Reconnaissance d'activités humaines à partir de séquences vidéo

Mouna Selmi

DOCTORAT EN CO-ACCREDITATION TELECOM SUDPARIS ET L'UNIVERSITE EVRY VAL D'ESSONNE

Spécialité : Informatique

Ecole doctorale : Sciences et Ingénierie

Présentée par :
Mouna Selmi

Pour obtenir le grade de
DOCTEUR DE TELECOM SUDPARIS

Reconnaissance d'activités humaines à partir de séquences vidéo

Thèse dirigée par Bernadette Dorizzi

Soutenue le 12/12/2014 devant le jury composé de :

Rapporteurs :	Mme.	Catherine Achard	Université Pierre et Marie Curie
	M.	Jean-Luc Dugelay	EURECOM
Directrice de thèse :	Mme.	Bernadette Dorizzi	Telecom SudParis
Encadrant :	M.	Mounim A. El Yacoubi	Telecom SudParis
Examinateurs :	M.	Laurent Lucat	CEA-List
	M.	Jean-Claude Martin	LIMSI-CNRS/CPU, Université Paris Sud

Thèse n° : 2014TELE0029

Remerciements

Tout d'abord, un immense merci et une reconnaissance éternelle à Benadette Dorizzi , directrice de thèse, qui m'a donné la chance de réaliser ce travail dans les meilleures conditions. Un très grand merci à Mounim A. El Yacoubi, encadrant de thèse. Merci pour votre excellent encadrement, pour votre disponibilité, pour vos précieuses suggestions, ainsi que pour vos nombreuses relectures et corrections de ce manuscrit. Merci aussi pour votre encouragements et soutien continus sans lesquels je n'aurai jamais réussi à aller jusqu'au bout.Ce fut un plaisir et un grand honneur de travailler sous votre direction.

Je tiens à remercier également Catherine Achard, maître de conférences à l'université Paris 6, et Jean-Luc Dugelay, professeur à EURECOM, pour l'honneur qu'elles m'ont fait d'avoir accepté de rapporter cette thèse et pour le temps consacré à la lecture et à l'évaluation de mon travail.

Je remercie sincèrement Laurent Lucat, Docteur à CEA-List, et Jean-Claude Martin, professeur à LIMSI-CNRS/CPU, Université Paris Sud, d'avoir accepté d'examiner mon travail et de faire partie de mon jury.

Merci aussi à tous les membres de l'équipe ainsi que les thésards et les stagiaires que j'ai rencontrés durant ces années pour leur gentillesse et leur bonne humeur.

Je remercie infiniment toutes mes chères amies d'avoir compris qu'une thèse demandait beaucoup de disponibilité et de m'avoir pardonner car j'étais loin d'elles dans plusieurs occasions. J'aimerais mentionner spécialement : Rim, Nadia, Meriem, Hana, Aicha, Gaberiela,...

Mes remerciements les plus chaleureux vont à mes parents et mes frères pour leur amour et leurs encouragements continus. Parce qu'ils m'ont inculqué le goût du travail et de la réussite, et parce que leurs prières ont toujours accompagné mes pas partout, je tiens à remercier particulièrement ma mère et mon père. C'est avec émotion que je leur dévoile le fruit de mes efforts. Je vous aime et je vous souhaite une longue vie ainsi qu'une santé de fer.

Un grand merci à mon mari Ali qui partage ma vie et mes idées. Merci pour ton soutien, ton encouragement et pour tes nombreux conseils tout le long de ma thèse. Un grand merci !

Je remercie finalement toutes les personnes qui m'ont aidé de près ou de loin, à la réalisation de ce travail.

Résumé

Cette thèse s'inscrit dans le contexte de la reconnaissance d'activités à partir de séquences vidéo, une des préoccupations majeures dans le domaine de la vision par ordinateur. Les domaines d'application pour les systèmes de vision sont nombreux. On peut citer notamment la vidéo surveillance, la recherche et l'indexation automatique de vidéos ou encore l'assistance aux personnes âgées et fragiles. Cette tâche reste problématique étant données les grandes variations dans la manière de réaliser les activités, l'apparence de la personne et les variations des conditions d'acquisition. L'objectif principal de cette thèse est de proposer une méthode de reconnaissance efficace qui soit robuste aux différents facteurs de variabilité.

Les représentations fondées sur les points d'intérêt ont montré leur efficacité dans les travaux de l'art ; elles ont été généralement couplées avec des méthodes de classification globales étant donné que ces primitives sont temporellement et spatialement désordonnées. Les travaux les plus récents atteignent des performances élevées en modélisant le contexte spatio-temporel des points d'intérêt ; par exemple en encodant leur voisinage sur plusieurs échelles. Dans cette thèse, nous proposons une méthode de reconnaissance d'activités, fondée sur un modèle hybride *Séparateur à Vaste Marge-Champ Aléatoire Conditionnel Caché* (SVM-HCRF) qui modélise explicitement l'aspect séquentiel des activités tout en exploitant la robustesse des points d'intérêt dans des conditions réelles. Nous commençons par l'extraction des points d'intérêt et nous montrons leur robustesse par rapport à l'identité de la personne par une analyse tensorielle multilinéaire. Ces primitives sont ensuite représentées comme une séquence de sacs de mots (BOW) locaux : la séquence vidéo est segmentée temporellement en utilisant la technique de fenêtre glissante et chacun des segments ainsi obtenu est représenté par le BOW des points d'intérêt lui appartenant. Le premier niveau de notre système de classification séquentiel hybride consiste à appliquer le SVM en tant que classifieur de bas niveau afin de convertir les BOWs locaux en des vecteurs de probabilités des classes d'activité. La séquence de vecteurs de probabilités ainsi obtenue est utilisée comme entrée du HCRF. Ce dernier permet de classifier d'une manière discriminante les séries temporelles tout en modélisant leurs structures internes via les états cachés.

Nous avons évalué notre approche sur des bases de données publiques ayant des caractéristiques diverses. Les résultats atteints sont compétitifs par rapport à ceux de l'état de l'art. Nous avons montré, en effet, que l'utilisation du classifieur de bas niveau permet d'améliorer la performance du système de reconnaissance vu que le classifieur séquentiel HCRF traite directement les informations sémantiques issues des BOWs locaux, à savoir la probabilité de chacune des activités relativement au segment local en cours. De plus, les vecteurs de probabilités ont une dimension faible, réduisant de manière significative, en conséquence, le risque de

sur apprentissage qui peut intervenir si la dimension des vecteur de caractéristiques est importante relativement au nombre des données ; c'est justement le cas d'ailleurs lorsqu'on utilise les BOWs, de dimension généralement très élevée. L'estimation les paramètres du HCRF dans un espace de dimension réduite permet aussi de réduire considérablement la durée de la phase d'entrainement.

Mots clés : Reconnaissance d'activités ; Points d'intérêt ; Points denses ; Analyse tensorielle multilinéaire, Classification de données séquentielles ; Séparateurs à vaste marge ; Champs aléatoires conditionnels cachés.

Abstract

Human activity recognition (HAR) from video sequences is one of the major active research areas of computer vision. There are numerous application HAR systems, including video-surveillance, search and automatic indexing of videos, and the assistance of frail elderly. This task remains a challenge because of the huge variations in the way of performing activities, in the appearance of the person and in the variation of the acquisition conditions.

The main objective of this thesis is to develop an efficient HAR method that is robust to different sources of variability.

Approaches based on interest points have shown excellent state-of-the-art performance over the past years. They are generally related to global classification methods as these primitives are temporally and spatially disordered. More recent studies have achieved a high performance by modeling the spatial and temporal context of interest points by encoding, for instance, the neighborhood of the interest points over several scales. In this thesis, we propose a method of activity recognition based on a hybrid model Support Vector Machine - Hidden Conditional Random Field (SVM-HCRF) that models the sequential aspect of activities while exploiting the robustness of interest points in real conditions. We first extract the interest points and show their robustness with respect to the person's identity by a multilinear tensor analysis. These primitives are then represented as a sequence of local "Bags of Words" (BOW) : The video is temporally fragmented using the sliding window technique and each of the segments thus obtained is represented by the BOW of interest points belonging to it.

The first layer of our hybrid sequential classification system is a Support Vector Machine that converts each local BOW extracted from the video sequence into a vector of activity classes probabilities. The sequence of probability vectors thus obtained is used as input of the HCRF. The latter permits a discriminative classification of time series while modeling their internal structures via the hidden states.

We have evaluated our approach on various human activity datasets. The results achieved are competitive with those of the current state of art. We have demonstrated, in fact, that the use of a low-level classifier (SVM) improves the performance of the recognition system since the sequential classifier HCRF directly exploits the semantic information from local BOWs, namely the probability of each activity relatively to the current local segment, rather than mere raw information from interest points. Furthermore, the probability vectors have a low-dimension which prevents significantly the risk of overfitting that can occur if the feature vector dimension is relatively high with respect to the training data size ; this is precisely the case when using BOWs that generally have a very high dimension. The estimation of the HCRF parameters in a low dimension allows also to significantly reduce the duration of the HCRF training phase.

Key words : Human Activity Recognition ; Interest Points ; Dense Points ; Mul-

tilinear tensor analysis ; Classification of sequential data ; Support Vector Machines ; Hidden Conditional Random Fields.

Table des matières

Table des figures

Liste des tableaux

Acronymes

ACP : Analyse en Composantes Principales
AMCP : Analyse Multilinéaire en Composantes Principales
ADL : Activité de la vie quotidienne - Activities of Daily Living
ASLP : Analyse Sémantique Latente Probabiliste
BOW : Sac de mots - Bag Of Words
CRF : Champs aléatoires conditionnels - Conditional Random Fields
DBN : Réseaux bayésiens dynamiques - Dynamic Bayesian Networks
FCRF : Champs aléatoires conditionnels factoriels - Factorial Conditional Random Field
GPS : Système de gocalisation par satellite - Global Positioning System
HCRF : Champs aléatoires conditionnels cachés - Hidden Conditional Random Field
HMM : Modèle de Markov à ts cachés - Hidden Markov Models
HOF : Histogramme de flux optique - Histogram of Optical Flow
HOG : Histogramme de gradients orientés - Histograms of Oriented Gradients
HOJ3D : Histogrammes des positions 3D des articulations
KLT : KanadeLucasTomasi
KNN : K plus proches voisins - K-Nearest Neighbors
LHMM : Modèle de Markov caché hiérarchique - Layered Hidden Markov Model
MAP : Maximum à Posteriori
MBH : Histogramme de frontière de mouvement - Motion Boundary Histograms
MCMC : Monte Carlo par chaîne de Markov - Monte Carlo Markov Chain
MEI : Image d'énergie du mouvement - Motion Energy Image
MEMM : Modèles de Markov d'Entropie Maximale - Maximum Entropy Markov Model
MHI : Image de l'historique du mouvement - Motion History Images
MHV : Volume de lhistorique du mouvement - Motion History Volume
ROI : Région d'intérêt - Region of Interest
RFID : Radio-identification - Radio Frequency Identification
SIFT : Scale Invariant Features Transform
STIP : Point d'intérêt spatio-temporel -Spatio-Temporel Interest Point
SURF : Speeded Up Robust Features
SVM : Séparateur à vaste marge

CHAPITRE 1

Introduction générale

Sommaire

1.1 Contexte

La vision par ordinateur est une thématique passionnante de recherche visant à doter des systèmes informatiques de capacités d'analyse et d'interprétation du contenu visuel d'une scène proches de celles de la vision humaine. L'un des objectifs majeurs en vision par ordinateur est de reconnaître et de comprendre le mouvement humain et notamment de permettre la classification des activités [1] humaines. La reconnaissance d'actions peut intervenir dans différents domaines, englobant :

- **La vidéo-surveillance automatique :** Les systèmes de vidéo-surveillance automatique aident à assurer au mieux la sécurité de personnes, de sites sensibles ou de lieux publics et privés. Elle permet d'analyser automatiquement des scènes et de détecter des comportements suspects. Les systèmes de reconnaissance d'activités permettent dans ce contexte la détection d'activités inhabituelles telles que l'agression ou l'intrusion.

- **L'indexation et recherche des vidéos :** L'indexation et la recherche des vidéos dans les archives de bases de données à grande échelle reposent essentiellement sur la description textuelle faite manuellement par l'humain. En effet, les moteurs de recherche pour vidéos sur le web reposent essentiellement sur les annotations textuelles afin de retrouver des vidéos pertinentes. La reconnaissance automatique des actions humaines permet d'améliorer l'indexation automatique des vidéos et ainsi d'améliorer la pertinence des réponses proposées aux utilisateurs.

1. Dans ce manuscrit, "action" et "activité" désignent la même entité.

- **Applications E-santé :** La reconnaissance des activités joue aussi un rôle important dans les applications de suivi médical et d'assistance aux personnes âgées que ce soit dans les habitats intelligents pour la santé, les hôpitaux ou dans les résidences pour personnes âgées. La reconnaissance des activités permet par exemple de détecter les évènements pouvant être dangereux pour la personne âgée et prévenir toute situation à risque.

Pour les applications d'e-santé, la reconnaissance automatique des activités permet d'améliorer la sécurité des personnes âgées souhaitant vivre le plus longtemps possible dans leurs domiciles. La reconnaissance des activités permet dans ce contexte de détecter les cas d'urgence tels que la chute et la non prise de médicaments. De plus, elle constitue une étape primordiale dans les systèmes de détection de perte d'autonomie de la personne âgée. En fait, la reconnaissance des activités de la vie quotidienne d'une personne âgée vivant seule permet d'observer et d'évaluer son état de dépendance. D'un point de vue clinique, l'entrée en dépendance se traduit par l'impossibilité de réaliser des activités élémentaires de la vie quotidienne. Les changements subtils dans des activités des personnes âgées pourraient aussi donner des signes importants sur la progression de certaines maladies. A titre d'exemple, des troubles de sommeil pourraient être causés par une insuffisance cardiaque ou encore par des maladies chroniques.

Plusieurs types de technologies ont été mis en oeuvre pour la reconnaissance des activités des personnes à domiciles à savoir les capteurs, les systèmes de reconnaissance sonore et les systèmes de reconnaissance vidéo. Dans le cadre du projet "Juliette" où notre tâche consiste à doter le robot humanoïde Nao de la capacité de reconnaitre des activités humaines des personnes âgées vivant seules dans leurs domiciles, nous abordons le problème de la reconnaissance des activités humaines à partir de séquences vidéo. Cette dernière a des atouts intéressants comme le caractère non- invasif et le faible coût. De plus, la vidéo a un contenu sémantiquement très riche et assez pertinent pour la reconnaissance des activités humaines.

1.2 Problématiques posées

Bien que le contenu vidéo soit assez informatif, la tâche de reconnaissance de l'action humaine reste problématique. Ceci est dû, d'une part, au grand degré d'occlusion dans les séquences réelles (dans notre contexte, grand degré d'occlusion à cause des meubles de maison). Ceci rend difficile l'extraction des postures et leur suivi de façon fiable et peut conduire à des difficultés dans l'interprétation de l'action au cours du temps. Une autre difficulté principale vient de la grande variabilité intra-classe à laquelle une classe d'action peut s'exposer. En effet, de grandes variations de style peuvent être observées dans la reproduction de la même action selon les habitudes, le genre, la taille, l'ethnicité des personnes et selon aussi le contexte dans lequel l'action se déroule. Par exemple, l'action "marcher" varie selon la vitesse et la taille de la personne. De plus, il faut prendre en compte l'ambiguïté interclasses puisqu'il existe des actions

similaires telles que les actions "courir lentement" (jogging) et "marcher" qui ne diffèrent que par la vitesse de réalisation (Figure 1.1). Ainsi un système de reconnaissance d'actions fiable doit avoir une grande capacité de généralisation pour prendre en compte les variations intra-classes tout en étant capable de discriminer les actions ayant une petite variation inter-classes.
Des difficultés additionnelles sont liées aux conditions d'enregistrement des séquences vidéo. Le changement des conditions d'illumination, le fond dynamique et la distance de la caméra peuvent influencer l'apparence des personnes effectuant l'action. De plus, l'apparence visuelle de l'action peut changer selon l'angle de vue d'enregistrement (Figure 1.1).

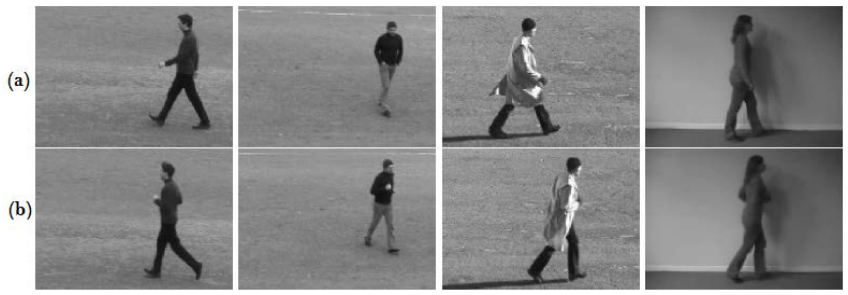


FIGURE *1.1: Illustration des facteurs de variabilité lors de la réalisation des actions "macher"(a) et "jogging" (b). [Schuldt et al., 2004].*

1.3 Contributions

Le but est de proposer une méthode de reconnaissance qui soit adaptée aux conditions réelles. Les contributions de cette thèse peuvent être résumées dans les points suivants :

- Nous proposons dans cette thèse une modélisation séquentielle des séquences d'activités en se basant sur les points d'intérêt comme descripteur. Ces primitives se sont dernièrement montrées pertinentes pour la reconnaissance des actions dans des scénarios réalistes. Ils ont été généralement couplés à des méthodes de classification globale étant donné qu'ils sont temporellement et spatialement désordonnés alors que les méthodes séquentielles nécessitent une séquence de primitives de taille fixe. Toutefois, les méthodes séquentielles sont bien plus adaptées aux problèmes de classification des séquences temporelles ; elles ont démontré leur efficacité grâce à leur capacité de modéliser la structure interne des séquences tout en étant discriminantes. Pour pouvoir utiliser les points d'intérêt avec les méthodes séquentielles, nous représentons les séquences vidéo en tant que séries temporelles de sacs de mots (BOW) locaux. Nous commençons par découper la vidéo en une séquence de segments temporels en utilisant la technique de fenêtre glissante. Chaque segment est représenté par un BOW des points d'intérêt qu'il contient.

- Notre modélisation séquentielle est basée sur un nouveau modèle hybride SVM-HCRF. Ce modèle permet de combiner le pouvoir de généralisation et de discrimination des séparateurs à vaste marge (SVM) pour la classification de données statiques (BOWs locaux) et la modélisation de la structure temporelle des actions via des modèles séquentiels conditionnels, en l'occurrence, les champs aléatoires conditionnels cachés (HCRF). SVM est appliqué en tant que classifieur de bas niveau pour extraire les vecteurs de probabilités conditionnelles des actions pour chacun des segments à partir des BOWs locaux. HCRF traite donc des informations sémantiques, à savoir les probabilités de chacune des activités au niveau local, ce qui permet d'améliorer l'efficacité de système de reconnaissance. L'utilisation d'un classifieur de bas niveau conduit aussi à une amélioration substantielle de la robustesse par rapport au problème de sur-apprentissage qui peut survenir si la dimension de vecteurs de primitives est élevée relativement au nombre des données d'entrainement. De plus, ce schéma permet un gain considérable en temps de calcul vu que le HCRF a comme entrée des vecteurs de dimension beaucoup moins élevée relativement à celle des BOWs.

- Nous avons effectué une analyse multi-tensorielle de la sensibilité des points d'intérêt par rapport à l'un des principaux facteurs de variabilité à savoir l'identité des personnes, permettant ainsi d'analyser leur corrélation avec le mouvement.

1.4 Plan de la thèse

Ce manuscrit est organisé en 5 chapitres :

Le chapitre 2 : Etat de l'art présente les travaux de l'état de l'art concernant les technologies d'assistance aux personnes âgées ainsi que les principaux travaux portant sur la reconnaissance des actions à partir des séquences vidéo.

Le chapitre 3 : Extraction des primitives présente les caractéristiques techniques des points d'intérêt que nous avons utilisés dans nos travaux ainsi qu'une analyse tensorielle de la robustesse des points d'intérêt par rapport aux facteurs nuisibles à la reconnaissance d'activités.
Nous proposons également une nouvelle modélisation des séquences vidéo qui est bien adaptée aux méthodes de reconnaissance séquentielle.

Le chapitre 4 : Modèle de reconnaissance hybride SVM-HCRF expose notre méthode de reconnaissance séquentielle hybride SVM-HCRF. Ce nouveau modèle de reconnaissance permet, d'une part, de modéliser d'une manière explicite l'aspect temporel des actions via HCRF et, d'autre part, de modéliser l'aspect statique des segments locaux de la vidéo par SVM qui est utilisé en tant que classifieur de bas niveau.

Le chapitre 5 : Expérimentations étudie la fiabilité de notre système de reconnaissance. Nous commençons par une présentation des bases de données utilisées dans nos expérimentations. Nous évaluons ensuite la pertinence de nos primitives décrivant les activités visuelles. Enfin, nous procédons à l'évaluation de la performance de notre système de reconnaissance.

Le chapitre 6 : Conclusion et perspectives rappelle nos principales contributions ainsi que nos perspectives de recherche.

CHAPITRE 2

Etat de l'art

Sommaire

2.1 Introduction

Avec le vieillissement rapide de la population, toute une réflexion est apparue autour des technologies apportant des réponses pratiques aux besoins des personnes âgées, en matière de maintien à domicile, avec un renforcement de leur sécurité et une amélioration de leur qualité de vie. Les technologies à base de reconnaissance d'actions à partir de séquences vidéo éveillent de plus en plus d'intérêt en raison de leur efficacité. Dans cette partie, nous allons tout d'abord présenter les données concernant la démographie française et mondiale, ce qui nous permettra de comprendre les enjeux et la nécessité des technologies d'assistance aux personnes âgées. Nous présenterons ensuite les différents types des technologies d'assistance aux personnes âgées. La section 2.5 sera consacrée à la présentation des différentes approches de reconnaissance des actions humaines à partir de séquences vidéo.

2.2 Contexte social

Ces dernières années, les évolutions dans le domaine médical ont entraîné une augmentation de l'espérance de vie de la population. Cette augmentation a pour conséquence un accroissement de l'âge moyen qui se traduit par un vieillissement de la population. Selon l'ONU [1], les personnes âgées de 60 ans ou plus représentaient, en 2007, près du cinquième de la population dans les pays développés et, d'ici à 2050, elles devraient en constituer le tiers, soit 2 milliards d'individus (Figure 2.1).

Au niveau national, le nombre de personnes de 60 ans et plus augmenterait de 10,4 millions entre 2007 et 2060 selon le scénario optimal. En 2060, 23,6 millions de personnes seraient ainsi âgées de 60 ans ou plus, ce qui représenterait une hausse de 80 % en 53 ans. L'augmentation est la plus forte pour les plus âgées : le nombre de personnes de 75 ans ou plus passerait de 5,2 millions en 2007 à 11,9 millions en 2060 (Figure 2.2).

Certes, l'allongement de l'espérance de vie de la population est un bienfait mais cela soulève des problèmes de caractère individuel et sociétal liés à la qualité de vie des personnes âgées. En effet, avec l'avancement de l'âge, les problèmes de santé deviennent plus nombreux et les problèmes liés à la perte d'autonomie se multiplient. Le haut risque de chutes et de troubles du comportement, par exemple, diminue considérablement la qualité de vie de celles qui en souffrent. Avec les avancées médicales, l'accès aux soins de maladies chroniques est devenu aisé, alors que les difficultés liées à la perte d'autonomie restent un problème majeur de santé publique et également de société. En effet, en France, aux alentours des années 2030, une personne sur quatre sera concernée par le problème de dépendance. En revanche, l'entrée en établissement d'hébergement pour les personnes âgées devient de plus en plus difficile en raison du nombre limité de places disponibles : en 2007, on estimait qu'il manquait de 30 000 à 40 000 places. Dans les années à venir, l'écart entre les besoins de prise en charge des

1. ONU : Organisation des Nations Unies

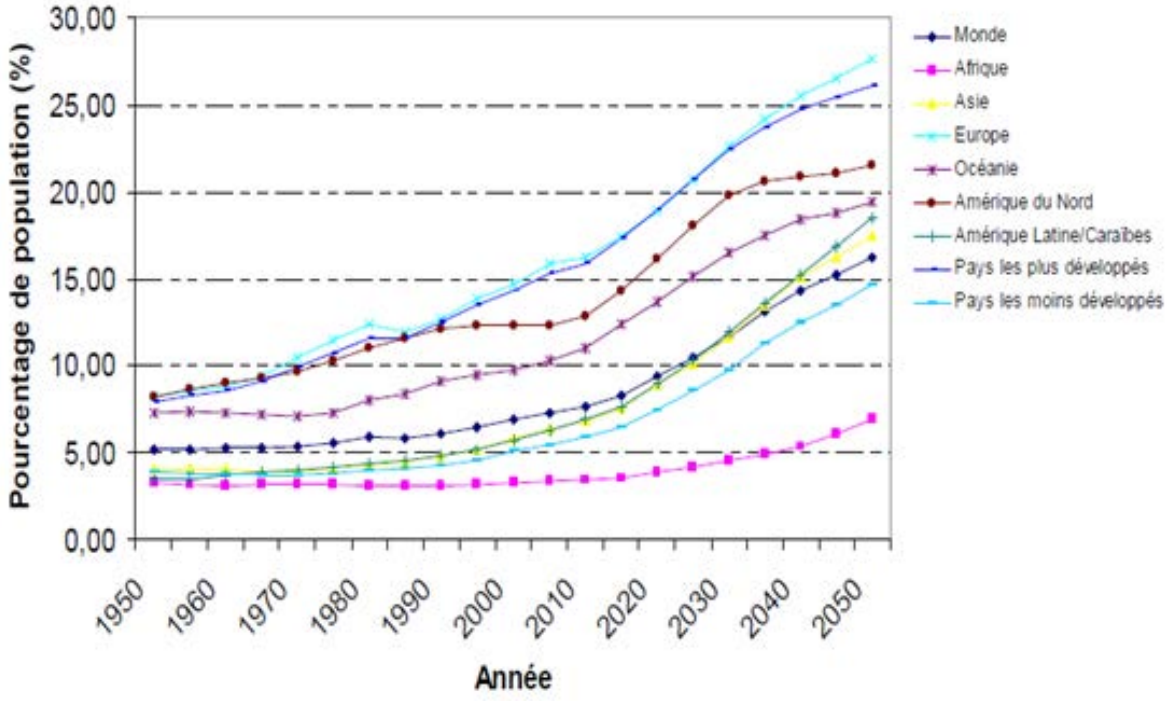


FIGURE *2.1: Pourcentage de la population âgée de plus de 65 ans (1950-2050), selon l'organisation des Nations Unies « United Nation, population Division ».*

personnes âgées et le nombre de places en établissements deviendra encore plus important, vu l'accroissement rapide de l'espérance de vie.

Pour diminuer cette pression, le maintien à domicile des personnes âgées, le plus longtemps possible, semble être une bonne solution. Cela permet, notamment, à la personne concernée, de préserver au maximum ses liens familiaux et sociaux ainsi que de conserver un environnement familier.

Dans ce contexte, un système de suivi et d'analyse des comportements des personnes âgées encore indépendantes, vivant seules à leur domicile est plus que nécessaire. Cela permettra de garantir leur sécurité, d'observer l'évolution de leur niveau de dépendance et d'émettre une alerte en cas de perte d'autonomie. De plus, il est bien connu que même des changements subtils dans le comportement des personnes âgées peuvent donner des signes importants quant à la progression de certaines maladies. Des troubles de sommeil, par exemple, pourraient être causés par une insuffisance cardiaque ou par des maladies chroniques. Les changements dans la démarche, d'autre part, peuvent être associés à des signes précoces de troubles neurologiques liés à plusieurs types de démence. Ces exemples soulignent l'importance de l'observation continue des changements de comportement chez les personnes âgées afin de détecter une détérioration de la santé avant que celle-ci ne devienne critique. Ainsi, la détection de ces situations à risque repose sur l'identification précoce des facteurs de « fragilité »

Année	Proportion au 1er janvier (en milliers)	Proportion (%) des					Solde naturel (en milliers)	Solde migratoire (en milliers)
		0-19 ans	20-50 ans	60-64 ans	65-74 ans	75 ans et +		
1960	45 465	32.3	51.0	5.1	7.3	4.3	298.9	140
1970	50 528	33.1	48.8	5.2	8.1	4.7	308.1	180
1980	53 731	30.6	52.4	3.0	3.8	5.7	253.3	44
1990	56 577	27.8	53.2	5.1	7.1	6.8	236.2	80
2000	58 858	25.6	53.8	4.6	8.8	7.2	243.9	70
2007	**61 759**	**24.8**	**53.8**	**4.9**	**8.1**	**8.5**	**263.9**	100*
2015	64 514	24.2	51.0	6.2	9.3	9.3	201.5	100
2020	65 962	23.9	49.6	6.0	11.0	9.4	173.2	100
2025	67 28S	23.5	48.4	6.1	11.1	10.9	154.1	100
2030	68 532	23.0	47.5	6.0	11.1	12.3	142.1	100
2035	69 705	22.6	46.7	5.9	11.1	13.6	120.0	100
2040	70 734	22.4	46.6	5.3	11.1	14.7	82.4	100
2050	72 275	22.3	45.9	5.6	10.2	16.0	31.9	100
2060	**73 557**	**22.1**	**45.8**	**5.4**	**10.5**	**16.2**	**+30.6****	**100**

FIGURE *2.2: Évolution de la population de la France métropolitaine de 1960 à 2060 (scénario central de projection)*

parmi lesquels figure, notamment, l'incapacité d'une personne à exécuter seule les activités de la vie quotidienne. Nous allons, dans la section suivante, présenter les différentes technologies d'assistance aux personnes âgées permettant un suivi de leur comportement.

2.3 Technologies d'assistance aux personnes âgées

Le suivi de comportement et, plus précisément, la capacité de la personne à exécuter les activités de la vie quotidienne sont des éléments importants pour détecter l'entrée en dépendance de la personne âgée et dépister des situations à risque. Nous allons présenter les différents types de technologies utilisées pour mesurer l'activité, à savoir les capteurs, la reconnaissance sonore et la reconnaissance à partir de séquences vidéo.

2.3.1 Les capteurs

Ces dernières années, la miniaturisation des composants ainsi que la facilité de leur intégration ont poussé plusieurs chercheurs à utiliser, dans leurs travaux de recherche, des capteurs pour mesurer l'activité de la personne. En effet, un système de capteurs qui est capable de reconnaître automatiquement les activités à la maison permettrait de nombreuses applications

potentielles dans le domaine de la santé. Les capteurs sont des dispositifs qui peuvent être utilisés pour détecter l'interaction entre une personne et son environnement. Il existe une grande variété de capteurs qui diffèrent les unes des autres en termes de prix, de facilité d'installation et de type de données en sortie [Fogarty et al., 2006]. Le tableau 2.1 illustre un ensemble de types de capteurs et certaines considérations de leur utilisation.

Capteurs	Considération	Type
Accéléromètres	Doit être porté	Activité
Mouvement	Incapacité à distinguer entre les sujets	Activité
RFID	Nécessité que le lecteur soit porté et les étiquettes installées	Activité
GPS	Intimité non assurée	Activité
Contact	Incapacité à distinguer entre les sujets	Activité
Eau	Incapacité à distinguer entre les sujets	Contexte
Lumière	Incapacité à distinguer entre les sujets	Contexte
Pression	Incapacité à distinguer entre les sujets	Activité, biomédical
Présence	Incapacité à distinguer entre les sujets	Activité
Température	Incapacité à distinguer entre les sujets	Contexte, biomédical
Vidéo	Intimité non assurée	Contexte, Activité
Audio	Intimité non assurée	Contexte, Activité
Fréquence cardiaque	Doit être porté	Biomédical
Oxymétre de pouls	Doit être porté	Biomédical

TABLE *2.1: Types de capteurs et examen de leur utilisation [Valentin, 2010].*

Chacun des capteurs décrits dans le tableau 2.1 peuvent être classés selon le type d'informations qu'ils récoltent [Valentin, 2010] : activité, biomédical, contexte.

Les capteurs d'activité peuvent être utilisés pour déduire l'activité ou les comportements d'une personne. Parmi ces capteurs, on peut trouver les capteurs cinématiques embarqués tels que les accéléromètres tri-axes qui calculent trois accélérations linéaires selon trois axes orthogonaux de l'objet sur lequel ils sont fixés. L'actimétrie embarquée peut également utiliser des magnétomètres qui mesurent le champ magnétique perçu. On trouve aussi les gyroscopes qui délivrent une mesure de la vitesse angulaire instantanée autour d'un axe. On trouve également les capteurs de détermination de position tels que le système de localisation mondial (Global Positioning System : GPS) et la radio-identification (radio frequency identification : RFID) qui donnent une indication de l'endroit où une personne se trouve. En effet, si le récepteur GPS est porté par la personne, il est facile de détecter le mouvement de la personne autour de sa maison bien que la connaissance préalable de la scène soit obligatoire. Un capteur RFID pourrait effectuer une tâche similaire en utilisant des étiquettes RFID fixées sur des objets qui interagissent avec la personne. Cette catégorie inclut aussi les capteurs fixés sur l'ameublement de la maison. Il s'agit, notamment, des capteurs de contact sur les portes des armoires et des réfrigérateurs qui indiquent qu'elles ont été ouvertes, des capteurs de pression qui indiquent si une personne est assise dans un lit ou sur une chaise, et des capteurs électriques qui indiquent si un poêle a été allumé. La deuxième catégorie de capteurs, "capteurs de contexte", tels que les capteurs de lumière, les capteurs d'eau et les capteurs de température permet de recueillir des informations de contexte sur la scène .

Les capteurs biomédicaux sont conçus pour mesurer les signes vitaux des personnes. Ce type de capteurs joue un rôle important dans le suivi de l'état de santé des personnes âgées. Comme nous pouvons le voir dans le tableau ci-dessus, les capteurs biomédicaux peuvent mesurer la fréquence cardiaque, la pression artérielle, ou la température cutanée.

Projets de recherche industrielle et de surveillance des activités de la vie quotidienne basés sur les capteurs

L'un des meilleurs moyens de détecter les problèmes de santé physique et mentale est la recherche de changements dans les activités de la vie quotidienne (ADL) telles que la préparation des aliments, la prise des repas, le ménage, le bain ou la douche, l'habillement, l'utilisation des toilettes, la lessive et la gestion des médicaments. Plusieurs systèmes de surveillance des activités basés sur différents types de capteurs ont été proposés tels que les systèmes QuietCare [Qui, 2002], Physilog et Trident.

Le système QuietCare utilise des capteurs sans fil de mouvement et de température pour surveiller des personnes âgées vivant seules dans leurs domiciles Figure 2.3. Ces capteurs sont installés dans les pièces où on peut accueillir le plus d'information sur les activités telles que la chambre à coucher, la salle de bain, la cuisine et l'endroit où les médicaments sont rangés. Ceci permet par exemple de mesurer la durée des séjours dans la salle de bains, l'utilisation de médicaments et le nombre de sorties du lit pendant la nuit. Ainsi s'il y a un changement par rapport au schéma de mouvement quotidien, une alerte est envoyée à la personne en charge de la surveillance. La limitation principale de ce système est qu'il fournit une analyse d'activité limitée en détectant juste la présence d'un mouvement dans une chambre sans aucune identification de nature de l'activité effectuée.

FIGURE *2.3: Le système QuietCare*

Le système Physilog, développé par le laboratoire LMAM de l'EPFL, est un système portable de suivi à long terme de mouvements. Il mesure le mouvement à partir des signaux enregistrés à travers un ou plusieurs capteurs cinématiques fixés sur le patient, soit par l'intermédiaire d'une ceinture, soit en étant directement collés sur la peau (Figure 2.4).

Le CEA-LETI et sa startup Movea ont développé un dispositif de mesure de mouvement en 3D. Ce dispositif a été utilisé dans le projet CAPAMETRIM afin de caractériser les mouvements

FIGURE *2.4: Le système Physilog*

durant les crises d'épilepsie. Movea a développé également une montre MotionPod, couplée avec le logiciel Bioval, permettant une mesure fiable des amplitudes des articulations. Ce dispositif a été initialement utilisé dans le projet ACTIDOM (ACTImétrie à DOMicile) afin de surveiller l'activité des personnes âgées et fragiles dans leur domicile [Bonnet et al., 2004, Noury et al., 2004].

2.3.2 La reconnaissance sonore

Afin d'effectuer la reconnaissance automatique de la parole, des modèles acoustiques adaptés à la situation sont nécessaires. La voix des personnes âgées diffère généralement de la notre puisque elle a été modifiée avec le temps, en raison de la vieillesse ou des maladies. Des modèles acoustiques ont donc été construits pour cette population cible et testés afin de montrer qu'il est utile de réaliser ce type de modèles [Baba et al., 2004].

Fezari et Bousbia-Salah [Fezari and Bousbia-Salah, 2007] ont proposé une méthode de contrôle d'une chaise roulante - utilisée par les personnes handicapées ou bien âgées - via certaines commandes vocales. Ce type de commande utilise des moteurs de reconnaissance de parole existants qui sont adaptés pour cette utilisation. De son côté, l'École Nationale Supérieure des Télécommunications de Paris a travaillé sur un système possédant peu de commandes et s'adaptant au fur et à mesure à la voix de la personne [S. Renouard, 2003]. Ce système possède une classe de rejet.

Le laboratoire CLIPS10 [Istrate et al., 2006, Vacher et al., 2006] a proposé un système de reconnaissance des situations de détresse à l'intérieur de l'habitat. À partir de la reconnaissance de sons de la vie courante, le système proposé permet la reconnaissance des ADL et aussi la détection des mots-clés de détresse.

Medjahed et al, Virone et Istrate [Medjahed et al., 2008, Virone and Istrate, 2007] utilisent la reconnaissance sonore pour fournir une application de télévigilance médicale des personnes âgées ou fragiles au domicile.

De plus, afin de faciliter l'accès aux nouvelles technologies pour les personnes âgées, Ku-

miko *et al.* [KUMIKO, 2004] proposent un projet qui sert à aider les personnes âgées peu habituées à la saisie au clavier, dans l'utilisation de l'ordinateur. Cela s'effectue à l'aide de l'utilisation des commandes vocales. Des recherches ont également été menées sur la création d'un dialogue cohérent pour les personnes âgées [Takahashi et al., 2003]. Ce type de recherche trouve ses applications dans les robots de compagnie des personnes âgées.

2.3.3 La reconnaissance à partir de séquences vidéo

Toujours dans l'optique d'assistance aux personnes âgées, des solutions à base de reconnaissance d'images vidéo sont proposées par Microsoft dans son environnement EasyLiving [Krumm et al., 2000] et par le projet GERHOME [Zouba et al., 2008]. Ces solutions utilisent un ensemble de caméras disposées dans l'environnement à surveiller et permettent un suivi des personnes.

Zouba *et al.* [Zouba et al., 2008] ont proposé une approche de reconnaissance d'un ensemble d'activités de la vie quotidienne réalisées par des personnes âgées vivant seules à leur domicile à partir d'une séquence vidéo, afin d'être capable de détecter des événements reflétant l'état de la personne âgée. En effet, les auteurs ont modélisé trente-quatre événements vidéo, des événements simples comme « une personne qui est debout » et d'autres complexes tels que « une personne qui a des sentiments de faiblesse ». La méthode de reconnaissance des activités s'appuie sur la reconstruction 3D de la posture humaine.

Joumier *et al.* [Joumier et al., 2011] ont étudié la capacité d'un système de reconnaissance d'activité vidéo à détecter si l'activité est réalisée par des personnes âgées avec ou sans démence. En effet, un total de 28 volontaires (11 sujets âgés en bonne santé, 17 patients atteints de la maladie d'Alzheimer) ont participé à une expérimentation clinique. L'étude proposée a montré la possibilité de distinction entre les deux profils de participants en fonction des paramètres de l'activité motrice tels que la vitesse de marche calculée à partir du système de reconnaissance de l'activité vidéo automatique proposé. Au sein du projet « The AwareHome Research Initiative » [Abowd et al., 2002], une maison destinée aux personnes âgées est équipée de caméras vidéo pour la reconnaissance d'activités et aussi pour la détection de chute [Moore and Essa, 2002].

D'autres travaux se sont penchés sur la détection de chute, véritable problème de santé publique, en se basant sur des informations vidéo. [Lee and Mihailidis, 2005] et [Nait-Charif and McKenna, 2004] détectent la chute en mesurant la vitesse de l'ellipse qui approxime la silhouette. Les auteurs utilisent une caméra placée au plafond. Ainsi le champ de vision de la caméra est limité et il est donc très difficile de mesurer une vitesse verticale. Anderson *et al.* [D. Anderson, 2006] utilisent des silhouettes acquises de plusieurs caméras afin de détecter des évènements anormaux et, en particulier, la chute. Thome et Miguet [Thome et al., 2008] intègrent les résultats d'une classification de la position de la silhouette provenant de plusieurs caméras à l'aide d'une fusion basée sur la logique floue. Rougier et al. [Rougier et al., 2011] présentent une méthode qui utilise une classification par un modèle de mixture de gaussiennes appliqué au contour de la personne suivi dans le temps afin de détecter la chute. La détection

de la chute se fait par une méthode de vote sur le résultat de chacune des caméras. Lee et Mihailidis [Lee and Mihailidis, 2005] définissent d'abord des zones d'inactivité habituelle dans la pièce (telles que le lit, le canapé), puis analysent la silhouette de la personne et sa vitesse en utilisant des seuils caractéristiques selon que la zone correspond à une activité ou inactivité. Pelissier et al. [Pélissier, 2003] ont proposé un système de télésurveillance dédié à la prévention des chutes. Cette approche repose sur l'évaluation en continu de la signature de la marche afin d'y détecter d'éventuelles altérations, dans le but de déclencher une action préventive. Cette signature est caractérisée par un ensemble d'attributs pertinents calculés par l'observation, via une caméra vidéo, du comportement locomoteur d'une personne âgée à domicile.

2.3.4 Acceptabilité des technologies d'assistance

Nous avons présenté une analyse des différents types de technologies déjà utilisées pour l'assistance aux personnes âgées. Les systèmes basés sur l'utilisation de capteurs portables fournissent un diagnostic fiable et objectif. Toutefois, l'utilisation des capteurs portables ne présente pas une solution convenable puisque les personnes suivies, surtout les plus âgées, changent souvent d'habits et risquent de ne pas remettre les capteurs, par oubli ou par manque de volonté. Certaines personnes ne sont pas toujours suffisamment habillées pour porter les capteurs sur elles ; les malades d'Alzheimer, par exemple, se déshabillent souvent et abandonnent ainsi les capteurs [Burgio, 1997]. De plus, les sujets âgés sont très sensibles à tout changement, même petit, dans leur environnement de vie [Burgio et al., 2001] et se sentent gênés et encombrés par les câbles et les capteurs.

L'utilisation de systèmes sonores possède des avantages comme leur intrusivité limitée, leur coût faible et leur installation facile. Cependant, les conditions d'enregistrement du son à traiter peuvent être fortement perturbées. Le système doit, par exemple, reconnaître le mot ou le son dans un environnement caractérisé par des bruits ambiants forts.

Les systèmes d'assistance basés sur les séquences vidéo sont confortables puisque le sujet ne porte aucun capteur. Notons aussi que la vidéosurveillance est polyvalente car elle peut servir à la détection de plusieurs types d'activités (chute, prise de médicaments, absence prolongée, sommeil trop long, localisation dans une pièce, etc.). Cependant, la question du respect de l'intimité et de la vie privée de la personne âgée se pose lors de l'utilisation de cette technologie.

2.4 Objectif de la thèse

L'objectif de cette thèse est le développement d'un système de reconnaissance d'activités humaines à partir de séquences vidéo dans un contexte d'assistance aux personnes âgées ou en perte d'autonomie. Un tel système permettrait par exemple la détection de cas d'urgence comme la chute ou l'analyse précoce de changement de comportement. Pour mettre en œuvre une plateforme d'assistance (robot, etc) qui soit robuste, il est fondamental de doter la plate-

forme de capacités sensorielles naturelles : par exemple, le robot devrait être capable d'assurer une interaction avec l'humain par la voix, la vision, le toucher, etc. Alors qu'un nombre considérable de travaux sur la reconnaissance par la voix ont été effectués, les efforts dédiés à la reconnaissance d'actions par la vision ont été bien moins nombreux et restreints à des tâches limitées. Ce décalage peut être expliqué par les capacités de processeurs de la génération précédente qui étaient plus adaptées à des données de type voix, bien moins gourmandes que les images et surtout les séquences vidéo. Les algorithmes d'apprentissage statistiques étaient (et sont toujours) aussi bien plus matures pour la reconnaissance de parole que pour la reconnaissance d'objets et de mouvements humains dans des scènes complexes. La voix et la vision, toutefois, sont conjointement nécessaires en raison de leur complémentarité avérée : L'exploitation de la voix est moins robuste en présence de bruits sonores environnants provenant de la télé ou d'autres machines. La vidéo peut complémenter la voix de manière efficace dans de telles situations. La reconnaissance par la vision permet également de détecter des évènements qui ne font pas intervenir le son ou la voix, par exemple une personne allongée sur un canapé pour une longue durée ou une personne en sommeil. D'un autre coté, la reconnaissance par la vision, permet d'autres applications importantes telles que la communication par le geste pour des personnes ayant des troubles de la parole et du langage, ou l'interaction homme-robot pour la stimulation d'exercices physiques, la réhabilitation et le divertissement. Il faut également noter que l'acquisition par la vision a l'avantage de ne pas exiger des utilisateurs une coopération stricte contrairement par exemple à des capteurs embarqués sur la personne. Ces derniers peuvent être oubliés ce qui peut limiter leur efficacité. Toutefois, pour une meilleure acceptabilité de la vidéo, des mécanismes assurant l'anonymisation des utilisateurs devront être implémentés.

Dans la section suivante, nous présentons l'état de l'art sur la reconnaissance des activités à partir de séquences vidéo, en faisant une catégorisation des différentes approches proposées selon les primitives et les stratégies de classification. Après une synthèse de ces approches, nous présenterons les grandes lignes de l'approche proposée.

2.5 Approches proposées pour la reconnaissance des actions

La reconnaissance des actions se compose généralement d'une étape d'extraction des primitives et d'une étape de classification. L'extraction des primitives consiste à identifier des caractéristiques distinctives à partir d'une séquence vidéo tout en étant robuste au bruit. Dans l'étape de classification, on s'intéresse à la possibilité d'identifier les actions à l'aide des méthodes d'apprentissage automatique, tout en prenant en compte la grande variabilité à laquelle une classe d'action peut s'éxposer, en particulier si elle est exécutée par différents sujets de genre et de taille différents, et avec une vitesse et une manière différentes. Dans cette section, on va exposer les principales approches d'extraction des primitives et les stratégies de classification

utilisées dans l'état de l'art.

2.5.1 Catégorisation des approches selon les primitives

L'une des tâches principales pour la reconnaissance des actions à partir de séquences vidéo est l'extraction des primitives qui caractérise le mouvement dans ces séquences. Généralement, on peut distinguer trois types d'approches d'extraction de primitives : **Approches basées sur la reconstruction du modèle du corps humain**, **Approches holistiques** basées sur la dynamique globale du corps humain et, enfin, les **Approches locales** qui caractérisent l'action à partir des points d'intérêt.

2.5.1.1 Approches basées sur un modèle du corps humain

Cette catégorie d'approches se base sur une étude psychophysique sur l'interprétation visuelle du mouvement biologique [Johansson, 1973] qui montre que les humains sont capables de reconnaître les actions uniquement à partir de quelques points de repère en mouvement «Moving light displays » attachés au corps humain (Figure 2.5). Plusieurs travaux de reconnaissance des actions se sont inspirés de cette étude et ont utilisé une représentation similaire basée sur les trajectoires des points repères sur le corps humain, tels que la tête, les mains, les jambes, etc. Cette étude a été également à l'origine des travaux qui se basent sur des modèles de mouvement du corps.

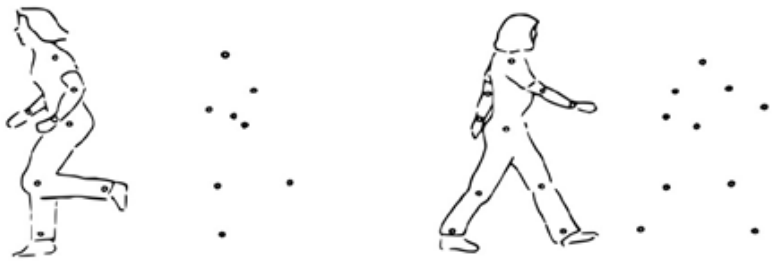
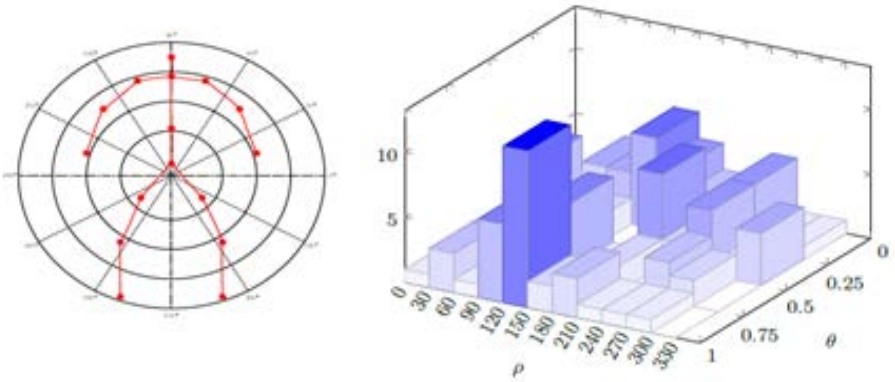


FIGURE *2.5: Illustration du mouvement des points repères [Johansson, 1973]*

Une reconstitution du corps humain peut être réalisée à partir d'un modèle cinématique des articulations de corps humain. Ainsi, la reconnaissance est déterminée par l'évolution et la trajectoire des articulations. La difficulté majeure de cette modélisation est le grand nombre de degrés de liberté possibles des différentes articulations du corps humain et les grandes variabilités des formes de corps humain. Yacoob et Black [Yacoob and Black, 1998] ont proposé une approche utilisant une segmentation du corps pour en extraire les articulations principales, ainsi qu'une réduction de dimension par analyse en composantes principales (ACP). Chacune des activités à reconnaître est modélisée par un exemplaire. La reconnaissance se fait par un

alignement temporel des parties du corps. Han *et al.* , [Han et al., 2010] ont exploité la capture de mouvements afin de représenter les trajectoires de chacune des articulations dans un espace manifold. Leur méthode nécessite de grandes bases d'actions pour l'apprentissage, car elle utilise une simplification hiérarchique du squelette d'animation, au lieu de l'intégralité de celui-ci. Dès lors, des ambiguïtés peuvent apparaître entre des mouvements ainsi représentés. Un processus complexe et coûteux en termes de temps de calcul est nécessaire à la bonne classification des actions similaires. Fujiyoshi et Lipton [Fujiyoshi and Lipton, 1998] ont extrait le contour de la personne pour en calculer un squelette simplifié. Ils introduisent ainsi le concept de star skeletonization ou star skeleton. Il s'agit d'une représentation du corps humain par l'intermédiaire d'un squelette à cinq branches au maximum, représentant les extrémités des jambes et des bras, ainsi que la tête. Ali *et al.* [Ali et al., 2007] ont utilisé la théorie des systèmes chaotiques et la recherche d'invariants dans cet espace dynamique pour catégoriser des actions par l'intermédiaire de l'algorithme k-plus proche voisins (kNN). Bien que cette approche soit efficace pour reconnaître des actions issues d'images, elle est tributaire de la qualité de l'extraction faite, notamment pour l'extraction 2D du squelette. Ziaeefard et Ebrahimnezhad [Ziaeefard and Ebrahimnezhad, 2010] ont proposé le calcul d'un histogramme polaire des positions des articulations au cours du temps. Il s'agit d'un histogramme dont les coordonnées s'expriment en (ρ ; θ), comme présenté sur la figure 2.6.



(a) Calcul de l'histogramme polaire pour un squelette d'animation. (b) Exemple de représentation de l'histogramme polaire d'une action.

FIGURE *2.6: Histogramme polaire du squelette d'animation [Ziaeefard and Ebrahimnezhad, 2010]*

Sheikh *et al.* [Sheikh et al., 2005] ont appliqué une projection affine aux trajectoires des articulations afin de reconnaître des actions, de façon invariante à l'angle de vue, par la mesure des angles entre les articulations dans l'espace projeté. Dans des travaux plus récents, Xia *et al.* [Xia et al., 2012] ont proposé des histogrammes des positions 3D des articulations (HOJ3D) qui encodent principalement l'occupation spatiale des articulations par rapport au centre de la silhouette (hanche). En effet, les articulations sont projetées dans un espace sphérique partitionné en n-bins (Figure 2.7). Ensuite, une quantification vectorielle est réalisée à l'aide de

k-means pour construire les vecteurs de primitives. Un modèle de Markov caché est utilisé pour la classification d'actions.

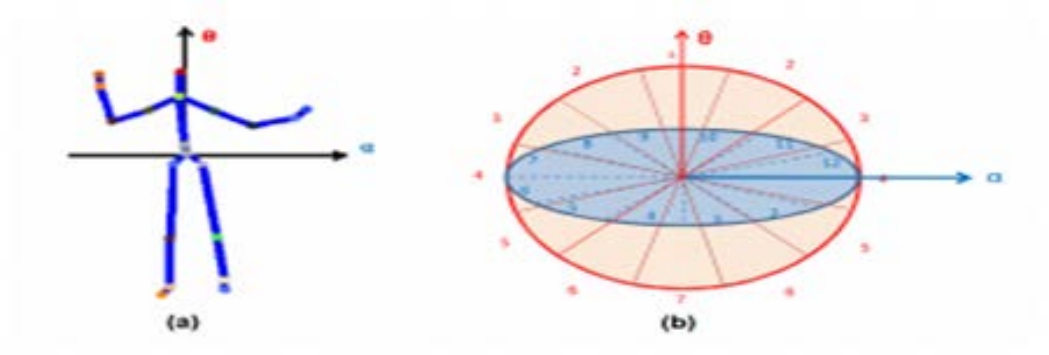


FIGURE *2.7: (a) les coordonnées de référence de HOJ3D. (b) système de coordonnées sphérique [Xia et al., 2012]*

D'autres travaux utilisent, pour leur modélisation du corps humain, des formes géométriques de trois dimensions [Gavrila and Davis, 1995, Ke et al., 2005, Brand, 1997]. En effet, Marr et Nishihara [Marr and Nishihara, 1978] ont proposé une représentation théorique du modèle humain. Cette représentation est basée sur des cylindres généralisés pour modéliser les formes articulées réelles où chaque partie de l'objet peut, à son tour, être décomposée en un ensemble de sous-parties, de façon hiérarchique. S'inspirant de ce modèle, Gavrila et Davis [Gavrila and Davis, 1995] ont proposé une approche combinant la capture de mouvements 3D, issus de plusieurs caméras, afin d'avoir un squelette simplifié. L'algorithme d'alignement de séquences temporelles, Dynamic Time Warping, est utilisé pour reconnaître des actions humaines. Brand et Kettnaker [Brand and Kettnaker, 2000] utilisent des Modèles de Markov Cachés pour estimer l'orientation 3D du corps humain à partir de silhouettes en basse résolution. Huang et Trivedi [Huang and Trivedi, 2005] ont présenté le concept d'histogramme cylindrique, basé sur une représentation voxelique, pour effectuer une reconnaissance par des chaînes de Markov cachées. Une méthode similaire a été proposée par Parameswaran et Chellappa [Parameswaran and Chellappa, 2006] pour traiter des données de capture de mouvements avec des marqueurs 3D dans un espace de mouvement invariant aux actions, pour reconnaître chacune d'entre elles. Les auteurs soulignent qu'il n'y a pas d'invariance 3D dans l'espace de mouvement. Les auteurs ont montré l'importance de la capture de mouvements 3D du corps humain dans le contexte de la reconnaissance d'actions.
En outre, Ke *et al.* [Ke et al., 2010] proposent une méthode de reconstruction de pose 3D en temps réel (Figure 2.8). L'apparence, la couleur et l'information temporelle sont conjointement utilisées pour effectuer un suivi des parties du corps humain en 2D. Ensuite, l'algorithme d'optimisation « downhill simplex » est utilisé pour faire l'alignement entre les primitives 2D suivies et des modèles humains 3D prédéfinis. Cette méthode est invariante par rapport aux angles de vue qui est l'un des défis les plus importants dans la reconnaissance d'actions. Cependant, elle est sujette aux erreurs précoces dans l'étape de suivi. Ramanan et Forsyth [Ramanan and Forsyth,

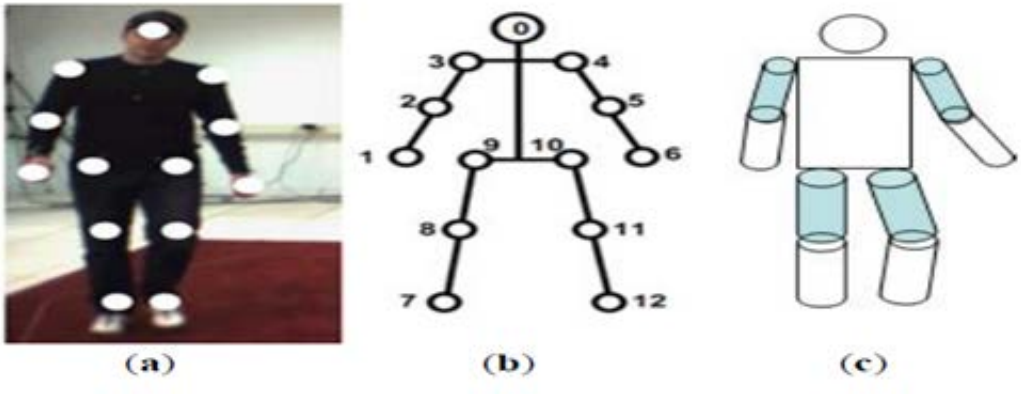


FIGURE *2.8: Modèles du corps humain (a) l'image originale de corps, (b) modèle cinématique basé sur 13 articulations, (c) le modèle 3D [Ke et al., 2010].*

2003] ont aussi utilisé un apprentissage de capture de mouvements 3D et une reconnaissance de mouvements se basant sur une capture de mouvements 2D. Une phase d'annotation manuelle des données 3D est nécessaire mais cela permet ensuite de reconnaître les actions avec une acquisition vidéo 2D. Une modélisation par chaînes de Markov et le classifieur de maximum a posteriori (MAP) sont utilisés pour la reconnaissance d'actions.

2.5.1.2 Approches holistiques

Les approches holistiques utilisent la structure et la dynamique globale du corps pour représenter les actions humaines. L'idée clé est que, étant donnée une région d'intérêt centrée sur le corps humain, une dynamique globale est assez discriminante pour pouvoir caractériser les actions humaines. Par rapport à des approches qui utilisent explicitement un modèle du corps ou des informations sur les parties du corps, les représentations holistiques sont beaucoup plus simples car elles modélisent seulement le mouvement et l'apparence globale du corps humain. Donc, généralement, ces primitives sont calculées d'une manière plus robuste et efficace. Ces approches peuvent être classifiées en deux sous-catégories : celles qui sont basées sur la silhouette et celles qui se basent sur le flux optique ou bien le gradient.

- **Méthodes basées sur la silhouette**

L'une des utilisations les plus anciennes de la silhouette remonte à Yamato et al. [Yamato et al., 1992]. Leur approche porte sur la classification de mouvements de tennis en s'appuyant sur des techniques de quantification vectorielle des séquences d'images binaires. Cette quantification est réalisée en divisant la région d'intérêt correspondant à la silhouette en une grille où chaque cellule est représentée par le ratio des nombres de pixels noirs et blancs. Ce vecteur de primitives est utilisé comme l'entrée des modèles de Markov à états cachés, Hidden Markov Models (HMM).

En outre, Bobick et Davis [Bobick and Davis, 2001] ont proposé le concept de « Motifs Spatio-temporels » (Spatio-Temporal Template). Les auteurs procèdent à l'extraction des silhouettes à partir d'une seule vue et, ensuite , ils regroupent les différences entre les images d'une séquence représentant une action. Un volume spatio-temporel est ainsi construit, constituant l'Image d'Energie du Mouvement (MEI : Motion Energy Image) et aussi l'Image d'Historique du Mouvement (MHI : Motion History Image) dont l'intensité des pixels représente une trace des mouvements de la silhouette (Figure 2.9(a)). Cette modélisation simple permet de reconnaître des actions comme le fait de s'asseoir, agiter les bras ou s'accroupir en temps réel. Un inconvénient inné à cette méthodologie est l'auto-occultation du mouvement lorsque des mouvements sont effectués sur la même zone.

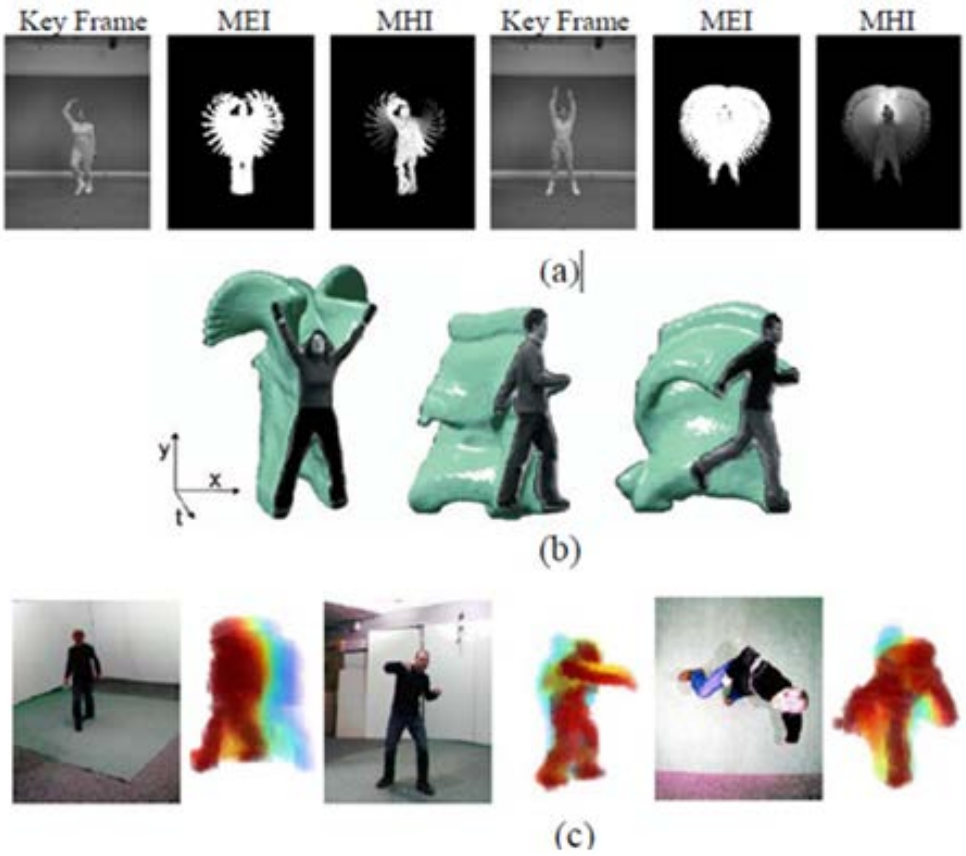


FIGURE *2.9: Illustration de modèle holistique (a) Images d'Energie du Mouvement (MEI) et images de l'histoirique du mouvement (MHI) ; (b) volume spatio-temporel ; (c) volume de l'histoire du mouvement (MHV)*

Au lieu de modéliser l'historique de mouvement dans une seule image, Blank *et al.*, [Blank et al., 2005, Gorelick et al., 2007] empilent d'abord les silhouettes à partir d'une séquence donnée pour former un volume spatio-temporel (Figure 2.9 (b)). Puis ils appliquent l'équation de Poisson pour déduire les caractéristiques de saillance et d'orientation d'un pixel par rapport à

son voisinage. Des descripteurs globaux sont ainsi obtenus pour un intervalle temporel donné en additionnant les moments relatifs aux descripteurs locaux. Dès lors, une classification par kNN leur permet de reconnaître les actions.

Weinland *et al.* [Weinland et al., 2006] combinent les silhouettes issues de plusieurs caméras pour construire une représentation basée sur des voxels (pixels en 3D). Ils utilisent le Volume d'Historique du Mouvement (MHV : Motion History Volume) qui est une extension de MHI en 3D (Figure 2.9 (c)). Cette représentation est assez informative mais elle nécessite l'extraction de silhouette et aussi un réseau de caméras calibrées. Ragheb *et al.* [Ragheb et al., 2008] ont proposé de transformer le MHV dans domaine de Fourier. Les auteurs commencent par construire les MHVs et, ensuite, ce volume est divisé en sous-volumes. Le vecteur de primitives correspond à la moyenne des fréquences obtenues. La classification se base sur un appariement (matching) utilisant une distance euclidienne pondérée. La méthode s'est montrée robuste par rapport au changement de vue et aussi au bruit.

- **Méthodes basées sur le calcul de flux optique et de gradient**

La seconde catégorie des approches holistiques utilise le flux optique calculé à partir d'images consécutives. Cette catégorie ne nécessite pas de soustraction de fonds qui est généralement une étape difficile. Parmi les travaux qui ont utilisé le flux optique, Efros *et al.* [Efros et al., 2003] ont proposé de diviser le flux optique en quatre champs scalaires différents (correspondant à la composante négative et positive, horizontale et verticale de flux), voir Figure 2.10 (a), et de les utiliser séparément pour l'appariement. Cette représentation a également été utilisée dans [Robertson and Reid, 2005, Wang et al., 2007]. Zhang *et al.* [Zhang et al., 2008] calculent des masques de forme en s'appuyant sur la détection des zones contenant un mouvement significatif. Ensuite, un descripteur de contexte est calculé dans des régions constantes de mouvement en utilisant une grille polaire (Figure 2.10 (b)). Chaque cellule de la grille est décrite avec un histogramme sur les descripteurs SIFTs quantifiés. Le descripteur final pour une séquence vidéo correspond à la somme des descripteurs des différentes cellules. La classification est effectuée en utilisant un SVM ainsi que des modèles pour l'Analyse Sémantique Latente Probabiliste (ASLP).

L'un des inconvénients majeurs des approches basées sur le flux optique est qu'elles s'appuient sur l'hypothèse selon laquelle les différences d'images peuvent être expliquées comme une conséquence d'un mouvement, alors qu'elles peuvent correspondre aux changements dans les propriétés des objets, de fonds, d'éclairage, etc. De ce fait, des approches basées sur des gradients ont été proposées. Ce type d'approche partage de nombreuses caractéristiques avec celles qui sont basées sur le flux optique. En particulier, elles ne dépendent pas de la soustraction du fond, mais elles sont aussi sensibles aux propriétés du matériau, à la texture, à l'éclairage, etc. [Klaser et al., 2008]. À la différence du flux optique, les gradients sont discriminatifs à la fois pour les régions qui sont en mouvement et celles qui ne le sont pas, ce qui est avantageux dans certaines situations mais pas dans d'autres.

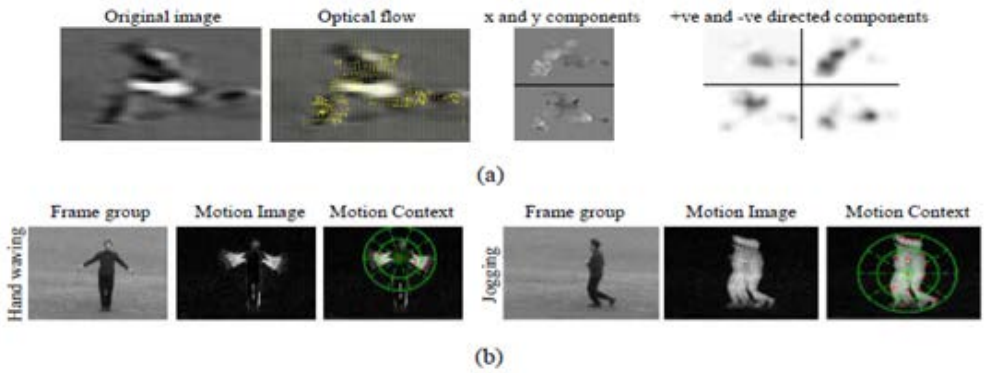


FIGURE *2.10: Illustration de méthodes basées sur le flux optique. (a) descripteur de mouvement en divisant le flux optique en quatre champs scalaires différents ; (b) descripteur de contexte calculé en utilisant une grille polaire.*

L'histogramme de gradients orientés est parmi les représentations les plus populaires basées sur le gradient. Ce descripteur a été initialement appliqué avec succès pour la détection de personnes et d'objets [Dalal and Triggs, 2005]. [Lu and Little, 2006] ont proposé une méthode de suivi et de reconnaissance des actions basée sur le descripteur histogrammes de gradients orientés (HOG). Ce descripteur est calculé globalement pour chacune des trames. L'ACP est ensuite appliquée pour réduire la dimension du descripteur. Un HMM est utilisé pour modéliser les actions telles que la course, le patinage, etc. Thurau et Hlavac [Thurau and Hlavác, 2010] ont étendu le descripteur HOG. En effet, au lieu de calculer un seul histogramme par trame, les auteurs divisent chaque trame en des grilles qui se chevauchent. Un descripteur HOG est calculé dans chacune des cellules. La reconnaissance des actions se fait par la méthode du plus proche voisin.

2.5.1.3 Approches basées sur des caractéristiques locales

Les approches appartenant à cette catégorie décrivent les observations comme une série de descripteurs locaux ou patchs. Il n'est pas nécessaire de procéder à une étape de prétraitement, par exemple la suppression de fonds et le suivi de silhouette. Ceci évite ainsi la propagation des erreurs produites en phase de prétraitement, surtout dans le cas de scènes réelles ayant des fonds dynamiques. Ces primitives sont généralement robustes, voire invariantes aux changements d'angle, à l'apparence des personnes et aux occlusions partielles. Les premiers travaux utilisant ce type de représentation pour la reconnaissance d'actions humaines ont été motivés par le succès des points d'intérêt pour la reconnaissance d'objets. Ces points correspondent à un ensemble de pixels présentant une singularité que ce soit au niveau du gradient ou du contour.

- **Détecteurs de points d'intérêt spatio-temporels**

Parmi les premiers travaux proposés pour extraire les points d'intérêts spatio-temporels (STIPs) figure celui de Laptev et Lindeberg [Laptev and Lindeberg, 2003]. Les auteurs ont étendu le détecteur de coins de Harris [Harris and Stephens, 1988] en lui rajoutant la dimension temporelle. Ce détecteur spatio-temporel, communément appelé Harris 3D, leur permet d'extraire des motifs de mouvement. Ces points d'intérêt spatio-temporels correspondent aux points dont le voisinage local est soumis à une variation spatiale et temporelle significative. L'échelle spatiale et temporelle du voisinage est automatiquement sélectionnée. Ce travail a été amélioré par Laptev *et al.* [Laptev et al., 2007] pour compenser les mouvements relatifs aux caméras.
Cependant, selon Dollar *et al.* [Dollar et al., 2005], le nombre des points d'intérêt vérifiant le critère Harris3D est relativement faible par rapport aux zones contenant des mouvements significatifs. Ainsi, les auteurs ont proposé des nouveaux points d'intérêt spatio-temporels plus denses. Ils appliquent un filtre de Gabor à une dimension au niveau spatial et temporel séparément. Le nombre de points d'intérêt est ajusté en changeant la dimension spatiale et temporelle du voisinage dans lequel les minimas locaux sont sélectionnés. Bregonzio *et al.* [Bregonzio et al., 2009] ont étendu cette approche en appliquant un filtre de Gabor 2D au volume spatio-temporel composé par la différence entre les trames adjacentes.
Willems *et al.* [Willems et al., 2008] ont proposé une extension du détecteur Hessian au domaine spatio-temporel. Le principe de cette méthode est de détecter des points d'intérêt denses, invariants à l'échelle avec un algorithme robuste et efficace d'un point de vue calculatoire. Zhen et Shao [Zhen and Shao, 2013] ont introduit un détecteur compact décomposant une séquence vidéo avec une pyramide de Laplace, dans lequel les caractéristiques spatio-temporelles de différentes tailles sont localisées sur plusieurs échelles et orientations. Wang *et al.* , [Wang et al., 2011a] ont proposé les points d'intérêt denses. Ces points sont régulièrement échantillonnés dans l'espace et le temps.

- **Descripteurs locaux**

Pour caractériser les points d'intérêt, des descripteurs locaux sont calculés par rapport au volume spatio-temporel extrait autour d'un point d'intérêt. Il s'agit de capturer l'information portée par un pixel dans l'espace spatio-temporel. La dimension spatiale et temporelle du patch est généralement déterminée par l'échelle du point d'intérêt. Laptev *et al.* [Laptev et al., 2007] ont été les pionniers de ce type de description pour les séquences vidéo. Les auteurs ont proposé les histogrammes de flux optique (HOF), les HOGs. Les auteurs divisent une région cuboïde de taille $M \times M \times M$ en une grille de cellules. Pour chaque cellule, on calcule des histogrammes de 4 types d'orientations du gradient (HOG) et des histogrammes de 5 types d'orientations de flux optique (HOF). Dans [Dollar et al., 2005] les gradients calculés pour chaque pixel dans une région cuboïde sont concaténés en un seul vecteur. On utilise ensuite l'ACP pour projeter les vecteurs sur un espace de dimension plus faible.
Kläser *et al.* [Klaser et al., 2008] ont étendu HOG au domaine 3D. Les auteurs utilisent une représentation de vidéo intégrale, extension de l'idée d'image intégrale en 3D, pour calculer

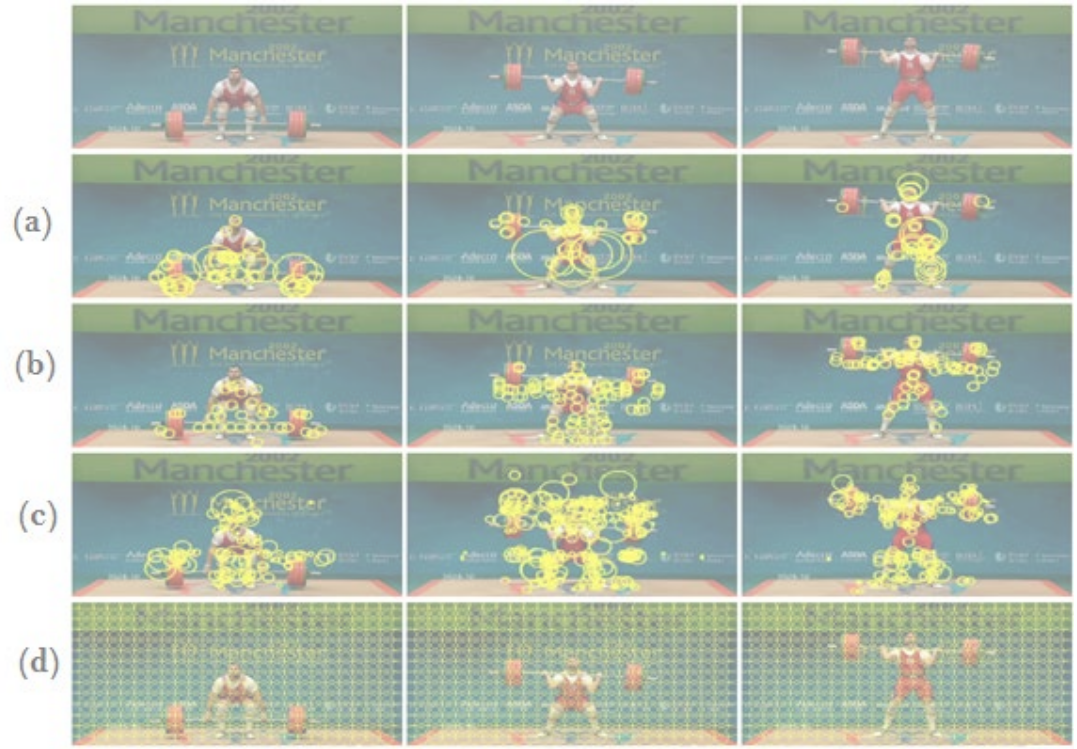


FIGURE *2.11: Visualisation de points d'intérêt détectés par différents détecteurs sur des trames successives d'une séquence vidéo. Harris3D (ligne a), Gabor (ligne b), Hessian3D (ligne c) et l'échantillonnage dense (ligne d) [Ullah, 2012].*

de façon rapide et efficace les gradients 3D à plusieurs échelles spatiales et temporelles. Des polyèdres réguliers sont ensuite utilisés pour quantifier les orientations 3D du gradient de façon uniforme.

Le descripteur "Speeded Up Robust Features" (SURF) [Bay et al., 2006] a été aussi adapté à la représentation 3D par Willems *et al.* [Willems et al., 2008] sous le nom de eSURF (SURF étentdu). Les cuboïdes 3D sont divisés en une grille de cellules. Chaque cellule est représentée par une somme pondérée de réponses d'ondelettes de Haar uniformément échantillonnées. Scovanner *et al.* [Scovanner et al., 2007] ont créé une version 3D de l'algorithme SIFT de Lowe [Lowe, 1999] pour exploiter le volume spatio-temporel. Plus récemment Wang et al. [Wang et al., 2011a] ont proposé les histogrammes de frontière de mouvement « Motion Boundary Histograms» (MBH) calculés autour des trajectoires. MBH représente donc le gradient de flux optique : il s'agit de calculer la dérivée des composantes horizontale et verticale de flux optique séparément (Figure 2.12). Ainsi, ce descripteur permet d'éliminer le mouvement constant tel que le mouvement constant de caméra. D'autre part, MBH met en valeur le changement de flux optique (correspondant au changement au sein de mouvement), d'où son nom "Frontière de mouvement". Ce descripteur est très discriminant et combiné avec les vecteurs de déplacement, HOG et HOF, il a donné des taux de reconnaissance relevant de l'état de l'art.

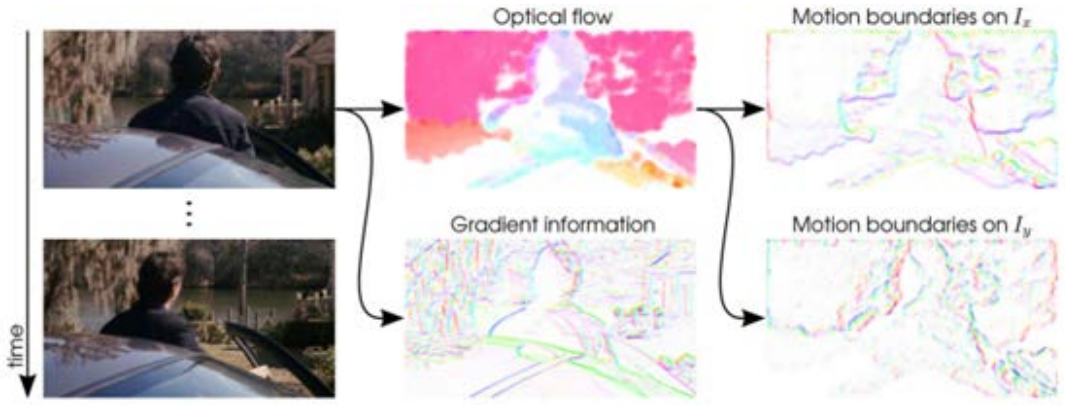


FIGURE *2.12: Histogramme de gradient (HOG), de flux optique (HOF) et frontière de mouvement (MBH) [Wang et al., 2011a].*

2.5.1.4 Synthèse

La reconstruction de modèle humain donne une estimation précise de la posture afin de reconnaître des actions humaines. Elle fournit une description riche et des primitives de haut niveau. Cependant, il n'est pas rare que certains membres du corps soient occultés et ainsi ne soient pas suivis correctement. De plus, si plusieurs personnes sont présentes dans la scène la reconstruction du modèle humain pour chacune des personnes devient beaucoup plus difficile.

Contrairement à cette approche, les méthodes holistiques ne nécessitent pas la localisation de parties du corps. La structure globale du corps humain et sa dynamique sont utilisées pour représenter les actions humaines. Les représentations holistiques sont plus simples et plus robustes que les approches qui utilisent explicitement un modèle du corps. Donc, généralement, ces primitives sont calculées d'une manière plus efficace. Cependant, les représentations holistiques ne sont pas généralement invariantes aux angles de vue de caméra (direction). Cela doit être pris en compte, soit par l'apprentissage de plusieurs modèles pour les différents angles de vues (frontaux, latéraux, arrières), ou bien en fournissant une quantité suffisamment importante de données d'apprentissage. De plus, les corps humains peuvent apparaître à différentes échelles (vue lointaine, vue rapprochée) de telle sorte que certaines parties du corps peuvent ne pas être visibles dans l'image. Les points d'intérêts spatio-temporels permettent de surmonter ces problèmes puisqu'ils sont invariants aux changements d'angle, à l'apparence des personnes et aux occlusions partielles. De plus, ils sont plus adaptés aux scènes réelles puisqu'il n'est pas nécessaire de procéder à une étape de prétraitement, par exemple la suppression de fond et le suivi de silhouette. Ceci évite la propagation des erreurs produites dans cette étape. Néanmoins, lorsque l'on s'intéresse à l'analyse de l'activité d'un être humain, l'identification de l'acteur peut être utile. Cela permet, d'une part de réduire la complexité de calcul en se limitant à la région dans laquelle se trouve la personne et, d'autre part, d'éviter le traitement des points d'intérêts qui sont détectés sur des fonds dynamiques. Dans le contexte de la reconnaissance des activités dans des conditions réelles, les occlusions des parties du corps de la

personne constituent un problème récurrent pour la reconnaissance. Par conséquent, l'utilisation des approches holistiques ou bien des approches basées sur le modèle humain est peu robuste dans les conditions réelles. De ce fait, nous considérons la représentation par points d'intérêt qui s'est révélée comme la plus prometteuse pour la description de mouvements humains. Ces primitives ont donné les meilleurs taux de reconnaissance dans la littérature sur des bases difficiles contenant des scènes réelles.

2.5.2 Stratégies de classification

Dans cette section, nous allons présenter les approches utilisées pour la classification des actions. On peut distinguer trois types d'approches : *les approches de classification globales* qui utilisent des méthodes de classification statique comme le kNN ou SVM, se basant sur des représentations de la séquence entière sans considération de l'aspect temporel des séquences vidéos ; *les approches de classification séquentielles* qui modélisent explicitement l'évolution temporelle des actions, généralement en utilisant des modèles stochastiques. Il est également possible d'utiliser des *approches de classification par trames clés* qui utilisent un sous-ensemble des trames de la séquence.

2.5.2.1 Classification globale

Les approches de classification globales ne prêtent pas une attention au domaine temporel. En effet, pour classifier les actions, un représentant unique et global par séquence vidéo (un seul vecteur de primitives) est construit. Ainsi l'ordre temporel des trames est négligé. Cette représentation permet l'utilisation des classifieurs discriminants tels que SVM, Adaboost et les réseaux de neurones. Cette méthode de classification a été appliquée sur les primitives holistiques et aussi sur les primitives locales. La section suivante présente les méthodes de classification globales appliquées aux primitives holistiques et celles extraites après la reconstruction de modèle humain. Ensuite, nous présenterons les méthodes globales utilisées avec les primitives locales.

- **Classification globale basée sur des primitives holistiques et des modèles humains**

Pour avoir un vecteur de primitives de taille fixe pour des séquences de longueurs différentes, les approches globales représentent la dynamique des séquences par l'extraction de caractéristiques holistiques, à partir d'un volume spatio-temporel composé d'images empilées d'une séquence vidéo ou bien à partir des modèles de corps humain reconstruits.
Par exemple, Bobick et Davis regroupent les différences entre les régions d'intérêts (Region of Interest : ROI) d'une séquence en une seule image afin de construire le MHI et ensuite extraient les moments de Hue 2D à partir de cette présentation. Blank *et al.* [Blank et al., 2005], Gorelick *et al.* [Gorelick et al., 2007] et Yilmaz et Shah [Yilmaz, 2005] construisent des volumes spatio-temporels 3D en empilant les ROI correspondant aux silhouettes sous forme d'une seule représentation volumétrique, et ensuite ils extraient un ensemble de caractéristiques locales

et globales de ce volume pour représenter les actions. Par exemple, Gorelick *et al.* [Gorelick et al., 2007] calculent les propriétés de la solution de l'équation de Poisson à partir du volume spatio-temporel des silhouettes pour la caractérisation de l'action.
En obtenant un vecteur de primitives de taille fixe, des classifieurs statistiques discriminant peuvent être utilisés. En effet, la classification par kNN a été largement utilisée pour la reconnaissance des actions [Blank et al., 2005, Batra et al., 2008, Sullivan and Carlsson, 2002, Lin et al., 2009, Weinland and Boyer, 2008]. Il s'agit d'un classifieur non paramétrique où une nouvelle observation est classée dans la classe d'appartenance de l'échantillon d'apprentissage qui lui est la plus proche, au regard des covariables utilisées. La détermination de leur similarité est basée sur des mesures de distance. Il s'agit d'une approche simple, qui a néanmoins donné de bonnes performances. La classification par kNN a été mise en œuvre avec une distance euclidienne par Batra *et al.* [Batra et al., 2008], Blank *et al.* [Blank et al., 2005]. Bobick et Davis [Bobick and Davis, 2001] utilisent la distance de Mahalanobis entre les moments de Hue 2D pour comparer deux actions.
D'autres travaux [Ahmad and Lee, 2008, Ziaeefard and Ebrahimnezhad, 2010, Zhang et al., 2008] utilisent des classifieurs discriminants tels que des SVMs. Par exemple, Ziaeefard et Ebrahimnezhad [Ziaeefard and Ebrahimnezhad, 2010] ont appliqué le classifieur SVM aux histogrammes polaires des positions des articulations au cours du temps. ZHANG *et al.* [Zhang et al., 2008] proposent d'utiliser SVM pour classifier les actions en se basant sur les descripteurs SIFTs quantifiés. Shao et Chen [Shao and Chen, 2010] ont proposé une reconnaissance utilisant directement les silhouettes comme des vecteurs caractéristiques d'une action. Ils construisent un sac de mots visuels à l'aide de l'algorithme k-moyenne. Chaque silhouette est projetée dans un espace de dimension inférieure, à l'aide d'une régression spectrale, puis un histogramme des occurrences de chacun des mots visuels est construit par association au voisin le plus proche.

- **Classification globale basée sur des primitives locales**

Les points d'intérêt se sont dégagés comme les primitives les plus prometteuses pour la description du mouvement humain. De ce fait, la plupart des travaux de recherche récents se sont focalisés sur ces primitives. Étant donné que le nombre de STIPs extrait dans chacune des vidéos est variable, la technique du sac de mots a été largement utilisée pour construire un vecteur de primitives de taille fixe. Cette technique a été initialement proposée dans le domaine de la recherche documentaire, où chaque document est décrit par la fréquence d'apparition de ses mots (BOW : Bag of Words). Par analogie, la séquence de mouvement est représentée par la fréquence de ses mots visuels. Les principales étapes de classification par sac de mots (Figure 2.13) sont les suivantes :

(a) **Extraction de primitives :** La première étape consiste à calculer des représentations locales qui décrivent les observations. Pour calculer les représentations locales, les points d'intérêt spatio-temporels sont d'abord détectés (exemple, Harris-3D, Hessian-3D), puis

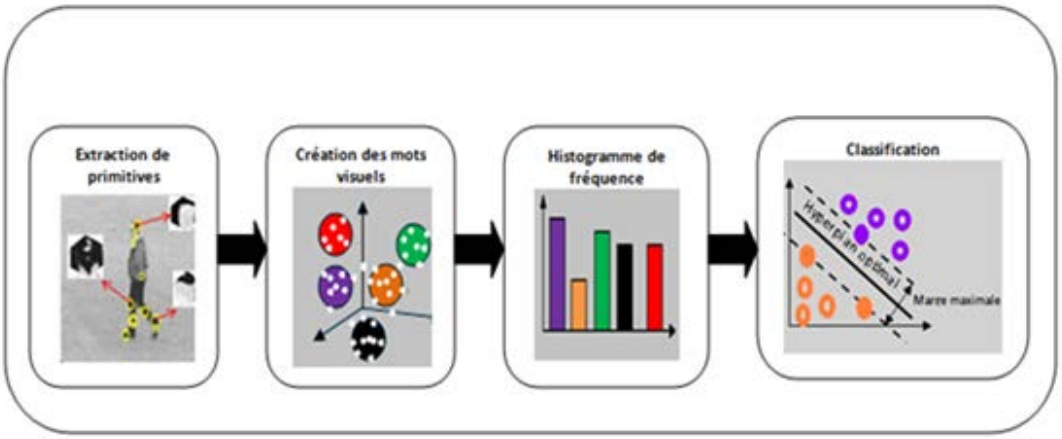


FIGURE *2.13: Principales étapes de classification par sac de mots.*

des descripteurs locaux comme le HOG/HOF sont calculés dans un patch autour de ces points.

(b) **Construction de vocabulaire visuel :** Dans cette étape, on s'intéresse à la construction d'un vocabulaire visuel pour modéliser au mieux les descripteurs locaux issus des séquences vidéo. L'enjeu le plus important est la recherche d'un vocabulaire donnant une représentation compacte et discriminante des actions présentes dans les vidéos. L'idée principale est de partitionner l'espace de descripteurs locaux en un ensemble de régions informatives ; chacune des régions est appelée mot visuel. L'algorithme des k-moyennes (k-means) est l'algorithme de partitionnement de données le plus utilisé dans la littérature pour construire le vocabulaire visuel. Il permet de créer le dictionnaire visuel de façon non-supervisée. Étant donné un ensemble d'observations $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, où chaque observation est un vecteur de dimension d, k-means vise à partitionner les n observations dans k ensembles $S = \{S_1, S_2, \ldots, S_k\} (k \leq n)$ afin de minimiser la distance entre les points à l'intérieur de chaque partition. L'optimum S^* est donc :

$$S^* = \arg\min_{S} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \left\| \mathbf{x}_j - \mu_i \right\| \quad (2.1)$$

où μ_i est la moyenne des points dans S_i.

(c) **Histogramme de mots visuels :** Une quantification des primitives locales est effectuée en affectant chacun des descripteurs au mot visuel le plus proche. Les séquences sont représentées en tant qu'histogrammes de fréquence de chaque mot visuel (sac de mots).

(d) **Classification :** Pour classer les histogrammes d'occurrences de mots visuels, les SVMs ont été largement utilisées avec la technique de sac de mots. En effet, SVM est une méthode de classification supervisée discriminative. Intuitivement, les SVM sont basées

sur deux principes. Le premier est de maximiser la marge du classifieur, c'est-à-dire la distance entre la frontière de décision et les échantillons les plus proches. Le deuxième est l'utilisation d'une fonction noyau (kernel trick). Dans ce cas, les caractéristiques sont projetées dans un autre espace en utilisant la fonction noyau. Dans cet espace, si le noyau est bien choisi, les caractéristiques deviennent linéairement séparables, et il est possible d'appliquer la classification par SVM afin de trouver l'hyperplan optimal de séparation. Les SVM présentent de bonnes propriétés de généralisation. En effet, ils sont capables de construire des modèles sans sur-apprentissage même dans le contexte de peu d'exemples d'apprentissage représentés par des vecteurs de grande dimension. L'algorithme est decrit en détail dans l'annexe A.

Dans la plupart des travaux, un SVM non-linéaire à noyau χ^2 a été utilisé pour classer les histogrammes d'occurrences de mots visuels.

$$K(\mathbf{H}_i, \mathbf{H}_j) = \exp\left(-\frac{1}{A}\chi^2(\mathbf{H}_i, \mathbf{H}_j)\right) \tag{2.2}$$

où χ^2 est la distance appliquée pour comparer deux histogrammes $\mathbf{H}_i = [\mathbf{H}_i(1)...\mathbf{H}_i(k)]^T$ et $\mathbf{H}_j = [\mathbf{H}_j(1)...\mathbf{H}_j(k)]^T$:

$$\chi^2(\mathbf{H}_i, \mathbf{H}_j) = \frac{1}{2}\sum_{l=1}^{k}\frac{\big(\mathbf{H}_i(l) - \mathbf{H}_j(l)\big)^2}{\mathbf{H}_i(l) + \mathbf{H}_j(l)} \tag{2.3}$$

A est la moyenne des distances χ^2 calculées pour toutes les paires des histogrammes des données d'entraînement.

Cette technique a été utilisée avec succès pour la classification des actions dans des scènes réelles en se basant sur des points d'intérêt. Laptev *et al.* [Laptev, 2005] ont été les premiers à adapter la technique de BOW à la reconnaissance d'action en s'appuyant sur des points d'intérêt spatio-temporels. Les auteurs construisent des BOWs des points d'intérêt "Harris 3D" qui sont décrits par des HOG/HOF. Un SVM à noyau χ^2 a été utilisé pour classifier les actions. Ce travail a donné des résultats très intéressants par rapport à l'état de l'art. L'efficacité de cette technique a été confirmée par les travaux de Dollar *et al.* [Dollar et al., 2005]. Ces derniers ont également eu des taux de reconnaissance élevés en appliquant la technique de BOW aux points d'intérêt "Cuboid" décrits par des gradients calculés pour chaque pixel dans une région cuboïde. Ces gradients sont concaténés en un seul vecteur. On utilise ensuite l'ACP pour projeter les vecteurs sur un espace de dimension plus faible.
Étant donné le succès de cette approche, un intérêt croissant s'est manifesté pour améliorer la représentation de base des BOWs qui ne tient pas compte de la structure géométrique des positions des STIPs. En effet, des travaux récents ont appliqué les modèles des sacs de mots à la reconnaissance d'actions en prenant en compte les relations spatio-temporelles entre les points d'intérêt. Laptev *et al.* [Laptev et al., 2008] utilisent une modélisation géométrique simple

pour construire des grilles de corrélations entre les points d'intérêt. Les auteurs calculent des histogrammes de fréquence des mots visuels pour toutes les séquences vidéo et aussi dans des sous-séquences de la vidéo définies par des grilles spatio-temporelles. En effet, chaque vidéo est divisée en un ensemble de 24 grilles spatio-temporelles ayant des niveaux de chevauchement temporel et spatial variés. Et, pour chacune de ces grilles, un descripteur vidéo, appelé canal, correspondant au BOWs, est construit. Ainsi, chaque séquence vidéo est décrite par 24 histogrammes (24 canaux). Pour classer les actions, on utilise un SVM avec le noyau χ^2 et le noyau multi-canal [Zhang et al., 2007] qui permet de combiner les différents canaux d'une manière robuste. Ce noyau est défini comme suit :

$$K(\mathbf{H}_i, \mathbf{H}_j) = \exp\left(-\sum_{c \in C} \frac{1}{A_c} \chi^2(\mathbf{H}_i^c, \mathbf{H}_j^c)\right) \tag{2.4}$$

où $\mathbf{H}_i^c$ est la i^{me} composante de $\mathbf{H}_i$ correspondant au canal c, $\mathbf{H}_i = [\mathbf{H}_i^1, \mathbf{H}_i^2, ..., \mathbf{H}_i^c]$, et A_c c'est la moyenne des distances χ^2 entre les paires d'histogrammes des données d'entrainement correspondant au canal c. Toutes les combinaisons possibles entre les différents canaux ont été testées pour en trouver la meilleure.

Pour enrichir la description des points d'intérêt, d'autres travaux se sont basés sur les trajectoires des points d'intérêt spatiaux tels que les SIFTs et les points d'intérêt spatio- temporels tels que « Harris-3D ». En effet, les points d'intérêt sont suivis dans le temps afin d'encoder leurs évolutions temporelles. Matikainen *et al.* [Matikainen et al., 2009, Matikainen et al., 2010] ont extrait des trajectoires de points d'intérêt de longueur fixe. Des BOW des vecteurs de vitesse de ces trajectoires et aussi de leur transformation affine sont utilisées pour modéliser les vidéos. La classification a été faite par SVM.

Une approche hiérarchique basée sur les trajectoires des SIFT a été proposée par Sun *et al.* [Sun et al., 2009]. Les points d'intérêt sont suivis avec l'algorithme KLT "Kanade–Lucas–Tomasi". Les auteurs utilisent le contexte spatio-temporel des trajectoires à plusieurs niveaux : (i) contexte au niveau local qui encode le voisinage spatial du point d'intérêt par la moyenne de descripteurs SIFT, (ii) contexte intra-trajectoire qui modélise les transitions au sein des trajectoires ; (iii) contexte inter-trajectoire qui capture la relation entre les trajectoires adjacentes (Figure 2.14). La dynamique de deux derniers niveaux de contexte est extraite par l'évaluation de la distribution stationnaire d'une chaîne de Markov. Des histogrammes d'occurrence des trajectoires et un SVM multi-canaux ont été utilisés pour la reconnaissance d'actions.

Yuan *et al.* [Yuan et al., 2012] combinent les BOWs des trajectoires des points d'intérêt, d'apparence et de forme pour modéliser les actions. La qualité ainsi que la quantité des trajectoires extraites par l'algorithme "KLT" ne sont pas toujours suffisantes. De plus ces trajectoires sont relativement parcimonieuses. Pour dépasser ces limites, Wang *et al.* [Wang et al., 2011a] ont utilisé les trajectoires des points denses. Ces trajectoires sont décrites par le vecteur de déplacement, les histogrammes HOG, HOF et MBH calculés dans un volume qui entoure la trajectoire (Figure 2.15). Ainsi, les contextes spatial et temporel de chacune de ces trajectoires sont modélisés par ces histogrammes. Cette méthode a donné des taux de reconnaissance

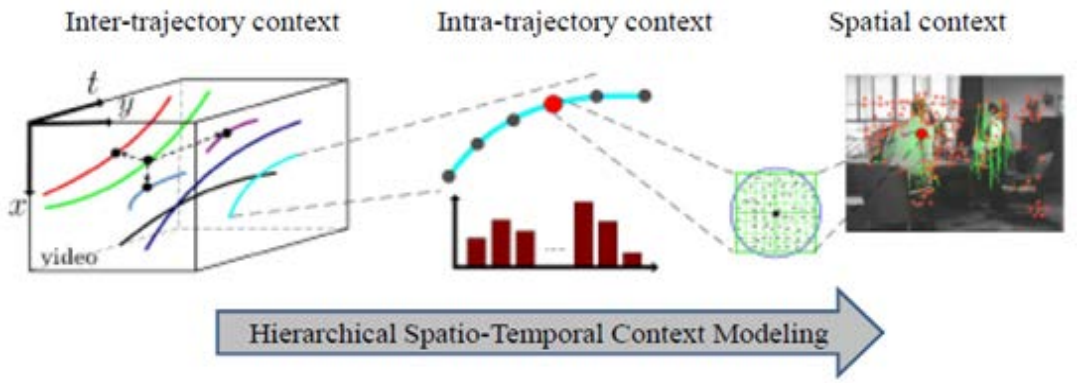


FIGURE *2.14: Modélisation hiérarchique de contexte spatio-temporel [Sun et al., 2009].*

élevés par rapport aux travaux dans l'état de l'art.

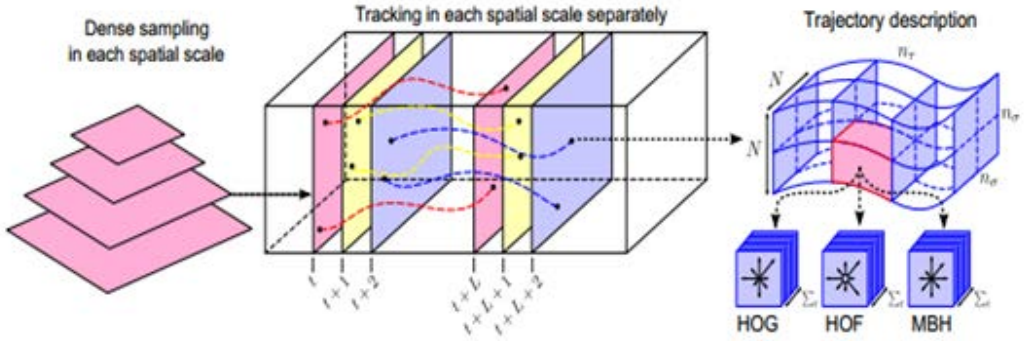


FIGURE *2.15: Illustration des trajectoires denses et des descripteurs (HOG, HOF, MBH) calculés le long de la trajectoire [Wang et al., 2011a].*

D'autres méthodes ont opté pour la modélisation du contexte spatio-temporel des points d'intérêt sans avoir recours à la méthode de BOWs. Gilbert *et al.* [Gilbert et al., 2008] ont proposé une modélisation de contexte spatio-temporelle des points d'intérêt « Harris3D ». Pour chaque point, la répartition spatiale relativement aux points d'intérêt appartenant à la même grille spatio-temporelle est calculée. Il en résulte donc un très grand nombre de primitives. Des techniques de fouille de données sont ensuite utilisées pour sélectionner les combinaisons de primitives les plus informatives pour chacune des classes. Pour améliorer leur modèle, Gilbert *et al.* [Gilbert et al., 2009] ont introduit une approche hiérarchique pour modéliser les relations spatio-temporelles entre les points d'intérêt "Harris3D". Les combinaisons des primitives les plus fréquentes sont apprises et ensuite modélisées d'une manière hiérarchique. Bilinski *et al.* [Bilinski and Brémond, 2012] ont modélisé le contexte des points d'intérêt via la fréquence de co-occurrence des paires de mots visuels. L'ordre spatio-temporel entres les paires des mots visuels a été pris en considération dans leur modèle. Dans [Messing et al., 2009], les points

d'intérêt Harris sont suivis à l'aide de la méthode KLT. Des trajectoires de longueurs différentes sont construites et quantifiées comme un historique des magnitudes des vecteurs de mouvement (Figure 2.16). L'activité globale est ensuite représentée via une mixture de chaînes de Markov.

FIGURE *2.16: Exemples de trajectoires des points Harris [Messing et al., 2009].*

Pour conclure, il faut noter que les approches basées sur les BOWs en utilisant un SVM multi-canaux restent les plus performantes et, plus précisément la méthode de Wang *et al.* [Wang et al., 2011a] qui se base sur des BOWs des descripteurs MBH, HOG, HOF et vecteur de déplacement calculés autour des trajectoires des points denses en utilisant un SVM multi-canaux comme classifieur.

2.5.2.2 Classification séquentielle

Une fois que les actions sont modélisées comme des séries temporelles des caractéristiques locales, les approches séquentielles permet une classification efficace puisqu'elle a été spécialement conçue pour classifier des séquences temporelles. En effet, les classifieurs séquentiels exploitent la dimension temporelle des actions et, plus particulièrement, les relations temporelles entre les caractéristiques locales. Étant donné que ces modèles nécessitent un vecteur de taille fixe à chaque instant, ce type de modèle de classification a été généralement utilisé pour la reconnaissance en se basant sur des primitives holistiques ou issues de reconstruction de modèles humains. Cela vient du fait que le nombre des STIPs extraits dans chacune des trames est variable, d'où l'impossibilité d'avoir un vecteur de primitives de taille fixe à chaque instant. Les approches séquentielles peuvent être classées en deux sous-catégories : approches génératives et approches discriminatives.

- **Approches génératives**

Les modèles séquentiels génératifs cherchent dans un premier lieu à modéliser la distribution jointe de la séquence d'observation $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T)$ et la séquence de labels (classes) $\mathbf{Y}$ par un modèle $P(\mathbf{X}; \mathbf{Y}, \theta)$; où θ est le vecteur de paramètres à optimiser dans la phase

d'apprentissage pour maximiser cette probabilité. On en déduit ensuite la règle de classification grâce à la règle de Bayes.

Parmi les modèles génératifs les plus répandus, on trouve les HMMs. Les HMMs sont des automates stochastiques à états finis permettant de déduire des séquences d'états non observables (ou états cachés) à partir de séquences de données observées (observables).
Yamato *et al.* [Yamato et al., 1992] utilisent un HMM discret pour représenter des séquences de tennis en utilisant comme primitives une quantification de silhouette. Lv et Nevatia [Lv and Nevatia, 2006] ont utilisé un modèle 3D basé sur les angles des articulations et ont entraîné plusieurs HMMs chacun spécialisé sur un sous-ensemble des articulations. Les sorties des HMMs sont considérées comme des primitives pour le classifieur final Adaboost. Ahmad et Lee [Ahmad and Lee, 2008] ont essayé de prendre en considération plusieurs angles de vue en utilisant un HMM multidimensionnel. Dans [Nguyen et al., 2005], les auteurs essaient de reconnaître les activités liées à la prise de repas telles que «prendre un repas rapide (sandwich)», «prendre un repas complet» en se basant sur la trajectoire de la personne et sur des points de repère indiquant la nature de l'emplacement (cuisine, salon, etc.). Les auteurs ont utilisé un modèle de Markov caché hiérarchique (LHMMs) à deux couches. La couche inférieure est destinée à la reconnaissance des actions simples (se déplacer d'un point de repère A à un point de repère B) en faisant correspondre les modèles HMMs des actions avec la séquence de vecteurs caractéristiques extraite de la vidéo. La couche supérieure considère les actions atomiques reconnues comme des observations. Autrement dit, les activités sont considérées comme une séquence d'actions qui vont être l'entrée du LHMM. Cependant, ce travail se base sur une simplification du modèle d'activités en le réduisant à une trajectoire entre quelques points de repère.
Il faut noter que le modèle HMM met une forte hypothèse sur l'indépendance des observations. Par conséquent, il ne peut pas modéliser les données dont la dépendance est importante. Plus précisément, il se base sur deux hypothèses d'indépendance. La première est l'hypothèse de Markov selon laquelle les transitions à un état dépendent uniquement de l'état précédent et non pas de tout l'historique. Dans la deuxième hypothèse, on suppose que les observations ne sont conditionnées que par l'état actuel, et sont donc conditionnellement indépendantes.
Les approches hiérarchiques utilisant les réseaux Bayésiens dynamiques (DBN) ont été aussi étudiées pour la reconnaissance des activités complexes. Un DBN peut contenir de multiples niveaux d'états cachés, ce qui suggère qu'ils peuvent modéliser les activités d'une façon hiérarchique. Dai *et al.* [Dai et al., 2008] ont proposé un DBN pour reconnaître les activités dans un environnement semblable à une salle de conférence [Veeraraghavan and Roy-chowdhury, 2006]. Ils ont essayé de reconnaître les activités de présentation, pause, discussions, etc. Damen et Hogg [Damen and Hogg, 2009] ont construit des réseaux Bayésiens en utilisant une méthode de Monte Carlo par chaîne de Markov (MCMC) pour l'analyse hiérarchique des activités relatives à la bicyclette. Ils ont utilisé les réseaux Bayésiens pour modéliser les relations entre les actions atomiques. Ces réseaux sont mis à jour itérativement en utilisant la MCMC pour rechercher la structure qui explique le mieux les observations en cours. Bien que les modèles séquentiels génératifs permettent une classification efficace des séquences temporelles, ils ont recours à la recherche de la probabilité conjointe de $P(\mathbf{X}, \mathbf{Y})$ pour résoudre un problème

conditionnel alors que les observations sont déjà données pendant l'étape de test. Donc, le modèle génératif essaie de résoudre un problème conditionnel en utilisant un problème génératif plus complexe comme étape intermédiaire.

- **Approches discriminatives**

Pour éviter les limitations des approches génératives évoquées dans la section précédente, une attention croissante a été donnée aux approches discriminantes qui permettent de modéliser une distribution conditionnelle des labels sachant les observations $P(\mathbf{Y}|\mathbf{X})$. Elles sont donc capables de discriminer directement les différentes actions au lieu d'apprendre à modéliser chaque classe individuellement, comme c'est le cas dans les modèles génératifs. De plus, les modèles discriminatifs ne posent aucune hypothèse d'indépendance entre les observations. Parmi ces approches, on peut citer les champs aléatoires conditionnels (Conditional Random Fields ou CRFs) [Lafferty et al., 2001] qui ont été récemment utilisés dans des travaux de reconnaissance des actions et qui ont donné des résultats encourageants. Sminchisescu *et al.* [Sminchisescu et al., 2006] ont utilisé un CRF linéaire, où les dépendances entre les états sont de premier ordre pour la reconnaissance des actions simples comme la marche, le saut, etc. Ils ont comparé le CRF avec les HMM et les MEMM (modèles de Markov d'Entropie Maximale). Ce dernier est aussi un modèle discriminatif, mais il souffre d'un problème de biais sur le label "label bias problem" c.à.d qu'il privilégie les états ou labels à faible entropie et, dans certains cas, l'observation peut même être ignorée si l'état courant n'a qu'un seul successeur. Sminchisescu et *et al.* [Sminchisescu et al., 2006] ont montré que la performance de CRF dépasse celle de MEMMs et de HMM. Ces résultats sont, en partie, confirmés par Mendoza et Perez de la Blanca [Mendoza and de la Blanca, 2008], qui obtiennent de meilleurs résultats par CRF par rapport au HMM en utilisant des primitives holistiques. Des extensions de CRF ont été utilisées aussi pour la reconnaissance des actions. Par exemple, Zhang et Gong [Zhang and Gong, 2010] utilisent des CRF cachés (Hidden CRFs ou HCRF, [Quattoni et al., 2007]) pour classifier des séquences d'actions. Les états cachés des HCRFs sont analogues à ceux des HMMs. Finalement, Banerjee et Nevatia [Banerjee and Nevatia, 2011] ont utilisé un HCRF dans une approche basée sur les STIPs pour la reconnaissance des actions simples. Wang et Suter [Wang and Suter, 2007] utilisent des CRF Factoriels (FCRF) qui permettent de modéliser des interactions complexes entre les états. Kjllstrom *et al.* [Kjellström et al., 2011] ont aussi utilisé un FCRF pour pouvoir exploiter les différentes relations de corrélation entre les objets utilisés pour la réalisation des activités (verre, récipient, livre, etc.) et les actions simples (ouvrir, verser, etc.) afin de classifier les activités. En effet, les interactions entre les objets et les actions simples peuvent être très utiles pour l'identification des activités. Mais il est très difficile de modéliser et d'identifier manuellement ces relations et ces interactions. Ainsi le FCRF a été un moyen efficace pour les modéliser. Dans Kjllstrom *et al.* [Kjellström et al., 2011], les auteurs ont utilisé un FCFR à trois niveaux : le premier est destiné à la reconnaissance des actions à partir de mouvements des mains, le deuxième est consacré à la reconnaissance des objets et le dernier exploite les actions et les objets reconnus pour identifier l'activité réalisée. Mais l'une des limites

de ce travail est que les auteurs se bornent à des activités simples basées seulement sur le mouvement des mains (par exemple ouvrir un livre, verser de l'eau dans un récipient, verser de l'eau dans un verre, etc.). Les auteurs n'utilisent ainsi qu'une simple méthode de détection des mains basée sur la couleur de la peau et sur des modèles pré-calculés de la main. Le modèle de la main est ensuite considéré comme l'observation principale utilisée par la couche destinée à la reconnaissance des actions. L'un des rares travaux qui ait essayé de reconnaître des activités quotidiennes effectuées par une seule personne et sans aucune simplification est celui de Sung *et al.* [Sung et al., 2011]. Ils ont travaillé avec une base de données contenant 12 activités Kinect, par exemple cuisiner, écrire dans un tableau, utiliser un ordinateur, se brosser les dents, boire de l'eau, etc. L'approche de reconnaissance de ces activités est basée sur un MEMM hiérarchique à deux couches. La première couche reconnaît les sous-tâches des activités qui vont être l'entrée de la deuxième couche destinée à la reconnaissance des activités. Cette méthode réalise de bonnes performances. Cependant, il faut noter que les auteurs utilisent des primitives pré-calculées par la Kinect, comme le modèle de squelette avec la précision des angles joints et la longueur de différents segments, les images RGB et des images de profondeur où la profondeur de chaque point est donnée. De plus, il faut noter que le modèle MEMM souffre généralement du problème du biais.

2.5.2.3 Classification par trame clés

La méthode de trames clés permet de caractériser l'action seulement avec un petit sous-ensemble de trames de la vidéo. Il s'agit d'un ensemble de trames caractéristiques d'une action. Cette approche est parciminieuse par rapport aux approches exploitant toute la séquence vidéo. Carlsson et Sullivan [Carlsson and Sullivan, 2001] ont été les premiers à introduire l'utilisation des trames clés pour reconnaître les coups droits et les coups de revers dans des séquences de match de tennis. En effet, une seule trame clé par classe a été sélectionnée manuellement et une méthode d'appariement basée sur la silhouette a été utilisée pour faire la classification. D'autres travaux ont opté pour la sélection automatique des trames. Dans [Zhao and Elgammal, 2008] les trames sont triées selon des heuristiques basées sur des primitives holistiques. Le premier quart de ces trames est sélectionné pour participer à une classification par vote. L. Liu *et al.* [Liu et al., 2013] utilisent Adaboost pour sélectionner automatiquement les trames clés. Seulement entre 13% et 20% des trames de la séquence sont gardées. Les auteurs font une étape de prétraitement pour la localisation spatiale et temporelle des personnes. Dans des travaux récents, [Raptis and Sigal, 2013, Vahdat et al., 2011] une action est modélisée comme une séquence de trames clés décrites par des poselets distinctives. Dans [Raptis and Sigal, 2013], les trames clés sont considérées comme des variables latentes apprises via une approche discriminative à marge maximale. Cependant, cette approche nécessite la définition des poselets et leur annotation d'une façon manuelle. Les poselets sont des parties d'images qui décrivent des parties du corps. Les auteurs ont utilisé 28 différents types de poselets comme attributs (Figure 2.17).

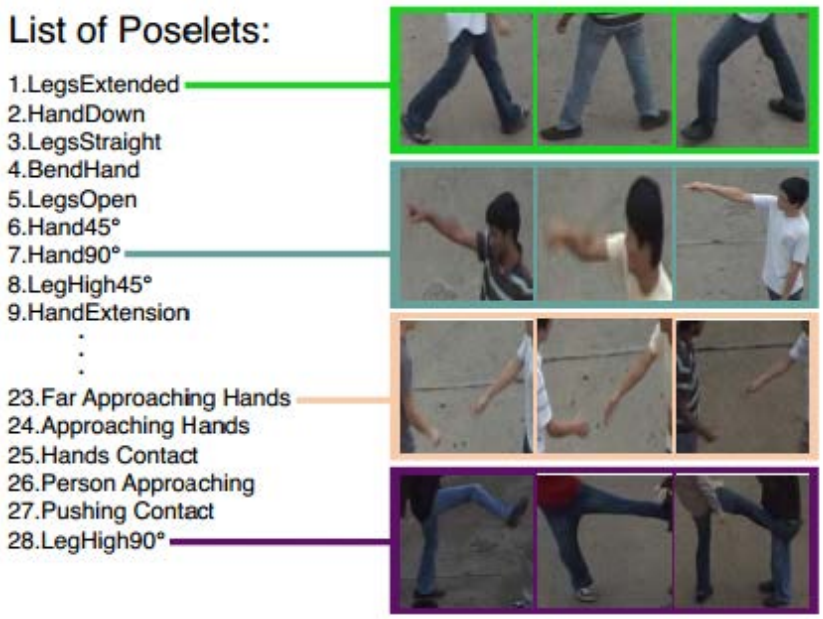


FIGURE *2.17: Exemples de poselets [Raptis and Sigal, 2013].*

2.5.2.4 Autres méthodes

D'autres approches de reconnaissance s'appuient sur les poses clés. Il s'agit de trouver les poses ; positions particulières des personnes représentées par exemple sous forme de silhouettes ou sekeletons ; possédant un maximum d'informations localement afin qu'elles puissent être de bonnes représentantes des actions (Figure 2.18). Pour extraire les poses clés, certains travaux utilisent des exemplaires comme Weinland et Boyer [Weinland and Boyer, 2008] ; d'autres utilisent une mesure d'énergie, comme Gong et al. [Gong et al., 2010], Lv et Nevatia [Lv and Nevatia, 2007] ; ou encore une distance géométrique, comme Barnachon et al. [Barnachon et al., 2012]. Dans la majorité des cas, l'extraction de poses clés sert de phase intermédiaire entre les données en entrée (silhouettes, séquence de capture de mouvements, etc.) et la classification de l'action, par des classifications globales, séquentielles [Lv and Nevatia, 2007], ou par des trames clés [Raptis and Sigal, 2013].

2.5.2.5 Synthèse

La plupart des travaux de l'état de l'art se sont basés sur la technique de BOWs des points d'intérêt, vue son efficacité et sa robustesse dans des conditions réelles. Ces méthodes ont donné de meilleures performances que les méthodes séquentielles dans des scènes réelles. Cela est dû au fait que les méthodes séquentielles utilisent principalement des primitives holistiques ou celles issues de modèle humain. De plus, la méthode des BOWs par points d'intérêt reste beaucoup plus efficace que les approches basées sur les trames clés. En effet, bien que

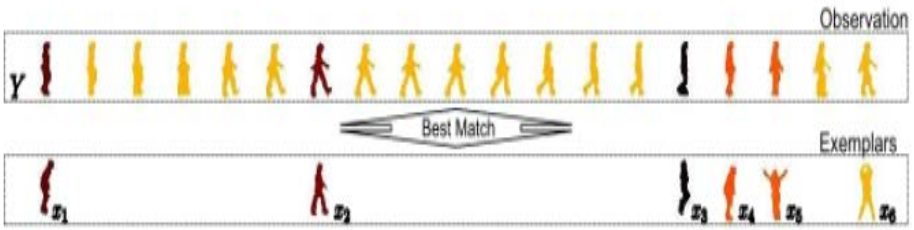


FIGURE *2.18: Exemple de poses clés sous forme de silhouettes en reconnaissance d'actions. [Weinland and Boyer, 2008].*

les approches de trames clés soient parcimonieuses par rapport aux approches exploitant toute la séquence vidéo, cette méthode perd en flexibilité en limitant la représentation de l'action à quelques trames. De plus, l'efficacité de cette méthode n'était testée qu'avec des bases moins difficiles que celles utilisées par les BOWs des points d'intérêt et aussi par les méthodes séquentielles.

Malgré la performance des approches basées sur les BOWs des points d'intérêt, il faut cependant noter que cette technique n'exploite pas l'aspect séquentiel des activités qui est modélisé par les approches séquentielles. Dans cette thèse, nous proposons de prendre en compte aussi bien la robustesse des BOWs des points d'intérêt que la capacité des méthodes séquentielles à modéliser la dynamique temporelle des activités.

2.6 Approche proposée

De l'état de l'art précédent, nous pouvons constater que les points d'intérêt spatio-temporels se sont dernièrement montrés très efficaces pour la reconnaissance des actions dans des scénarios réalistes. Ces primitives sont néanmoins désordonnées et ont été ainsi généralement couplées à des méthodes de classification globale. Ces dernières, contrairement aux méthodes séquentielles, ne permettent pas une modélisation de la dynamique temporelle des actions.

Dans cette thèse, nous proposons une nouvelle modélisation des séquences permettant de tirer profit à la fois de la robustesse des points d'intérêt et de la capacité de modéliser la dynamique interne des actions par des approches séquentielles.

Ces dernières ont comme entrée une séquence de vecteurs de primitives de dimension fixe à chaque instant alors que le nombre de STIPs par trame est variable. Pour résoudre ce problème, nous représentons la séquence vidéo comme une série temporelle de BOWs locaux des STIPs. Ceci permet d'avoir une séquence de vecteurs de primitives de dimension fixe. En effet, nous commençons par une segmentation temporelle des séquences vidéo à l'aide de la technique de fenêtre glissante. Chacun des segments ainsi obtenu est représenté par un BOW de descripteurs des points d'intérêt qui lui appartiennent. La phase de reconnaissance est basée sur un modèle hybride SVM-HCRF. Ce modèle permet d'exploiter les avantages de SVM

et HCRF à savoir le pouvoir de généralisation et discriminant de SVM et la modélisation de structure temporelle des actions via le modèle séquentiel discriminatif HCRF. SVM est appliqué en tant que classifieur de bas niveau permettant de transformer les BOWs locaux en des vecteurs de probabilités conditionnelles des actions. De ce fait, le HCRF traite des informations de haut niveau générées par SVM ce qui permet d'améliorer la performance de système de reconnaissance. De plus, étant donné que les BOWs sont généralement de dimension importante l'utilisation de SVM permet de réduire considérablement la dimension des vecteurs d'entrée au HCRF et ainsi de réduire le temps d'entrainement. Notre modèle hybride permet aussi d'éviter le problème de sur-apprentissage au niveau du HCRF. En effet, l'utilisation des séquences de BOWs de dimension importante comme entrée du HCRF peut engendrer le problème de sur-apprentissage vu que la dimension des vecteurs de primitives devient beaucoup plus importante que le nombre des données d'entrainement. Notons aussi que le pouvoir de généralisation de SVM permet de surmonter ce problème au niveau local et l'utilisation de séquences de vecteurs de probabilité de dimension réduite permet de le surmonter au niveau global. La motivation pour notre approche peut être récapitulée ainsi :

- Dans notre travail, nous évitons l'utilisation des approches basées sur le modèle de corps humain ou bien les modèles holistiques. Comme cela a été mentionné précédemment, ces modèles nécessitent généralement l'extraction des silhouettes du fond et leur suivi, ce qui reste un problème difficile dans les conditions réelles. Nous pensons que la reconnaissance des actions à partir des points d'intérêt donne plus de robustesse à notre approche. Ainsi que nous allons le montrer dans la section suivante par une étude tensorielle, ces primitives encodent principalement les actions indépendamment de l'identité de la personne.

- Notre approche est basée sur un modèle séquentiel discriminatif qui est bien adapté à la reconnaissance des séquences temporelles. Nous avons aussi utilisé un SVM comme classifieur local permettant un gain en temps de calcul, en robustesse par rapport au problème de sur-apprentissage et aussi en efficacité au niveau de la precision de reconnaissance.

CHAPITRE 3

Extraction des primitives

Sommaire

3.1 Introduction

Le jeu de caractéristiques extraites des vidéos que nous avons considérées est fondé sur les points d'intérêt spatio-temporels. Les STIPs sont particulièrement adaptés pour plusieurs raisons. D'abord, ils permettent de caractériser principalement les mouvements et sont invariants par rapport au changement d'échelle et de vitesse. De plus, ils ne nécessitent pas la suppression du fond et ne requièrent pas un suivi et/ou un modèle explicite du corps humain, ce qui est très difficile pour les tâches où la personne se trouve de dos ou de profil par rapport à la caméra et où des occultations ont lieu. Enfin, ils permettent une représentation parcimonieuse des primitives tout en étant discriminants.

Etant donné que ces primitives sont désordonnées, elles étaient généralement utilisées avec des méthodes de classification globales. Dans notre travail, nous proposons une représentation séquentielle des actions basée sur les STIPs. Celle-ci permet de modéliser l'aspect temporel des actions tout en exploitant les avantages de ce type de primitives.

Nous avons considéré, dans un premier temps, une représentation séquentielle basée sur les points d'intérêt spatio-temporels Harris-3D [Laptev and Lindeberg, 2003]. Ces primitives ont donné des résultats relevant de l'état de l'art lors de la réalisation de nos travaux. De plus, nous avons confirmé via une analyse tensorielle que ces primitives permettent de caractériser principalement les actions indépendamment de l'identité de la personne tout en étant parcimonieuses. Bien que leur parcimonie garantisse une simplicité de mise en œuvre et un gain au niveau du temps de calcul, dans le cas des actions de faible amplitude de mouvement, un nombre très réduit de points d'intérêt est extrait. Dans ce cas, les points extraits peuvent être insuffisants pour le codage de l'activité ce qui peut fortement influencer la performance de système de reconnaissance. De ce fait, nous avons aussi étudié l'utilisation des points d'intérêt denses. Ces points ont été récemment proposés par Wang *et al.* [Wang et al., 2011a] et se sont relevés être des extracteurs efficaces. Ces primitives permettent d'avoir une représentation assez informative même pour les activités de faible amplitude.

Dans la première section de ce chapitre, nous commençons par une présentation des points d'intérêt Harris-3D et par une étude de leur robustesse par rapport aux facteurs nuisibles à la reconnaissance d'activités. Dans la deuxième section, nous présentons les points d'intérêt denses. Enfin, nous exposons notre représentation séquentielle des vidéos d'activités.

3.2 Les points d'intérêt Harris-3D

Il existe plusieurs types de détecteurs [Laptev and Lindeberg, 2003, Dollar et al., 2005, Oikonomopoulos et al., 2005, Willems et al., 2008, Wong and Cipolla, 2007] de points d'intérêt spatio-temporels. Pour caractériser le mouvement et l'apparence de ces derniers, plusieurs types de descripteurs sont proposés [Laptev et al., 2008, Klaser et al., 2008, Laptev and Lindeberg, 2004, Scovanner et al., 2007]. Nous nous intéressons dans cette section aux points d'intérêt Harris-3D qui correspondent à des points ayant des variations significatives en temps et en espace. Ces points d'intérêt sont robustes par rapport au changement d'échelle et à la rotation. Nous avons aussi étudié leur robustesse à l'un des principaux facteurs de variabilité, en l'occurrence, l'identité de la personne via une analyse multi-dimensionnelle fondée sur les tenseurs. Ceci permet de vérifier le degré de leur corrélation au mouvement. Dans ce qui suit, nous exposons le principe d'extraction des points Harris-3D, ensuite nous étudions la robustesse de ces points par rapport à l'identité de la personne.

3.2.1 Principe d'extraction des points Harris-3D

Le détecteur Harris-3D proposé par Laptev et Lindeberg [Laptev and Lindeberg, 2003] est une extension au domaine spatio-temporel de la méthode d'Harris de détection de coins. Les auteurs calculent la matrice spatio-temporelle du moment d'ordre 2 en chaque point de la vidéo comme suit :

$$\mu(.;\sigma,\tau) = G(.;s\sigma,s\tau) * \begin{pmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{pmatrix} \tag{3.1}$$

en utilisant des échelles spatiale σ et temporelle τ différentes, une fonction gaussienne lissante G et un paramètre s qui relie G et les échelles locales σ, τ. La dérivée de premier ordre de séquence vidéo v est définie comme :

$$\begin{aligned} L_x(.;\sigma,\tau) &= \partial_x(G \times v), \\ L_y(.;\sigma,\tau) &= \partial_y(G \times v), \\ L_t(.;\sigma,\tau) &= \partial_t(G \times v). \end{aligned} \tag{3.2}$$

Les points d'intérêt spatio-temporels sont localisés aux maxima locaux de :

$$H = \det(\mu) - ktrace^3(\mu), > 0. \tag{3.3}$$

Les auteurs ont aussi proposé un mécanisme optionnel de sélection d'échelles spatio-temporelles. Comme dans [Laptev et al., 2008], nous n'avons pas utilisé la sélection automatique ; néanmoins, nous avons utilisé des points d'intérêt extraits pour des échelles multiples en se basant sur un échantillonnage régulier des paramètres d'échelle σ,τ. Nous utilisons l'implémentation disponible en ligne et les valeurs standard des paramètres : $k = 0.0005, \sigma^2 = 4;8;16;32;64;128; \tau^2 = 4,8$ pour extraire les points Harris-3D. Un exemple des points est illustré dans la figure 3.1.

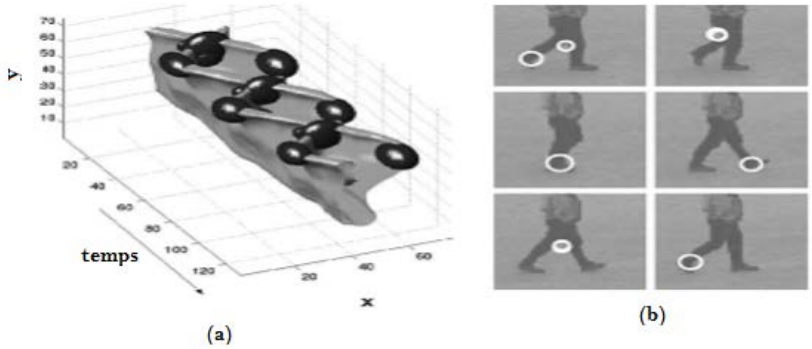


FIGURE *3.1: Détection des points d'intérêt spatio-temporels du mouvement des jambes d'une personne qui marche. (a) : modèle 3-D du mouvement de la jambe : les points d'intérêt détectés sont illustrés par des ellipsoides, (b) : les points d'intérêt spatio-temporels dans des images 2D [Laptev and Lindeberg, 2003]*

Pour chaque point extrait Harris-3D $(x;y;t;\sigma;\tau)$, un descripteur local est calculé dans un volume centré en $(x;y;t)$. Les dimensions spatiales $\Delta_x(\sigma)$ et $\Delta_y(\sigma)$ de ce volume 3D sont calculées en fonction de l'échelle spatiale σ ; sa longueur temporelle $\Delta_t(\tau)$ est fonction de l'échelle temporelle τ. Les points Harris-3D sont souvent décrits par des HOGs et de HOFs. En effet, la région cuboïde est divisée en une grille de $n_x \times n_y \times n_t$ cellules. Pour chaque cellule, on calcule des histogrammes à 4 composantes d'orientations du gradient et des histogrammes à 5 composantes d'orientation de flux optique. Les cellules des histogrammes sont normalisées via la norme L_2 et ensuite concaténées pour former les descripteurs HOG et HOF.

Comme illustré dans la Figure 3.2, dans nos expériences nous utilisons le descripteur HOG/-HOF qui est obtenu par la concaténation des HOG et HOF. Ces derniers sont calculés dans un volume de taille $\Delta_x(\sigma) = \Delta_y(\sigma) = 18\sigma$ et $\Delta_t(\tau) = 8\tau$. La division de ce cuboïde est faite selon les paramètres $n_x, n_y = 3$, et $n_t = 2$. Ainsi le descripteur HOG/HOF a une dimension de 162.

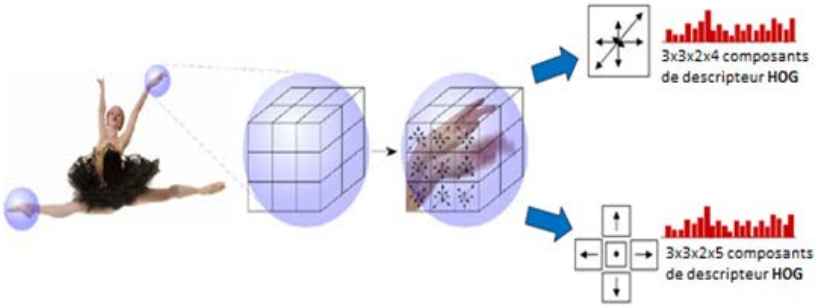


FIGURE *3.2: Illustration du descripteur HOG / HOF. Une région d'intérêt est décrite par un volume divisé en une grille de cellules. Pour chaque cellule, un HOG ainsi que l'HOF sont calculés. Le descripteur final est la concaténation des histogrammes HOG et HOF correspondant à chaque cellule de la grille [Laptev et al., 2008].*

3.2.2 Modélisation tensorielle pour l'évaluation de la robustesse des Harris-3D par rapport à l'identité de la personne

Les Harris-3D offrent une représentation discriminante des actions tout en contournant les problèmes liés aux occlusions et de suivi dans les conditions réelles. Comme les actions sont le résultat de l'interaction de plusieurs facteurs tels que le type d'action, l'identité de la personne, l'illumination, les vêtements, etc., il est intéressant de modéliser explicitement l'aspect multimodal des actions afin de vérifier la robustesse des Harris-3D à ces différents facteurs de variabilité. Une analyse tensorielle fondée sur l'algèbre multi-linéaire est adaptée à cette tâche de modélisation. Comme illustré dans la Figure 3.3, M. Vasilescu et D. Terzopoulos [Vasilescu and Terzopoulos, 2002] ont utilisé des tenseurs dans le cadre de la reconnaissance des visages afin de modéliser les différents facteurs de variabilité (personne, luminosité, pose, expression,. . . ..).

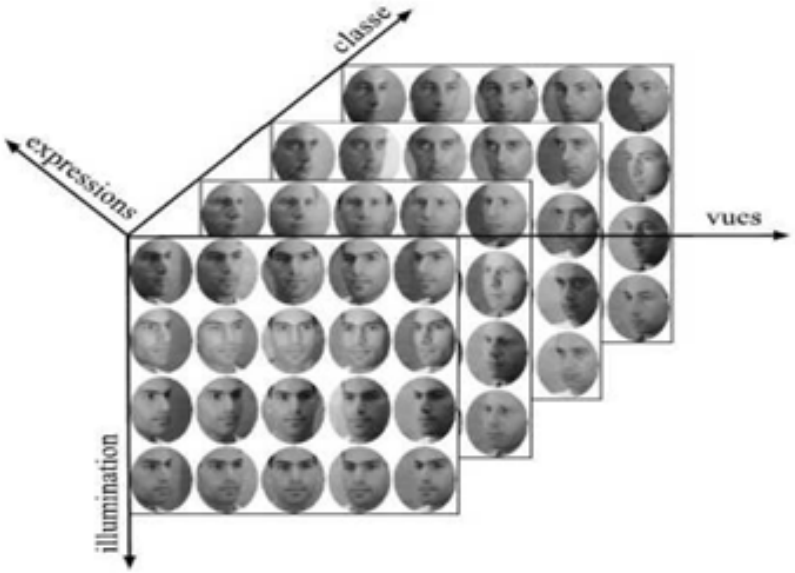


FIGURE *3.3: Représentation tensorielle du visage : les données sont rangées sous la forme d'un tenseur, les quatre dimensions de celui-ci représentent la classe d'appartenance, la vue considérée, les conditions d'illumination et l'expression faciale. Ici, seul le sous-tenseur correspondant à une expression faciale neutre est montré [Vasilescu and Terzopoulos, 2002].*

Dans notre travail, nous avons considéré une modélisation tensorielle des actions représentées par un BOW de points Harris-3D pour étudier la corrélation de ces derniers au mouvement (type d'action), d'une part, et à l'identité de la personne, d'autre part.

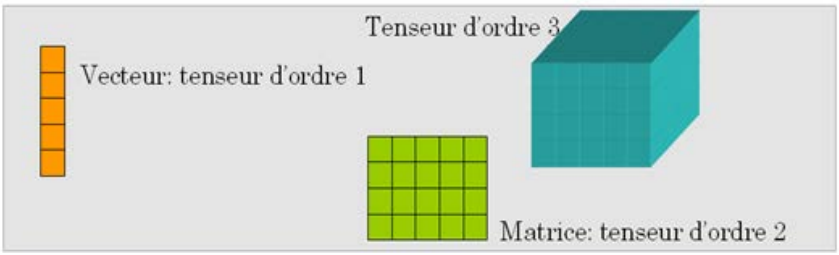


FIGURE *3.4: Illustration des tenseurs de différents ordres.*

3.2.2.1 Analyse tensorielle multilinéaire

L'analyse multilinéaire est une méthode mathématique efficace pour analyser des données multimodales représentées par des tenseurs. Un tenseur, aussi connu comme un tableau de n-dimensions ou matrice n-mode, est une généralisation d'ordre n d'un vecteur (tenseur 1-mode) et d'une matrice (tenseur 2-Mode). Le tenseur $\mathcal{D} \in \mathbb{R}^{\mathbf{I}_1 \times \mathbf{I}_2 .. \times \mathbf{I}_n .. \times \mathbf{I}_N}$ a un ordre N (Figure 3.4). L'ordre d'un tenseur indique le nombre de ses dimensions et chaque dimension d'un tenseur

correspond à un mode. Le mode n d'un tenseur correspond aux vecteurs $\mathbf{I}_n$ de la matrice $\mathbf{D}_{(n)} \in \mathbb{R}^{\mathbf{I}_n \times (\mathbf{I}_1 \times \mathbf{I}_2 .. \times \mathbf{I}_{n-1} \times \mathbf{I}_{n+1} .. \times \mathbf{I}_N)}$. C'est le résultat de l'aplatissement du tenseur $\mathcal{D}$ selon le mode n (Figure 3.5). Plus précisément, les vecteurs $\mathbf{D}_{(n)}$ sont obtenus à partir de $\mathcal{D}$ en faisant varier l'indice n tout en gardant tous les autres indices fixes.

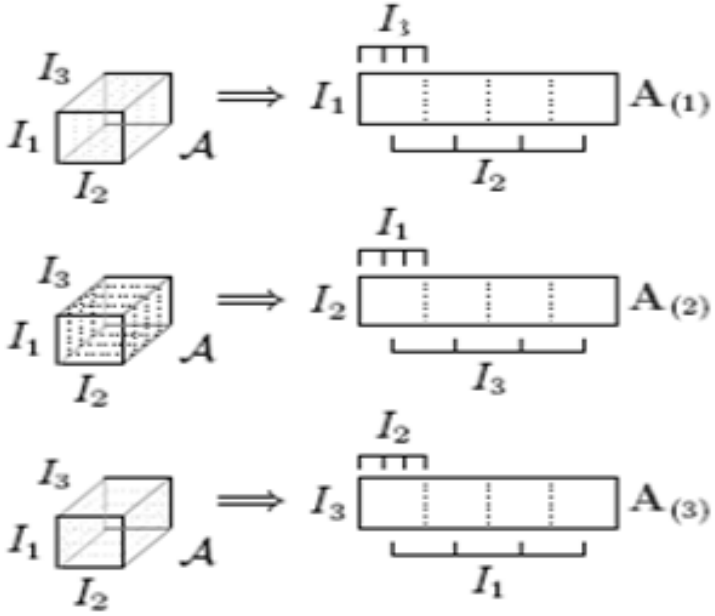


FIGURE *3.5: Exemple de l'aplatissement d'un tenseur d'ordre 3.*

Pour représenter et séparer les différents modes (ou facteurs) du tenseur, on utilise la méthode de décomposition SVD N-mode qui est une généralisation de SVD classique d'une matrice au tenseur. Elle permet de décomposer le tenseur en un produit « n-mode » de N espaces orthogonaux :

$$\mathcal{D} = \mathcal{Z} \times_1 \mathbf{U}_1 ... \times_n \mathbf{U}_n ... \times_N \mathbf{U}_N \tag{3.4}$$

où $\mathcal{Z}$ est le "tenseur Coeur" qui traduit les interactions entre les différents modes. Ces modes sont représentés par les matrices $\mathbf{U}_n$.

Dans les bases de données utilisées dans nos expériences, les séquences de mouvements sont le résultat de l'interaction de trois modes : les personnes, les actions et les primitives. Par conséquent, nous représentons les séquences de mouvement comme un tenseur $\mathcal{D}$ de dimension $M \times P \times F$, où M est le nombre d'actions, P est le nombre de personnes et F est la taille du vecteur des primitives. Ainsi, on peut appliquer l'algorithme N-mode-SVD pour décomposer le tenseur $\mathcal{D}$ comme suit :

$$\mathcal{D} = \mathcal{Z} \times_1 \mathbf{U}_{actions} \times_2 \mathbf{U}_{personnes} \times_3 \mathbf{U}_{primitives} \tag{3.5}$$

où le tenseur $\mathcal{D}$ est le produit de trois matrices orthogonales et où le tenseur $\mathcal{Z}$ régit l'interaction entre les trois modes. $\mathbf{U}_{actions}$ est une matrice de dimension $M \times M$ qui couvre l'espace d'actions $(a = 1...M)$; chaque ligne a de cette matrice encode une représentation invariante de l'action a (signature de l'action a) indépendamment de l'identité des personnes. La matrice $\mathbf{U}_{personnes}(P \times P)$ représente l'espace des personnes ; chaque ligne p de la matrice $\mathbf{U}_{personnes}$ encode une représentation invariante de la personne p par rapport aux différentes actions. La matrice $\mathbf{U}_{primitives}$ représente l'espace de primitives ; les colonnes de cet espace correspondent aux vecteurs propres des descripteurs des points Harris-3D (EigenHarris3D) qui peuvent être calculés par une ACP conventionnelle.
Peu de travaux appliquent les tenseurs dans le contexte de la reconnaissance des actions. Ces approches utilisent des primitives holistiques pour représenter le mouvement, et elles sont confrontées aux problèmes liés à la segmentation et au suivi de la personne, ce qui n'est pas évident dans des scènes complexes et réelles. Par exemple, dans le travail proposé par Berhrouz*et al.* [Khadem and Rajan, 2009], les primitives sont obtenues par une simple concaténation de toutes les silhouettes – chacune caractérisée par une représentation en pixels – qui apparaissent durant le cycle de réalisation de l'action. Pour avoir des vecteurs de primitives de taille fixe, les auteurs ont choisi un nombre de trames égal au nombre de trames du plus long cycle de réalisation des actions. En plus des problèmes de segmentation, le principal inconvénient des méthodes du type [Khadem and Rajan, 2009] est la dimensionnalité élevée du vecteur de primitives (pour la base de Weizmann, par exemple, la taille du vecteur de primitives est égale au nombre de frames $\times S$ où $S = 3072$ pixels). Une autre méthode qui utilise des tenseurs en reconnaissance d'actions a été proposée par [Zhang et al., 2010]. Cette méthode représente l'action par une image d'historique de mouvement, ce qui garantit une représentation de dimension fixe. Cependant, comme pour [Khadem and Rajan, 2009], cette approche nécessite une étape de soustraction de fond et de suivi, et elle utilise aussi un vecteur de primitives de dimensionnalité élevée. L'utilisation de BOW de Harris-3D permet d'analyser des scènes complexes sans avoir besoin d'une étape de soustraction de fond, ni de suivi de personnes, évitant ainsi les erreurs éventuelles liées à ces deux stades préliminaires. En plus, les BOWs peuvent être appliqués avec un nombre réduit de mots visuels, donnant ainsi une représentation tensorielle parcimonieuse. En effet, la dimensionnalité des tenseurs utilisant les BOW est beaucoup plus réduite que celle générée par une représentation tensorielle basée sur des méthodes holistiques.
Donc le tenseur $\mathcal{D}$ (Figure 3.6) peut être décomposé ainsi :

$$\mathcal{D} = \mathcal{Z} \times_1 \mathbf{U}_{actions} \times_2 \mathbf{U}_{personnes} \times_3 \mathbf{U}_{BOWs} \tag{3.6}$$

Puisque nous nous intéressons à la reconnaissance des actions, on définit le tenseur d'action $\mathcal{A}$ tel que

$$\mathcal{A} = \mathcal{Z} \times_2 \mathbf{U}_{personnes} \times_3 \mathbf{U}_{BOWs} \tag{3.7}$$

Le tenseur $\mathcal{D}$ peut être obtenu par cette équation :

$$\mathcal{D} = \mathcal{A} \times_1 \mathbf{U}_{actions} \tag{3.8}$$

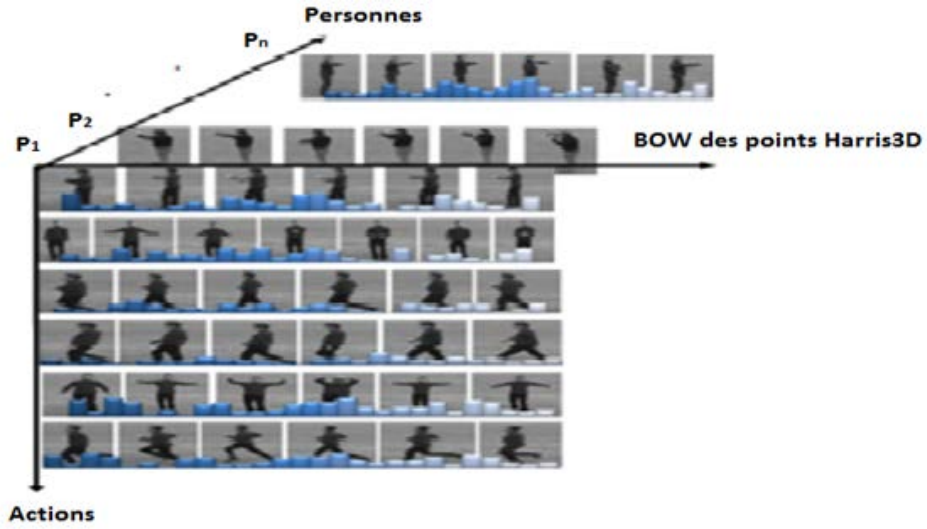


FIGURE *3.6: Illustration de tenseur d'action $\mathcal{D}$.*

Chaque sous-tenseur $\mathcal{A}_i$ contient les informations relatives aux actions effectuées par la personne i. Supposons que $\mathcal{D}_{pa}$ soit un BOW associé à l'action a effectuée par la personne p. On a donc :

$$\mathcal{D}_{pa} = \mathcal{A}_p \times_1 \mathbf{U}_{actions(a)} \tag{3.9}$$

Donc la signature $\mathbf{U}_{actions(a)}$ de l'action a effectuée par une personne p est égale à :

$$\mathbf{U}_{actions(a)} = \mathcal{A}_p^T \times_1 \mathcal{D}_{pa} \tag{3.10}$$

Pour reconnaître l'action d'une nouvelle séquence $\mathbf{d}$ dont on ne connaît pas la personne associée, on doit :

- calculer pour chaque personne p des données d'apprentissage la projection $\mathbf{a}$ de $\mathbf{d}$ dans l'espace d'action :

$$\mathbf{a} = \mathcal{A}_p^T \times_1 \mathbf{d} \tag{3.11}$$

- comparer le vecteur $\mathbf{a}$ aux lignes de la matrice $\mathbf{U}_{actions}$ et retenir l'action a^* correspondant au meilleur appariement (matching). En d'autres termes, on doit retenir le vecteur $\mathbf{U}_{actions(a^*)}$ qui donne la distance minimale $\|\mathbf{a} - \mathbf{U}_{actions(a^*)}\|$.

3.2.2.2 Expérimentation

A. Évaluation sur la base Weizmann

Nous avons évalué cette modélisation tensorielle sur la base de données « Weizmann » [Blank et al., 2005] qui comprend des vidéos de réalisation de 10 actions effectuées par 9 per-

sonnes. Les actions sont : course, marche, saut américain, marche avec un pied, salut de main (1 et 2 mains), saut avec 2 pieds (en se déplaçant et sur place), marche latérale et courbement.

Nous avons utilisé le protocole de test « leave-one-out » pour calculer le taux de reconnaissance. En effet, ce taux de reconnaissance reflète la capacité les vecteurs propres des descripteurs des Harris-3D (EigenHarris3D) de modéliser l'action indépendamment de l'identité de la personne. Dans nos expériences, nous avons utilisé plusieurs tailles de BOW. Pour chaque taille, nous avons effectué l'algorithme k-means à plusieurs reprises et nous avons calculé la moyenne des taux de reconnaissance obtenus en utilisant les codebook générés. Le BOW de taille 500 a donné le résultat optimal (Tableau 3.1). Ainsi, le tenseur de dimension $10 \times 8 \times 500$ a été retenu. Ce tenseur a une dimension beaucoup plus faible que ceux utilisés dans des travaux utilisant des primitives holistiques [Khadem and Rajan, 2009, Zhang et al., 2010]. Nous avons obtenu, avec notre modélisation tensorielle des points Harris-3D, un taux de reconnaissance de 86,7%. Ce taux est plus important que celui obtenu par une modélisation tensorielle basée sur les silhouettes [Khadem and Rajan, 2009] qui a obtenu un résultat de 80,25% sur la même base de données « Weizmann ». Ce résultat montre que, par rapport aux représentations holistiques utilisées dans [Khadem and Rajan, 2009], les points Harris-3D sont plus discriminants. De plus, comme le montre le Tableau 3.1, la modélisation par tenseurs (qui est, en fait, une ACP Multiple ou ACPM) a donné un taux de reconnaissance meilleur que celui donné par une ACP classique. Cela montre que, bien que les points Harris-3D caractérisent principalement les actions, ils sont liés légèrement à l'identité des personnes.

Pour confirmer cette observation, nous avons utilisé, pour reconnaître les personnes, la même modélisation tensorielle que pour les actions, en considérant cette fois le mode personne $\mathbf{U}_{personnes}$. Nous avons obtenu un taux de reconnaissance de 25%. Ce taux de reconnaissance est le double de celui qu'on aurait obtenu par un classifieur aléatoire (11% environ), ce qui montre que les points Harris-3D peuvent être sensibles à l'identité et aux caractéristiques des personnes. Cependant, comme le taux indiqué ci-dessus est faible, on peut conclure que les points Harris-3D ne codent que légèrement ces dernières.

Taille de codebook	ACPM	ACP
400	84.1	81.9
500	86.7	81.1
600	84.1	79.3
700	83.7	72.6
800	84.4	74.8
900	85.5	72.2

TABLE *3.1: Taux de reconnaissance obtenu pour la base Weizmann*

B. Évaluation sur la base KTH

La base de données KTH [Schuldt et al., 2004] contient 6 actions simples (marche, course, jogging, boxe, applaudissement et salut la main) effectuées par 25 personnes dans 4 scénarios différents. Nous avons retenu 16 personnes pour la phase d'apprentissage et 9 pour le test, en considérant le scénario des actions effectuées à l'extérieur. Le tableau 3.2 présente les ré-

sultats obtenus pour différentes tailles de codebook. Le résultat optimal est obtenu avec une ACPM pour une taille de codebook de 800. Le tenseur retenu a donc une taille de $6 \times 16 \times 800$.

Comme pour la base Weizmann, la modélisation tensorielle est légèrement plus performante que l'ACP. Cela confirme que les points d'intérêt spatio-temporels n'encodent que légèrement l'identité de la personne. Dans l'ensemble, on constate que les résultats obtenus sur la base de données KTH sont similaires et confirment ceux obtenus sur la base de données Weizmann.

Taille de codebook	ACP	ACPM
100	78.4	80.9
200	85.8	87.7
300	80.9	88.2
400	87.7	90.8
500	88.9	91.4
600	88.9	90.1
700	88.9	90.1
800	86.4	91.4
900	80.2	90.1
1000	82.1	90.1

TABLE *3.2: Taux de reconnaissance obtenu pour la base KTH*

3.3 Points denses

Comme décrit dans [Wang et al., 2011a], pour extraire les points denses, on considère une grille de points espacés de w pixels l'un de l'autre. Cet échantillonnage est réalisé d'une façon régulière pour des échelles spatiales variables d'une manière séparée. Cela garantit que les points extraits couvrent toutes les positions spatiales et aussi les échelles sélectionnées. Les points situés dans des régions n'ayant aucune structure, c.-à-d. des régions homogènes, sont supprimés. Ces points correspondant à des matrices d'auto-corrélation de faibles valeurs propres. Un seuil T sur les valeurs propres est défini pour chaque trame I par :

$$T = 0.001 \times \max_{i \in I} \min(\lambda_i^1, \lambda_i^2) \tag{3.12}$$

$(\lambda_i^1, \lambda_i^2)$ sont les valeurs propres du point i dans l'image I. Comme le montre la Figure 3.7, la plupart des points appartenant à des régions homogènes sont supprimés.

3.3.1 Trajectoires des points denses

Les points denses sont suivis pour chaque échelle spatiale séparément. Pour chaque trame I_t, le flux optique dense $w_t = (u_t, v_t)$ est calculé par rapport à la trame I_{t+1} , u_t et v_t étant les composantes horizontale et verticale du flux optique. Étant donné un point $P_t = (x_t, y_t)$ dans

FIGURE *3.7: Visualisation des points denses après l'application du critère de suppression des points appartenant aux régions homogènes [Wang et al., 2011a].*

une trame I_t, la position de son successeur dans la trame I_{t+1} est obtenue par l'application d'un filtre médian au flux optique w_t :

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * w_t)|_{(x_t, y_t)} \tag{3.13}$$

où M est le noyau du filtre médian. La taille de ce noyau est de 3x3 pixels. L'utilisation du filtre médian permet un lissage qui, à son tour, garantit un suivi relativement robuste des types de mouvements rapides et irréguliers. Les points des différents trames sont ensuite concaténés pour former les trajectoires $(P_t, P_{t+1}, P_{t+2}, \dots)$. Comme ces derniers ont tendance à diverger de leurs positions initiales durant le processus de suivi, on limite la longueur des trajectoires à L trames. Empiriquement, on fixe la longueur de la trajectoire à 15 trames. Pour chaque trame, s'il n'y a pas de points à suivre dans le voisinage $W \times W$, un nouveau point est échantillonné et ajouté au processus de suivi. Notons aussi que les trajectoires statiques sont supprimées dans une étape de post-traitement puisque elles ne comprennent pas d'information sur le mouvement. Les trajectoires avec des déplacements brusques sont aussi supprimées. En effet, si le déplacement entre deux trames consécutives est plus important de 70% que la majorité des déplacements des différentes trajectoires, on le considère comme déplacement brusque.

3.3.2 Descripteur des trajectoires denses

Pour décrire les trajectoires des points denses, on utilise un descripteur de forme de trajectoire et aussi des informations d'apparence et de mouvement extraites à partir d'un volume entourant le voisinage d'une trajectoire. Nous détaillons le calcul des différents descripteurs dans ce qui suit.

Descripteur de forme de trajectoire : Étant donné une trajectoire de longueur L, on décrit le modèle de déplacement des trajectoires par la séquence $(\Delta P_t, ..., \Delta P_{t+L-1})$ où $\Delta P_t = (P_{t+1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t)$. Ce vecteur est normalisé par la somme des amplitudes de vecteurs

de déplacement :

$$\mathbf{T} = \frac{(\Delta P_t, ..., \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} \|\Delta P_j\|} \tag{3.14}$$

Descripteur d'apparence et de mouvement : Outre les informations de forme de trajectoire, des descripteurs d'apparence et de mouvement sont calculés dans un volume 3D aligné avec la trajectoire. La taille du volume est $N \times N \times L$. Ce dernier est divisé en une grille spatio-temporelle de taille $n_\sigma \times n_\sigma \times n_\tau$. On utilise les mêmes valeurs des paramètres que Wang et al. [Wang et al., 2011a] : $N = 32; n_\sigma = 2; n_\tau = 3$. Pour chacune des cellules, on calcule les descripteurs d'apparence et de mouvement.

Pour encoder l'apparence, on calcule les histogrammes HOGs . Ce descripteur se base sur une quantification de 8 directions de gradient. Les histogrammes HOFs et MBHs sont utilisés pour encoder le mouvement (Figure 3.8). Le HOF capture les informations du mouvement local. Ce type d'histogramme est constitué de 9 composants : 8 orientations de flux optique et une composante supplémentaire qui représente les pixels dont l'ampleur du flux optique est très petite (inférieure à un seuil donné). Vu que le flux optique représente le mouvement absolu entre deux images, ce descripteur encode ainsi le mouvement de plusieurs sources, c'est-à-dire le mouvement de l'objet de premier plan et le mouvement de caméra de fond. Si le mouvement de la caméra est considéré comme mouvement d'action, il peut corrompre la classification d'action. Pour remédier à ce problème, on utilise les MBHs. En effet, les MBHs ont été proposés par Dalal et al. [Dalal et al., 2006] pour la détection humaine en calculant séparément les dérivés des composantes horizontales et verticales du flux optique. Ces derniers sont ensuite quantifiés par rapport à différentes orientations. On obtient ainsi deux histogrammes de 8 composantes correspondant respectivement aux MBHx et MBHy. Ces deux composantes de l'histogramme sont ensuite normalisées séparément avec la norme L_2. Etant donné que le MBH représente le gradient du flux optique, le mouvement de la caméra et les mouvements constants sont supprimés alors que les changements au niveau du flux optique (mouvements frontière) sont maintenus. Ainsi, le MBH est plus robuste au mouvement de la caméra que le HOF et il est donc plus discriminant pour la reconnaissance de l'action. Une comparaison expérimentale des différents descripteurs est présentée dans le chapitre 5.3.2.

3.3.3 Comparaison avec les Harris-3D

Les points d'intérêt Harris-3D correspondent à des points ayant des variations locales significatives en temps et en espace. Ces représentations locales gèrent les occlusions de façon efficace et elles ont été utilisées avec succès dans de nombreuses tâches de reconnaissance de mouvement dans des scènes complexes et réelles [Laptev et al., 2008]. L'une de leurs principales caractéristiques est qu'ils sont peu denses (parcimonieux). Ainsi, pour les activités comprenant peu de mouvement, le nombre des points d'intérêt extraits est relativement faible. En revanche, l'échantillonnage des points denses produit un grand nombre de points, typique-

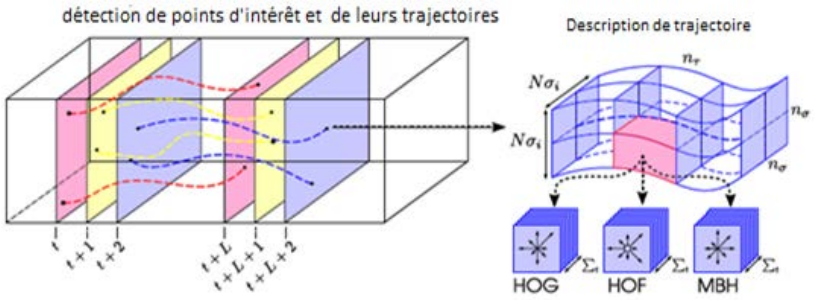


FIGURE *3.8: Descripteurs d'apparence et de mouvement calculés dans un volume 3D aligné avec les trajectoires des points denses [Wang et al., 2011a].*

ment 15 à 20 fois plus que les points Harris-3D. Bien qu'on soit amené à traiter une grande quantité de points denses relativement au nombre réduit de points Harris-3D, les points denses permettent d'avoir une représentation assez informative même pour les activités de faible amplitude. De plus, la considération des points denses et du descripteur MBH donne des résultats assez satisfaisants pour des bases de données assez variées.

3.4 Extraction des primitives à partir des séquences vidéo

Étant donné que les actions humaines sont des séries temporelles de caractéristiques locales, il nous a paru naturel d'utiliser des modèles séquentiels pour les reconnaître. Ainsi l'aspect temporel des actions et, plus particulièrement, les relations temporelles entre les caractéristiques locales peuvent être exploités. Pour ce faire, il faut utiliser un vecteur de primitives de taille fixe à chaque instant ; or, le nombre de nos primitives de base, i.e. les points d'intérêt varie d'une trame à l'autre. Pour contourner ce problème, nous divisons la séquence vidéo en segments locaux de taille fixe L. Chacun de ces segments est représenté par un BOW de ses primitives locales ; ensuite, un classifieur SVM de bas niveau est appliqué à ces BOWs locaux pour les convertir en des vecteurs de probabilités conditionnelles des actions. De cette manière, nous représentons chacune des actions comme une suite temporelle de segments décrits par des vecteurs de probabilités, obtenus via le classifieur de bas niveau. Chaque segment porte des informations concernant la classe d'action à laquelle il appartient. Mais la quantité d'informations diffère d'un segment à l'autre en fonction de la quantité de mouvement véhiculée par le segment en question. De ce fait, on peut trouver des segments très informatifs où l'information mutuelle entre les segments et les classes est assez importante. Il s'agit des segments à faible entropie. Mais on peut aussi trouver des segments beaucoup moins informatifs qui se caractérisent par une forte entropie. Pour notre représentation, si le segment est fortement discriminant, la probabilité conditionnelle de la classe associée est importante alors que les probabilités des autres classes sont faibles. Ainsi, ce type de segment est représenté par des vecteurs de probabilité de faible entropie. En revanche, si un segment ne porte aucune information sur la classe

en question, les probabilités des différentes classes sont similaires. Prenons comme exemple explicatif le cas des activités de la vie quotidienne réalisées pendant une durée raisonnable. Pour ce type d'activités, il peut y avoir des segments où aucun mouvement ne se produit. Ces segments n'apportent donc aucune information sur l'activité réalisée et nous voulons que cela se traduise par la génération d'un vecteur de probabilités qui ne favorise aucune classe. Ainsi, cette ambiguïté se manifeste à travers un vecteur de probabilité de forte entropie. Il est important de noter que, malgré le faible pouvoir discriminant de ces segments, ceux-ci peuvent être pertinents pour la reconnaissance de l'activité. En effet, leurs occurrences, ainsi que leurs relations temporelles avec les autres segments, diffèrent d'une classe à une autre. Grace à ces deux caractéristiques, les segments de ce type sont assez précieux pour la prise de décision au niveau de la séquence. De ce fait, contrairement aux approches à base de trame clés, les segments contenant des trames qui ne portent pas assez d'informations sur la nature de l'activité ne sont pas négligés parce que leur apparition dans le temps est, en soi, un facteur discriminant pour la reconnaissance de l'activité. De plus, si on prend l'exemple des segments contenant un mouvement de lever du bras, le vecteur de probabilité de sortie favorise des activités telles que boire ou appeler à l'aide d'un téléphone cellulaire et sanctionne les activités qui n'incluent pas un tel mouvement. Ce type de vecteurs est discriminant car il transmet un niveau d'ambiguïté plus petit.

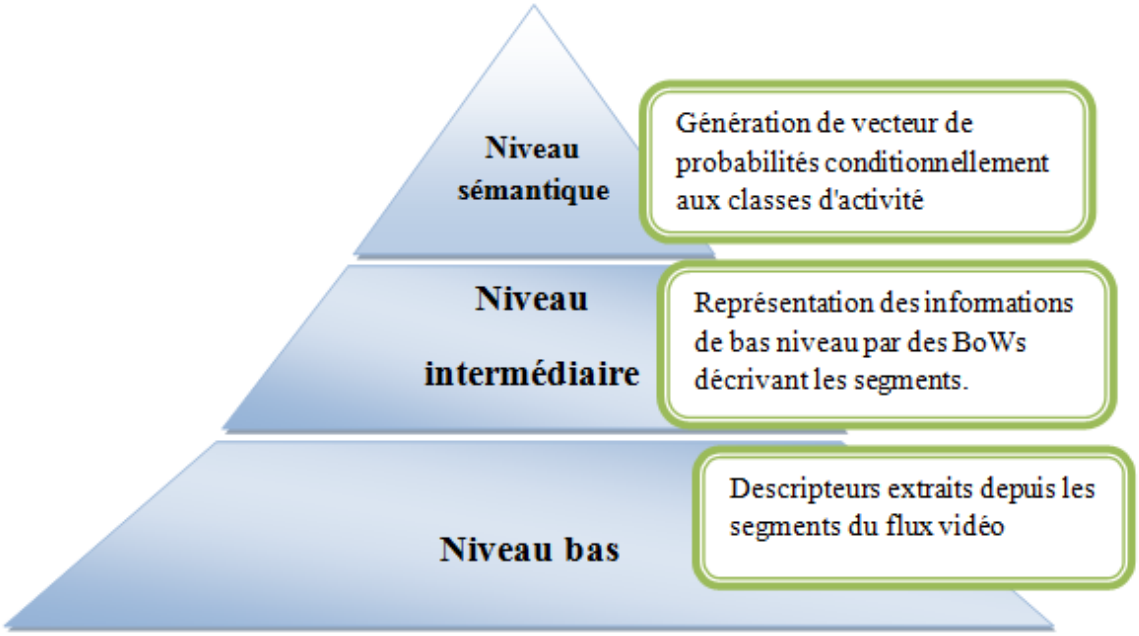


FIGURE *3.9: Structure générale de l'approche proposée.*

Globalement, après avoir segmenté temporellement la séquence vidéo, nous suivons une démarche pyramidale en trois niveaux, telle qu'elle est illustrée dans la Figure 3.9, pour extraire nos primitives :

- **Niveau bas** : La première étape porte sur l'extraction des points d'intérêt. Ces points sont par la suite suivis et leurs trajectoires sont extraites.
- **Niveau intermédiaire** : les BOWs des points d'intérêt associés aux différents segments temporels sont construits.
- **Niveau sémantique** : Les probabilités locales de différentes activités sont obtenues grâce au SVM.

3.4.1 Extraction des primitives de bas niveau

Nous utilisons une fenêtre glissante pour diviser la vidéo en segments de taille fixe avec chevauchement. Nous considérons une taille de fenêtre de 30 trames et un taux de chevauchement de 50% pour tenir compte de la corrélation entre les trames. Ainsi, le nombre de segments est proportionnel à la longueur de la séquence vidéo. Pour chaque segment, nous extrayons des primitives basées sur les points d'intérêt.

Nous avons utilisé, dans un premier temps, les trajectoires des points Harris-3D comme primitives de base pour reconnaître des activités dans des vidéos à basse résolution (160x120). Ces travaux ont été motivés par les bons résultats donnés par ces primitives lors de leur utilisation pour la reconnaissance des mouvements dans des scènes complexes et réelles. Les points Harris-3D sont généralement décrits par des histogrammes de HOG et de HOF. Dans notre travail, ces points d'intérêt sont décrits à travers leurs trajectoires. Ces derniers fournissent une riche description spatio-temporelle locale pour la reconnaissance des activités.
Après avoir extrait les points Harris-3D des séquences vidéo, on effectue un suivi de ces points à l'aide de l'algorithme de $Kanade-Lucas-Tomasi$ (KLT) pendant L trames. Comme décrit dans [Baker and Matthews, 2004], cet algorithme cherche pour chaque point son correspondant sur l'image suivante I. Chaque point est caractérisé par une texture T extraite d'une fenêtre carrée centrée sur la position du point suivi. La méthode optimise les paramètres $\mathbf{p}$ dans le but de minimiser la fonction de coût suivante :

$$\sum_{\mathbf{x}} [I(\mathbf{W}(\mathbf{x};\mathbf{p})) - T(\mathbf{x})]^2 \tag{3.15}$$

où $\mathbf{x} = (x,y)$ est un point appartenant à une fenêtre centrée sur la position du point suivi ; c'est la fenêtre dans laquelle la texture est calculée, $\mathbf{W}$ est la fonction de déformation (dans ce cas, la translation) qui donne la correspondance entre les points des deux fenêtres (images) et $\mathbf{p}$ représente les paramètres de cette fonction. Pour optimiser l'équation 3.15, on l'exprime tout abord comme un problème de mise à jour itérative en remplaçant $\mathbf{p}$ par $\mathbf{p}+\Delta\mathbf{p}$. L'équation peut être ainsi réécrite sous forme de série de Taylor :

$$\sum_{\mathbf{x}} [I(\mathbf{W}(\mathbf{x};\mathbf{p}+\Delta\mathbf{p})) - T(\mathbf{x})]^2 \approx \sum_{\mathbf{x}} \left[I(\mathbf{W}(\mathbf{x};\mathbf{p})) + \nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}} - T(\mathbf{x}) \right]^2 \tag{3.16}$$

où $\nabla I = (\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y})$.
Considérons maintenant la dérivée partielle de l'équation 3.16 par rapport à $\Delta\mathbf{p}$, on obtient :

$$\sum_{\mathbf{x}} \left[\nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}}\right]^T \left[I(\mathbf{W}(\mathbf{x};\mathbf{p})) + \nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}} \Delta\mathbf{p} - T(\mathbf{x})\right] \tag{3.17}$$

en réduisant cette équation à zéro et en résolvant pour le paramètre $\partial\mathbf{p}$, on obtient :

$$\Delta\mathbf{p} = H^{-1} \sum_{\mathbf{x}} \left[\nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}}\right]^T [T(\mathbf{x}) - I(\mathbf{W}(\mathbf{x};\mathbf{p}))] \tag{3.18}$$

où $H^{-1} = \sum_{\mathbf{x}} \left[\nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}}\right]^T \left[\nabla I \frac{\partial \mathbf{W}}{\partial \mathbf{p}}\right]$ est la matrice Hessienne. En mettant à jour d'une façon itérative le paramètre $\Delta\mathbf{p}$ selon l'équation 3.18, on peut trouver le déplacement optimal qui minimise la fonction de coût 3.15.
Le résultat de l'application de cet algorithme aux points Harris-3D est illustré par la figure 3.10.

FIGURE *3.10: Exemple des trajectoires des Harris-3D obtenues par l'application de l'algorithme KLT.*

Ces trajectoires sont ensuite décrites par les séquences temporelles composées par :

- Les coordonnées spatiales des points : $\mathbf{P}_t = < x_t, y_t >$
- L'angle du déplacement : $\theta_t = \arctan\left(\frac{y_t - y_{t-1}}{x_t - x_{t-1}}\right)$
- L'amplitude de déplacement : $\mathbf{V}_t = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2}$ où $t = 1..T$ est l'index temporel.

Nos premières expérimentations de classification (globale) fondées sur ces primitives, effectuées sur une base de données de basse résolution, ont donné des résultats assez compétitifs. Toutefois, étant donné l'aspect parcimonieux du détecteur Harris-3D, le nombre de points extraits pour des activités de faible amplitude de mouvement s'avère insuffisant pour le codage du

mouvement. En effet, en analysant les vidéos non reconnues, nous avons constaté que, dans le cas d'activités comprenant peu de mouvement, le nombre des points Harris-3D est faible. Dans ce cas, l'activité ne peut être codée de manière efficace à partir de cette petite quantité d'informations extraites. Pratiquement, on a remarqué que, pour quelques vidéos, une dizaine de points d'intérêt seulement ont été détectés, ce qui est insuffisant pour avoir une bonne représentation de l'activité.

Pour remédier à ce problème, nous avons opté pour l'utilisation des points denses décrits dans la section 3.3. Ces points d'intérêt permettent une représentation assez dense des activités. De plus, comme cela est démontré dans [Wang et al., 2011a], la qualité des trajectoires obtenues par la méthode de suivi dense est meilleure que celle obtenue par l'algorithme KLT. Ces points d'intérêt denses correspondent à un ensemble de points échantillonnés d'une manière uniforme pour différentes échelles. Dans nos expériences, nous avons considéré 8 échelles spatiales. Pour chaque échelle, les points d'intérêt sont suivis sur 15 trames en appliquant le filtre médian aux flux optiques denses. Ce dernier est obtenu par l'algorithme décrit dans [Farnebäck, 2003]. Une fois les trajectoires générées, les descripteurs de mouvement sont calculés dans un volume spatio-temporel autour de chacune des trajectoires. Quatre descripteurs sont généralement utilisés dans les travaux de l'art qui utilisent les points denses, à savoir les HOG, HOF, MBH et les vecteurs de déplacement. Comme le montrent les expériences présentées dans le chapitre 5, MBH s'est avéré comme le descripteur le plus discriminant pour la reconnaissance d'actions. En effet, MBH représente le gradient de flux optique. Ainsi, les mouvements constants tels que ceux d'une camera sont éliminés, alors que les mouvements correspondant au changement de flux optique sont maintenus. De ce fait, MBH est assez fiable par rapport aux mouvements de caméra, ce qui renforce son aspect discriminatif. Nous optons donc pour l'utilisation du descripteur MBH calculé autour de chacune des trajectoires pour différentes échelles comme primitives de base de nos segments temporels.

3.4.2 Extraction des primitives de niveau intermédiaire

Après l'extraction de ces primitives, nous construisons les BOWs locaux pour chacun des segments. Nous construisons d'abord un dictionnaire visuel avec l'algorithme de clustering k-means. Pour limiter la complexité, nous appliquons k-means à un sous-ensemble de données d'entraînement sélectionnées aléatoirement. Ensuite, pour chacun des segments temporels, chaque descripteur est assigné au mot visuel le plus proche en utilisant la méthode du plus proche voisin. Les histogrammes d'occurrence de ces mots visuels sont utilisés comme des représentants locaux de segments. La séquence vidéo est ainsi représentée par une séquence de T observations, où la i-ème observation correspond au BOW local extrait du segment S_i, et T est le nombre total des segments.

3.4.3 Extraction des primitives de haut niveau

L'idée principale est d'appliquer un classifieur discriminant pour fournir les probabilités attribuées à chacune des activités conditionnellement au BOW local extrait du segment en cours. L'utilisation d'un classifieur de bas niveau permet de réduire considérablement la dimension du vecteur de primitives, tout en renforçant l'aspect discriminant du classifieur final.
Nous avons opté pour l'utilisation de SVM, vue sa forte capacité de généralisation et de discrimination. Les SVMs reposent sur deux notions : celle de marge maximale et celle de la fonction noyau. Ils cherchent une fonction de forme linéaire (un hyperplan, droite dans le cas de deux dimensions) qui sépare le mieux deux classes. Ils sont donc des systèmes d'apprentissage qui utilisent un espace d'hypothèses de fonctions linéaires dans un espace de caractéristiques à haute dimension. Dans le cas d'un problème non linéaire, on utilise une fonction noyau pour projeter les données dans un espace de plus grande dimension où elles seront linéairement séparables.
Nous utilisons le classifieur SVM pour convertir le BOW local de chacun des segments en un vecteur de probabilités conditionnelles des classes d'activités. Cela permet de passer d'un vecteur de primitives de dimension importante (généralement supérieur à 1000) à un vecteur de primitives de dimension réduite qui est égale au nombre de classes d'activité considérées. On a opté pour un SVM non-linéaire avec un noyau χ^2 exponentiel.

$$K(\mathbf{H}_i,\mathbf{H}_j) = \exp(-\frac{1}{A}\chi^2(\mathbf{H}_i,\mathbf{H}_j)) = exp(-\frac{1}{2A}\sum_{k=1}^{D}\frac{(\mathbf{H}_{ik}-\mathbf{H}_{jk})^2}{\mathbf{H}_{ik}+\mathbf{H}_{jk}}) \tag{3.19}$$

où A est la moyenne des distances χ^2 entre les différents couples de BOWs associés aux segments du corpus d'entraînement. Nous considérons un SVM probabiliste pour fournir les probabilités conditionnelles des classes au lieu des scores. En effet, avoir des primitives entre 0 et 1 est plus adéquat pour une initialisation aléatoire des paramètres du HCRF dans l'intervalle [0, 1]. Nous allons montrer, dans le chapitre 5, l'impact des deux configurations SVM (scores versus probabilités) sur la performance globale de notre système de reconnaissance. Enfin, grâce à l'application du SVM à chaque segment d'entrée, une vidéo est finalement convertie en une séquence de T vecteurs d'observation de dimension C où C est le nombre de classes.

3.5 Conclusion

Nous avons proposé des primitives de haut niveau adaptées à l'analyse du comportement humain d'une manière séquentielle. Ces primitives sont basées sur les STIPs comme information de bas niveau. Nous avons commencé par la présentation des techniques pour l'extraction des points d'intérêts spatio-temporels utilisés dans nos travaux pour caractériser les actions.

Dans la deuxième section, nous avons étudié l'efficacité des points d'intérêts spatiaux temporels et leur capacité à caractériser principalement le mouvement indépendamment des différents facteurs de variabilité qui peuvent affecter les actions et les activités. Les STIPs semblent

être une technique efficace de représentation de mouvement dans des conditions réelles. Nous avons aussi introduit dans la section 3.4 le processus de génération de nos primitives de haut niveau à savoir la séquence de vecteurs de probabilités conditionnelles générées par SVM. Ce dernier a été utilisé comme classifieur de bas niveau.

Dans le cadre pratique d'une application de reconnaissance d'activités, le chapitre suivant expose l'approche de classification.

CHAPITRE 4

Modèle de reconnaissance hybride SVM-HCRF

Sommaire

4.1 Introduction

À l'issue des étapes de détection et de description, nous disposons d'une représentation des séquences vidéo par des séquences de vecteurs de probabilités. Ces vecteurs sont issus de la classification des BOW locaux par SVM. Concernant les primitives de base, les points d'intérêt denses ont montré leur efficacité pour la représentation de mouvement dans des conditions réelles. Ces points sont décrits par les histogrammes MBH calculés autour de leurs trajectoires.

Nous nous intéressons, dans ce chapitre, à la phase de classification de ces séquences temporelles. Nous avons choisi de nous appuyer sur les modèles graphiques probabilistes. Cette famille de modèles présente une solution pratique pour traiter les données séquentielles. Plusieurs modèles graphiques probabilistes ont été proposés, parmi lesquels on trouve les HCRF. Cette approche séquentielle discriminative a récemment montré des performances relevant de l'état de l'art dans de très nombreux domaines. En se basant sur cette méthode, nous proposons une méthode de classification hybride "SVM-HCRF" pour classifier les données séquentielles. SVM est utilisé comme classifieur local qui fournit en sortie une séquence de vecteurs de probabilités conditionnelles des activités. Cette dernière est considérée comme l'entrée au HCRF.

Après avoir introduit la méthode de référence "champs aléatoires conditionnels" en section 4.2, nous présenterons les fondements théoriques des champs aléatoires conditionnels cachés

dans la section 4.2.2. Nous exposerons, dans la deuxième partie, notre approche de reconnaissance. Nous comparerons ensuite notre approche avec les travaux récents de l'état de l'art portant sur la classification séquentielle pour la reconnaissance des actions.

4.2 Modèles graphiques probabilistes pour la classification de séquences

4.2.1 Les champs aléatoires conditionnels

Les champs aléatoires conditionnels [Lafferty et al., 2001] sont des modèles graphiques non dirigés, ayant pour objectif d'étiqueter et de segmenter les séquences à partir d'une approche probabiliste conditionnelle. Pour différents domaines, les CRFs ont montré leur supériorité par rapport aux modèles génératifs, tels que les HMMs traditionnellement utilisés pour ce genre de problème. Le principal avantage des CRF est leur capacité à modéliser directement une probabilité discriminante de l'étiquetage sachant les observations alors que les modèles génératifs sont amenés à résoudre un problème plus difficile que le problème original : apprendre une probabilité jointe plutôt que d'apprendre la probabilité de l'étiquetage sachant les données observées. Ces modèles conditionnels permettent aussi de relaxer les hypothèses d'indépendance conditionnelle des observations. De plus, ils se sont montrés plus performants que le modèle graphique discriminatif MEMM. En effet, les CRFs permettent d'éviter le problème du "label bias" rencontré avec les MEMM.

Supposons donnée une séquence d'entrée $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_T\}$ ainsi qu'une séquence de labels (étiquettes) à prédire $Y = \{y_1, ..., y_T\}$. Les modèles graphiques non orientés, dont CRF est un cas particulier, peuvent être définis comme suit : Soit $G = (V, E)$ un graphe non dirigé où V est l'ensemble des nœuds et E l'ensemble des arcs. Chaque nœud $v \in V$ est associé à une variable aléatoire y_v dans Y. Ce graphe est appelé graphe d'indépendance. On dit que $(\mathbf{X}, Y)$ est un champ aléatoire conditionnel si chaque variable aléatoire y_v respecte la propriété de Markov suivante :

$$P(y_v|\mathbf{X}, y_w, w \neq v) = P(y_v|\mathbf{X}, y_w, w \sim v) \tag{4.1}$$

où $w \sim v$ signifie que w et v sont voisins dans G. Ainsi, chaque variable aléatoire y_v ne dépend que de $\mathbf{x}$ et de ses voisins dans le graphe d'indépendance. En respectant cette condition d'indépendance, le théorème de Hammersley-Clifford [Hammersley and Clifford, 1971] permet d'exprimer la probabilité conditionnelle comme un produit de fonctions potentiel $\psi_c(y_c, \mathbf{X})$sur tous les sous-graphes complètement connectés, appelés cliques, du graphe d'indépendance.

$$P(Y|\mathbf{X}) = \frac{1}{Z(\mathbf{X})} \prod_{c \in C} \psi_c(y_c, \mathbf{X}) \tag{4.2}$$

où C est l'ensemble des cliques de G et $Z(\mathbf{X})$est un terme de normalisation défini comme suit :

$$Z(\mathbf{X}) = \sum_{Y} \prod_{c \in C} \psi_c(y_c, \mathbf{X}) \tag{4.3}$$

Pour les CRFs, Lafferty [Lafferty et al., 2001] ont proposé de définir la forme de ces fonctions de potentiel comme l'exponentielle d'une somme pondérée de fonctions f_k appelées « primitives » du modèle :

$$\psi_c(y_c, \mathbf{X}) = \exp(\sum_{k} \lambda_k f_k(y_c, \mathbf{X}, c)) \tag{4.4}$$

les λ_k étant les poids associés à chacune de ces fonctions de primitives. Ainsi, comme l'illustre la figure 4.1, les CRFs sont définis avec différents niveaux de dépendances.

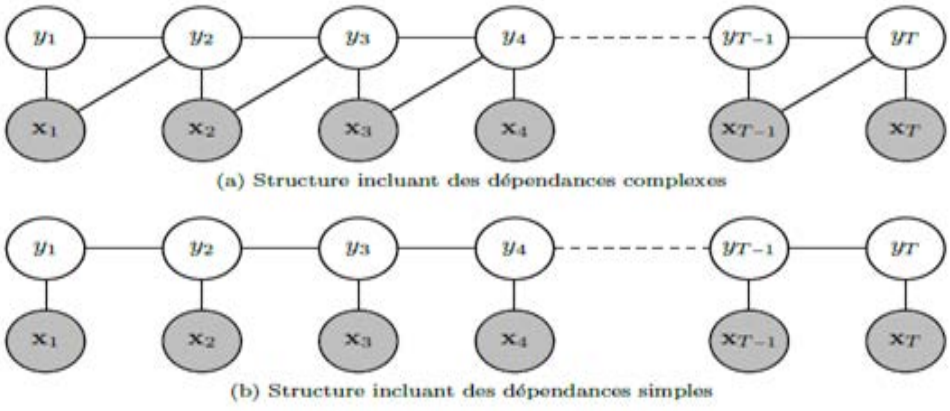


FIGURE *4.1: CRF ayant plusieurs niveaux de dépendance [Vinel, 2013].*

Dans le cas de la structure de dépendance la plus simple (Figure 4.1b), il existe deux types de cliques (Figure 4.2) :

- Les cliques locales qui relient l'observation $\mathbf{x}_t$ à son étiquette y_t pour lesquelles nous notons les fonctions de potentiel s.
- Les cliques de transition qui connectent deux étiquettes successives y_{t-1} et y_t pour lesquelles nous notons les fonctions de potentiel g .

La probabilité conditionnelle de séquence de labels peut donc s'exprimer par l'équation suivante :

$$P(Y|\mathbf{X}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{t,k} \sigma_k g_k(y_{t-1}, y_t, \mathbf{x}, t) + \sum_{t,k} \mu_k s_k(y_t, \mathbf{x}, t)\right) \tag{4.5}$$

où μ et σ sont les vecteurs de poids relatifs aux fonctions de potentiel s et g. Ces paramètres sont estimés lors de la phase d'apprentissage du modèle.

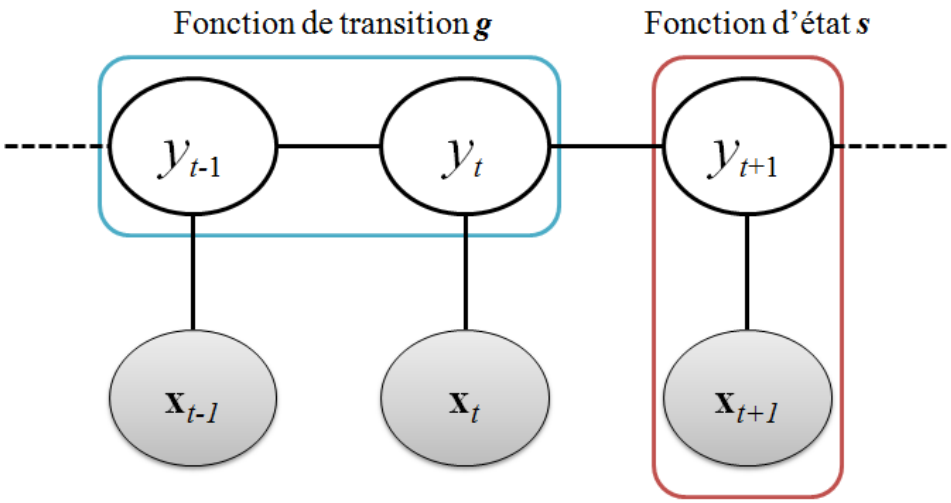


FIGURE *4.2: Illustration des fonctions potentielles de CRF.*

4.2.2 Les champs aléatoires conditionnels à états cachés

Les CRF permettent d'affecter un label à chaque observation de la séquence, et ne sont donc pas adaptés au problème de classification des séquences entières auxquelles on veut affecter un seul label, comme c'est le cas du problème de classification des séquences vidéos contenant une seule action par séquence. De plus, les CRFs ne permettent pas la modélisation de la structure interne des séquences. Ces limitations ont motivé la proposition des champs aléatoires conditionnels cachés [Quattoni et al., 2004]. Les HCRF sont des modèles discriminants affectant un label par séquence. A l'instar des HMM, ces modèles sont capables de modéliser la structure interne de séquences en utilisant des états cachés. La prise en compte d'états cachés permet ainsi d'introduire une structure temporelle dans les entités (caractère, phonème, action élémentaire) à reconnaître. L'objectif est ainsi d'apprendre une fonction d'appariement des observations $\mathbf{x} \in \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m\}$ à un des labels $y \in \mathcal{Y}$, en introduisant des états cachés $h \in \mathcal{H}$ pour représenter de manière compacte la distribution des observations.

La distribution de probabilité conditionnelle $P(y|\mathbf{x};\theta)$ d'un label de classe y étant donnée une séquence d'observations $\mathbf{x}$ avec le vecteur de paramètres θ proposée pour le modèle HCRF est comme suit :

$$P(y|\mathbf{x},\theta) = \sum_h P(y,h|\mathbf{x},\theta) = \frac{1}{Z}\sum_h \exp(\theta.\Phi(y,h,\mathbf{x})) \tag{4.6}$$

Dans l'équation ci-dessus, Z désigne le facteur de normalisation défini par :

$$Z = \sum_{y' \in \mathcal{Y}} \sum_h \exp(\theta.\Phi(y',h,\mathbf{x})) \tag{4.7}$$

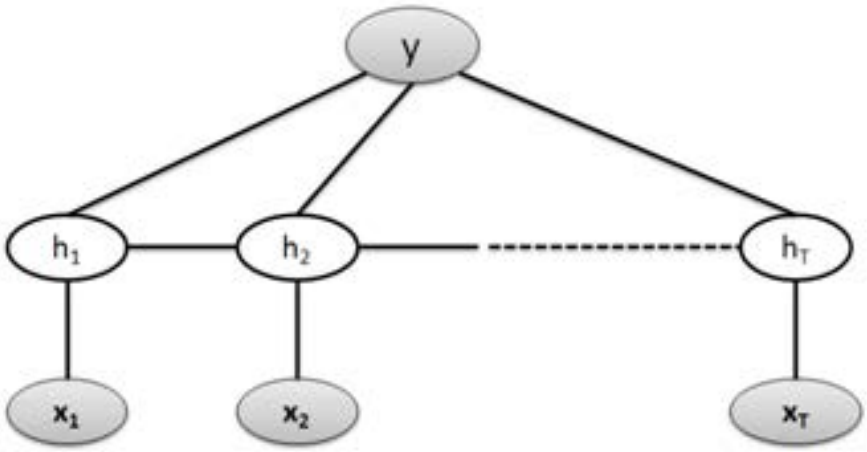


FIGURE *4.3: Illustration du modèle HCRF.*

où $\theta.\Phi(y,h,\mathbf{x})$ est la fonction potentielle définie comme suit :

$$\theta.\Phi(y,h,\mathbf{x}) = \sum_{j\in V} \alpha.\phi(\mathbf{x}_j,h_j) + \sum_{j\in V} \beta.\varphi(y,h_j) + \sum_{(j,k)\in E} \gamma.\psi(y,h_j,h_k) \tag{4.8}$$

Le vecteur de poids θ relatif à la fonction potentielle Φ peut être décomposé en 3 composantes telles que $\theta = [\alpha,\beta,\gamma]$. $\phi(.),\varphi(.)$ et $\psi(.)$ sont les fonctions potentielles qui définissent les "primitives" du modèle (Figure 4.4). La fonction potentielle $\alpha.\phi(\mathbf{x}_j,h_j)$ mesure le degré de compatibilité entre l'observation $\mathbf{x}_j$ et l'état caché h_j, le vecteur de poids α est de dimension $|\mathcal{H}|.\|\mathbf{x}_j\|$. D'une manière similaire $\beta.\varphi(y,h_j)$ modélise la compatibilité entre l'étiquette y et l'état caché h_j, le vecteur de poids β est donc d'une dimension $|\mathcal{Y}|.|\mathcal{H}|$. La fonction potentielle $\gamma.\psi(y,h_j,h_k)$ mesure la compatibilité entre la classe y et le couple d'états cachés h_j et h_k ; où $(j,k) \in E$. E représente les arcs du graphe de dépendance G modélisant la structure de HCRF. Ainsi cette fonction potentielle respecte la structure du graphe G. La dimension du vecteur de poids γ est $|\mathcal{Y}|.|\mathcal{H}|^2$.

Apprentissage du modèle HCRF

Étant donné un corpus d'apprentissage constitué de N séquences étiquetées $\left\{\mathbf{x}^{(n)},y^{(n)}\right\}_{n=1}^{N}$ supposées statistiquement indépendantes, la phase d'apprentissage consiste à estimer les paramètres $\theta = (\theta_1;...;\theta_k)$ qui maximisent la log-vraisemblance conditionnelle des classes connaissant les observations à laquelle est ajoutée une fonction de régularisation. La fonction de coût est définie par :

$$L(\theta) = \sum_n logP(y_n|\mathbf{x}_n,\theta) - \frac{1}{2\sigma^2}\|\theta\|^2 \tag{4.9}$$

Ainsi, on cherche à trouver le vecteur de poids optimal $\theta^* = \arg\max_\theta L(\theta)$ à partir des don-

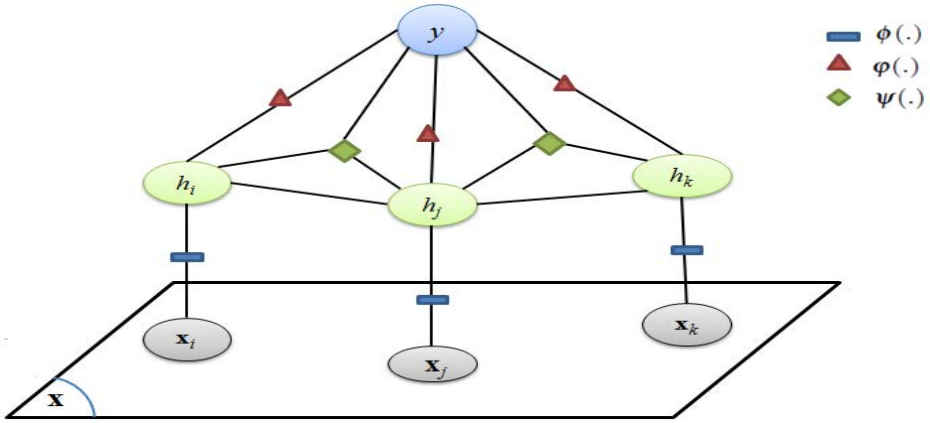


FIGURE *4.4: Illustration des fonctions potentielles de HCRF.*

nées d'entraînement. Cet optimum correspond aux paramètres pour lesquels le gradient s'annule. Toutefois, les paramètres optimaux ne pouvant pas être calculés de façon analytique, des méthodes de descente de gradient sont utilisées. H.Wallach [Wallach, 2003] a montré que la méthode la plus performante, dans ce contexte, est l'algorithme BFGS à mémoire limitée (L-BFGS) [Liu and Nocedal, 1989] pour lequel il est nécessaire de calculer les dérivés partielles de $L(\theta)$ par rapport à chaque paramètre du HCRF. Cependant, comme pour les autres modèles basés sur les états cachés tels que HMM, l'utilisation des états cachés engendre une fonction objective $L(\theta)$ non convexe (contrairement au CRF). Ainsi, l'optimisation ne garantit qu'une convergence vers un maximum local dépendant de l'initialisation du θ.

Dans ce qui suit, nous détaillons le calcul de gradient de $L(\theta)$. On désigne par $L_n(\theta)$ la log-vraisemblance du n^{me} exemple d'entraînement

$$\begin{aligned} L_n &= \log P(y_n|\mathbf{x},\theta) \\ &= \log \frac{\sum_h \exp(\theta.\Phi(\mathbf{x}_n,h,y_n))}{\sum_{y'\in\mathcal{Y}}\sum_h \exp(\theta.\Phi(\mathbf{x}_n,h,y'))} \\ &= \log \sum_h \exp(\theta.\Phi(\mathbf{x}_n,h,y_n)) - \log \sum_{y'\in\mathcal{Y}}\sum_h \exp(\theta.\Phi(\mathbf{x}_n,h,y')) \end{aligned} \tag{4.10}$$

$L(\theta)$ et son gradient par rapport à θ peuvent donc s'écrire ainsi :

$$\begin{aligned} L(\theta) &= \sum_n L_n(\theta) - \frac{1}{2\sigma^2}\|\theta\|^2 \\ \frac{\partial L(\theta)}{\partial \theta} &= \sum_n \frac{\partial L_n(\theta)}{\partial \theta} - \frac{\theta}{\sigma^2} \end{aligned} \tag{4.11}$$

Comme le paramètre θ est composé des vecteurs de poids α, β et γ associés respectivement aux fonctions potentielles $\phi(.)$, $\varphi(.)$et $\psi(.)$ d'une façon linéaire, le gradient de la fonction

$L(\theta)$ est calculé respectivement par rapport à chacun des paramètres du modèle. On commence par calculer $\frac{\partial L_n(\theta)}{\partial \alpha}$:

$$\frac{\partial L_n(\theta)}{\partial \alpha} = \frac{\sum_h \exp(\theta.\Phi(\mathbf{x}_n,h,y_n)).\frac{\partial \theta.\Phi(\mathbf{x}_n,h,y_n)}{\partial \alpha}}{\sum_h \exp(\theta.\Phi(\mathbf{x}_n,h,y_n))} - \frac{\sum_{y'\in\mathcal{Y}}\sum_h \exp(\theta.\Phi(\mathbf{x}_n,h,y')).\frac{\partial \theta\Phi(\mathbf{x}_n,h,y')}{\partial \alpha}}{\sum_{y'\in\mathcal{Y}}\sum_h \exp(\theta\Phi(\mathbf{x}_n,h,y'))}$$

$$= \sum_h P(h|y_n,\mathbf{x}_n,\theta).\frac{\partial \theta.\Phi(\mathbf{x}_n,h,y_n)}{\partial \alpha} - \sum_{y'\in\mathcal{Y}}\sum_h P(y',h|\mathbf{x}_n,\theta)\frac{\partial \theta.\Phi(\mathbf{x}_n,h,y')}{\partial \alpha}$$
$$= \sum_h P(h|y_n,\mathbf{x}_n,\theta).\sum_{j\in V}\phi(\mathbf{x}_{n,j},h_j) - \sum_{y'\in\mathcal{Y}}\sum_h P(y',h|\mathbf{x}_n,\theta)\sum_{j\in V}\phi(\mathbf{x}_{n,j},h_j). \quad (4.12)$$

où

$$P(h|y_n,\mathbf{x}_n,\theta) = \frac{P(y_n,h|\mathbf{x}_n,\theta)}{P(y_n|\mathbf{x}_n,\theta} = \frac{\exp(\theta.\Phi(\mathbf{x}_n,h,y_n))}{\sum_h \exp(\theta.\Phi(\mathbf{x}_n,h,y_n))} \quad (4.13)$$

et

$$P(y_n,h|\mathbf{x}_n,\theta) = \frac{P(y_n,h|\mathbf{x}_n,\theta)}{P(y_n|\mathbf{x}_n,\theta)} = \frac{\exp(\theta.\Phi(\mathbf{x}_n,h,y_n))}{\sum_{y'\in\mathcal{Y}}\sum_h \exp(\theta.\Phi(\mathbf{x}_n,h,y'))} \quad (4.14)$$

Ce gradient peut ainsi se réecrire comme une somme des espérances sous la loi conditionnelle :

$$\frac{\partial L_n(\theta)}{\partial \alpha} = \sum_{j\in V}\left[\mathbb{E}_{P(h_j|y_n,\mathbf{x}_n,\theta)}\phi(\mathbf{x}_{n,j},h_j) - \mathbb{E}_{P(h_j,y|\mathbf{x}_n,\theta)}\phi(\mathbf{x}_{n,j},h_j)\right] \quad (4.15)$$

On peut calculer d'une façon similaire le gradient de $L_n(\theta)$ par rapport aux paramètres β et γ

$$\frac{\partial L_n(\theta)}{\partial \beta} = \sum_{j\in V}\left[\mathbb{E}_{P(h_j|y_n,\mathbf{x}_n,\theta)}\varphi(y_n,h_j) - \mathbb{E}_{P(h_j,y|\mathbf{x}_n,\theta)}\varphi(y,h_j)\right] \quad (4.16)$$

$$\frac{\partial L_n(\theta)}{\partial \gamma} = \sum_{(j,k)\in E}\left[\mathbb{E}_{P(h_j,h_k|y_n,\mathbf{x}_n,\theta)}\psi(y_n,h_j,h_k) - \mathbb{E}_{P(h_j,h_k,y|\mathbf{x}_n,\theta)}\psi(y,h_j,h_k)\right] \quad (4.17)$$

Le calcul de ces espérances est généralement effectué par un algorithme dit Forward-Backward. Ce dernier nécessite plusieurs balayages des séquences d'entrainement. Chaque balayage étant associé à une dimension du vecteur de primitives. De ce fait, le temps d'entrainement du HCRF est relativement long.

4.3 Modèle hybride SVM-HCRF pour la classification

Les modèles HCRFs ont démontré leur performance, dans les problèmes de reconnaissance des séries temporelles, grâce à leur bonne capacité de modélisation des dépendances temporelles qui existent entre les observations et leur aspect discriminant. Ce modèle discriminatif a été appliqué avec succès sur des données réelles issues de plusieurs domaines d'appli-

cation tels que la reconnaissance de la parole [Gunawardana et al., 2005], la reconnaissance d'objets [Quattoni et al., 2004], et aussi la reconnaissance du mouvement humain [Wang et al., 2006,Zhang and Gong, 2010,Chikkanna and Guddeti, 2013]. Pour la reconnaissance du mouvement humain, le HCRF a été principalement utilisé pour la reconnaissance des gestes [Wang et al., 2006, Chikkanna and Guddeti, 2013]. Wang et al. [Wang et al., 2006] ont appliqué le HCRF pour la reconnaissance des gestes de main et de tête. Dans ce travail, la supériorité des HCRF par rapport au HMM et au CRF pour le cas des gestes déjà segmentés a été démontrée. En outre, la capacité des états cachés à modéliser des parties cachées de mouvement a été mise en évidence (Figure 4.5). En revanche, il est important de noter que les auteurs ont utilisé une base de données avec des fonds simples et statiques afin de pouvoir extraire les primitives holistiques utilisées dans ce travail comme entrée du HCRF.

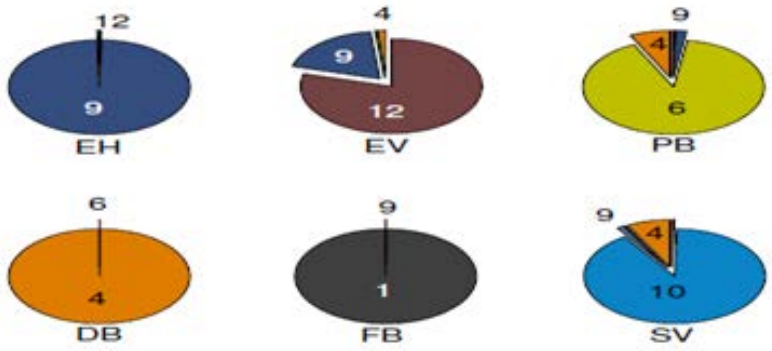


FIGURE *4.5: Distribution des états cachés pour chacun des gestes. Le nombre dans chacune des partitions désigne le label de l'état caché [Wang et al., 2006].*

Dans un travail plus récent [Chikkanna and Guddeti, 2013], le HCRF a été utilisé pour la reconnaissance de gestes plus complexes, à savoir ceux du langage des signes, en temps réel. Les auteurs utilisent des primitives issues de la Kinect.

Il faudrait noter que ces travaux traitent un problème de difficulté moindre que le nôtre, vu qu'ils se limitent à la reconnaissance des actions effectuées par seulement une partie du corps (mains et tête) et réalisées d'une manière conditionnée (on donne des consignes sur la façon de les réaliser).

En effet, dans le cadre de reconnaissance des actions, très peu de travaux ont utilisé l'approche séquentielle. Bien que les CRFs ont été proposés pour la segmentation et reconnaissance conjointe des séquences, il existe des travaux qui ont utilisé le CRF pour la reconnaissance des actions dans des séquences déjà segmentées. Par exemple, dans [Sminchisescu et al., 2006] le CRF a été appliqué à deux types de séquences de primitives : (1) holistiques où, pour chacune des trames, des primitives 2D sont extraites à partir de la silhouette ; (2) fondées sur le modèle cinématique où, pour chacune des frames, on extrait les angles des articulations 3D.

Des travaux plus récents [Wang and Mori, 2008, Liu and Jia, 2008, Wang and Mori, 2009, Zhang et al., 2010] ont appliqué le HCRF qui est bien adapté à la problématique de reconnaissance des séquences temporelles déjà segmentées. Dans [Zhang et al., 2010] par exemple, HCRF a été combiné avec HMM afin d'optimiser les performances. Dans tous ces travaux, les HCRF ont été utilisés avec des primitives holistiques qui permettaient d'avoir des vecteurs de primitives de taille fixe à chaque frame. Ainsi, afin de surmonter les limites de ce type de primitives, à savoir la difficulté de soustraction de fond dans des conditions réelles, ces études se limitent au cas où le fond est relativement facile. De plus, les vecteurs de primitives générées par les approches holistiques sont généralement de dimensionnalité élevée ce qui engendre, dans les cas du modèle HCRF, un temps d'entrainement très important à cause de l'utilisation de l'algorithme Forward-Backward pour le calcul du gradient de $L(\theta)$ pour chaque dimension du vecteur de primitives.

Bien que les STIPs permettent de surmonter les problèmes liés aux fonds complexes, ces primitives n'ont pas été utilisées avec HCRF vu que le nombre de STIPs générés par trame est variable et non structuré. Ils ont été, en revanche, utilisés avec succès dans plusieurs travaux récents abordant des scènes très complexes. Ces travaux sont fondés, dans la plupart des cas, sur les SVMs. Ces derniers sont des classifieurs discriminants avec un bon pouvoir de généralisation. De plus, ils permettent de traiter des vecteurs de dimensionnalité élevée d'une manière efficace à l'aide des fonctions de noyau. Cependant, les SVM ne permettent pas de représenter l'évolution temporelle des observations ; or, cette évolution est informative pour la discrimination des séries séquentielles telles que celles issues des séquences de mouvement. L'information sur l'évolution temporelle réside, dans ce cas, dans la modélisation des actions élémentaires des activités et les transitions entre elles. HCRF est capable de coder ces informations.

Le but est donc d'exploiter les avantages des HCRF et des SVM en un seul modèle hybride SVM-HCRF. Ce modèle permet d'exploiter les avantages des SVM et des HCRF et de compenser leurs points faibles : Exploiter la modélisation des dépendances temporelles par les HCRF et le pouvoir discriminant et généralisant des SVM. Ces derniers sont utilisés comme un classifieur de bas niveau afin d'extraire les probabilités conditionnelles des actions pour chacun des segments locaux. De ce fait, HCRF traite directement des informations de haut niveau générées par SVM à partir des BOWs locaux. HCRF a donc comme entrée des vecteurs de dimension beaucoup moins élevée relativement à celle des BOWs ce qui permet ainsi de réduire considérablement le temps d'entrainement de HCRF. De plus, grâce à cette modélisation hybride HCRF exploite directement des informations sémantiques, à savoir les probabilités de chacune des activités au niveau local, ce qui aide à améliorer la performance de notre système de reconnaissance. Notre modèle permet aussi d'éviter le problème de sur-apprentissage au niveau de HCRF. En effet, l'utilisation des séquences de BOWs de dimension importante comme entrée du HCRF peut engendrer le problème de sur-apprentissage. En effet, le sur-apprentissage peut survenir si la dimension de vecteurs de primitives est élevée relativement au nombre des données d'entrainement. Notons aussi qu'au niveau local, ce problème est surmonté grâce au bon

aspect de généralisation de SVM. La figure 4.6 illustre l'architecture globale de notre système.

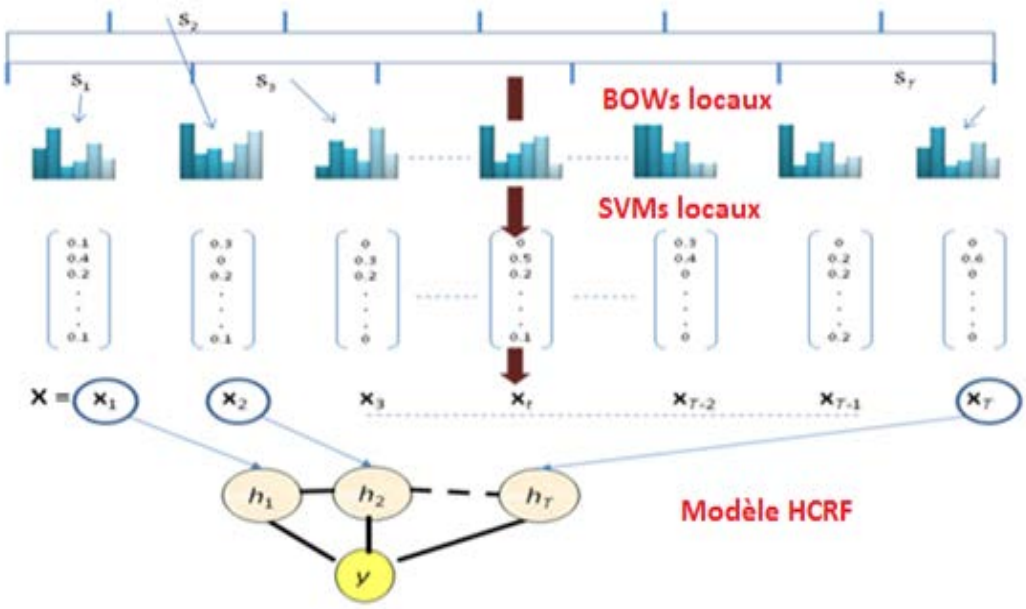


FIGURE *4.6: Architecture de modèle de reconnaissance hybride SVM-HCRF.*

On commence par segmenter la séquence vidéo en des segments de taille fixe, tout en considérant un chevauchement de 50% entre ces segments locaux. Ce chevauchement permet de modéliser implicitement le contexte temporel entre les STIPs appartenant aux segments adjacents. Cette segmentation permet de représenter la séquence vidéo en tant que série temporelle. Puisque le nombre des STIPs inclus dans chacun des segments est variable, nous avons opté pour l'application de l'algorithme de BOW au niveau local "segments". Cette modélisation nous permet d'avoir un vecteur de primitives de dimension fixe au niveau des segments. De plus, les BOWs donnent généralement de bons résultats lorsqu'ils sont utilisés avec le classifieur SVM.

Notre méthode de classification hybride a donc, comme entrée, une séquence de BOWs qui ont généralement une dimension importante. On applique, dans un premier temps, le classifieur SVM à ces BOWs locaux. SVM est ainsi appliqué comme un classifieur de bas niveau "niveau segment" convertissant chaque BOW en un vecteur de probabilités conditionnelles des classes d'activités. Le SVM agit donc comme un extracteur de primitives discriminatives de haut niveau. La dimension des vecteurs de probabilités générés est égale au nombre de classes des activités à reconnaître où chaque composante i de ce vecteur correspond à la probabilité que le segment appartient à la classe i. Chaque vecteur de probabilités locales devrait donc refléter la classe à laquelle appartient le segment correspondant. En effet, l'aspect généralisant de SVM compense l'effet de variabilités existant dans les segments appartenant à la même classe. Ce-

pendant, dans certains cas, par exemple celui des segments ne contenant pas de mouvement ou bien contenant une action élémentaire partagée par de multiples classes, les vecteurs de probabilités ne favorisent pas une classe particulière. Bien que ces vecteurs apportent des ambigüités entre les classes, ils restent pertinents pour la reconnaissance au niveau séquence. En effet, leurs occurrences temporelles ainsi que leurs relations avec les autres segments varient selon la classe d'activité. Pour représenter ces relations temporelles, nous utilisons le HCRF. Ce modèle séquentiel permet de modéliser les éventuelles dépendances entre les actions élémentaires représentées par les vecteurs de probabilités. En utilisant notre système hybride, HCRF a comme entrée une séquence de vecteurs de probabilité au lieu de la séquence des BOWs locaux. De ce fait, le modèle HCRF utilise directement l'information sémantique des BOWs locaux, à savoir la probabilité de chacune des activités relativement au segment en question. Cela permet, d'une part, de réduire la dimension de l'entrée du HCRF et donc de réduire sa durée d'entraînement et, d'autre part, d'exploiter l'aspect discriminatif du SVM. De plus, le fait d'estimer les paramètres du HCRF dans un espace de dimension réduite, celui des primitives de haut niveau, permet de réduire le risque de sur-apprentissage. Comme nous le verrons dans la section 5.4, le modèle hybride SVM-HCRF permet d'améliorer les performances aussi bien en termes de taux de reconnaissance qu'en termes de durée d'entraînement par rapport au modèle HCRF pur utilisant les séquences des BOWs locaux.

4.4 Comparaison avec l'état de l'art

Dans l'état de l'art, il existe d'autres approches qui se basent sur la décomposition de séquences vidéo en segments. Dans [Niebles et al., 2010], les auteurs décomposent la vidéo en des segments en considérant plusieurs échelles. On obtient donc des segments de longueur variable incluant même la séquence entière (Figure 4.7). Un classifieur SVM basé sur la similarité entre les images et la position du segment dans la séquence est défini pour chaque échelle. En combinant les scores des différents segments, on obtient un score global de correspondance entre chacune des classes d'activités et la séquence vidéo. Contrairement à notre modèle, cette approche ne modélise pas les relations temporelles entre les différents segments locaux et leurs corrélations spatio-temporelles.

K.Tang et al. [Tang et al., 2012] ont aussi proposé une approche fondée sur la segmentation temporelle des séquences vidéo. Comme c'est le cas dans notre approche, les auteurs décomposent la séquence vidéo en des segments de taille fixe et chaque segment est représenté par un BOW local. Ces séquences des BOWs locaux sont considérées comme entrée du modèle de reconnaissance. Le modèle semi-Markovien caché, utilisé ici comme classifieur, a, comme entrée, des séquences de primitives de dimensionnalité bien plus élevée que celle de notre vecteur de probabilités (la taille de BOWs est généralement de 4000). La dimensionnalité élevée des vecteurs de primitives augmente considérablement le risque de sur-apprentissage. Notons que, pour notre système, ce problème est évité grâce à l'utilisation de SVM comme classifieur

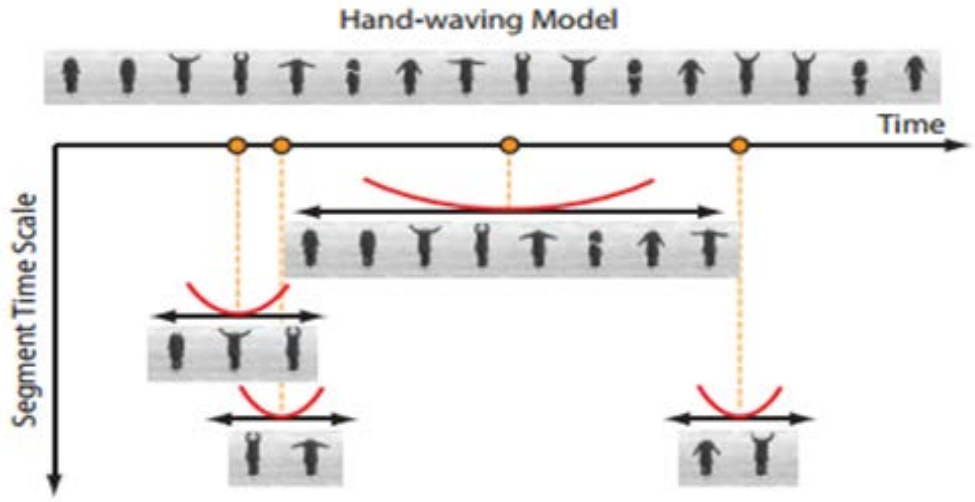


FIGURE *4.7: Segmentation temporelle des séquences en utilisant plusieurs échelles temporelles [Niebles et al., 2010].*

de bas niveau. Cela permet de réduire la dimension des vecteurs de primitives à une dimension égale au nombre de classes à reconnaître (dans l'état de l'art actuel, 50 classes au maximum). Il permet aussi de profiter de la capacité de généralisation de SVM tout en utilisant un modèle séquentiel discriminant au niveau séquence.

4.5 Conclusion

Dans ce chapitre, nous avons proposé une méthode de classification séquentielle hybride. Cette méthode permet d'améliorer le modèle séquentiel de référence HCRF en ajoutant un classifieur de bas niveau SVM connu pour sa bonne capacité de généralisation et de discrimination. L'exploitation de la complémentarité de ces deux classifieurs permet d'améliorer les performances aussi bien en termes de précision de reconnaissance qu'en termes de temps de calcul. La performance de ce classifieur sera démontrée expérimentalement dans le chapitre suivant.

CHAPITRE 5

Exprimentation

Sommaire

5.1 Introduction

Nous avons proposé, pour la reconnaissance des activités, une approche séquentielle fondée sur un modèle de classification hydride SVM-HCRF prenant comme primitives d'entrée des descripteurs extraits des trajectoires des points d'intérêt. Dans ce chapitre, nous montrerons l'efficacité de notre modèle sur un ensemble de bases de données qui seront décrites dans la première section. Dans la deuxième section, nous étudierons la performance des primitives de bas niveau explorées dans nos travaux. Ensuite, nous procéderons au choix des principaux paramètres de notre modèle de reconnaissance. Dans la section 5.4, nous évaluerons la

performance de notre système de reconnaissance SVM-HCRF, en le comparant avec les performances des travaux de l'état de l'art. Nous montrons dans la section 5.5 l'apport de la prise en compte du contexte à travers l'identification des objets dans la scène ainsi que la flexibilité de notre système à prendre en compte plusieurs types de primitives.

5.2 Description des bases de données

Dans cette section, nous présentons une description détaillée des bases de données utilisées pour valider notre système de reconnaissance. Nous présentons la base de données d'actions KTH, qui a été largement utilisée dans la littérature. L'ensemble de données, cependant, est composé d'actions simples avec un fond homogène. Relativement difficiles et vastes, les ensembles de données Rochester, UT-Interaction et CAD-120 sont décrits respectivement dans la section 5.2.3, la section 5.2.2 et la section 5.2.4. Ces bases de données offrent relativement une grande variation de réalisation des actions humaines. Les actions sont réalisées d'une façon réaliste et sans aucune contrainte.

5.2.1 KTH

La base de données KTH a été introduite par Schuldt et *al*. [Schuldt et al., 2004] et a été largement utilisée dans la littérature [Laptev et al., 2008, Gilbert et al., 2011, Bilinski and Brémond, 2012]. Cette base de données de faible résolution (images en niveau de gris de 160 $\times$ 120 pixels) regroupe 6 types d'actions : marcher (walking), courir lentement (jogging), courir rapidement (running), boxer (boxing), faire un mouvement circulaire des bras (hand waving) et applaudir (clapping). Ces actions ont été effectuées plusieurs fois par 25 acteurs selon quatre scénarios : scènes en extérieur (s1), changement d'échelle (s2), changement de vêtements (s3) et scènes en intérieur (s4) (Figure 5.1). Dans la plupart des séquences, le fond est homogène ; toutefois, la présence de l'ombre peut rendre la soustraction de fond relativement difficile. En outre, il existe plusieurs facteurs de variations tels que les zooms pour le deuxième scénario, l'angle de prise de vue, les conditions d'éclairage et la durée des actions. Nous utilisons dans nos expériences le même protocole que celui suggéré par Schuldt et *al*. [Schuldt et al., 2004], à savoir l'utilisation des séquences des 9 personnes avec l'indice 2, 3, 5, 6, 7, 8, 9, 10, et 22 pour la phase de test et les séquences des 16 personnes restantes pour la phase d'apprentissage.

5.2.2 Ut-Interaction

Ryoo et *al*. [Ryoo and Aggarwal, 2010] ont proposé une base de données divisée en deux ensembles "set1" et "set2", regroupant chacun des séquences de réalisation de 6 types d'interactions humaines : handshake, hug, kick, point, punch and push (Figure 5.2). Ces activités ont été réalisées par plus de 15 personnes, dans des conditions de vidéosurveillance réelles dans un parc de stationnement. Plusieurs facteurs de variabilité, tels que le degré de luminosité, la

FIGURE *5.1: Illustration de données KTH. Échantillon pour les six classes d'actions (colonne) enregistrées sous différentes scénarios (par rangée).*

variation d'échelle et le changement de tenue vestimentaire des individus, sont pris en considération. Les séquences vidéo enregistrées ont une résolution de 720 $\times$ 480 et un nombre de trames par seconde égal à 30. Nous allons, dans nos expériences, considérer les séquences du "set1". Cet ensemble est composé de 10 longues séquences ; chacune contient des réalisations appartenant aux différentes actions. Afin qu'il y ait adaptation à la problématique de reconnaissance des actions déjà segmentées, pour chaque séquence, 6 sous-séquences vidéo, contenant chacune une seule action, ont été extraites. Pour évaluer notre méthode, nous utilisons le protocole d'évaluation décrit dans [Ryoo and Aggarwal, 2010] qui consiste à réaliser 10 "leave-one-out" où les sous-séquences extraites de la même longue séquence sont maintenues pour la phase de test et les autres sont utilisées pour la phase d'entraînement. Le taux de reconnaissance global est la moyenne des taux de reconnaissance obtenus pour chacun des leave-one-out.

5.2.3 Base Rochester

Notre modèle pour la reconnaissance des activités a été testé sur une base publique de tâches domestiques "Rochester" [Messing et al., 2009] pour la reconnaissance des activités complexes de la vie quotidienne. Cette base comprend 10 activités de la vie quotidienne : répondre au téléphone, composer un numéro de téléphone, chercher un numéro dans l'annuaire, écrire un numéro de téléphone sur un tableau, boire un verre d'eau, manger banane, découper banane, peler banane, manger avec couverts et grignoter (Figure 5.3). Ces activités ont été réalisées par 5 personnes et chacune les a répétées 3 fois. Ces personnes sont d'âge, de taille, de genre et d'ethnie différents ce qui assure une grande variance dans le corpus de données. Cette base de vidéos est de haute définition (1280×720 pixels) et a un nombre de trame par seconde égale à 30. Nous avons utilisé le protocole de test "leave-one-out" pour calculer le taux

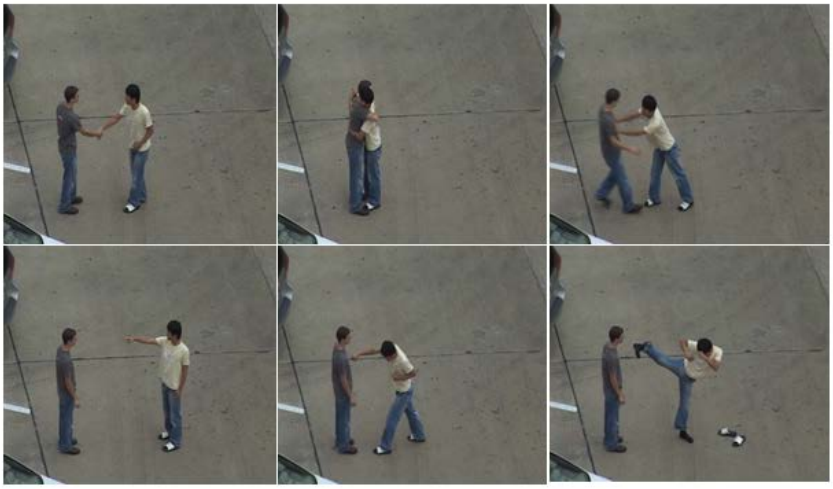


FIGURE *5.2: Illustration des actions de la base de données UT-Interaction : Hand Shaking, Hugging, Kicking, Pointing, Punching, Pushing.*

FIGURE *5.3: Illustration des actions de la base de données Rochester : répondre au téléphone, découper banane, composer un numéro de téléphone, boire un verre d'eau, manger banane, grignoter, chercher un numéro dans l'annuaire, peler banane, manger avec couverts et écrire un numéro de téléphone sur un tableau.*

de reconnaissance. Pour cela, à chaque fois on retient toutes les vidéos effectuées par une personne pour la phase de test et on utilise le reste pour la phase d'apprentissage.

5.2.4 Base CAD-120

La base de données Cornell-120 (Cornell Activity Dataset : CAD-120) [Koppula et al., 2013] contient 120 vidéos-3D de réalisation de 10 activités de la vie quotidienne (Figure 5.4) à savoir «Préparer des céréales, «Prendre des médicaments», «Empiler des objets», «Dépiler», «Mettre des aliments dans micro-onde», «Ramasser des objets», «Nettoyer des objets», «Prendre un

repas», «Ranger des objets» et «Préparer un repas». Ces activités sont réalisées par quatre personnes. Chaque personne répète chacune des activités trois fois en utilisant des objets différents. Cette base contient un total de 61 585 images RGB-D. Les réalisations des activités varient selon le genre, le style de la personne, l'ordre de réalisation de sous-activités, etc. Les labels des sous-activités sont fournis pour chacune des séquences. Dans cette base, on considère 10 sous-activités : «Atteindre», «Déplacer», «Verser», «Manger», «Boire», «Ouvrir», «Placer», «Fermer», «Laver» et «Null».

FIGURE *5.4: Illustration des activités de la base de données CAD-120.*

Comme pour la base Rochester, nous avons utilisé le protocole de test "leave-one-out" pour calculer le taux de reconnaissance. A chaque fois on retient toutes les vidéos effectuées par une personne pour la phase de test et on utilise le reste pour la phase d'apprentissage.

5.3 Evaluation des primitives de bas niveau

Dans notre travail, nous avons exploré deux types de primitives de bas niveau fondés sur les points d'intérêt. Dans un premier temps, nous avons considéré les points d'intérêt spatio-temporels Harris-3D qui sont bien adaptés au traitement des séquences spatio-temporelles. Ces points sont l'extension des points d'intérêt spatiaux Harris au domaine spatio-temporel. Ces points sont pertinents dans la mesure où ils correspondent surtout au mouvement, ils sont peu denses et sont discriminants. Une fois que ces points sont extraits, ils sont généralement décrits par des HOGs et des HOFs. Cependant, ces descripteurs ne tiennent pas suffisamment compte du contexte dans lequel ils sont détectés, information qui peut être pertinente pour la reconnaissance des activités. Pour éviter cette limitation, quelques travaux récents [Laptev et al., 2008, Bilinski and Brémond, 2012, Gilbert et al., 2009] ont exploré l'utilisation de la distribution spatio-temporelle des STIPs. Néanmoins, ces méthodes n'exploitent pas l'évolution spatio-temporelle de chaque STIP dans la séquence vidéo. D'autre part, les approches utilisant les trajectoires des points d'intérêt spatiaux ont donné des résultats relevant de l'état de l'art. Dans nos travaux de thèse, nous avons proposé d'utiliser les trajectoires des points d'intérêt

spatio-temporels "Harris-3D". Ces trajectoires permettent de capturer l'évolution temporelle des points Harris-3D. Ce nouveau descripteur a été décrit dans la section 3.2.

Nous nous sommes intéressés ensuite à l'utilisation des trajectoires des points denses qui ont été proposées par Wang et *al*. [Wang et al., 2011a]. Il s'agit d'extraire les trajectoires des points uniformément échantillonnés dans l'espace pour des échelles variables. Contrairement aux points Harris-3D, les points denses ne prennent pas en considération l'aspect temporel des séquences vidéo. Des points statiques ne modélisant pas le mouvement peuvent ainsi être extraits et suivis dans le temps. Pour remédier à ce problème, les trajectoires statiques sont supprimées dans une étape de post-traitement. Dans ce qui suit, nous présentons une évaluation de ces deux descripteurs sur les bases KTH et Rochester. Pour comparer les différentes primitives, nous nous sommes appuyés sur une modélisation par BOW où chaque séquence de mouvement est représentée par la fréquence de mots visuels. Un SVM non-linéaire de noyau χ^2 exponentiel est appliqué pour la phase de classification.

5.3.1 Trajectoires des points Harris-3D

Dans cette section, nous évaluons l'efficacité des trajectoires des points Harris-3D en tant que primitives pour la description du mouvement. Ces points d'intérêt ont été extraits pour des échelles spatiales et temporelles multiples en utilisant l'implémentation disponible en ligne de Laptev et *al*. [Laptev et al., 2008]. Les valeurs des paramètres d'échelle considérés dans nos expérimentations sont $\sigma^2 = 4;8;16;32$ et $\tau^2 = 4;8$. Après avoir extrait les points Harris-3D, nous effectuons le suivi de ces points en utilisant l'algorithme KLT pendant 15 trames. À partir de ces trajectoires, nous générons des séquences temporelles composées par :

- Les coordonnées spatiales des points : $\mathbf{P}_t =< x_t, y_t >$
- L'angle du déplacement : $\theta_t = \arctan(\frac{y_t - y_{t-1}}{x_t - x_{t-1}})$
- L'amplitude de déplacement : $\mathbf{V}_t = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2}$

où $t = 1..T$ est l'indice temporel.

Pour évaluer la performance de ces descripteurs, nous les avons comparés aux trajectoires des SURF [Bay et al., 2006]. Ces derniers correspondent aux coins détectés sur l'image. Nous avons également comparé les trajectoires des Harris-3D avec des trajectoires de SURFs qui appartiennent à la région d'intérêt (ROI). La détection de ROI est basée sur le calcul de vecteurs de mouvement entre deux images consécutives. On retient seulement les régions ayant une amplitude de mouvement importante afin d'ignorer les régions correspondant au bruit et à l'arrière-plan. Nous avons aussi comparé ces descripteurs aux travaux de l'état de l'art basés sur les trajectoires des points Harris-2D.

Pour établir une comparaison entre les performances des trajectoires des Harris-3D et celles des SURF, nous avons choisi de faire nos expériences en utilisant les vidéos qui correspondent au scénario « outdoor scenario » (vidéos réalisées en dehors de l'appartement) de la base

KTH. Nous avons retenu 16 personnes pour la phase d'apprentissage et 9 pour le test. Dans nos expériences, nous avons testé plusieurs tailles de BOW. Le BOW de taille 100 a donné le résultat optimal pour les trajectoires Harris-3D. Comme le montre le Tableau 5.1, les trajectoires des Harris-3D sont plus efficaces que celles des SURFs. Cela est principalement dû au fait que la plupart des SURFs extraits des trames appartiennent au fond. De ce fait, l'utilisation des trajectoires des SURFs qui appartiennent seulement aux régions en mouvement améliore considérablement le taux de reconnaissance, étant donné que le nombre des points SURFs appartenant au fond a diminué. Néanmoins, les trajectoires des Harris-3D restent beaucoup plus performantes que les trajectoires des SURFs incluses dans les ROI. En effet, le suivi des points Harris-3D revient à ne suivre que les points ayant initialement un mouvement significatif et caractérisant essentiellement le mouvement.

Méthode	Taux de reconnaissance
Trajectoires des Harris-3D	**92,59%**
Trajectoires des SURFs	61.11%
Trajectoires des SURFs dans ROI	83,33%

TABLE *5.1: Comparaison des trajectoires des Harris-3D et des SURFs en considérant un scénario de KTH.*

Nous avons aussi comparé les trajectoires des Harris-3D avec quelques travaux de l'état de l'art qui utilisent les trajectoires des points d'intérêts spatiaux, à savoir les points Harris-2D ; pour la description de mouvement. Comme le montre le Tableau 5.2, en réalisant l'évaluation sur toute la base KTH (4 scénarios), nous obtenons, avec les trajectoires des points Harris-3D, de meilleures performances que celles obtenues en utilisant l'historique des vitesses des trajectoires des Harris qui a été proposée par Messing et *al.* [Messing et al., 2009]. De plus, l'utilisation des trajectoires des Harris-3D est plus efficace que l'utilisation des HOF/HOG comme descripteur bien que ces derniers fournissent de riches informations sur la texture.

Méthode	Taux de reconnaissance
Trajectoires des Harris-3D	**84.25%**
Velocity Histories [Messing et al., 2009]	74%
BOW de HOG/HOF [Laptev, 2005]	80%

TABLE *5.2: Comparaison des primitives sur la base KTH.*

Ces résultats ont été confirmés en testant sur la base Rochester (Tableau 5.3). En effet, les trajectoires des points Harris-3D sont plus performantes que celles des points Harris-2D utilisés dans les travaux reportés dans [Messing et al., 2009, Matikainen et al., 2010]. Les trajectoires des Harris-3D sont beaucoup plus performantes que les descripteurs HOG/HOF.

La Figure 5.5 montre la matrice de confusion obtenue dans le cadre d'utilisation des trajectoires des Harris-3D pour la reconnaissance des actions de la base Rochester. Nos primitives permettent une discrimination assez nette entre des actions comprenant des trajectoires "géométriquement " éloignées. Ainsi, la distinction entre la classe «écrire sur un tableau» et toutes

Méthode	Taux de reconnaissance
Trajectoires des Harris-3D	**74%**
Velocity Histories [Messing et al., 2009]	67%
SCM-Traj [Matikainen et al., 2010]	70%
BOW de HOG/HOF [Messing et al., 2009]	59%

TABLE *5.3: Comparaison des primitives sur la base Rochester.*

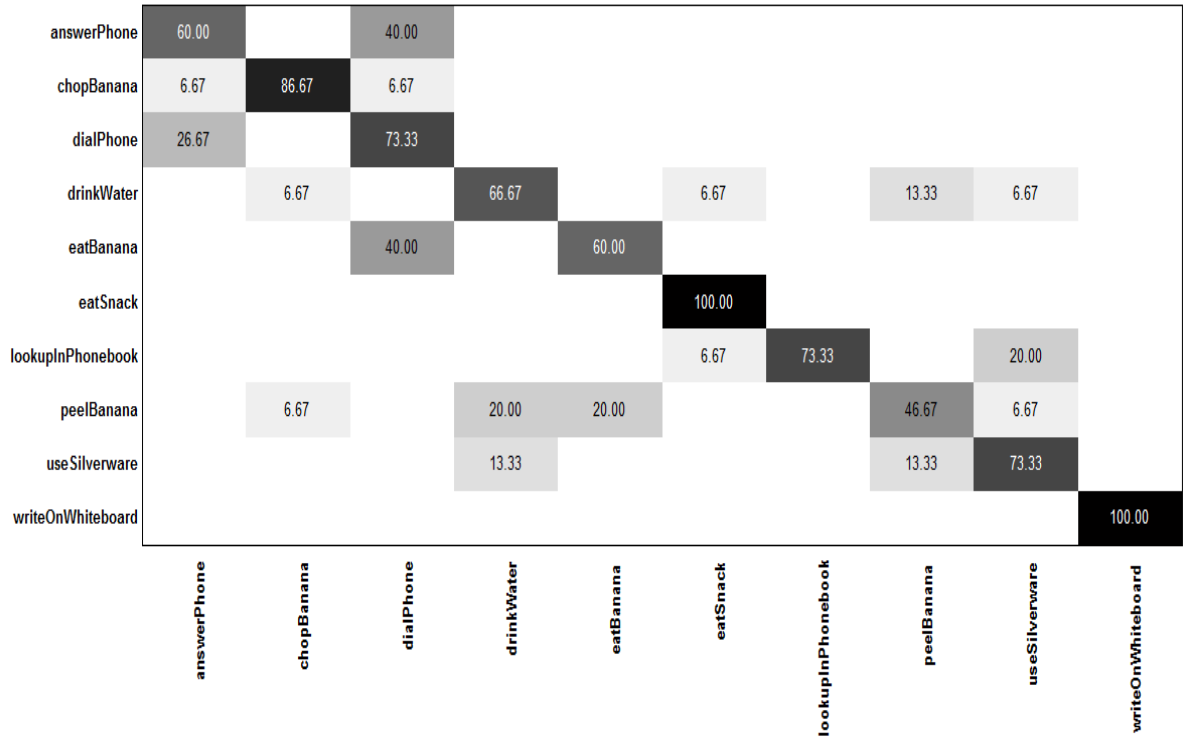


FIGURE *5.5: Matrice de confusion pour la base Rochester.*

les autres classes est très importante. Toutefois, les actions mettant en œuvre des trajectoires similaires de la main, comme «répondre au téléphone» et «appeler via le téléphone», sont vues comme semblables. De plus, on peut observer d'après la Figure 5.6 une grande variabilité interne du nombre des points Harris-3D détectés au sein des classes les moins reconnues telles que «peel banana» (46,67%) et «eat banana» (60%). Cela s'explique, premièrement, par la grande variation, d'une exécution à une autre, dans la réalisation d'une activité et, deuxièmement, par le fait que la méthode d'extraction Harris-3D ne détecte que les points ayant un mouvement significatif. Ainsi, pour les réalisations comprenant peu de mouvements, le nombre de points extraits reste insuffisant pour modéliser l'action de manière correcte et compenser les variations dans la façon de réaliser cette action.

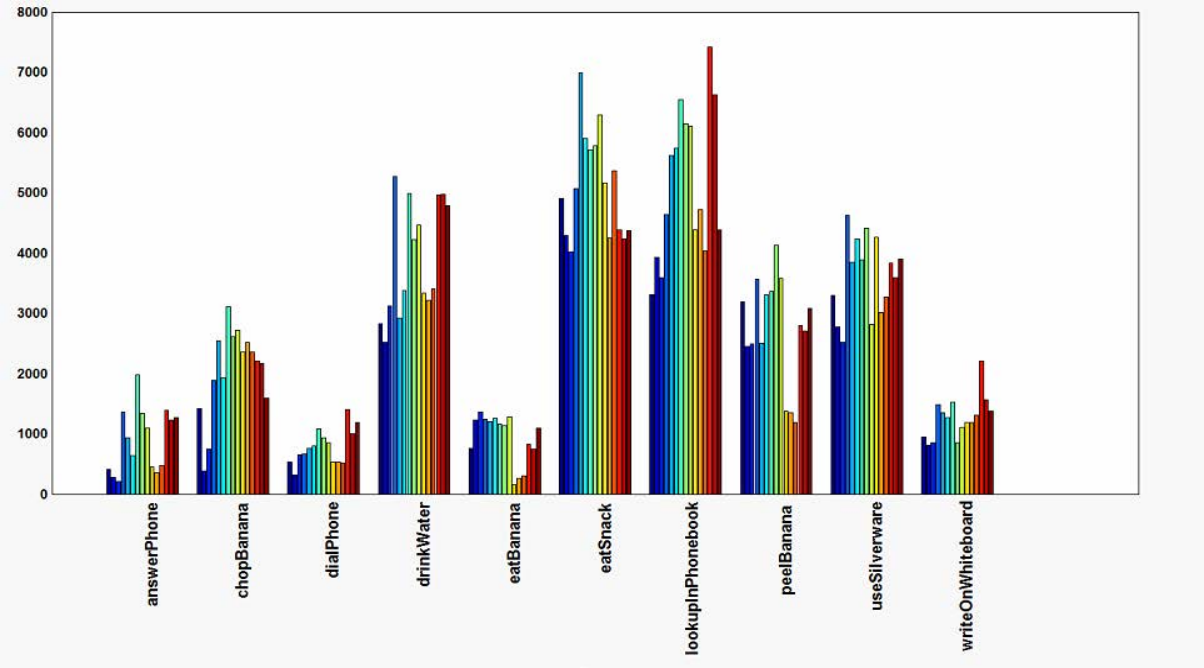


FIGURE *5.6: Nombre de points Harris-3D détectés pour chacune des réalisations regroupées par classe d'action.*

5.3.2 Trajectoires des points denses

Le deuxième descripteur considéré dans notre étude comparative, correspond aux trajectoires des points denses (Figures 5.7, 5.8) , proposées par Wang et *al.* [Wang et al., 2011a]. Nous rappelons brièvement ici le principe d'extraction de ces trajectoires qui a été détaillé dans la section 3.3. Nous commençons par un échantillonnage régulier pour 8 échelles spatiales en considérant, pour chacune des échelles, une grille de points espacés de 5 pixels. Les points appartenant à des zones homogènes sont supprimés. Les points denses ainsi extraits sont suivis durant 15 trames. Les trajectoires statiques et celles qui comprennent de larges déplacements sont supprimées.

Chacune des trajectoires est décrite par un vecteur de déplacement. De plus, les descripteurs HOG, HOF et MBH sont calculés dans un volume spatio-temporel aligné avec la trajectoire. Ce volume a une dimension de $32\times32\times15$ et est divisé en grilles de $2\times2\times3$ cellules. Pour chacune des cellules, on calcule les différents histogrammes. Les descripteurs obtenus, à savoir les vecteurs de déplacement, HOG, HOF et MBH, ont respectivement une dimension de 30, 96, 108 et 192. Pour évaluer les trajectoires denses, on a utilisé le même protocole que celui qui a servi pour l'évaluation des trajectoires des Harris-3D.

Nous nous intéressons, tout d'abord, à l'évaluation des performances des différents descripteurs des trajectoires denses sur Rochester, que ce soit au niveau global - un BOW par séquence - ou au niveau local - après la segmentation temporelle des séquences vidéos et en considérant un taux de chevauchement de 50%, on construit un BOW par segment. Cette

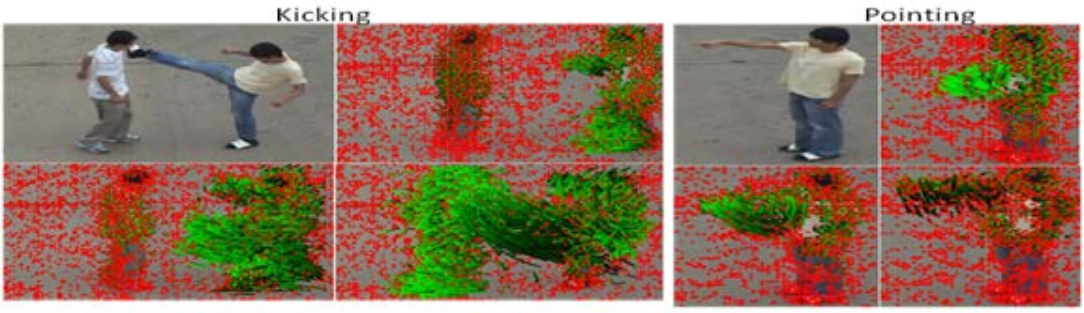


FIGURE *5.7: Illustration des trajectoires des points denses (en vert) calculées pour les actions «Kicking » et « Pointing » de la base Ut-Interaction .*

FIGURE *5.8: Illustration des trajectoires des points denses (en vert) calculées pour les actions «Drink Water» et « Eat Banana » de la base Rochester .*

évaluation est effectuée en utilisant un BOW de taille 2000. Le descripteur MBH montre son efficacité par rapport aux autres descripteurs aux niveaux global et local, comme le montrent les tableaux 5.4. Pour cette raison, dans le reste de cette évaluation, nous nous limitons uniquement sur une représentation des trajectoires denses fondée sur les MBHs.

Descripteur	Taux au niveau global	Taux au niveau local
Trajectoire	74%	44.1%
HOF	82%	50.7%
MBH	**85.3%**	**59.6%**
HOG	78%	56.2%

TABLE *5.4: Comparaison des performances des différents descripteurs pour la base Rochester.*

Pour comparer les trajectoires Harris-3D et celles des points denses, nous testons, tout d'abord, la performance de ces derniers sur la base KTH avec la taille de codebook optimisant le résultat pour les trajectoires Harris-3D, à savoir une taille de 100. Le tableau 5.5 montre que le taux de reconnaissance obtenu avec les trajectoires des points denses est plus important que

celui obtenu avec les trajectoires des points Harris-3D.
Le résultat obtenu sur la base Rochester (Tableau 5.6) confirme que les trajectoires des points

Méthode	Taux de reconnaissance
Trajectoires denses décrites par MBH	**89.81%**
Trajectoires des Harris-3D	84.25%

TABLE *5.5: Comparaison des résultats pour la base KTH obtenus avec les trajectoires des points denses.*

denses sont plus discriminantes que celles des trajectoires des Harris-3D. Pour les trajectoires denses, le taux de reconnaissance est obtenu en utilisant un codebook de taille 2000.

Méthode	Taux de reconnaissance
Trajectoires denses décrites par MBH	**85.3%**
Trajectoires des Harris-3D	74%

TABLE *5.6: Comparaison des résultats pour la base Rochester obtenus avec les trajectoires des points denses.*

Pour aller plus loin dans cette étude comparative, nous nous intéressons aussi à l'étude de la densité des points d'intérêt extraits. En effet, comme on l'a évoqué dans la section précédente, le nombre de points d'intérêt extraits pour le cas des Harris-3D risque d'être problématique dans le cas des réalisations comprenant peu de mouvements. On peut observer d'après le tableau 5.7 que le nombre des points denses extraits pour la base Rochester est 13 fois plus important que celui extrait avec la méthode Harris-3D. En prenant le cas de la réalisation la plus problématique, comprenant la quantité de mouvements minimale, le nombre de points denses extraits est assez suffisant pour modéliser le mouvement de manière correcte. Ce nombre s'élève à 2735 points denses alors qu'avec la méthode Harris-3D, on ne peut extraire que 148 points, soit 18 fois moins que le nombre de points denses.

	Points denses	Points Harris-3D
Nombre total de points d'intérêt extraits	5 107 911	396 245
Nombre minimal de points d'intérêt extraits par réalisation	2735	148

TABLE *5.7: Nombre de points d'intérêt extraits pour la base Rochester.*

5.3.3 Synthèse

Les trajectoires des points Harris-3D permettent de capturer l'évolution temporelle des points d'intérêt spatio-temporels Harris-3D. Ces nouveaux descripteurs se sont avérés plus performants que ceux utilisés dans l'état de l'art pour les points d'intérêts Harris-2D et SURF, car ils sont pertinents dans la mesure où ils correspondent surtout au mouvement, sont peu denses et sont discriminants. Néanmoins, cette faible densité peut être contraignante dans le cas des séquences comprenant peu de mouvement. En effet, comme on l'a constaté dans les expériences

faites sur la base de Rochester, les actions incluant des réalisations de faible amplitude de mouvement sont généralement mal reconnues. Les trajectoires des points denses se sont montrées plus robustes par rapport à cette problématique. L'utilisation des histogrammes des MBH permet aussi plus de robustesse par rapport au mouvement constant appartenant au fond. De plus, d'après l'étude comparative des performances des deux descripteurs sur les bases KTH et Rochester, il s'avère que la modélisation des actions avec les trajectoires des points denses décrite par MBH aboutisse à de meilleurs résultats que les trajectoires des points Harris-3D. Pour ces raisons, dans le reste de notre travail de thèse, la représentation des actions sera fondée sur les trajectoires des points denses.

5.4 Évaluation du modèle hybride SVM-HCRF

Après avoir choisi notre descripteur de bas niveau, nous nous intéressons, dans cette section, à l'évaluation de notre système de reconnaissance hybride SVM-HCRF. Comme on l'a indiqué dans la section 4.3, les séquences temporelles sont, tout d'abord, temporellement segmentées en utilisant la technique de fenêtre glissante. Les segments temporels ainsi obtenus sont localement décrits par des BOWs des trajectoires des points denses. En se basant sur ces descripteurs locaux, on utilise SVM pour avoir les scores de probabilités pour chacune des activités. Ces scores (vecteurs de dimension égale au nombre de classes) constituent l'entrée du modèle HCRF.

Dans ce qui suit, nous allons étudier l'influence des paramètres de chacune des étapes de notre système de classification - à savoir ceux de la segmentation temporelle et les paramètres du classifieur de bas niveau SVM. Ces évaluations seront effectuées sur les bases Rochester et KTH. Ensuite, nous présenterons une évaluation de notre système global en le comparant avec les travaux de l'état de l'art.

5.4.1 Évaluation des paramètres de segmentation

Pour utiliser une méthode de classification séquentielle, la séquence vidéo doit être représentée en tant que série temporelle de primitives. Le nombre de points denses extraits par trame est variable, d'où l'impossibilité d'avoir un vecteur de primitives de taille fixe par trame. De ce fait, nous optons pour une segmentation temporelle de la séquence vidéo en des segments de L trames, où chaque segment est représenté par un BOW local.

Comme le chevauchement entre les segments temporels permet de modéliser la corrélation temporelle entre les points d'intérêt appartenant à des trames adjacentes, nous allons, dans cette section, étudier l'apport du chevauchement sur le taux de reconnaissance local pour les bases Rochester. Cette évaluation est effectuée en considérant une taille de codebook 2000 et un SVM de noyau χ^2 exponentiel (sans optimisation des paramètres de SVM). Nous avons considéré une fenêtre de segmentation de 30 trames ; pour la base Rochester, cela correspond à une durée d'une seconde.

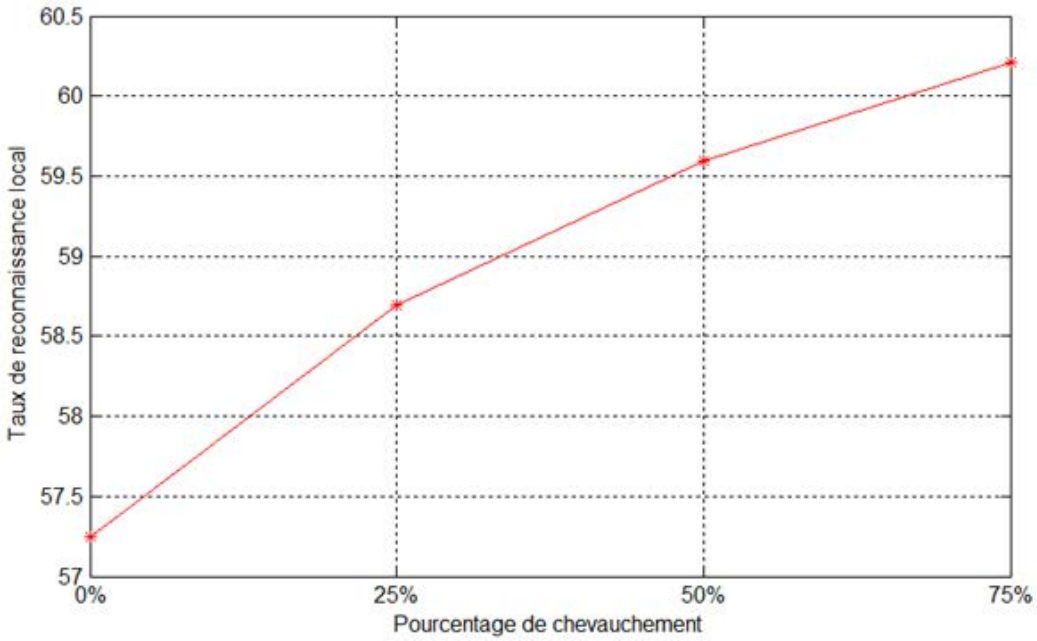


FIGURE *5.9: Taux de reconnaissance local selon le taux de chevauchement.*

Comme le montre la figure 5.9, un chevauchement de 50% permet un gain significatif en taux de reconnaissance locale, gain de 0.89 % par rapport à un taux de chevauchement de 25% et de 2.34% si l'on ne prend pas en considération le chevauchement. Un taux de chevauchement de 75% permet une petite amélioration du taux de reconnaissance par rapport à celui de 50%, tout en générant un nombre de segments plus élevé ce qui entraînera un temps de calcul plus important. De ce fait, nous avons choisi de maintenir un taux de chevauchement de 50% pour la suite de nos expérimentations.

5.4.2 Évaluation des paramètres du SVM-local

Après la segmentation temporelle et la construction des BOW locaux, nous nous intéressons à la conversion des BOWs qui sont généralement de taille importante en des vecteurs de probabilités conditionnelles de chacune des classes d'actions. Comme on l'a indiqué dans la section 3.4.3, nous utilisons un SVM non-linéaire de noyau χ^2 exponentiel. La plupart des bases de données contiennent plusieurs classes d'actions à reconnaître et, ainsi, on est dans le cas de classification multi-classes. Pour SVM multi-classes, deux stratégies de décomposition sont envisageable, à savoir un-contre-un et un-contre-reste :

- **un-contre-reste** : il s'agit d'apprendre n fonctions (n est le nombre de classes) de décision f_i, $i = 1..n$ permettant de faire la discrimination entre chaque classe de toutes les autres classes. L'affectation d'un nouveau point $\mathbf{x}$ à une classe C_k se fait en général par la relation :

$$k = \arg\max_{i=1..n} f_i(\mathbf{x})$$

- **un-contre-un** : Au lieu d'apprendre n fonctions de décision, ici, chaque classe est distinguée d'une autre. Ainsi $n \times (n-1)/2$ fonctions sont apprises et chacune d'entre elles constitue la réponse d'un classifieur binaire. L'affectation d'un nouvel individu $\mathbf{x}$ est réalisée par la combinaison des classifieurs, généralement faite par la règle de vote majoritaire.

Comme le montre le Tableau 5.8, la stratégie de décomposition un-contre-reste donne un taux de reconnaissance locale plus important que celui donné par la stratégie un-contre-un, sur la base Rochester et en considérant une taille de codebook 2000.

Stratégies de décomposition	Taux de reconnaissance local
Un-contre-reste	**59.6%**
Un-contre-un	51.4%

TABLE *5.8: Comparaison des stratégies de décomposition sur la base Rochester au niveau local.*

N'étant pas un modèle probabiliste, les SVMs utilisent les scores comme mesures de confiance pour la classification. Par ailleurs, plusieurs travaux permettent une transformation des scores SVM en des estimations de probabilité [Platt, 1999, Grandvalet et al., 2006, Wu et al., 2004]. Parmi les approches les plus utilisées dans le cas d'estimation de probabilités pour la classification binaire, on peut citer celle de Platt et *al.* [Platt, 1999]. Cette approche utilise une régression logistique sur les données d'entraînement. La transformation des scores SVM en probabilités se fait alors par une fonction sigmoïde. Se basant sur la stratégie de décomposition un-contre-un, Wu et *al.* [Wu et al., 2004] ont proposé un modèle exponentiel pour construire une sortie probabiliste.

Pour évaluer l'apport de l'utilisation de SVM comme classifieur de bas niveau et aussi l'utilisation d'une sortie probabiliste au lieu de score, nous avons effectué une étude comparative en termes de performance et de durée d'entraînement des modèles suivants :

- **BOWs-HCRF** : Utilisation directe des BOWs locaux comme entrée de HCRF.
- **SVM-Score-HCRF** : Génération des scores de chacune des classes d'activités conditionnellement aux BOWs locaux. Le HCRF a comme entrée une séquence des scores des classes d'activités.
- **SVM-Prob-HCRF** : Génération des probabilités de chacune des classes d'activités conditionnellement aux BOWs locaux. Le HCRF a comme entrée une séquence des vecteurs de probabilités des classes d'activités.

Cette étude comparative a été menée sur les bases Rochester et KTH. Les tailles de codebook utilisées pour ces deux bases sont respectivement 2000 et 100. HCRF a été entraîné avec 15

Méthode	Taux de reconnaissance	Durée d'entrainement (s)
BOW-HCRF	90.28%	33350.1 $\sim$ 9h
SVM-Score-HCRF	93.98%	23755.3 $\sim$ 6h
SVM-Prob-HCRF	**93.98%**	**16683.8** $\sim$ 4h

TABLE *5.9: Taux de reconnaissance et durée d'entrainement des différents modèles obtenus sur la base KTH.*

Méthode	Taux de reconnaissance	Durée d'entrainement (s)
BOW-HCRF	72%	jours
SVM-Score-HCRF	73.7%	heures
SVM-Prob-HCRF	**93.34%**	heures

TABLE *5.10: Taux de reconnaissance et durée d'entrainement des différents modèles obtenus sur la base Rochester.*

états cachés et une initialisation aléatoire entre 0 et 1. Les résultats obtenus sont reportés dans les tableaux 5.9 et 5.10.

Le modèle hybride SVM-HCRF s'est montré plus performant que BOW-HCRF pour les deux bases de données en termes de taux de reconnaissance et en temps de calcul. Par exemple, pour la base Rochester, nous avons passé d'un temps d'entraînement de l'ordre des jours à des heures. Cela peut évidemment s'expliquer par l'importante réduction de la dimension des vecteurs d'entrée du HCRF grâce à l'utilisation du classifieur de bas niveau SVM ; nous avons passé, dans le cas de la base Rochester, d'une séquence des BOWs de dimension 2000 à une séquences des vecteurs de dimension 10 (de 100 à 6 pour la base KTH). En termes de taux de reconnaissance, l'utilisation de SVM permet d'avoir un gain de 3.7% pour KTH et de 21.34% pour la base Rochester en considérant une sortie probabiliste et de 1.7% en considérant les scores.

Comme le montrent les résultats présentés dans les tableaux 5.9 et 5.10, l'utilisation des vecteurs de probabilités au lieu des scores permet un gain en termes de durée d'entraînement. Cela peut s'expliquer par le fait que l'initialisation des paramètres HCRF entre 0 et 1 est mieux adaptée aux séquences des vecteurs de probabilités. Notons aussi que, pour le cas de la base de données Rochester, l'utilisation des vecteurs de probabilités permet d'avoir un gain important en taux de reconnaissance (19.64%) par rapport à l'utilisation des scores.

On peut conclure, d'après ces expérimentations, que l'utilisation de SVM en tant que classifieur de bas niveau générant des vecteurs de probabilités permet un gain considérable en termes de taux de reconnaissance et en termes de temps de calcul par rapport au modèle HCRF. Ainsi, nous retenons le modèle SVM-Prob-HCRF dans le reste de nos expérimentations.

5.4.3 Evaluation de modèle SVM-HCRF

Dans cette section, nous allons procéder à une évaluation de notre système de reconnaissance hydride SVM-HCRF qui a, comme entrées, des histogrammes MBHs calculés à partir des trajectoires des points denses. Pour effectuer cette évaluation, on le compare avec la méthode

de reférence BOWs-SVM et aussi avec les travaux de l'art sur les bases de données décrites dans la première section de ce chapitre.

Nous commençons par la comparaison de notre modèle avec la méthode de BOWs qui a été largement utilisée dans la littérature [Laptev et al., 2007]. Pour ce faire, nous utilisons, comme pour les sections précédentes, les bases de données Rochester et KTH avec des tailles de codebook respectivement de 2000 et 100.

Les résultats obtenus sur les deux bases de données ; Tableaux 5.11 et 5.12 ; montrent que notre méthode de reconnaissance séquentielle est plus performante que la méthode de reconnaissance globale BOW-SVM. Cela permet de renforcer notre hypothèse sur l'importance d'utiliser des méthodes de reconnaissance séquentielles afin d'exploiter l'aspect temporel des séquences associées aux activités humaines. Notre méthode de reconnaissance SVM-HCRF est capable de modéliser implicitement les sous-activés via les états cachés de HCRF.

Méthodes	**Taux de reconnaissance**
BOW-SVM	89.8%
SVM-HCRF	**94**%

TABLE *5.11: Comparaison de BOW-SVM et SVM-HCRF sur la base KTH.*

Méthodes	**Taux de reconnaissance**
Global SVM	85.3%
SVM-HCRF	**93.3**%

TABLE *5.12: Comparaison de BOW-SVM et SVM-HCRF sur la base Rochester.*

Notre méthode s'est montrée bien compétitive par rapport aux travaux de l'état de l'art sur les bases KTH (Tableau 5.13), UT-Interaction (Tableau 5.14) et Rochester (Tableau 5.15). Ces résultats sont obtenus avec les tailles de codebook optimales à savoir 2000 pour KTH et Rochester et 500 pour UT-Interaction.

Méthodes	**Année**	**Taux de reconnaissance**
Laptev et *al.* [Laptev et al., 2008]	2008	91.8%
J. Yuan et *al.* [Yuan et al., 2009]	2009	93.3%
Kovashka et *al.* [Kovashka and Grauman, 2010]	2010	94.5%
Wang et *al.* [Wang et al., 2011a]	2011	94.2%
Gilbert et *al.* [Gilbert et al., 2011]	2011	94.5%
Zhang et *al.* [Zhang et al., 2012]	2012	94.1%
Kaaniche et *al.* [Kaâniche and Bremond, 2012]	2012	94.7%
Bilinski et *al.* [Bilinski et al., 2013]	2013	94.9%
Notre modèle hybride SVM-HCRF	2013	95.8%

TABLE *5.13: Comparaison des performances sur la base KTH.*

Notre méthode se place au premier rang pour les bases KTH avec un taux de reconnaissance de 95.8%. Comme le montre la matrice de confusion présentée dans la Figure 5.10,

Méthodes	Année	Taux de reconnaissance
Ryoo et *al*, Cuboid+SVMs [Ryoo and Aggarwal, 2010]	2010	85%
Ryoo, BP+SVM [Ryoo, 2011]	2011	83.3%
Yao et *al*. [Yao et al., 2010]	2010	88%
Vahdat et *al*. [Vahdat et al., 2011]	2011	93.3%
Zhang et *al*. [Zhang et al., 2012]	2012	95%
Kong et *al*. [Kong et al., 2012]	2012	88.3%
Raptis et *al*. [Raptis and Sigal, 2013]	2013	93.3%
Yuan et al. [Yuan et al., 2012]	2012	87%
Notre modèle hybride SVM-HCRF	2013	95.8%

TABLE *5.14: Comparaison des performances sur la base UT-Interaction.*

Méthodes	Année	Taux de reconnaissance
Messing et *al*. [Messing et al., 2009]	2009	89%
Satkin et *al*. [Satkin and Hebert, 2010]	2010	80%
Benabbas et *al*. [Benabbas et al., 2010]	2010	81%
Raptis et *al*. [Raptis and Soatto, 2010]	2010	82.7%
Matikainen et *al*. [Matikainen et al., 2010]	2010	70%
Wang et *al*. [Wang et al., 2011b]	2011	96% (93.8% sur KTH)
Piotr et *al*. [Bilinski and Brémond, 2012]	2012	93.3%
F. Yuan et *al*. [Yuan et al., 2012]	2012	92%
Prest et *al*. [Prest et al., 2013]	2013	92%
Notre modèle hybride SVM-HCRF	2013	93.3% (95.8% sur KTH)

TABLE *5.15: Comparaison des performances sur la base Rochester.*

notre système assimile parfois les actions « Jogging » à « Running ». Cela est dû au fait que ces actions correspondent à un comportement semblable et ne diffèrent que légèrement par la vitesse alors que dans la phase d'entraînement la vitesse de réalisation de ces actions varie beaucoup selon la personne. Il existe aussi une légère confusion entre les actions « Hand Clapping » et « Hand Waving ». Cela peut s'expliquer par le fait que ces deux actions comprennent des mouvements similaires des bras.

La plupart des travaux ayant des performances relevant de l'état de l'art sur la base KTH utilisent la méthode de BOW en se basant sur différents types de descripteurs. Par exemple, Bilinski et *al*. [Bilinski et al., 2013] calculent les trajectoires des points denses pour des échelles multiples. Pour décrire ces trajectoires, les auteurs calculent leurs vecteurs de déplacement et aussi les vecteurs de déplacement relativement à la position de la tête. Un SVM multi-noyau ayant comme entrée les BOWs de ces deux descripteurs et aussi ceux de HOG-HOF est utilisé pour la phase de classification. Dans ce travail, les auteurs se limitent aux cas dans lesquels une seule personne est présente dans la scène afin qu'ils puissent déterminer la position de la tête et ainsi encoder la position des trajectoires par rapport à cette dernière. De ce fait, cette méthode n'est pas adaptée aux situations dans lesquelles plus d'une personne est présente, telles que la base UT-Interaction. Plusieurs modélisations du contexte spatio-temporel ont également été explorées dans les travaux de l'état de l'art, utilisant KTH dans leurs expérimentations [Wang et al., 2011a, Gilbert et al., 2011]. Les auteurs utilisent généralement cette modélisation comme

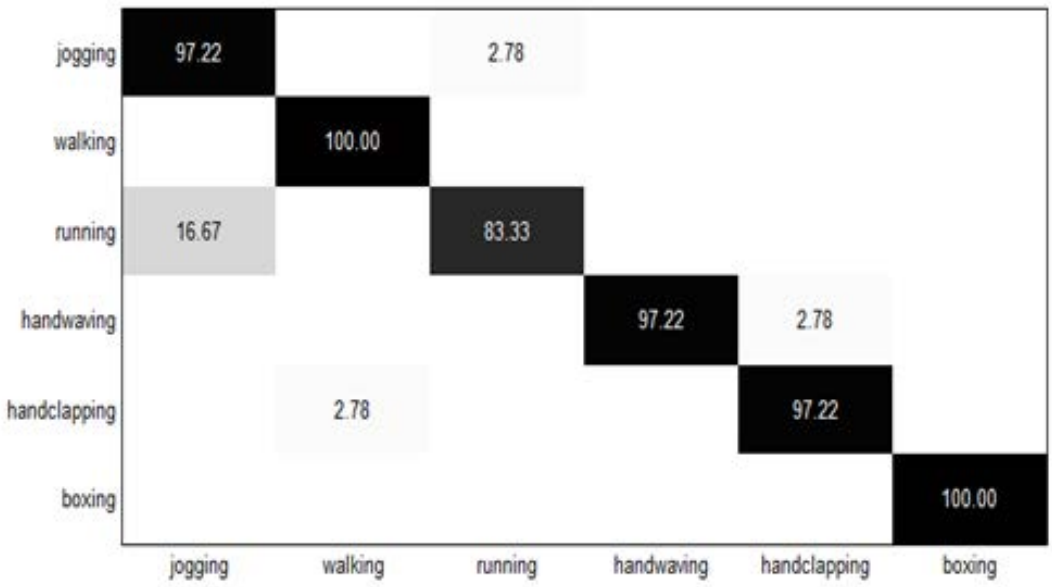


FIGURE *5.10: Matrice de confusion pour la base KTH.*

l'entrée de la méthode de BOW de référence. Comme le montre le tableau 5.13, c'est notre approche qui est la plus performante.

Nous aboutissons aussi à de meilleurs résultats que les travaux de l'art sur la base UT-interaction. Comme le montre la figure 5.11, notre approche de reconnaissance produit seulement une confusion entre les actions « punch » et « push ». Cela est dû au fait que des actions ont des sous-actions communes telles que le fait de tendre la main (dans un cas pour pousser et dans l'autre pour frapper), le recul de la personne en face ; ainsi, ces deux actions génèrent des trajectoires semblables (Figure 5.12).

Comme le montre le tableau 5.14, notre approche est plus performante que celle de Raptis et *al*. [Raptis and Sigal, 2013]. Ces derniers ont défini une approche de reconnaissance basée seulement sur un ensemble de trames clés qui sont décrites par des poselets En effet, bien que cette approche soit assez parcimonieuse et qu'elle exploite des poselets définis manuellement, il s'avère que l'utilisation de notre approche séquentielle SVM-HCRF permet une meilleure modélisation des actions. En effet, notre approche permet une modélisation explicite de l'aspect temporel des actions.

Concernant la base Rochester, nous nous plaçons au deuxième rang après le travail de Wang et *al* [Wang et al., 2011b] qui utilisent une approche basée sur la modélisation du contexte spatio-temporel des points d'intérêt à plusieurs échelles. Toutefois, notre méthode est plus performante que la leur sur la base KTH.

Comme on peut le constater d'après la matrice de confusion présentée dans la Figure 5.13, 33.34% des erreurs viennent de la confusion entre les actions « Answer Phone » et « Dial Phone ». Comme le montre la Figure 5.14 qui illustre les trajectoires extraites pour une réalisation de

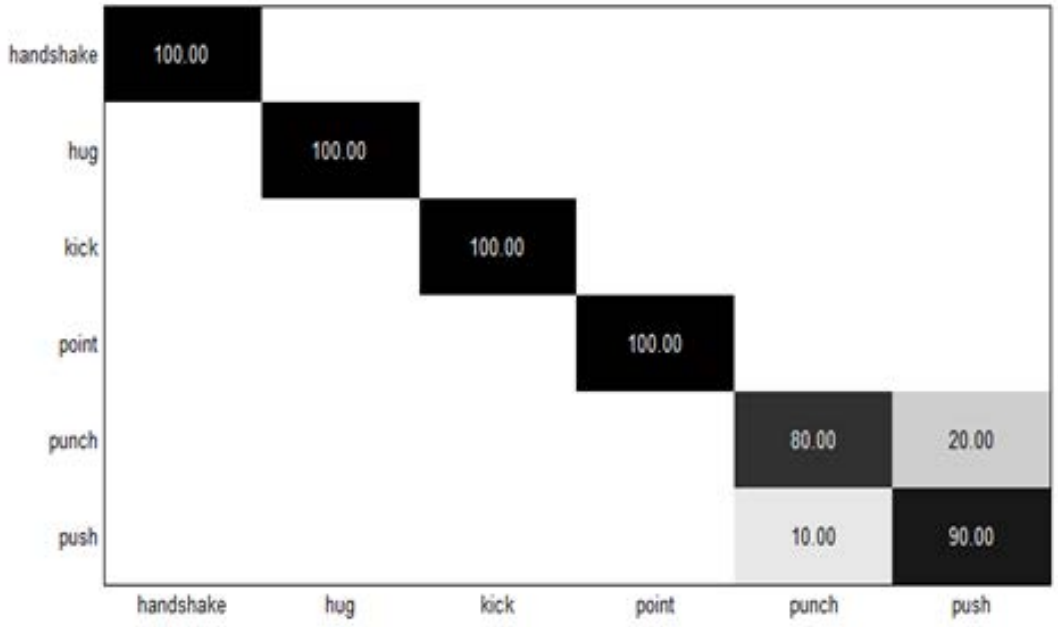


FIGURE *5.11: Matrice de confusion pour la base UT-Interaction.*

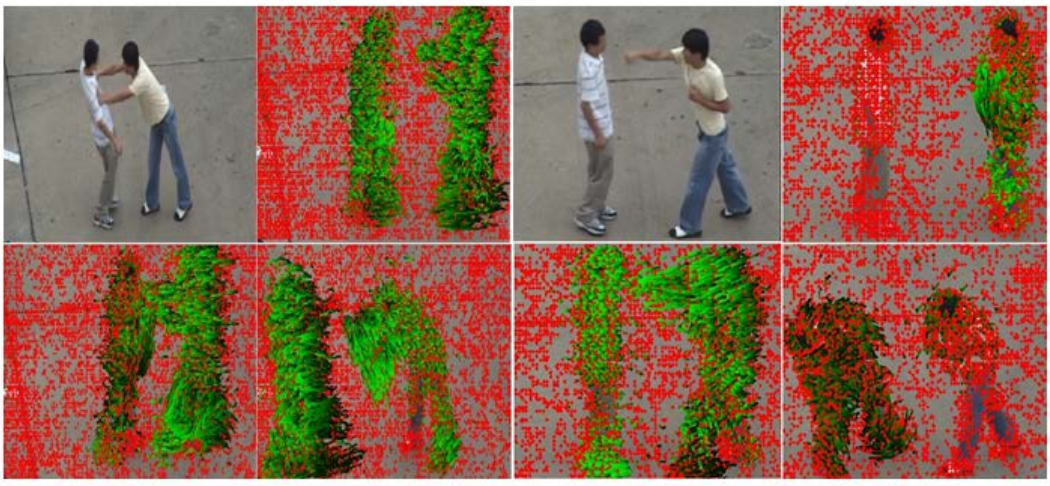


FIGURE *5.12: Trajectoires générées pour les actions push (à gauche) et punch (à droite).*

l'action « Dial Phone », ces deux actions sont très similaires. En effet, l'action « Dial Phone » inclut toutes les étapes de réalisation de l'action « Answer Phone ». En effet, après la composition du numéro de téléphone, la personne lève le télephone pour parler alors que pour l'action « répondre au téléphone», la personne remonte le téléphone et parle. Cela peut évidemment expliquer le taux d'erreur généré par notre système de reconnaissance pour ces deux classes d'activités.

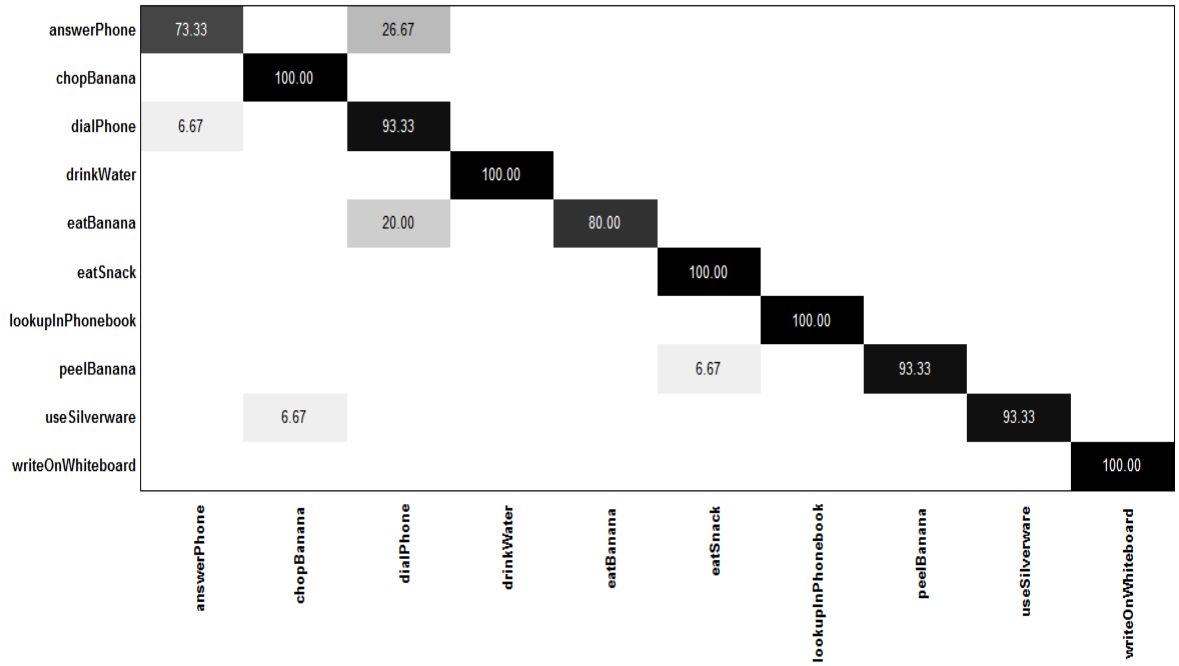


FIGURE *5.13: Matrice de confusion pour la base Rochester.*

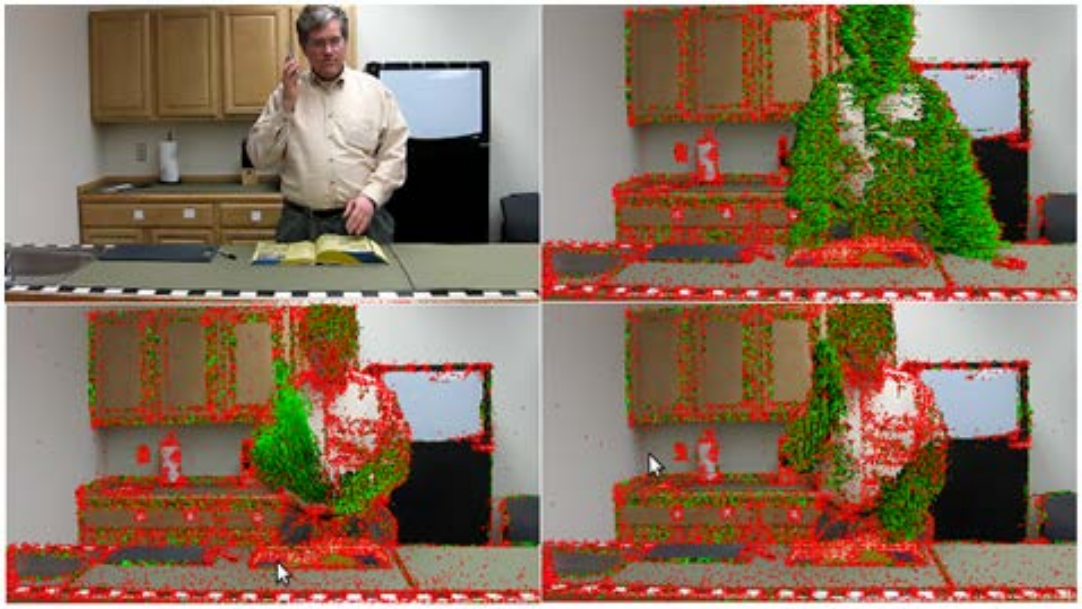


FIGURE *5.14: Illustration des trajectoires des points denses extraites pour une réalisation de l'action « Dial Phone ».*

De plus, l'action «Peel Banana» peut être confondue avec l'action «Eat Snack» car elle a un comportement initial similaire qui consiste à ramener un objet avant de l'utiliser.

Cette étude comparative sur les trois bases permet de mettre en valeur la performance et la

robustesse de notre méthode qui a été testée sur des bases de données ayant des caractéristiques variées. En effet, notre approche s'est montrée bien compétitive que se soit par rapport à des méthodes se basant sur des méthodes de reconnaissance globale [Laptev et al., 2008, Bilinski et al., 2013] et que celles basées sur les trames clés [Raptis and Sigal, 2013]. Notons aussi que ces travaux utilisent des primitives basées sur les trajectoires des points d'intérêts [Messing et al., 2009, Bilinski et al., 2013] et la modélisation du contexte spatio-temporel des points d'intérêt [Gilbert et al., 2011, Bilinski and Brémond, 2012, Laptev et al., 2008].

5.5 Exploitation du contexte

La modélisation du contexte de réalisation des actions s'est avérée bénéfique pour l'amélioration de l'efficacité des systèmes de reconnaissance [Gupta et al., 2009, Yao and Fei-Fei, 2010]. L'interdépendance entre la nature des objets manipulés et les actions permet de lever l'ambigüité lors de la reconnaissance d'actions similaires. Par exemple, les actions « répondre au téléphone » et « boire » ont un début de mouvement similaire qui est l'action élémentaire « lever la main ». Ainsi savoir si l'objet manipulé est un "verre" ou un "téléphone" permet de lever cette ambigüité. Notre modèle de reconnaissance permet d'inclure d'une façon fluide la modélisation du contexte. En effet, la façon la plus simple consiste à concaténer le vecteur de probabilités calculé au niveau segment avec le vecteur modélisant le contexte. Un SVM multi-noyaux pourrait aussi être utilisé afin de combiner les primitives de base à savoir les BOWs locaux et les vecteurs relatifs au codage du contexte pour obtenir via le SVM le vecteur de probabilités des classes d'activités par segment. Nous allons dans cette section montrer l'apport de la concaténation des vecteurs de primitives : vecteurs de probabilités de chacune d'activités conditionnellement aux BOWs locaux et les vecteurs de codage du contexte.

Pour encoder le contexte, nous avons exploré l'utilisation de la fréquence d'apparition de chacun des objets. Chaque segment temporel de la séquence vidéo $S = (s_1, s_2, ..., s_t, \ldots, s_T)$ est ainsi décrit par un vecteur $s_t = (a_1, a_2, \ldots a_i, \ldots, a_A, o_1, o_2, \ldots o_j, \ldots, o_O)$; où A est le nombre d'activités, O est le nombre d'objets, a_i est la probabilité de l'activité i conditionnellement au BOW local de segment t et o_j indique la fréquence d'apparition de l'objet j dans les trames comprises dans le segment.

Nous avons évalué notre modèle sur la base CAD-120 décrite dans la section 5.2.4. Nous commençons par une évaluation de notre modèle de base SVM-HCRF sans modélisation de contexte. Pour se mettre dans les mêmes conditions d'évaluation, nous nous comparons tout d'abord aux travaux n'utilisant que des primitives liées au mouvement. Comme le montre le Tableau 5.16, nous obtenons un taux de reconnaissance beaucoup plus important que celui obtenu par [Koppula et al., 2013] en utilisant les primitives extraites à partir des squelettes (skeletons) (Figure 5.15) obtenus par la Kinect et un SVM structurel pour la phase de classification. Bien que les sekeletons extraits par la Kinect fournissent une riche description de mouvements,

ils restent peu robustes par rapport aux occlusions. De plus, notre modèle SVM-HCRF permet une meilleure modélisation de l'aspect temporel des activités.

FIGURE *5.15: Illustration des skeletons [Koppula et al., 2013].*

Le Tableau 5.17 montre les résultats obtenus en exploitant les objets. Comme l'activité est composée d'une séquence temporelle de sous activités, certains des travaux de l'état de l'art avec lesquels nous nous comparons exploitent la vérité terrain sur la segmentation temporelle des sous-activités ; par exemple l'activité «nettoyer un objet» est composée de sous activités «atteindre objet», «ouvrir» et «laver». Pour chaque segment temporel donné par la vérité terrain, Koppula et *al.* [Koppula et al., 2013] par exemple modélisent l'interaction entre les objets identifiés et les primitives extraites à partir des skeletons à travers un Champ Aléatoire de Markov afin d'identifier la nature des sous-activités. Un SVM structurel est ensuite utilisé pour reconnaitre l'activité à partir des sous activités.

Méthodes	Taux de reconnaissance
Skeleton [Koppula et al., 2013]	27.4%
SVM-HCRF	**73,38**%

TABLE *5.16: Comparaison des performances sur la base CAD-120 sans considération des objets.*

Méthodes	Taux de reconnaissance
Vérité terrain sur la segmentation temporelle donnée	
Koppula et *al.*, IJRR 2013. [Koppula et al., 2013]	84.7%
Koppula, Saxena, ICML 2013. [Koppula and Saxena, 2013]	**93.5**%
Sans vérité terrain sur la segmentation temporelle	
Koppula et *al.*, IJRR 2013. [Koppula et al., 2013]	80.6%
Koppula, Saxena, ICML 2013. [Koppula and Saxena, 2013]	83.1%
SVM-HCRF + fréquence objets	**90.32**%

TABLE *5.17: Comparaison des performances sur la base CAD-120.*

Comme le montre le tableau 5.17 notre méthode est plus performante que les travaux de l'état de l'art n'utilisant pas la vérité terrain sur la segmentation temporelle des sous activités. De plus, elle permet de donner le deuxième meilleur taux de reconnaissance en la comparant aussi avec

les travaux utilisant la vérité terrain sur la segmentation temporelle des sous activités. En effet, Koppula et *al* [Koppula and Saxena, 2013] utilisent un CRF pour modéliser l'aspect temporel des activités et utilisent des primitives extraites à partir des skeletons donnés par la kineck. Bien que leur méthode donne de bonnes performances, elle n'est pas toutefois généralisable aux cas de vidéo 2-D.

5.6 Conclusion

Dans ce chapitre, nous avons montré l'efficacité de notre système de reconnaissance en le testant sur plusieurs bases publiques. Nous avons commencé par une étude de nos descripteurs de base, les trajectoires des points denses décrits par les MBH qui se sont montrés plus efficaces que les trajectoires des points Harris-3D. Notre méthode de classification séquentielle hybride SVM-HCRF s'est avérée performante pour la classification des activités. HCRF est bien adapté à la classification des séquences temporelles. De plus, l'utilisation du SVM comme classifieur de bas niveau a permis d'améliorer la performance du système de reconnaissance puisqu'il fournit au classifieur de haut niveau, HCRF, des informations sémantiques et il permet une réduction importante de la dimension. Nous avons aussi montré l'extensibilité de notre système de classification à d'autres types de primitives telles que l'identité d'objets utilisés lors de la réalisation des activités.

CHAPITRE 6

Conclusion et perspectives

Sommaire

6.1 Synthèse de mes contributions

Dans ce travail, nous nous sommes intéressés à la reconnaissance d'actions à partir de séquences vidéo. Pour traiter ce problème, nous avons développé une approche de reconnaissance séquentielle fondée sur un modèle hybride SVM-HCRF en suivant une démarche hiérarchique à trois niveaux. Le premier niveau concerne l'extraction de primitives de bas niveau à partir des séquences vidéo. Le niveau intermédiaire permet d'extraire des descripteurs de niveau intermédiaire plus pertinents et ayant plus de sémantique. Le dernier niveau a pour but de classifier les actions en utilisant l'approche séquentielle par les modèles conditionnels HCRF.

Dans l'étape d'extraction des primitives de bas niveau, nous avons utilisé les points d'intérêt spatio-temporels qui se sont montrés prometteurs et robustes dans les conditions réelles. En effet, ces points sont robustes aux occlusions partielles et leur extraction n'exige pas une étape de prétraitement telle que l'extraction de fond, ou le suivi de la silhouette. De plus, ils encodent principalement le mouvement comme on l'a montré par une analyse tensorielle. Etant donné que les points d'intérêt sont désordonnés, nous avons utilisé des BOWs locaux encodant les segments obtenus par la méthode de fenêtre glissante. On obtient ainsi une séquence de vecteurs de primitives de dimension fixe ce qui est bien adapté aux méthodes de reconnaissances séquentielles.

L'étape intermédiaire permet d'extraire des descripteurs de niveau intermédiaire plus pertinents et ayant plus de sémantique. SVM est utilisé comme classifieur local afin de convertir les BOWs locaux en des vecteurs de probabilités conditionnelles des classes d'activités ce qui résume d'une manière concise et discriminante l'information de bas niveau. Cette stratégie permet une importante réduction de dimensionnalité des vecteurs de primitives en entrée au HCRF garantissant ainsi un gain en temps de calcul considérable. Elle permet, d'autre part, une forte

diminution du risque de sur apprentissage, la dimension des BOWs étant beaucoup plus importante que le nombre des données d'entrainement, au contraire du nombre de classes en sortie du SVM local. Comme on l'a montré dans nos travaux, cette représentation permet aussi un gain en termes de précision, ce qui est cohérent avec le fait que le HCRF induit est moins sujet au problème de sur apprentissage.

Notre système de reconnaissance hybride SVM-HCRF s'appuie sur le HCRF comme classifieur de haut niveau ayant comme entrée les vecteurs de probabilités conditionnelles des classes fournis par le SVM. Ce classifieur séquentiel permet de prendre en considération l'aspect temporel des séquences vidéo tout en étant discriminant.

Notre modèle de reconnaissance offre également une grande flexibilité par rapport à l'ajout de nouvelles sources d'information encodant le contexte ou l'identité des objets manipulés. Comme on l'a montré dans la section 5.5, une première exploration de l'ajout de l'identité des objets a permis un gain considérable des taux de reconnaissance sans changement significatif dans notre modèle.

6.2 Perspectives

A l'issue de ce travail, de nombreuses perspectives s'ouvrent sur les divers sujets traités. Les principales perspectives à cours terme incluent :

- Au niveau des descripteurs des trajectoires des points denses, nous nous sommes limités au cours de nos travaux à l'utilisation du descripteur MBH. Une première amélioration possible est de le combiner avec les autres descripteurs des trajectoires des points denses (HOG, HOF, Vecteur déplacement) en utilisant par exemple un SVM multi-noyau. Notant aussi que cette combinaison des descripteurs ne va pas changer notre modélisation puisque la sortie du SVM restera la même mais les primitives seront beaucoup plus informatives.

- Dans l'étape d'extraction des descripteurs de niveau intermédiaire, dans nos travaux, nous avons généré via SVM des vecteurs de probabilité des classes d'activités. On n'est pas obligé, néanmoins, de se limiter aux identifiants des classes comme sorties du SVM. Nous proposons par exemple d'explorer l'augmentation du nombre de sorties du SVM en générant d'une manière non-supervisée, les sous-activités associées à une activité donnée, afin d'avoir une description plus fine des segments en fonction de leurs structures internes et de leurs positions temporelles dans la séquence vidéo. Cette extension du nombre de classes au niveau intermédiaire devrait permettre une description discriminante plus riche des segments et, en conséquence, une amélioration de la précision du HCRF.

- Au niveau de modélisation, notre méthode fondée sur le modèle SVM-HCRF est extensible à la segmentation et la reconnaissance conjointes d'activités dans un flux vidéo continu. Une première investigation possible est d'appliquer les champs aléatoires conditionnels dynamiques cachés (LDCRF : Latent Dynamic Conditional Random Fields) [Morency et al., 2007] comme classifieur de haut niveau de séquences d'activités. LDCRF est une extension de HCRF permettant la reconnaissance et la segmentation de séquences de labels dans une série temporelle tout en utilisant des états cachés pour modéliser la structure interne des activités (sous activités ou actions élémentaires). Nous avons déjà exploré cette stratégie vers la fin de cette thèse en implémentant le modèle SVM-LDRF. Les performances préliminaires obtenues sur les bases TUM [Tenorth et al., 2009] et KTH sont satisfaisantes dans l'ensemble même si elles peuvent être significativement améliorées en optimisant les différents paramètres du modèle.

- Nous proposons aussi d'utiliser des bases plus grandes et difficiles telles que Hollywood2 [Marszałek et al., 2009] et HMDB [Kuehne et al., 2011]. De plus une extension de nos travaux vers la localisation et détection des actions sur ces bases est envisageable.

A plus long terme

- Pour les applications d'assistance aux personnes âgées, il serait intéressant d'étendre nos travaux à la prévention de l'entrée en dépendance de la personne. Il s'agit de détecter une dégradation lorsque la personne effectue les activités de la vie quotidienne en se basant sur l'historique de réalisation des activités de la même personne. Ceci nécessite donc le stockage d'un historique relativement long de la même personne sur une longue période de temps. Cependant, au niveau de bases de données actuellement disponibles on est encore loin d'avoir cette grande quantité de données. De plus, ce type d'application nécessite l'analyse d'une énorme quantité de données et requiert ainsi une grande capacité de calcul.

- Nous proposons aussi de fusionner les informations issues de la vidéo avec d'autres sources d'informations hétérogènes associées à des capteurs et à des sources sonore. Ceci permettrait d'enrichir l'information en entrée du système de classification et d'améliorer significativement la précision de reconnaissance.

ANNEXE A

Machine à vecteurs de support

Les séparateurs à vaste marge sont une classe de méthodes d'apprentissage supervisées initialement définies pour la classification binaire. Elle a été ensuite adaptée à la classification multi-classes. Cet algorithme cherche l'hyperplan de marge maximale qui sépare le mieux les deux classes des données d'apprentissage. Cette marge est la distance minimale entre les données d'apprentissage et cet hyperplan.
Dans ce qui suit nous commençons par le cas le plus simple, à savoir le principe de classification par SVM des données linéairement séparables. Ensuite, nous présentons le cas de classification non-linéaire binaire ainsi que le cas de la classification multi-classes.

A.1 Cas linéaire

Dans le cas des données linéairement séparables, une fonction de discrimination linéaire f est conventionnellement liée à un hyperplan séparateur d'équation :

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + \beta = 0,\ (\mathbf{x}, \mathbf{w} \in \mathbb{R}^d, \beta \in \mathbb{R}) \tag{A.1}$$

où $\mathbf{w}$ est un vecteur normal à l'hyperplan. Pour prédire la classe y d'une observation $\mathbf{x}$, on utilise la fonction de décision $g(\mathbf{x})$ à valeurs dans $-1, 1$ définie par :

$$g(\mathbf{x}) = signe(f(\mathbf{x})). \tag{A.2}$$

Si les données d'apprentissage $\mathbf{x}_i, y_i$ sont linéairement séparables, il existe une infinité d'hyperplans qui séparent correctement ces données. L'idée donc est de chercher l'hyperplan le plus robuste en considérant celui qui maximise la marge tout en garantissant que $y_i.f(\mathbf{x}_i) \geq 1$. Etant donné que la distance normalisée entre l'exemple $\mathbf{x}$ et l'hyperplan est égale à $\frac{|f(\mathbf{x})|}{\|\mathbf{w}\|}$, la marge normalisée est égale à $\frac{2}{\|\mathbf{w}\|}$(Figure A.1)
Finalement, chercher l'hyperplan optimal revient à minimiser la fonction objective suivante :

$$\min_{\mathbf{w},\beta} P(\mathbf{w}, \beta) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \mathcal{L}(y_i f(\mathbf{x}_i)) \tag{A.3}$$

Le premier terme de cette équation est relatif à la maximisation de la marge ; alors que le deuxième est relatif à la minimisation de somme pondérée des erreurs d'apprentissage. Le

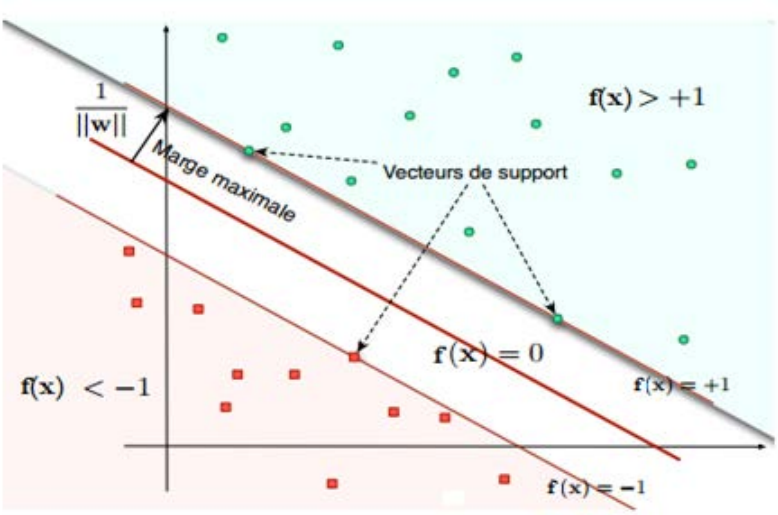


FIGURE *A.1: Recherche d'un hyperplan de séparation optimal au sens de la marge maximale.*

méta-paramètre C permet d'ajuster un compromis entre une marge vaste et le nombre d'erreurs de classification tolérées sur l'ensemble d'apprentissage. La fonction de coût $\mathcal{L}$ permet de pénaliser les séparatrices admettant des données d'apprentissage qui ne sont pas du bon côté des marges, sans cependant interdire une telle possibilité. La fonction de coût la plus utilisée, connue sous le nom de "hinge loss", est définie comme $\mathcal{L}(y, f(\mathbf{x})) = \max(0, 1 - yf(\mathbf{x}))$. En introduisant des variables dites « ressorts » (slack variables) généralement dénotées ξ_i, la fonction objective peut être réécrite comme :

$$\min_{\mathbf{w},\beta} P(\mathbf{w},\beta) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \; avec \begin{cases} \forall i \; \xi_i \geq 0 \\ \forall i \; y_i f(\mathbf{x}_i) \geq 1 - \xi_i \end{cases} \tag{A.4}$$

Si les données d'apprentissage sont linéairement séparables, on peut trouver une solution pour laquelle tous les ξ_i sont nuls.

A.2 Cas non linéaire

Les problèmes de classification sont souvent non linéaires alors que la détermination d'une fonction non linéaire est très difficile, voire impossible. Pour remédier à cette difficulté, la solution consiste à projeter les données dans un espace d'une dimensionnalité plus importante où cette fonction devient linéaire, l'idée étant qu'en augmentant la dimensionnalité du problème, on se retrouve dans le cas des SVMs basés sur la séparation linéaire vue précédemment. La

transformation d'espace est réalisée par la fonction :

$$\begin{cases} \mathbb{R}^d \rightarrow \mathbb{R}^D \\ \mathbf{x} \rightarrow \Phi(\mathbf{x}) \end{cases} \tag{A.5}$$

Ce changement va conduire à passer d'un produit scalaire dans l'espace d'origine $\langle \mathbf{x}_i.\mathbf{x}_j \rangle$ au produit $\langle \Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j) \rangle$ dans le nouvel espace. Pour éviter de calculer ce produit scalaire dans cet espace de re-description, on utilise une fonction de noyau k telle que :

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j) \rangle \tag{A.6}$$

Il existe de nombreuses fonctions noyau prédéfinies ; parmi les plus utilisées, on peut citer :

Le noyau polynomial :

$$k(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T \mathbf{z} + p)^l ; l \in \mathbb{N}$$

Le noyau gaussien :

$$k(\mathbf{x}, \mathbf{z}) = \exp\left\{-\gamma \sum_{i=1}^{D} \|\mathbf{x}_i^a - \mathbf{z}_i^a\|^b\right\}, a \in \mathbb{R}^+, b \in [0, 2]$$

Le noyau d'intersection :

$$k(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{D} \min(\mathbf{x}_i, \mathbf{z}_i)$$

Le noyau chi-square :

$$k(\mathbf{x}, \mathbf{z}) = exp\left\{-\gamma \chi^2(\mathbf{x}, \mathbf{z})\right\} o \chi^2 = \sum_{i=1}^{D} \frac{(\mathbf{x}_i - \mathbf{z}_i)^2}{\mathbf{x}_i + \mathbf{z}_i}$$

D est égale à la dimension de $\mathbf{x}$.

A.3 SVMs multiclasses

Les machines à vecteur support ont été conçues au départ pour la classification binaire. Elles ont été étendues pour la classification multi-classes. Ces extensions réduisent le problème multi-classes à une composition de plusieurs hyperplans biclasses permettant de tracer les frontières de décision entre les différentes classes [Hamel, 2009, Abe, 2010]. Ces méthodes décomposent l'ensemble des exemples en plusieurs sous-ensembles représentant chacun un problème de classification binaire. Pour chaque problème, un hyperplan de séparation est déterminé par la méthode SVM binaire.

A.3.1 Une-contre-reste (1vsR)

Cette méthode de décomposition consiste à utiliser pour chaque classe un classifieur binaire pour la séparer de toutes les autres classes. Le k-ième classifieur est destiné à distinguer la classe d'indice k de toutes les autres. Ainsi, pour un problème à K classes, il en résulte K SVM binaires (Figure A.2).

Pour affecter un exemple $\mathbf{x}$ à une classe, la décision s'obtient en application du principe "winner-takes-all" : l'étiquette retenue est celle associée au classifieur ayant renvoyé le score plus élevé. Il convient cependant de souligner qu'elle implique d'effectuer des apprentissages aux répartitions entre catégories très déséquilibrées, ce qui soulève souvent des difficultés pratiques.

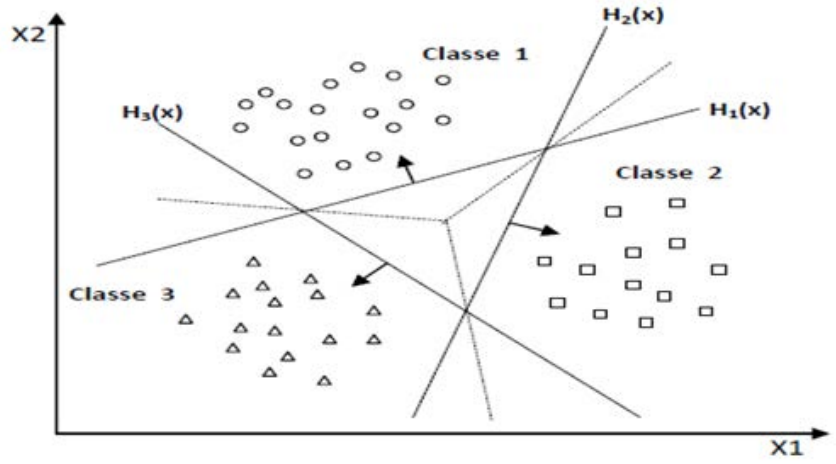


FIGURE *A.2: Classification multi-classes par la méthode une-contre-reste.*

A.3.2 Une-contre-une (1vs1)

Une autre méthode de décomposition est la méthode "une contre une". Cette méthode consiste à utiliser un classifieur par paire de classes (Figure A.3). Pour K classes, on apprend C_K^2 SVM binaires.

La classification d'un nouvel exemple s'obtient en effectuant un vote majoritaire ("max wins voting"), chaque classifieur émettant un vote associé à la classe qu'il a reconnue. Le vote de chaque classifieur peut être pondéré par une fonction du score de la sortie calculée et le nouvel exemple est affecté à la classe la plus votée.

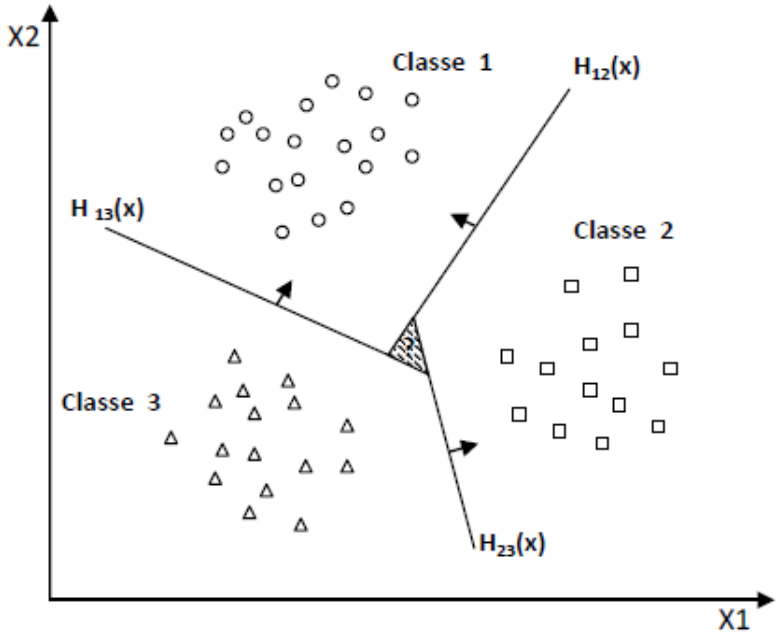


FIGURE *A.3: Classification multi-classes par la méthode une-contre-une.*

Bibliographie

[Qui, 2002] (2002). Quiet care systems generation health and elder care company.

[Abe, 2010] Abe, S. (2010). *Support Vector Machines for Pattern Classification*. Advances in Computer Vision and Pattern Recognition. Springer.

[Abowd et al., 2002] Abowd, G. D., Mynatt, E. D., and Rodden, T. (2002). The human experience. *IEEE Pervasive Computing*, 1(1) :48–57.

[Ahmad and Lee, 2008] Ahmad, M. and Lee, S.-W. (2008). Human action recognition using shape and clg-motion flow from multi-view image sequences. *Pattern Recogn.*, 41(7) :2237–2252.

[Ali et al., 2007] Ali, S., Basharat, A., and Shah, M. (2007). Chaotic invariants for human action recognition. In *ICCV*, pages 1–8. IEEE.

[Baba et al., 2004] Baba, A., Yoshizawa, S., Yamada, M., Lee, A., and Shikano, K. (2004). Acoustic models of the elderly for large-vocabulary continuous speech recognition. *Electronics and Communications in Japan (Part II : Electronics)*, 87(7) :49–57.

[Baker and Matthews, 2004] Baker, S. and Matthews, I. (2004). Lucas-kanade 20 years on : A unifying framework. *Int. J. Comput. Vision*, 56(3) :221–255.

[Banerjee and Nevatia, 2011] Banerjee, P. and Nevatia, R. (2011). Learning Neighborhood Cooccurrence Statistics of Sparse Features for Human Activity Recognition. In *2011 8th IEEE International Conference on Advanced Video and Signal-Based Surveillancei (AVSS)*, page 6.

[Barnachon et al., 2012] Barnachon, M., Bouakaz, S., Boufama, B., and Guillou, E. (2012). Human actions recognition from streamed motion capture. In *ICPR'12*, pages 3807–3810.

[Batra et al., 2008] Batra, D., Chen, T., and Sukthankar, R. (2008). Space-time shapelets for action recognition. *Motion and Video Computing, IEEE Workshop on*, 0 :1–6.

[Bay et al., 2006] Bay, H., Tuytelaars, T., and Gool, L. V. (2006). Surf : Speeded up robust features. In *ECCV*, pages 404–417.

[Benabbas et al., 2010] Benabbas, Y., Lablack, A., Ihaddadene, N., and Djeraba, C. (2010). Action recognition using direction models of motion. In *ICPR*, pages 4295–4298. IEEE.

[Bilinski and Brémond, 2012] Bilinski, P. and Brémond, F. (2012). Statistics of pairwise cooccurring local spatio-temporal features for human action recognition. In Fusiello, A., Murino, V., and Cucchiara, R., editors, *ECCV Workshops (1)*, volume 7583 of *Lecture Notes in Computer Science*, pages 311–320. Springer.

[Bilinski et al., 2013] Bilinski, P., Corvée, E., Bak, S., and Brémond, F. (2013). Relative dense tracklets for human action recognition. In *FG*, pages 1–7. IEEE.

[Blank et al., 2005] Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). Actions as space-time shapes. *Computer Vision, IEEE International Conference on*, 2 :1395–1402.

[Bobick and Davis, 2001] Bobick, A. F. and Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(3) :257–267.

[Bonnet et al., 2004] Bonnet, S., Couturier, P., Favre-Reguillon, F., and Guillemaud, R. (2004). Evaluation of postural stability by means of a single inertial sensor. *Conf Proc IEEE Eng Med Biol Soc*, 3 :2275–8.

[Brand, 1997] Brand, M. (1997). Coupled hidden markov models for modeling interacting processes. Technical report.

[Brand and Kettnaker, 2000] Brand, M. and Kettnaker, V. (2000). Discovery and segmentation of activities in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8) :844–851.

[Bregonzio et al., 2009] Bregonzio, M., Gong, S., and Xiang, T. (2009). Recognising action as clouds of space-time interest points.

[Burgio, 1997] Burgio, L. D. (1997). Audiotape intervention for agitation. Technical Report R01NR02988, National Institute of Nursing Research.

[Burgio et al., 2001] Burgio, L. D., Scilley, K., Hardin, J. M., and Hsu, C. (2001). Temporal patterns of disruptive vocalization in elderly nursing home residents. *Int J Geriatr Psychiatry*, 16(4) :378–86.

[Carlsson and Sullivan, 2001] Carlsson, S. and Sullivan, J. (2001). Action recognition by shape matching to key frames. In *IEEE Computer Society Workshop on Models versus Exemplars in Computer Vision*.

[Chikkanna and Guddeti, 2013] Chikkanna, M. and Guddeti, R. M. R. (2013). Kinect based real-time gesture spotting using hcrf. In *ICACCI*, pages 925–928. IEEE.

[D. Anderson, 2006] D. Anderson, J. Keller, M. S. X. C. e. Z. H. (2006). Recognizing falls from silhouettes. In *Proceedings of the 28th Conference of the IEEE. EMBS Annual International Conference*, pages 6388–6391.

[Dai et al., 2008] Dai, P., Di, H., Dong, L., Tao, L., and Xu, G. (2008). Group interaction analysis in dynamic context. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 39(1) :34–42.

[Dalal and Triggs, 2005] Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 886–893, Washington, DC, USA. IEEE Computer Society.

[Dalal et al., 2006] Dalal, N., Triggs, B., and Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *European Conference on Computer Vision*.

[Damen and Hogg, 2009] Damen, D. and Hogg, D. (2009). Recognizing linked events : Searching the space of feasible explanations. In *CVPR*, pages 927–934. IEEE.

[Dollar et al., 2005] Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *Proceedings of the 14th International Conference on Computer Communications and Networks*, ICCCN '05, pages 65–72, Washington, DC, USA. IEEE Computer Society.

[Efros et al., 2003] Efros, A. A., Berg, A. C., Mori, G., and Malik, J. (2003). Recognizing action at a distance. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, pages 726–, Washington, DC, USA. IEEE Computer Society.

[Farnebäck, 2003] Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *Proceedings of the 13th Scandinavian Conference on Image Analysis*, SCIA'03, pages 363–370, Berlin, Heidelberg. Springer-Verlag.

[Fezari and Bousbia-Salah, 2007] Fezari, M. and Bousbia-Salah, M. (2007). Speech and sensor in guiding an electric wheelchair. *Automatic Control and Computer Sciences*, pages 39–43.

[Fogarty et al., 2006] Fogarty, J., Au, C., and Hudson, S. E. (2006). Sensing from the basement : A feasibility study of unobtrusive and low-cost home activity recognition. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology*, UIST '06, pages 91–100, New York, NY, USA. ACM.

[Fujiyoshi and Lipton, 1998] Fujiyoshi, H. and Lipton, A. J. (1998). Real-time human motion analysis by image skeletonization. In *In Proceedings of IEEE WACV98*, pages 15–21.

[Gavrila and Davis, 1995] Gavrila, D. M. and Davis, L. S. (1995). Towards 3-d model-based tracking and recognition of human movement : a multi-view approach. In *In International Workshop on Automatic Face- and Gesture-Recognition. IEEE Computer Society*, pages 272–277.

[Gilbert et al., 2008] Gilbert, A., Illingworth, J., and Bowden, R. (2008). Scale invariant action recognition using compound features mined from dense spatio-temporal corners. In Forsyth, D. A., Torr, P. H. S., and Zisserman, A., editors, *ECCV (1)*, volume 5302 of *Lecture Notes in Computer Science*, pages 222–233. Springer.

[Gilbert et al., 2009] Gilbert, A., Illingworth, J., and Bowden, R. (2009). Fast realistic multi-action recognition using mined dense spatio-temporal features. In *ICCV*, pages 925–931. IEEE.

[Gilbert et al., 2011] Gilbert, A., Illingworth, J., and Bowden, R. (2011). Action recognition using mined hierarchical compound features. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5) :883–897.

[Gong et al., 2010] Gong, W., Bagdanov, A. D., Roca, F. X., and Gonzàlez, J. (2010). Automatic key pose selection for 3d human action recognition. In *Proceedings of the 6th International Conference on Articulated Motion and Deformable Objects*, AMDO'10, pages 290–299, Berlin, Heidelberg. Springer-Verlag.

[Gorelick et al., 2007] Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12) :2247–2253.

[Grandvalet et al., 2006] Grandvalet, Y., Mariéthoz, J., and Bengio, S. (2006). A probabilistic interpretation of svms with an application to unbalanced classification. *NIPS*, pages 467–474.

[Gunawardana et al., 2005] Gunawardana, A., Mahajan, M., Acero, A., and Platt, J. C. (2005). Hidden conditional random fields for phone classification. In *in Interspeech*, pages 1117–1120.

[Gupta et al., 2009] Gupta, A., Kembhavi, A., and Davis, L. S. (2009). Observing human-object interactions : Using spatial and functional compatibility for recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(10) :1775–1789.

[Hamel, 2009] Hamel, L. (2009). *Knowledge Discovery with Support Vector Machines*. Wiley Series on Methods and Applications in Data Mining. Wiley.

[Hammersley and Clifford, 1971] Hammersley, J. M. and Clifford, P. E. (1971). Markov random fields on finite graphs and lattices. Unpublished manuscript.

[Han et al., 2010] Han, L., Wu, X., Liang, W., Hou, G., and Jia, Y. (2010). Discriminative human action recognition in the learned hierarchical manifold space. *Image and Vision Computing*, 28(5) :836 – 849. Best of Automatic Face and Gesture Recognition 2008.

[Harris and Stephens, 1988] Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference*, pages 147–151.

[Huang and Trivedi, 2005] Huang, K. S. and Trivedi, M. M. (2005). 3d shape context based gesture analysis integrated with tracking using omni video array. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 0 :80.

[Istrate et al., 2006] Istrate, D., Vacher, M., and Serignat, J.-F. (2006). Generic implementation of a distress sound extraction system for elder care. In *28th IEEE EMBS Annual International Conference*, pages 3309–3312, New York City, USA.

[Johansson, 1973] Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14 :201–211.

[Joumier et al., 2011] Joumier, V., Romdhane, R., Bremond, F., Thonnat, M., Mulin, E., Robert, P. H., Derreumaux, A., Piano, J., and Lee, J., R. (2011). Video Activity Recognition Framework for assessing motor behavioural disorders in Alzheimer Disease Patients. In *International Workshop on Behaviour Analysis and Video Understanding (ICVS 2011)*, page 9, Sophia Antipolis, France.

[Kaâniche and Bremond, 2012] Kaâniche, M.-B. and Bremond, F. (2012). Recognizing Gestures by Learning Local Motion Signatures of HOG Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. ST Microelectronics Rousset which supported this work under PS26/27 Smart Environment project financed by Conseil Régional Provence-Alpes-Côte d'Azur.

[Ke et al., 2005] Ke, Y., Sukthankar, R., and Hebert, M. (2005). Efficient visual event detection using volumetric features. In *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 - Volume 01*, ICCV '05, pages 166–173, Washington, DC, USA. IEEE Computer Society.

[Ke et al., 2010] Ke, Y., Sukthankar, R., and Hebert, M. (2010). Volumetric features for video event detection. *International Journal of Computer Vision*, 88(3) :339–362.

[Khadem and Rajan, 2009] Khadem, B. S. and Rajan, D. (2009). Appearance-based action recognition in the tensor framework. In *CIRA*, pages 398–403. IEEE.

[Kjellström et al., 2011] Kjellström, H., Romero, J., and Kragić, D. (2011). Visual object-action recognition : Inferring object affordances from human demonstration. *Comput. Vis. Image Underst.*, 115(1) :81–90.

[Klaser et al., 2008] Klaser, A., Marszalek, M., and Schmid, C. (2008). A Spatio-Temporal Descriptor Based on 3D-Gradients. In Everingham, M., Needham, C., and Fraile, R., editors, *BMVC 2008 - 19th British Machine Vision Conference*, pages 275 :1–10, Leeds, United Kingdom. British Machine Vision Association.

[Kong et al., 2012] Kong, Y., Jia, Y., and Fu, Y. (2012). Learning human interaction by interactive phrases. In Fitzgibbon, A. W., Lazebnik, S., Perona, P., Sato, Y., and Schmid, C., editors, *ECCV (1)*, volume 7572 of *Lecture Notes in Computer Science*, pages 300–313. Springer.

[Koppula et al., 2013] Koppula, H., Gupta, R., and Saxena, A. (2013). Learning human activities and object affordances from rgb-d videos. *IJRR*, 32(8) :951–970.

[Koppula and Saxena, 2013] Koppula, H. and Saxena, A. (2013). Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation. In Dasgupta, S. and Mcallester, D., editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 792–800. JMLR Workshop and Conference Proceedings.

[Kovashka and Grauman, 2010] Kovashka, A. and Grauman, K. (2010). Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *CVPR*, pages 2046–2053. IEEE.

[Krumm et al., 2000] Krumm, J., Harris, S., Meyers, B., Brumitt, B., Hale, M., and Shafer, S. (2000). Multi-camera multi-person tracking for easyliving. In *Proceedings of the Third IEEE International Workshop on Visual Surveillance (VS'2000)*, VS '00, pages 3–, Washington, DC, USA. IEEE Computer Society.

[Kuehne et al., 2011] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). HMDB : a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

[KUMIKO, 2004] KUMIKO, O., M. M. A. E. S. S. e. R. T. (2004). Input support for elderly people using speech recognition. Technical report, Institute of Electronics, Information and Communication Engineers.

[Lafferty et al., 2001] Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Laptev, 2005] Laptev, I. (2005). On space-time interest points. *Int. J. Comput. Vision*, 64(2-3) :107–123.

[Laptev et al., 2007] Laptev, I., Caputo, B., Schüldt, C., and Lindeberg, T. (2007). Local velocity-adapted motion events for spatio-temporal recognition. *Computer Vision and Image Understanding*, 108(3) :207–229.

[Laptev and Lindeberg, 2003] Laptev, I. and Lindeberg, T. (2003). Space-time interest points. In *IN ICCV*, pages 432–439.

[Laptev and Lindeberg, 2004] Laptev, I. and Lindeberg, T. (2004). Local descriptors for spatio-temporal recognition. In *In First International Workshop on Spatial Coherence for Visual Motion Analysis*.

[Laptev et al., 2008] Laptev, I., Marszałek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Conference on Computer Vision & Pattern Recognition*.

[Lee and Mihailidis, 2005] Lee, T. and Mihailidis, A. (2005). An intelligent emergency response system : preliminary development and testing of automated fall detection. *Journal of telemedicine and telecare*, 11(4) :194–198.

[Lin et al., 2009] Lin, Z., Jiang, Z., and Davis, L. S. (2009). Recognizing actions by shape-motion prototype trees. In *ICCV*, pages 444–451. IEEE.

[Liu and Nocedal, 1989] Liu, D. C. and Nocedal, J. (1989). On the limited memory bfgs method for large scale optimization. *Math. Program.*, 45(3) :503–528.

[Liu and Jia, 2008] Liu, F. and Jia, Y. (2008). Human action recognition using manifold learning and hidden conditional random fields. In *ICYCS*, pages 693–698. IEEE Computer Society.

[Liu et al., 2013] Liu, L., Shao, L., and Rockett, P. (2013). Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition. *Pattern Recogn.*, 46(7) :1810–1818.

[Lowe, 1999] Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ICCV '99, pages 1150–, Washington, DC, USA. IEEE Computer Society.

[Lu and Little, 2006] Lu, W.-L. and Little, J. J. (2006). Simultaneous tracking and action recognition using the pca-hog descriptor. In *CRV*, page 6. IEEE Computer Society.

[Lv and Nevatia, 2006] Lv, F. and Nevatia, R. (2006). Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part IV*, ECCV'06, pages 359–372, Berlin, Heidelberg. Springer-Verlag.

[Lv and Nevatia, 2007] Lv, F. and Nevatia, R. (2007). Single view human action recognition using key pose matching and viterbi path searching. In *CVPR*. IEEE Computer Society.

[Marr and Nishihara, 1978] Marr, D. and Nishihara, H. K. (1978). Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes. *Proceedings of the Royal Society of LondonBiological Sciences*, 200(1140) :269–294.

[Marszałek et al., 2009] Marszałek, M., Laptev, I., and Schmid, C. (2009). Actions in context. In *IEEE Conference on Computer Vision & Pattern Recognition*.

[Matikainen et al., 2009] Matikainen, P., Hebert, M., and Sukthankar, R. (2009). Trajectons : Action recognition through the motion analysis of tracked features. In *Workshop on Video-Oriented Object and Event Classification, ICCV 2009*.

[Matikainen et al., 2010] Matikainen, P., Hebert, M., and Sukthankar, R. (2010). Representing pairwise spatial and temporal relations for action recognition. In Daniilidis, K., Maragos, P., and Paragios, N., editors, *ECCV (1)*, volume 6311 of *Lecture Notes in Computer Science*, pages 508–521. Springer.

[Medjahed et al., 2008] Medjahed, H., Istrate, D., Boudy, J., Baldinger, J.-L., Dorizzi, B., Belfeki, I., Martins, V., Steenkeste, F., and Andreao, R. (2008). A multimodal platform for database recording and elderly people monitoring. In Encarnação, P. and Veloso, A., editors, *BIOSIGNALS (2)*, pages 385–392. INSTICC - Institute for Systems and Technologies of Information, Control and Communication.

[Mendoza and de la Blanca, 2008] Mendoza, M. n. and de la Blanca, N. P. (2008). Applying space state models in human action recognition : A comparative study. In López, F. J. P. and Fisher, R. B., editors, *AMDO*, volume 5098 of *Lecture Notes in Computer Science*, pages 53–62. Springer.

[Messing et al., 2009] Messing, R., Pal, C., and Kautz, H. A. (2009). Activity recognition using the velocity histories of tracked keypoints. In *ICCV*, pages 104–111. IEEE.

[Moore and Essa, 2002] Moore, D. and Essa, I. (2002). Recognizing multitasked activities from video using stochastic context-free grammar. In *Eighteenth National Conference on Artificial Intelligence*, pages 770–776, Menlo Park, CA, USA. American Association for Artificial Intelligence.

[Morency et al., 2007] Morency, L. P., Quattoni, A., and Darrell, T. (2007). Latent-Dynamic Discriminative Models for Continuous Gesture Recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8. IEEE.

[Nait-Charif and McKenna, 2004] Nait-Charif, H. and McKenna, S. J. (2004). Activity summarisation and fall detection in a supportive home environment. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 4, pages 323–326. IEEE.

[Nguyen et al., 2005] Nguyen, N. T., Phung, D. Q., Venkatesh, S., and Bui, H. (2005). Learning and detecting activities from movement trajectories using the hierarchical hidden markov models. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02*, CVPR '05, pages 955–960, Washington, DC, USA. IEEE Computer Society.

[Niebles et al., 2010] Niebles, J. C., wei Chen, C., and Fei-fei, L. (2010). Modeling temporal structure of decomposable motion segments for activity classification. In *in Proc. 11th European Conf. Comput. Vision, 2010*, pages 392–405.

[Noury et al., 2004] Noury, N., Barralon, P., Couturier, P., Favre-Reguillon, F., Guillemaud, R., Mestais, C., Caritu, Y., David, D., Moine, S., Franco, A., Guiraud-By, F., Berenguer, M., and Provost, H. (2004). Actidom–a microsystem based on mems for activity monitoring of the frail elderly in their daily life. *Conf Proc IEEE Eng Med Biol Soc*, 5 :3305–8.

[Oikonomopoulos et al., 2005] Oikonomopoulos, A., Patras, I., and Pantic, M. (2005). Spatio-temporal saliency for human action recognition. In *Proceedings of IEEE Int'l Conf. Multimedia and Expo (ICME'05)*, pages 430–433, Amsterdam, The Netherlands.

[Parameswaran and Chellappa, 2006] Parameswaran, V. and Chellappa, R. (2006). View invariance for human action recognition. *Int. J. Comput. Vision*, 66(1) :83–101.

[Pélissier, 2003] Pélissier, L. (2003). Les nouvelles technologies au service des personnes âgées en perte d'autonomie. Stage A03-R-489 || pelissier03a. Diplôme de Recherche Technologique (DRT). Rapport de stage.

[Platt, 1999] Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press.

[Prest et al., 2013] Prest, A., Ferrari, V., and Schmid, C. (2013). Explicit modeling of human-object interactions in realistic videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(4) :835–848.

[Quattoni et al., 2004] Quattoni, A., Collins, M., and Darrell, T. (2004). Conditional random fields for object recognition. In *In NIPS*, pages 1097–1104. MIT Press.

[Quattoni et al., 2007] Quattoni, A., Wang, S., Morency, L.-P., Collins, M., and Darrell, T. (2007). Hidden conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10) :1848–1852.

[Ragheb et al., 2008] Ragheb, H., Velastin, S. A., Remagnino, P., and Ellis, T. (2008). In *ICIP*, pages 753–756. IEEE.

[Ramanan and Forsyth, 2003] Ramanan, D. and Forsyth, D. A. (2003). Automatic annotation of everyday movements. In Thrun, S., Saul, L. K., and Schölkopf, B., editors, *NIPS*. MIT Press.

[Raptis and Sigal, 2013] Raptis, M. and Sigal, L. (2013). Poselet key-framing : A model for human activity recognition. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 2650–2657, Washington, DC, USA. IEEE Computer Society.

[Raptis and Soatto, 2010] Raptis, M. and Soatto, S. (2010). Tracklet descriptors for action modeling and video analysis. In *Proceedings of the 11th European Conference on Computer Vision : Part I*, ECCV'10, pages 577–590, Berlin, Heidelberg. Springer-Verlag.

[Robertson and Reid, 2005] Robertson, N. M. and Reid, I. D. (2005). Behaviour understanding in video : A combined method. In *ICCV*, pages 808–815. IEEE Computer Society.

[Rougier et al., 2011] Rougier, C., Meunier, J., St-Arnaud, A., and Rousseau, J. (2011). Robust video surveillance for fall detection based on human shape deformation. *IEEE Trans. Circuits Syst. Video Techn.*, 21(5) :611–622.

[Ryoo, 2011] Ryoo, M. S. (2011). Human activity prediction : Early recognition of ongoing activities from streaming videos. In Metaxas, D. N., Quan, L., Sanfeliu, A., and Gool, L. J. V., editors, *ICCV*, pages 1036–1043. IEEE.

[Ryoo and Aggarwal, 2010] Ryoo, M. S. and Aggarwal, J. K. (2010). UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA). http ://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html.

[S. Renouard, 2003] S. Renouard, M. C. e. G. C. (2003). Vocal interface with a speech memory for dependent people. In *ICOST proceedings*, pages 15–21.

[Satkin and Hebert, 2010] Satkin, S. and Hebert, M. (2010). Modeling the temporal extent of actions. In Daniilidis, K., Maragos, P., and Paragios, N., editors, *ECCV (1)*, volume 6311 of *Lecture Notes in Computer Science*, pages 536–548. Springer.

[Schuldt et al., 2004] Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions : A local svm approach. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03*, ICPR '04, pages 32–36, Washington, DC, USA. IEEE Computer Society.

[Scovanner et al., 2007] Scovanner, P., Ali, S., and Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th International Conference on Multimedia*, MULTIMEDIA '07, pages 357–360, New York, NY, USA. ACM.

[Shao and Chen, 2010] Shao, L. and Chen, X. (2010). Histogram of body poses and spectral regression discriminant analysis for human action categorization. In *Proceedings of the British Machine Vision Conference*, pages 88.1–88.11. BMVA Press. doi :10.5244/C.24.88.

[Sheikh et al., 2005] Sheikh, Y., Sheikh, M., and Shah, M. (2005). Exploring the space of a human action. In *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 - Volume 01*, ICCV '05, pages 144–149, Washington, DC, USA. IEEE Computer Society.

[Sminchisescu et al., 2006] Sminchisescu, C., Kanaujia, A., and Metaxas, D. (2006). Conditional models for contextual human motion recognition. *Computer Vision and Image Understanding*, 104(2-3) :210–220.

[Sullivan and Carlsson, 2002] Sullivan, J. and Carlsson, S. (2002). Recognizing and tracking human action. In *Proceedings of the 7th European Conference on Computer Vision-Part I*, ECCV '02, pages 629–644, London, UK, UK. Springer-Verlag.

[Sun et al., 2009] Sun, J., Wu, X., Yan, S., Cheong, L. F., Chua, T.-S., and Li, J. (2009). Hierarchical spatio-temporal context modeling for action recognition. In *CVPR*, pages 2004–2011. IEEE.

[Sung et al., 2011] Sung, J., Ponce, C., Selman, B., and Saxena, A. (2011). Human activity detection from rgbd images. In *Plan, Activity, and Intent Recognition*, volume WS-11-16 of *AAAI Workshops*. AAAI.

[Takahashi et al., 2003] Takahashi, S., Morimoto, T., Maeda, S., and Tsuruta, N. (2003). Dialogue experiment for elderly people in home health care system. In Matousek, V. and Mautner, P., editors, *TSD*, volume 2807 of *Lecture Notes in Computer Science*, pages 418–423. Springer.

[Tang et al., 2012] Tang, K., Fei-fei, L., and Koller, D. (2012). Learning latent temporal structure for complex event detection. In *In CVPR*.

[Tenorth et al., 2009] Tenorth, M., Bandouch, J., and Beetz, M. (2009). The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition. In *IEEE International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS), in conjunction with ICCV2009*.

[Thome et al., 2008] Thome, N., Miguet, S., and sébastien Ambellouis (2008). A Real-Time, Multi-View Fall Detection System : a LHMM-Based Approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11) :1522–1532. Special Issue on Event Analysis in Videos.

[Thurau and Hlavác, 2010] Thurau, C. and Hlavác, V. (2010). Pose primitive based human action recognition in videos or still images. In *CVPR*. IEEE Computer Society.

[Ullah, 2012] Ullah, M. M. (2012). *Supervised Statistical Representations for Human Actions Recognition in Video*. PhD thesis, RENNES 1university, France.

[Vacher et al., 2006] Vacher, M., Serignat, J., Chaillol, S., Istrate, D., and Popescu, V. (2006). Speech and sound use in a remote monitoring system for health care. 4188/2006 :711–718.

[Vahdat et al., 2011] Vahdat, A., Gao, B., Ranjbar, M., and Mori, G. (2011). A discriminative key pose sequence model for recognizing human interactions. In *ICCV Workshops*, pages 1729–1736. IEEE.

[Valentin, 2010] Valentin, N. Z. (2010). *Multisensor Fusion for Monitoring Elderly Activities at Home*. PhD thesis, Nice-Sophia Antipolis university, France.

[Vasilescu and Terzopoulos, 2002] Vasilescu, M. A. O. and Terzopoulos, D. (2002). Multilinear analysis of image ensembles : Tensorfaces. In *IN PROCEEDINGS OF THE EUROPEAN CONFERENCE ON COMPUTER VISION*, pages 447–460.

[Veeraraghavan and Roy-chowdhury, 2006] Veeraraghavan, A. and Roy-chowdhury, A. K. (2006). The function space of an activity. In *in Proc. Comput. Vis. Pattern Recognit*, pages 959–968.

[Vinel, 2013] Vinel, A. (2013). *Champs Markoviens Conditionnels pour l'étiquetage de séquences*. PhD thesis, Université Pierre et Marie CURIE university, France.

[Virone and Istrate, 2007] Virone, G. and Istrate, D. (2007). Integration of an environmental sound module to an existing in-home activity simulator. *Conf Proc IEEE Eng Med Biol Soc*, 2007 :3810–3.

[Wallach, 2003] Wallach, H. (2003). Efficient training of conditional random fields. In *the 6th Annual Computational Linguistics U.K. Research Colloquium*, pages 1097–1104, Edinburgh, U.K.

[Wang et al., 2011a] Wang, H., Klaser, A., Schmid, C., and Liu, C.-L. (2011a). Action recognition by dense trajectories. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 3169–3176, Washington, DC, USA. IEEE Computer Society.

[Wang et al., 2011b] Wang, J., Chen, Z., and Wu, Y. (2011b). Action recognition with multiscale spatio-temporal contexts. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 3185–3192, Washington, DC, USA. IEEE Computer Society.

[Wang and Suter, 2007] Wang, L. and Suter, D. (2007). Recognizing human activities from silhouettes : Motion subspace and factorial discriminative graphical model. In *CVPR*. IEEE Computer Society.

[Wang et al., 2006] Wang, S. B., Quattoni, A., Morency, L.-P., and Demirdjian, D. (2006). Hidden conditional random fields for gesture recognition. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 1521–1527, Washington, DC, USA. IEEE Computer Society.

[Wang and Mori, 2008] Wang, Y. and Mori, G. (2008). Learning a discriminative hidden part model for human action recognition. In *Advances in Neural Information Processing Systems (NIPS) 21*.

[Wang and Mori, 2009] Wang, Y. and Mori, G. (2009). Max-margin hidden conditional random fields for human action recognition. In *CVPR*.

[Wang et al., 2007] Wang, Y., Sabzmeydani, P., and Mori, G. (2007). Semi-latent dirichlet allocation : A hierarchical model for human action recognition. In *In 2nd Workshop on Human Motion Understanding, Modeling, Capture and Animation*.

[Weinland and Boyer, 2008] Weinland, D. and Boyer, E. (2008). Action Recognition using Exemplar-based Embedding. In *CVPR 2008 - IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, Anchorage, États-Unis. IEEE Computer Society.

[Weinland et al., 2006] Weinland, D., Ronfard, R., and Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. *Comput. Vis. Image Underst.*, 104(2) :249–257.

[Willems et al., 2008] Willems, G., Tuytelaars, T., and Gool, L. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proceedings of the 10th European Conference on Computer Vision : Part II*, ECCV '08, pages 650–663, Berlin, Heidelberg. Springer-Verlag.

[Wong and Cipolla, 2007] Wong, S.-F. and Cipolla, R. (2007). Extracting spatiotemporal interest points using global information. In *ICCV*, pages 1–8. IEEE.

[Wu et al., 2004] Wu, T.-F., Lin, C.-J., and Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *J. Mach. Learn. Res.*, 5 :975–1005.

[Xia et al., 2012] Xia, L., Chen, C.-C., and Aggarwal, J. K. (2012). View invariant human action recognition using histograms of 3d joints. In *CVPR Workshops*, pages 20–27. IEEE.

[Yacoob and Black, 1998] Yacoob, Y. and Black, M. J. (1998). Parameterized modeling and recognition of activities. In *Proceedings of the Sixth International Conference on Computer Vision*, ICCV '98, pages 120–, Washington, DC, USA. IEEE Computer Society.

[Yamato et al., 1992] Yamato, J., Ohya, J., and Ishii, K. (1992). Recognizing human action in time-sequential images using hidden Markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR '92., 1992 IEEE Computer Society Conference on*, pages 379–385.

[Yao et al., 2010] Yao, A., Gall, J., and Gool, L. J. V. (2010). A hough transform-based voting framework for action recognition. In *CVPR*, pages 2061–2068. IEEE.

[Yao and Fei-Fei, 2010] Yao, B. and Fei-Fei, L. (2010). Modeling mutual context of object and human pose in human-object interaction activities. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 17–24. IEEE.

[Yilmaz, 2005] Yilmaz, A. (2005). Recognizing human actions in videos acquired by uncalibrated moving cameras. In *in : IEEE International Conference on Computer Vision*, pages 150–157.

[Yuan et al., 2012] Yuan, F., Xia, G.-S., Sahbi, H., and Prinet, V. (2012). Mid-level features and spatio-temporal context for activity recognition. *Pattern Recogn.*, 45(12) :4182–4191.

[Yuan et al., 2009] Yuan, J., Liu, Z., and Wu, Y. (2009). Discriminative subvolume search for efficient action detection. In *CVPR*, pages 2442–2449. IEEE.

[Zhang et al., 2010] Zhang, D., Wang, Y., and Bhanu, B. (2010). Ethnicity classification based on gait using multi-view fusion. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 108–115. IEEE.

[Zhang and Gong, 2010] Zhang, J. and Gong, S. (2010). Action categorization with modified hidden conditional random field. *Pattern Recogn.*, 43(1) :197–203.

[Zhang et al., 2007] Zhang, J., Marszałek, M., Lazebnik, S., and Schmid, C. (2007). Local features and kernels for classification of texture and object categories : a comprehensive study. *International Journal of Computer Vision*, 73(2) :213–238.

[Zhang et al., 2012] Zhang, Y., Liu, X., Chang, M.-C., Ge, W., and Chen, T. (2012). Spatio-temporal phrases for activity recognition. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part III*, ECCV'12, pages 707–721, Berlin, Heidelberg. Springer-Verlag.

[Zhang et al., 2008] Zhang, Z., Hu, Y., Chan, S., and Chia, L. (2008). Motion context : A new representation for human action recognition. In *Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV*, pages 817–829.

[Zhao and Elgammal, 2008] Zhao, Z. and Elgammal, A. (2008). Information theoretic key frame selection for action recognition. In *Proceedings of the British Machine Vision Conference*, pages 109.1–109.10. BMVA Press. doi :10.5244/C.22.109.

[Zhen and Shao, 2013] Zhen, X. and Shao, L. (2013). Spatio-temporal steerable pyramid for human action recognition. *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 0 :1–6.

[Ziaeefard and Ebrahimnezhad, 2010] Ziaeefard, M. and Ebrahimnezhad, H. (2010). Hierarchical human action recognition by normalized-polar histogram. In *ICPR'10*, pages 3720–3723.

[Zouba et al., 2008] Zouba, N., Boulay, B., Brémond, F., and Thonnat, M. (2008). Monitoring Activities of Daily Living (ADLs) of Elderly Based on 3D Key Human Postures. In Barbara Caputo, M. V., editor, *International Cognitive Vision Workshop*, volume 5329 of *Lecture notes in computer science*, pages 37–50, Santorini, Grèce. Springer Berlin / Heidelberg.