. State-transition systems representing complex scenario formed by multi-state scenario sequences a) $\{MS_1, MS_2, MS_3\}$ and b) $\{MS_1, MS_3, MS_2\}$.

Typical results are presented in Fig. 5. Fig. 5a shows a sequence of multi-state scenarios obtained from a video with medication intake. The results are 99.6% detection for both bottles. The complex scenario is defined by its second representation, shown in Fig. 4b.

Fig. 5b shows a sequence of complex scenarios involving three bottles. We can see that a detection of medication intake with the bottle 3 overlaps all the other scenarios. This has happened because the bottle was manipulated at the beginning of the sequence. The low probability shows the effectiveness of the penalization, which allows to reject the false detection.

Considering that a scenario is recognized if its probability is over 80%, results for all video sequences are summarized in Table I. In fact, most of the scenarios recognized had a probability over 90%.

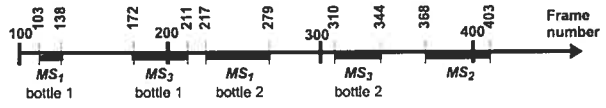
A major source of errors is occlusions between mobile objects. Indeed, if a mobile object passes in front of an other one, from the camera 2D point of view, the two objects are in (false) contact. Since contacts between hands and medication bottles form the core of states MS_1 and MS_3, errors can affect the whole recognition process. To limit this problem, the camera has been placed a little higher than the person's head, decreasing, this way, the number of occlusions.

In the actual system, the tracking part does not work in real time but it has not yet been optimized. Since the recognition part can be done after the person left the table, computation time is fast enough to be used in a continuous application.

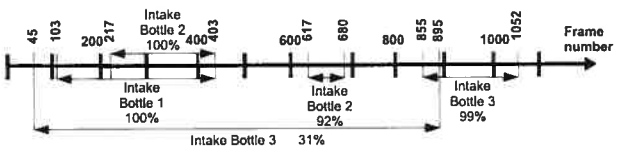
VI. Conclusion

Having described our experiments, we now discuss possible issues in less controlled environments. First, detection and tracking errors could be due to:

- low-quality of the webcam images resulting in high noise level and difficulties with pixel color classification,
- presence of objects similar to the medication bottles (same color and size),



a) Sequence of multi-state scenarios.



b) Sequence of complex scenarios.

Fig. 5. Typical results : a) sequence of multi-state scenarios forming the complex scenario $\{MS_{1_1}, MS_{3_1}, MS_{1_2}, MS_{3_2}, MS_2, MS_2\}$ and b) sequence of complex scenarios involving multiple bottles with corresponding probabilities.

TABLE I
Recognition results for 41 sequences.

Activity	Detected	Not Detected
Medication Intake	28	3
Other (lures)	1	9

- presence of occluding objects between the camera and the person.

Because activity recognition depends heavily on tracking of mobile objects, any enhancement of the tracking algorithm would be beneficial. For example, work could be done on finding a better skin color model, performing 3D positioning (to limit false contacts) and using a mouth-to-hand distance (instead of a head-to-hand distance) to test the S_3 scenario.

Second, changes in a person's normal behavior of medication intake could also impair recognition. For example, errors could happen if :

- the person touches his forehead instead of his mouth in the medication intake sequence (the mouth is not specifically tracked in our actual system),
- the person moves away and comes back between opening the medication bottle and swallowing his pills.

We have to mention here that, in this application, we consider that the person agrees to take his medication and that non-compliance is due to memory problems rather than non-acceptance.

Considering these possible sources of errors, we expect that adaptations should be needed for the medication intake monitoring system to work within real home environments. However, the presented algorithms have shown to be effective and promising.

We have tested the activity recognition process for medication intake, but it should be applicable to other everyday home activities.

References

[1] C. Lindsay. "A Portrait of Seniors in Canada," Third Edition. Ottawa: *Statistics Canada*, 1999.

[2] Committee of Officials (Seniors) for the Ministers Responsible for Seniors, "Working together on seniors medication use: a federal/ provincial/ territorial strategy for action," *Public Health Agency of Canada*, Ottawa, June 1996.

[3] C. Nugent et al., "Can technology improve compliance to medication?," in *Proc. 3rd Int. Conf. On Smart homes and health Telematic*, Sherbrooke, QC, Canada, 2005, pp. 65–72.

[4] D. Batz, M. Batz, N. da Vitoria Lobo and M. Shah, "A computer vision system for monitoring medication intake," in *Proc. IEEE 2nd Canadian Conf. on Computer and Robot Vision*, Victoria, BC, Canada, 2005, pp. 362–369.

[5] S. Hongeng, R. Nevatia and F. Bremond, "Video-based event recognition: activity representation and probabilistic recognition methods," *Computer Vision and Image Understanding*, vol.96, no. 2, Nov., pp. 129–162, 2004.

[6] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," in *Proc. IEEE Computer Vision and Pattern Recognition*, Santa Barbara, CA, USA, 1998, pp. 232–237.

[7] N. Habili, C. Lim and A. Moini, "Hand and face segmentation using motion and color cues in digital image sequences," in *Proc. IEEE Int. Conf. on Multimedia and Expo*, Tokyo, Japan, 2001, pp. 377–380.