Analyse et reconnaissance de séquences vidéos d'activités humaines dans l'espace sémantique

Cyrille Beaudry

UNIVERSITÉ DE LA ROCHELLE

ÉCOLE DOCTORALE SCIENCES ET INGÉNIERIE POUR L'INFORMATION

Laboratoire Mathématiques, Image et Applications

THÈSE présentée par :

Cyrille BEAUDRY

26 Novembre 2015

pour obtenir le grade de : **Docteur de l'université de La Rochelle**

Spécialité : **Automatique, Image et Signal**

Analyse et reconnaissance de séquences vidéos d'activités humaines dans l'espace sémantique

---

JURY :

Jenny BENOIS-PINEAU	Professeur, Labri, Bordeaux, Rapporteur
François BREMOND	Directeur de Recherche, INRIA, Sophia Antipolis Nice, Rapporteur
Mohamed DAOUDI	Professeur, Institut Mines-Télécom, Lille, Examinateur
Ewa KIJAK	Maître de conférences, IRISA-INRIA, Rennes, Examinateur
Laurent MASCARILLA	Maître de conférences, MIA, La Rochelle, Directeur de thèse
Renaud PÉTERI	Maître de conférences, MIA, La Rochelle, Co-directeur de thèse

Note au lecteur

Le domaine d'application de ce manuscrit étant l'analyse de vidéos de nombreuses illustrations ou résultats se présentent sous la forme de séquences d'images. Ces séquences d'images sont visionnables à partir des liens hypertextes situés dans les commentaires des figures concernées. Ces vidéos sont toutes disponibles sur le lien suivant : vimeo.com/user44277041. Le nom de chaque vidéo correspond à la référence de la figure qu'elle illustre.

Table des matières

Chapitre 1

Introduction

Chapitre 1

Introduction

Sommaire

Dans la première partie de ce chapitre, nous présentons le contexte général dans lequel intervient nos travaux. Nous discutons de l'accroissement de l'utilisation du support vidéo dans différents domaines : la télécommunication, la sécurité ainsi que l'assistance domestique. Par la suite, nous détaillons l'intérêt grandissant pour la reconnaissance des activités humaines dans des vidéos.

Dans la deuxième partie, nous présentons les différentes problématiques inhérentes à cette thématique, qui en font l'une des plus attractives mais aussi l'une des plus complexes dans la communauté de la vision par ordinateur.

Finalement, dans la troisième partie, nous détaillons l'ensemble des contributions associées à nos travaux qui seront présentées tout au long de ce manuscrit.

1.1 Contexte

1.1.1 Importance croissante du support vidéo

Cette thèse de doctorat traite de la détection et de la reconnaissance automatiques des activités humaines dans des vidéos. Ces dernières années, cette thématique a reçu une attention particulière dans la communauté de la vision par ordinateur et de la reconnaissance de forme. Nous présentons ici un ensemble d'éléments factuels permettant de comprendre les raisons de cet intérêt grandissant, les motivations, ainsi que les enjeux sous-jacents.

Domaine du grand public Notre quotidien a été amplement bouleversé par les avancées dans le domaine de l'informatique et l'expansion d'internet. Depuis ces quinze dernières années, ces deux domaines ont permis la création de nouveaux modes de communication, de travail et de consommation. Ces trois éléments de société ont été transformés par de nouveaux outils d'information dont l'efficacité est justifiée par l'utilisation de différents supports, notamment le texte, le son, l'image et la vidéo. Parmi ces vecteurs d'informations, le support vidéo est celui qui connait l'une des progressions les plus importantes.

Cette importance s'observe à la fois dans les moyens de communication et de travail, où des outils de visioconférence se sont largement popularisés. En effet, en 2013, l'application de visioconférence `Skype` concentrait à elle seule 1/3 des communications internationales [1], en 2014, 111 000 appels vidéos ont été passé chaque minute. Nous consommons également une énorme quantité d'information vidéos. De part l'émergence d'appareils électroniques mobiles munis de capteurs (appareils photos numériques, ordinateurs portables, smartphones, etc), d'un accès à internet rapide et démocratisé, ainsi que des capacités de stockage de plus en plus importantes, il est aujourd'hui simple de créer, partager et consommer du contenu vidéo.

Aujourd'hui, en une minute, 300 heures de vidéos sont visionnées sur `YouTube` ; 77 160 heures de vidéos sont vues sur `Netflix` et un million de vidéos ont été regardées sur l'application mobile vidéo `Vine`. En France, le support vidéo représentera 59% du trafic de l'internet ouvert en 2015, soit une croissance de 27% par an [2] et représentait déjà en 2014 60% du trafic web mondial [3]. Dès lors, on estime que le trafic web mondial sera, à 84% constitué de vidéos d'ici 2018 [4].

Domaine de la sécurité L'intérêt pour la vidéo ne touche pas seulement le secteur des télécommunications via internet. Le support vidéo est également en constant développement dans le domaine de la sécurité, notamment celui de la vidéo-surveillance. La nécessité d'assurer la sécurité via des systèmes globalisés oriente le choix de nombreux états et entreprises. Ce contexte donne une ampleur sans précédent à la vidéo-surveillance.

1. Rapport de l'Institut TeleGeography
2. Étude menée par l'IDATE
3. Étude de Shutterstock et Comscore
4. Étude de CISCO

Ces dernières années, le marché de la surveillance vidéo connaît des taux de croissance annuels à deux chiffres [5]. La `Cnil` fait état de 935 000 caméras de surveillance sur l'ensemble du territoire français (voie publique, transport en commun, commerces, entreprises, etc.) [6]. Cette accélération est accompagnée par un consentement grandissant de la population. En 2013, 75% des Français était favorables à la vidéo-surveillance [7].

Domaine de l'assistance Le marché des systèmes d'assistance domestique est également en pleine expansion. En effet, avec l'augmentation de l'espérance de vie ces dernières décennies, nous sommes confrontés à de nouvelles problématiques sociétales. Ces changements nous interrogent sur notre capacité à adapter notre société et assurer le maintien d'une bonne qualité de vie à un âge avancé. Une part grandissante de notre société peut désormais rester à domicile, cependant, ces personnes, souvent dépendantes, sont seules. Le taux de solitude explose chez les plus de 75 ans. D'après la `Fondation de France`, 27% des personnes âgées vivent seules (contre 16% en 2010) [8].

Les personnes âgées sont, avec les enfants en bas âge, les plus touchées par les accidents domestiques. La chute reste la principale cause de mortalité chez les plus de 65 ans. Pour répondre à cette urgence, les systèmes d'assistance domestiques se démocratisent. On dénombre plus de 9000 décès de séniors, chaque année, imputés à des chutes dans leur domicile [9]. Ces nouvelles technologies permettent d'assurer une surveillance et une assistance face aux risques du quotidien (chute, incendie, etc). Ces systèmes utilisent de plus en plus de caméras de surveillance, ou de webcams, afin de fournir un relai visuel à des tiers extérieurs. Les vidéos sont par la suite traitées et analysées via un centre de supervision ou via un utilisateur tiers (souvent membre de la famille) qui assure la surveillance du sujet.

1.1.2 Nécessité d'un traitement automatisé

Le premier constat que l'on peut faire à partir des éléments décrits précédemment (grand public, sécurité, assistance) est celui de la gestion de la quantité de données générées. Face à leur accroissement exponentiel, on est amené à se demander comment ces vidéos doivent être traitées, analysées et interprétées. Les chiffres des données vidéos téléversées sur internet mettent en évidence l'impossibilité du traitement manuel de ces dernières, que ce soit pour leur vérification (contenu adéquat à une charte de publication), mais également pour leur indexation.

Il en est de même pour les systèmes de vidéo-surveillance. L'un des arguments que l'on peut opposer à l'utilisation de cette technologie est l'incertitude sur la relation entre installation de vidéosurveillance et efficacité à long terme, notamment sur la baisse de la délinquance. Les études pointent avant tout le coût et la difficulté qu'impose le tri de l'information enregistrée par les opérateurs responsables de centres de supervisions.

5. Rapport de IHS Technology
6. Communiqué de presse de la Cnil
7. Sondage de l'institut BVA pour Le Figaro
8. Étude de la Fondation de France
9. Étude menée par l'INSERM

Les systèmes d'assistance domestique se basent également sur le principe de centres de supervision privés, à cette différence près qu'ils nécessitent moins de caméras et un nombre d'opérateurs plus faible. Dans les autres cas, lorsqu'elle n'est pas attribuée à un centre privé, la surveillance est laissée à un tiers extérieur, souvent membre de la famille. Le contrôle visuel est fait par un opérateur à une fréquence plus ou moins grande et est faite selon ses capacités d'attention.

Le traitement automatique de ces données est une évidence à la fois pour la gestion au quotidien du flux croissant de données issues des vidéos mais également pour améliorer la productivité et l'efficacité des différents domaines cités plus haut. Nous verrons par la suite comment la reconnaissance automatique d'actions humaines peut apporter des éléments de solution dans chacun de ces secteurs.

1.1.3 Les activités humaines comme élément discriminant

Les vidéos issues d'internet sont aujourd'hui majoritairement indexées à l'aide de descripteurs bas niveau, notamment par des éléments saisis par l'utilisateur. Cependant avec la croissance des données versées sur internet, indexer ces vidéos en analysant automatiquement leur contenu devient une nécessité. En effet, ces informations, si elles sont mal archivées, peuvent être considérées comme perdues. Il existe différents éléments pertinents capables de rendre compte du contenu d'une vidéo (scène, textures dynamiques, mouvement global, etc.). On constate, à travers le langage naturel, que la description d'une vidéo passe par la narration de quelques évènements clés se produisant dans celle-ci. Décrire une scène correspond à définir les évènements qui s'y produisent, les actions qui y sont exécutées. Le mouvement reste l'une des principales caractéristiques descriptives d'une séquence vidéo. De plus, les travaux de [Laptev, 2013] tendent à montrer qu'en moyenne, sur chaque vidéo issue du web, 35% des pixels représentent des humains. Cette donnée laisse à penser que décrire une vidéo à partir des mouvements et des activités humaines présents dans cette dernière est un des critères discriminatifs et pertinents pour l'indexation de vidéos issues du web.

La difficulté du traitement des données issues des caméras de vidéo-surveillance est le problème le plus souvent cité par ses détracteurs avec le non entretien des caméras. En Grande-Bretagne par exemple, on estime que 15% du temps passé par les opérateurs devant leurs écrans de contrôle relèverait du voyeurisme, 68% des personnes surveillées le sont en raison de leur couleur de peau, tout comme 86% des jeunes de moins de 30 ans, et 93% des hommes [10]. Assister ces opérateurs de technologies permettant la reconnaissance de mouvements d'intérêts (agression, vol, etc) à partir des images capturées permettrait à la fois de faciliter le travail de ces derniers mais également pérenniser l'efficacité de la surveillance.

Dans le cas de l'assistance domestique, le système de contrôle visuel est laissé aux opérateurs ou à un tiers extérieur. La détection d'un danger se fait visuellement, lorsqu'un tiers se connecte aux caméras du domicile. Sinon, il peut être alerté à l'aide d'un dispositif supplémentaire le plus souvent fixé sur la personne âgée. Un tel dispositif doit être porté constamment et peut

10. Propos de Mr. Noé Le Blanc dans Le Monde Diplomatique

représenter, à terme, une gêne pour le porteur. Dans les deux cas (tiers extérieur ou centre de supervision) se pose le problème de l'atteinte à la vie privée, d'autant plus que le sujet est observé à son domicile. L'intérêt dans ce domaine serait que le système de contrôle visuel effectue une reconnaissance automatique des mouvements afin de détecter les situations de danger, notamment les chutes, puis alerter les personnes concernées uniquement à ce moment. Ceci permettrait à la fois d'améliorer l'efficacité de la surveillance, car les dangers seraient automatiquement signalés, mais également de préserver l'intimité du sujet dans sa vie courante.

Ces différents éléments indiquent pourquoi la reconnaissance d'actions humaines dans des vidéos est l'une des thématiques de recherche les plus actives depuis ces dix dernières années dans le domaine de la vision par ordinateur. Le support vidéo est devenu aujourd'hui un des vecteurs d'information les plus utilisés. La nécessité d'organiser, traiter et exploiter, de façon efficace et automatique, le flux de données associé à ces vidéos est un des enjeux majeurs des acteurs de l'informatique de demain.

Cependant, l'interprétation des informations extraites de vidéos n'est pas chose simple et se heurte à de nombreux problèmes. Ces derniers font de ce sujet l'un des plus complexes en vision par ordinateur.

1.2 Enjeux et problématiques

Nous abordons ici un ensemble de problématiques inhérentes à la reconnaissance automatique d'actions humaines dans des vidéos. Nous rassemblons ces problématiques en quatre grandes thématiques. La *variabilité de la représentation* des actions ; La *complexité sémantique* quant à la description des actions ; l'ambiguité amenée par le *contexte visuel* et la difficulté de définir un *temps d'observation* adéquat pour distinguer un mouvement, une action ou une activité, qui sont respectivement effectués sur une durée de temps courte, moyenne et longue.

Variabilité de la représentation Un des problèmes majeurs auquel est confrontée la reconnaissance d'actions humaines est la grande variabilité de l'information visuelle. Une donnée, telle que l'apparence d'un objet dans une image ou le déplacement d'un sujet dans une vidéo, est dépendante de différents paramètres : le changement d'intensité lumineuse, la différence de point de vue, la déformation des objets non rigides, les occultations partielles de l'objet ou du sujet, les changements de fond, etc. La figure 1.1 montre, pour l'action `Course`, un exemple de variation de point de vue, pouvant compliquer la caractérisation des mouvements associés à cette action de façon unique. Tous ces paramètres, liés à l'acquisition et dûs au milieu extérieur, modifient la représentation visuelle d'un sujet. L'information liée à une action humaine est donc extrêmement fluctuante, d'autant que dans une séquence vidéo, ces contraintes visuelles peuvent survenir simultanément et également varier de façon imprévisible au cours du temps (rotation de caméra autour d'un sujet, extinction de lumière, etc.). On est donc amené à se demander comment rendre l'information liée à une action humaine la plus invariante possible.

FIGURE 1.1 – L'action `Course` sous différentes formes de représentation (randonnée, vidéo-surveillance, compétition sportive). Le contexte, l'éclairage et le point de vue sont différents pour chacun de ces exemples.

Complexité sémantique Au delà des difficultés propres aux informations extraites dans une vidéo, se pose également la problématique de l'interprétation de ses données. En effet, la sémantique est importante à la fois pour interpréter l'information mais également pour la catégoriser. Par exemple, dans une base de données d'images, en fonction de la sémantique employée, une voiture, un tracteur et un camion peuvent appartenir, ou non, à la même classe d'objets. De même, les actions `marcher` et `marcher avec un chien` sont sémantiquement distinctes mais illustrent la même notion à différents niveaux de description. Le vocabulaire employé pour étiqueter une vidéo est d'une grande variabilité et dépend de la "granularité" sémantique

de ce que l'on souhaite identifier. Ainsi, des actions telles que `ouvrir une porte` et `ouvrir un sac` peuvent être considérées comme deux classes différentes mais peuvent également illustrer le même concept dans le cas où l'objet de l'action effectuée importe peu. La figure 1.2 illustre cette ambiguité pour la classe d'action `Ouvrir`.

FIGURE 1.2 – Chaque exemple montre la difficulté d'illustrer le concept de l'action `Ouvrir`. Le manque de précision du vocabulaire utilisé peut être source d'ambiguïté.

Contexte visuel Le contexte visuel joue un rôle non-négligeable dans la reconnaissance des actions humaines. Il est intimement lié à la sémantique car il fournit des informations permettant une interprétation à un plus haut niveau de compréhension et qu'il influe également sur la façon dont une action est décrite. Par exemple, `Sauter d'un plongeoir` et `Sauter d'un building`, représentent la même action élémentaire `Sauter`. Cependant, le contexte visuel de ces deux actions implique un niveau de danger bien différent. L'information visuelle peut également être, dans certain cas, plus pertinente pour décrire une action que les mouvements effectués dans une vidéo. Ainsi, comme le montre la figure 1.3, les actions `Jouer du piano`, `Jouer de la guitare` ou `Jouer de la batterie` peuvent être discriminée uniquement par l'information visuelle relative aux instruments de musique et pas seulement par les gestes du sujet.

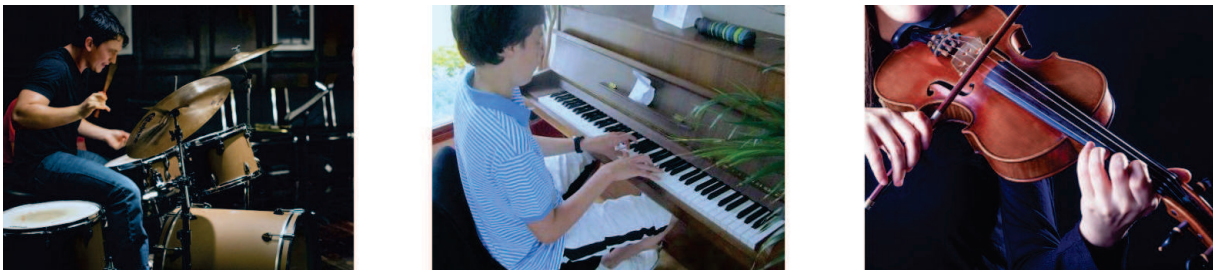

FIGURE 1.3 – On constate que pour chacun de ces exemples, l'information visuelle associée aux instruments de musique peut aisément être l'élément discriminant par rapport à l'action exécutée.

Temps d'observation Détecter et reconnaître une action humaine dans une vidéo dépend également de la fenêtre d'observation considérée pour définir une action. En effet, en fonction des actions exécutées, les limites temporelles de ces dernières ne sont pas toujours clairement définies. Le vocabulaire employé pour décrire une action participe également à cette variation de fenêtre temporelle. En fonction de ce temps d'observation, quand peut-on parler de mouvement,

d'action, ou d'activité ? Les actions `mouvement de la main`, `Serrer la main` et `préparer une salade`, comme illustrées sur la figure 1.4, ne nécessitent pas le même temps d'observation pour être détectées. Du point de vue du vocabulaire, la première peut être considérée comme un mouvement, la deuxième une action, et la troisième comme une activité. En fonction de l'information que l'on cherche à traiter, le temps d'observation peut drastiquement influencer la reconnaissance.



FIGURE 1.4 – Les actions `Mouvement de la main`, `Serrer la main` et `Préparer une salade` nécessitent toutes les trois des temps d'observation différents pour être reconnues.

Discussion Les problématiques liées à la reconnaissance d'actions humaines sont nombreuses et couvrent différents aspects : 1) la difficulté d'extraire une information stable liée aux actions ; 2) les ambiguïtés de langage liées aux descriptions de vidéos, notamment en termes de granularité sémantique ; 3) le contexte visuel, qui rentre en compte dans la caractérisation d'une action et dans la discrimination de ces dernières, 4) la taille de la fenêtre d'observation nécéssaire pour distinguer les actions étudiées. Ces difficultés imposent donc le développement de méthodes permettant une invariance à différents paramètres ainsi que la prise en compte d'informations pertinentes quant à la caractérisation des actions humaines exécutées dans les vidéos. Dans la section suivante, nous présentons l'ensemble des contributions, exposées dans ce manuscrit, et présentons notre approche pour la reconnaissance d'actions humaines répondant aux différentes problématiques citées plus haut.

1.3 Contributions

Dans cette thèse, nous nous intéressons à l'étude, la caractérisation et la reconnaissance des actions humaines contenues dans des vidéos. Les principales contributions de ce manuscrit sont les suivantes :

— une méthode de reconnaissance d'actions élémentaires basée sur l'estimation du mouvement dans les vidéos. Le champ vectoriel associé est utilisé afin d'extraire des points critiques, caractérisant de façon pertinente les mouvements présents dans une séquence vidéo,

— une méthode d'extraction multi-échelle des points critiques et de leur trajectoires, apportant une description robuste et précise, à différentes échelles de mouvements, des actions exécutées dans les vidéos. Cette extraction à différentes échelles précède une description fréquentielle des trajectoires, ainsi que la caractérisation locale des points critiques en termes d'orientation de mouvement et de variation locale du gradient,

— une méthode de compensation des mouvements de caméra, lors des acquisitions de scènes non-statiques, améliorant la reconnaissance d'actions dans des contextes réalistes et génériques,

— l'incorporation d'une méthode de boosting qui combine linéairement, dans le processus d'apprentissage, les différentes caractéristiques extraites (fréquence, orientation de mouvement, variation du gradient) en fonction de leur pertinence,

— une évaluation originale de la généricité de notre approche de reconnaissance d'actions élémentaires basées sur un apprentissage à partir d'un mélange de base de données. Cette évaluation met en évidence la capacité de notre approche à généraliser la représentation des actions, indépendamment des données utilisées pour l'apprentissage,

— une approche de caractérisation et de reconnaissance des activités humaines basée sur la décomposition des activités en séquences temporelles de probabilités d'actions élémentaires, ces activités sont projetées sur une variété statistique, en tant que trajectoires, afin d'être caractérisées et discriminées en respectant la géométrie de cet espace, et ainsi, passer d'un formalisme probabiliste à un formalisme déterministe,

— la construction de descripteurs de trajectoires d'activités, basée sur les coefficients de la transformée de Fourier de ces trajectoires, décrivant la fréquence de variation au cours du temps entre les actions élémentaires tout en respectant un ensemble de propriétés liées au fait que ces trajectoires évoluent sur une variété statistique.

Structure du manuscrit

Après avoir abordé dans le chapitre 1 le contexte dans lequel s'inscrit la reconnaissance d'actions humaines ainsi que les problématiques inhérentes à ce sujet, nous présentons, dans le chapitre 2, un ensemble de bases de données illustrant les difficultés liées à cette thématique. Elles fournissent aux chercheurs un moyen d'évaluer leurs méthodes sur des exemples concrets. Un ensemble non exhaustif des méthodes de la littérature est donc présenté afin d'obtenir une vision large des types d'approches développées ainsi que des avancées réalisées dans le domaine.

Le chapitre 3 introduit notre méthode de reconnaissance d'actions élémentaires. Nous revenons sur les différents éléments d'intérêts développés lors de nos travaux ainsi que le processus complet de cette méthode.

Les performances et la robustesse de cette méthode sont illustrés dans le chapitre 4 où nous évaluons cette dernière sur un ensemble de bases de données. L'information caractérisée par cette méthode permet d'obtenir des taux de reconnaissance parmi les meilleurs de la littérature sur ces bases de données de référence. Nous introduisons également une approche originale d'évaluation de notre méthode basée sur sa faculté à généraliser la représentation des actions élémentaires.

Cette méthode est réutilisée dans le chapitre 5 où nous nous intéressons à la caractérisation et la reconnaissance d'activités humaines. L'activité humaine correspond, dans notre modèle, à un niveau sémantique supérieur par rapport aux actions élémentaires. Nous revenons sur un ensemble de méthodes de la littérature développées pour ce sujet ainsi que sur l'approche que nous proposons. Cette dernière est basée sur notre approche de reconnaissance d'actions élémentaires et sa capacité à représenter de façon robuste et générique les actions élémentaires qui composent une activité humaine. Elle caractérise les activités comme des trajectoires dans un espace géométrique. Les résultats obtenus sont présentés dans le chapitre 6.

L'ensemble des éléments développés au cours de cette thèse sont résumés dans le chapitre 7. Nous revenons également sur les perspectives d'applications dans lesquelles s'inscrivent ces travaux, ainsi que sur les perspectives d'évolution de cette thématique dans les années à venir.

Chapitre 2

Étude et reconnaissance du mouvement humain

Chapitre 2

Étude et reconnaissance du mouvement humain

Sommaire

Introduction du chapitre Dans la première partie de ce chapitre, nous présentons un ensemble de bases de données de la littérature pour la reconnaissance d'actions humaines. Nous verrons comment ces différentes bases illustrent les problématiques liées à l'étude du mouvement humain, soulignées dans le chapitre précédent. Ces bases de données sont présentées chronologiquement de façon à entrevoir l'évolution de ces problématiques ainsi que l'évolution des approches de reconnaissance d'actions. Dans la deuxième partie de ce chapitre, nous présentons différentes méthodes proposées dans la littérature pour analyser et caractériser les mouvements humains dans des vidéos. Ces méthodes sont réparties selon plusieurs catégories qui seront décrites par la suite.

2.1 Bases de données d'actions humaines

Le but des bases de données d'actions humaines est de permettre aux chercheurs du domaine d'évaluer la robustesse et l'efficacité des méthodes développées. Les bases de données de la littérature se sont regroupées en différentes catégories au fil du temps. On peut distinguer les bases *contrôlées* possédant des vidéos avec de fortes contraintes d'acquisitions (caméra statique, fond homogène, etc.) ainsi que des bases de données *réalistes* constituées de vidéos génériques, avec de faibles contraintes d'acquisitions (changement de point de vue, occultations partielles, de couleurs, caméra non fixe, etc.). Ces vidéos illustrent les actions humaines dans des contextes du quotidien. Nous présentons ici un ensemble de bases très utilisées par la littérature dans l'ordre chronologique d'apparition.

2.1.1 Bases d'actions élémentaires contrôlés

Base de données KTH

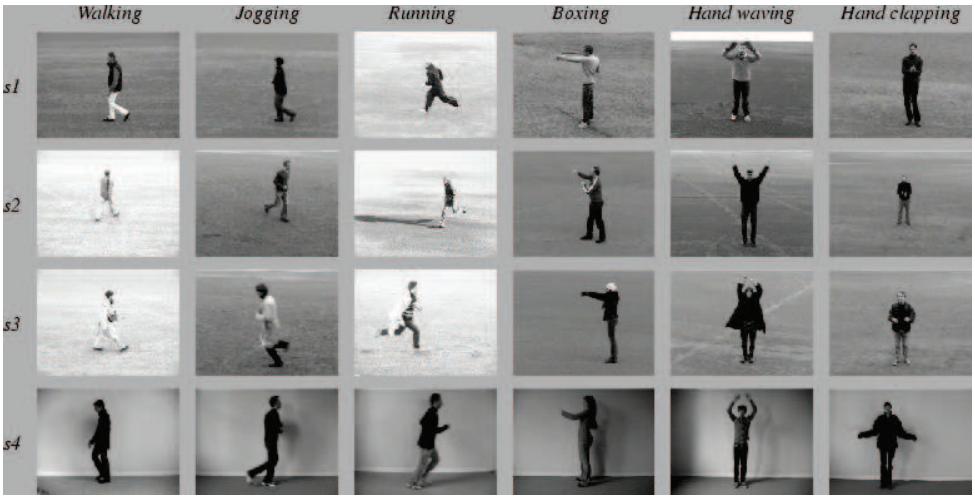

FIGURE 2.1 – Illustration de la base KTH. Les actions illustrées sont : *walking, jogging, running, boxing, hand waving, hand clapping.* (voir vidéo)

L'institut *KTH* (*Royal Institute of Technology* (Suède)) constitue en 2004 la base de données KTH [Schuldt et al., 2004]. Cette base de données est l'une des premières et l'une des plus importantes permettant d'étudier, d'analyser et de reconnaitre des actions humaines. De part les nombreuses méthodes qui lui ont été appliquées, elle est devenue avec le temps une base de référence. Elle fait partie des bases d'actions contrôlées car elle possède de forte contraintes d'acquisitions (caméra fixe, fond homogène, etc.) et chaque vidéo comporte une unique action jouée par un unique sujet. Ces actions sont effectuées de façon répétitives et non naturelles. Elle comporte 600 vidéos réparties en six classes d'actions. La figure 2.1 montre l'ensemble des classes d'actions de cette base de données.

Base de données Weizmann

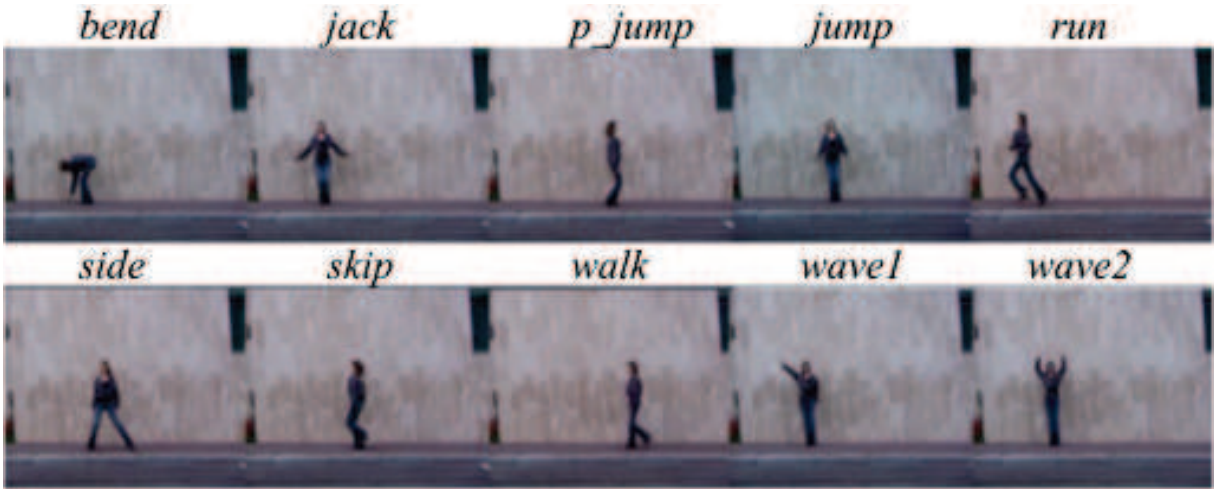

FIGURE 2.2 – Illustration de la base de données *Weizmann*. Les actions illustrées sont : *bend, jack, pjump, jump, run, side, skip, wave one hand, wave two hand* (voir vidéo).

La base de données `Weizmann` [Gorelick et al., 2007] a été constituée en 2005 par la *Weizmann Institute of Science* (Israël) dans l'intention d'étudier de nouveaux algorithmes de reconnaissance d'actions. Contrairement à la base de données `KTH`, celle-ci est moins volumineuse et comprend un total de 90 vidéos. Les contraintes d'acquisition y sont plus fortes (caméra statique, fond uniforme). Cependant la plupart des actions effectuées possèdent de forte similarités visuelles (`Jack, Run, Skip, Side`), ce qui reste un défi d'un point de vue de la reconnaissance. Elle comptabilise un total de 10 actions différentes. La figure 2.2 montre l'ensemble des classes d'actions de cette base de données.

2.1.2 Bases d'actions élémentaires réalistes

Base de données Hollywood Actions

FIGURE 2.3 – Illustration de la base de données `Hollywood Dataset`. On distingue les actions *Kiss, Answer the phone, Get out of the car.* respectivement sur la première deuxième et troisième ligne (voir vidéo).

La base `Hollywood` [Laptev et al., 2008] est réalisée en 2008 par le laboratoire *IRISA/INRIA* (France). Cette base de données est composée de vidéos extraites de films (69 au total) et présente des actions effectuées dans des contextes réalistes ("in the wild"), contrairement aux bases de données contrôlées. Les méthodes de reconnaissance d'actions humaines développées permettent à ce niveau de répondre à des problématiques plus complexes notamment les changements de point de vue, la gestion de la couleur, de forts mouvements de caméra, etc. Comme il a été vu dans le chapitre précédent, la difficulté de l'interprétation visuelle est prise en compte sur cette base de données. Des actions telles que `Handshake, HugAPerson, Kiss` sont parfois très similaires visuellement. Cette base de données comprend 8 actions. Une extension en 2009 (`Hollywood Dataset 2`) est réalisée avec un total de 12 actions [Marszalek et al., 2009]. La figure 2.3 illustre quelques images de cette base, notamment les actions `Kiss` sur la première ligne, `AnswerPhone` sur la deuxième et `GetOutCar` sur la troisième.

Base de données UCF-11

FIGURE 2.4 – Illustration de la base de données `UCF-11` (voir vidéo).

Le *Department of Electrical Engineering and Computer Science at University of Central Florida* (UCF, USA) fournit en 2009 la base de données `UCF-11 Dataset` [Liu et al., 2009] . Cette base est construite à partir de vidéos provenant de Youtube. Dans ce cadre, on est amené à traiter des actions dans des situations réalistes du quotidien, acquises majoritairement à partir d'appareils mobiles de basse résolution. Elle comptabilise un total de 11 actions. La figure 2.4 présente quelques images issues de cette base de données.

Base de données UCF-50

FIGURE 2.5 – Illustration de la base de données `UCF-50 Dataset` (voir vidéo).

La base `UCF-50 Dataset` [Reddy and Shah, 2013] est rendue publique en 2010. Elle est une extension de la précédente base de données `UCF-11 Dataset`. Les principales difficultés rencontrées ici sont la reconnaissance d'un très grand nombre de classes d'actions humaines ainsi que la grande variabilité des informations visuelles liées aux différentes actions. `UCF-50 Dataset` est constitué de 50 catégories d'actions issues de vidéo de Youtube. Cette base de données présente un véritable défi en terme de reconnaissance. En plus du nombre de classe à reconnaitre, le nombre d'instance par classe est également conséquent. Cette base comprend un ensemble de 6680 vidéos. La figure 2.5 montre l'ensemble des classes d'actions de cette base de données.

Base de données HMDB-51

FIGURE 2.6 – Illustration de la base de données `HMDB51 Dataset` (voir vidéo).

Le *Serre lab at Brown University* (USA) constitue en 2011 la base de données `HMDB 51 Dataset` [Kuehne et al., 2011]. Les vidéos de cette base de données sont issues de différentes sources (Youtube, Google vidéo, films, archives, etc). Cette base représente les actions humaines dans des situations très complexes et variées. Ces dernières peuvent être classées en cinq catégories : 1) actions faciales, actions faciales et interactions avec des objets, 2) mouvements du corps, 3) mouvements du corps et interactions avec objet, 4) mouvements du corps, 5) interactions entre humains. Cette base de données est l'une des plus complexes à ce jour et concentre l'ensemble des problématiques rencontrées dans la reconnaissance d'actions humaines. `HMDB 51 Dataset` contient un ensemble de 6849 vidéos pour 51 classes d'actions. La figure 2.6 montre l'ensemble des classes d'actions de cette base de données.

Tableau récapitulatif

Le tableau 2.1 dresse un récapitulatif des bases de données citées plus haut en présentant leur type, les sources, le nombre de vidéos ainsi que le nombre de classes d'actions et l'année de publication.

Nom	Type	Source	Nbre de classes	Nbre de vidéos	Année
KTH	contrôlée	vidéos filmées	6	600	2004
Weizmann	contrôlée	vidéos filmées	10	90	2005
Hollywood 2 Actions	Réaliste	extraits de films	12	1694	2009
UCF-11	Réaliste	vidéos issues du web	11	1100	2009
UCF-50	Réaliste	vidéos issues du web	50	6680	2010
HMDB-51	Réaliste	web, TV, films	51	6849	2011

TABLE 2.1 – Récapitulatif des bases de données de reconnaissance d'actions humaines de la littérature présentées dans ce chapitre.

Discussion Les bases de données avec contraintes d'acquisitions permettent de confronter les méthodes de la littérature aux problématiques les moins complexes de la reconnaissance d'action. Elles représentent une première étape d'évaluation et de comparaison. En effet, ces bases, de part leur date de publication, font partie des plus utilisées et plus citées dans la littérature. Les bases de données génériques offrent quand à elles, une plus grande complexité visuelle et permettent d'évaluer la robustesse des méthodes à différentes conditions (caméra non-statique, changement de point de vue, changement de contexte, nombre d'exemple, nombre de classe, etc.). Les bases de données d'actions humaines sont de plus en plus nombreuses dans l'état de l'art. À la fois pour garantir une illustration des problématiques variées de la reconnaissance humaines ([Yuan et al., 2011] mais également illustrer des situations beaucoup plus spécifiques [mex, 2015]. Après avoir vu comment les différentes bases de données ont permis l'illustration des défis posés par la caractérisation de mouvements humains dans des vidéos, nous verrons par la suite différentes méthodes de la littérature qui tentent de répondre à ces problématiques.

2.2 Reconnaissance d'actions : un état de l'art

Le chapitre précédent a mis en évidence les défis et problématiques qu'engendrent la reconnaissance d'actions humaines dans des vidéos. Dans ce chapitre nous revenons sur les principales méthodes de la littérature. De part leur nature, ces méthodes permettent d'étudier des actions relativement simples (exemple : *Marcher*, *Sauter*, *S'asseoir*, *Tirer*, *etc.*). Ces actions sont dites *élémentaires*. Ces méthodes offrent ainsi un premier niveau de représentation sémantique des mouvements humains dans des vidéos. En effet, les méthodes présentées ci-dessous considèrent une séquence d'images comme une classe particulière d'actions élémentaires humaines.
Les méthodes de la littérature sont très diverses et abordent la question de la reconnaissance sous différents angles. Nous faisons le choix de présenter ces méthodes en les classant en deux grandes catégories : les approches globales et les approches locales. La figure 2.7 fournit une représentation hiérarchique de ces différentes approches :

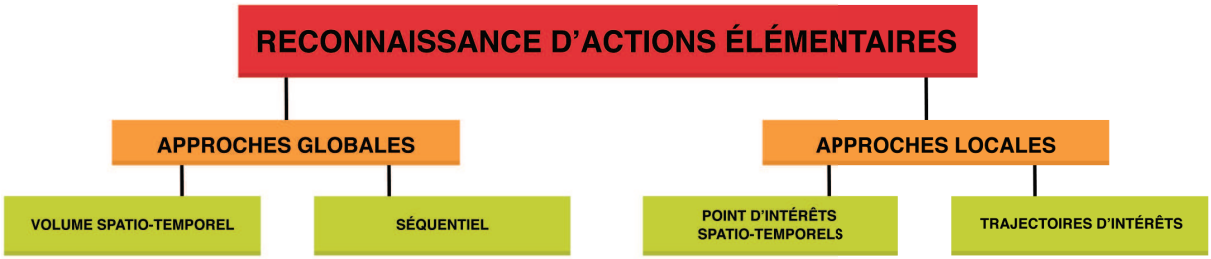

FIGURE 2.7 – Récapitutalif des méthodes de la littérature pour la reconnaissance d'actions humaines élémentaires.

— **Les approches globales** analysent les mouvements humains à partir d'un ensemble structuré spatialement et temporellement.

— **Les approches locales** représentent les actions humaines comme une collection d'éléments significatifs liés à ces actions.

Les approches globales ont été les premières à être développées pour la reconnaissance d'actions humaines dans des vidéos. Ci-après, nous dressons une liste non exhaustive des principales méthodes globales de la littérature.

2.2.1 Approches globales

La particularité des approches globales est qu'elles tentent de caractériser le lien temporel et spatial entre les différents éléments qui composent une action. L'étude de ces actions est basée sur une analyse précise des mouvements.
Ces dernières peuvent être distinguées en deux groupes. Les approches utilisant des volumes spatio-temporels, qui analysent en espace et en temps les caractéristiques liées à une séquence vidéo, et les approches séquentielles qui analysent la structure de l'exécution des mouvements composant une action à partir de similarité d'exemples déjà appris ou encore à l'aide de méthodes probabilistes.

2.2.1.1 Méthodes basées sur volume spatio-temporel

Les approches utilisant des volumes spatio-temporels considèrent une vidéo comme un volume 3D. Ces volumes correspondent communément à un empilement de silhouettes représentant l'évolution des mouvements du sujet au cours du temps. L'obtention de ces silhouettes est généralement effectuée à l'aide de pré-traitements tels que la suppression de fond. Le sujet est donc caractérisé par un ensemble structuré dans l'espace et dans le temps. La figure 2.8 montre un exemple de volume spatio-temporel à partir d'empilement de silhouettes des actions `lever les mains`, `marcher` et `courir` issues de la base de données `Weizmann`. La reconnaissance se fait en induisant une mesure de similarité entre les volumes obtenus.

FIGURE 2.8 – Exemple de volumes spatio-temporels obtenus sur des silhouettes de sujets de la base de données `Weizmann` (figure tirée de [Poppe, 2010]).

— [Bobick and Davis, 2001] ont proposé l'une des premières méthodes d'utilisation de volumes spatio-temporels construits à partir de silhouettes estimées au cours du temps après une étape de soustraction de fond. La méthode suppose que les mouvements dans une vidéo sont uniquement produits par le sujet étudié. Les silhouettes sont décrites à partir de deux composantes : la composante qui correspond au cumul des pixels associés aux mouvements, *Binary Motion Energy Image* (`MEI`) et la composante qui décrit les zones de forts mouvements de pixels, *Motion History Image* (`MHI`). Ces deux composantes sont ensuite concaténées et décrites statistiquement à l'aide de descripteurs basés sur les moments 2D de figures géométriques [Hu, 1962]. La méthode permet de reconnaître des actions telles que `s'asseoir`, `lever la main`, `s'accroupir`. La figure 2.9 montre des exemples de mouvements issus d'exercices d'aérobics caractérisés par la composante `MEI`.

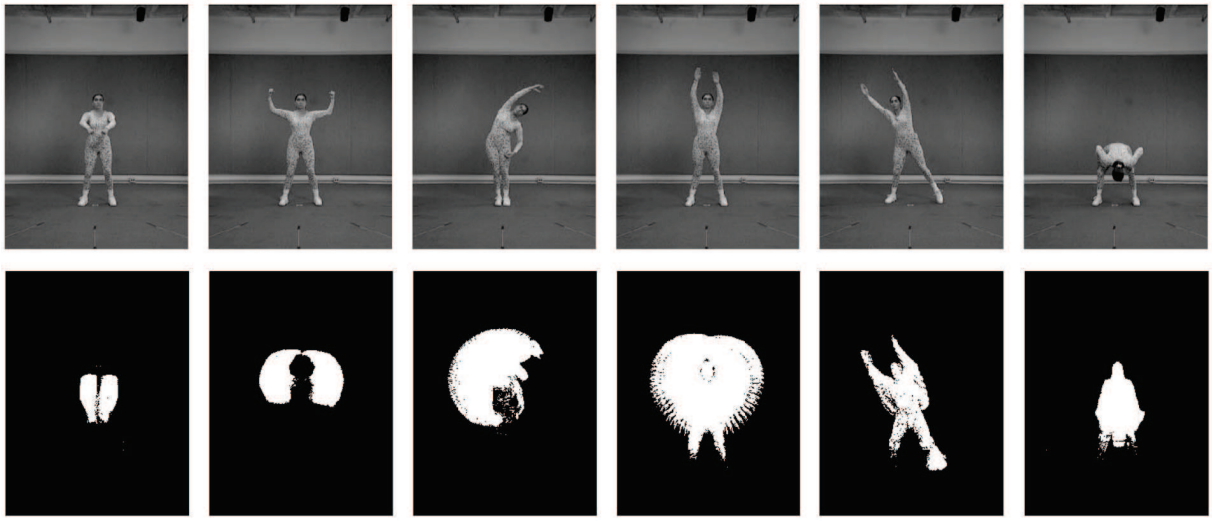

FIGURE 2.9 – Caractérisation de mouvement d'aérobics à partir du descripteur `MEI` de la méthode de F.Bobick *et al.*. (figure tirée de [Bobick and Davis, 2001]).

— [Gorelick et al., 2007] appliquent une suppression de fond sur des séquences vidéos et forment un volume spatio-temporel noté S en empilant les silhouettes obtenues au cours du temps. La solution de l'équation de Poisson de la forme $\Delta U(x, y, t) = -1$ avec $(x, y, t) \in S$ est ensuite utilisée pour obtenir plusieurs caractéristiques pertinentes au sens du mouvement. Les extrema de U permettent d'identifier le torse du sujet. À l'aide d'un seuillage adéquat, les éléments extérieurs au torse, possédant un mouvement rapide, sont détectés comme étant des régions saillantes. L'orientation de mouvements de ces régions saillantes sont également extraites à l'aide des magnitudes des valeurs propres λ_1, λ_2 et λ_3 de la Hessienne de la solution de Poisson.
$\lambda_1 \approx \lambda_2 \gg \lambda_3$ correspond aux zones quasi-fixes ("stickness"), $\lambda_1 \gg \lambda_2 \approx \lambda_3$ correspond aux zones de mouvements rapides ("plateness") et $\lambda_1 \approx \lambda_2 \approx \lambda_3$ correspond au zone n'ayant pas de direction particulière ("ballness"). La figure 2.10 illustre ces différentes caractéristiques.

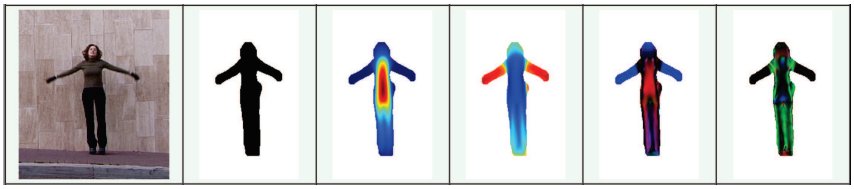

FIGURE 2.10 – De gauche à droite : image, soustraction de fond, solution équation de Poisson, région saillante, mesure région de forts mouvements ("Plateness"), mesure de région immobiles ("Stickness") (figure tirée de [Bobick and Davis, 2001]).

— [Weinland et al., 2006] étudient les actions humaines à partir de différentes caméras réparties dans une scène. Après suppression du fond, le volume obtenu sur chaque caméra est projeté dans un espace tri-dimensionnel pour former ce que les auteurs nomment le *Motion History Volume* qui est une extension 3D du *Motion History Image* (`MHI`) de Bobick *et al.* [Bobick and Davis, 2001]. Les appariements sont effectués en appliquant une transformée de Fourier sur les coordonnées cylindriques du volume comme le montre la figure **??**. Cette approche permet d'obtenir une invariance au changement de point de vue mais nécessite en contre partie les paramètres intrinsèques et extrinsèques des caméras filmant la scène.

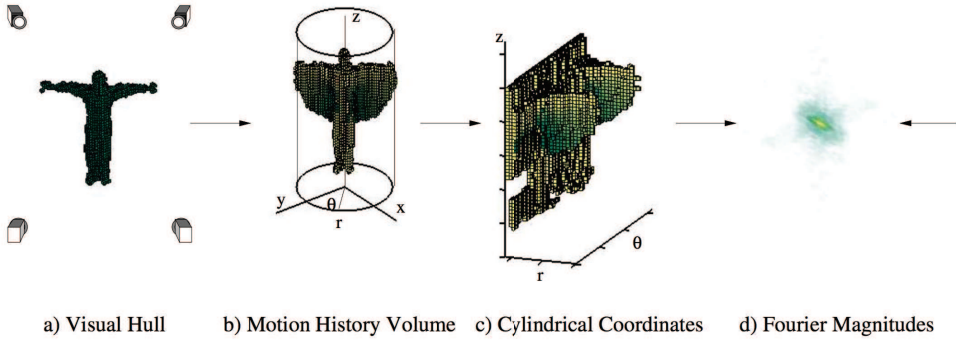

FIGURE 2.11 – a) capture à partir de plusieurs caméras, b) construction du volume 3D c) transformation en coordonnées cylindriques d) énergie de la transformée de Fourier du volume cylindrique. (figure tirée de [Weinland et al., 2006]).

— [Tabbone et al., 2006] appliquent une "$\mathcal{R}$-transformation", basée sur la transformée de Radon pour caractériser la forme des silhouettes humaines au cours du temps. La $\mathcal{R}$-transfomation assure l'invariance en échelle, en translation et en rotation comme le montre la figure 2.12. Un modèle de Markov caché est ensuite utilisé pour apprendre la variation de chaque silhouette en fonction des actions qu'elles représentent. Les actions élémentaires reconnues par cette méthode sont `se baisser, sauter, marcher, porter, se ruer.`

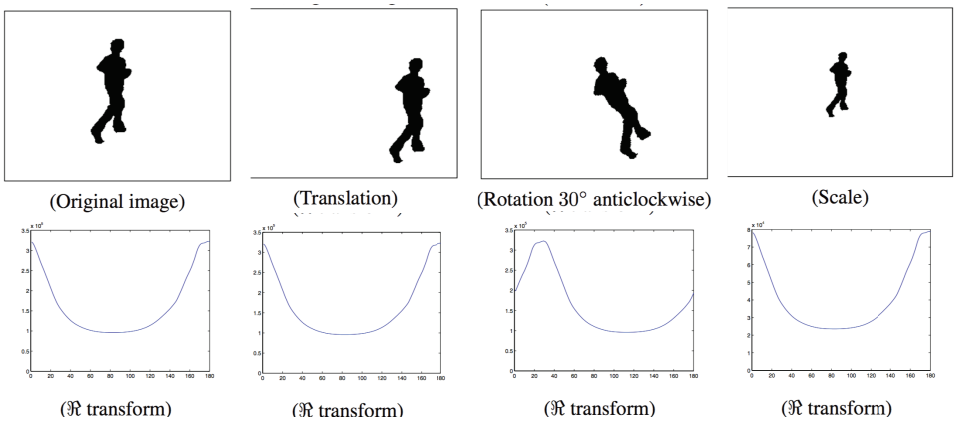

FIGURE 2.12 – Exemple de l'invariance en translation, rotation et échelle obtenue avec la $\mathcal{R}$-transformation. On constate que la réponse obtenue pour chaque transformation est la même à une translation près (figure tirée de [Tabbone et al., 2006]).

2.2.1.2 Méthodes séquentielles

Les méthodes séquentielles reconnaissent les actions élémentaires en analysant une séquence ordonnée d'éléments descriptifs. Ces éléments descriptifs sont le plus souvent issus d'extraction de silhouettes ou d'appariements de squelettes en fonction des parties du corps humain en mouvement dans la séquence vidéo. Les éléments descriptifs sont extraits au cours du temps afin d'estimer la pose du sujet à chaque image de la séquence. Ils sont considérés comme des observations traduisant l'évolution du statut du sujet. Les méthodes séquentielles considèrent une vidéo comme une séquence d'observations et en déduisent la présence d'une action élémentaire, préalablement apprise. La vraisemblance entre une séquence d'observations d'une vidéo requête avec les séquences d'observations, issues d'un ensemble d'entraînement, est estimée de façon à déterminer quelle action est exécutée dans la séquence.
On peut distinguer deux types de méthodes séquentielles.

— **Les méthodes basées exemples** qui reconnaissent les actions élémentaires comme une suite structurée d'exemples représentant une action élémentaire.

— **Les méthodes basées sur les états probabilistes** qui représentent les actions à partir de probabilités d'observation d'éléments caractéristiques au cours du temps. Ces probabilités sont produites à l'aide de modèles génératifs.

Ces deux types de méthodes sont explicitées ci-après avec quelques exemples de la littérature les illustrant.

Méthodes basées exemples Les méthodes basées exemples tentent de reconnaître des actions à partir de caractéristiques préalablement apprises, modélisant ces actions au cours du temps. Ces caractéristiques, ou exemples, modélisant ces actions peuvent être issus d'appariements de squelettes au cours du temps, de pose, de silhouette, etc. La structure spatio-temporelle de ces exemples caractérise les actions étudiées. Ces dernières sont apprises directement à partir d'éléments d'un ensemble d'entrainement. Les actions sont associées à un unique représentant par classe ou à un ensemble d'exemples préalablement appris. La similarité entre les exemples extraits d'une séquence vidéo requête et le modèle préalablement appris permet d'établir la reconnaissance des actions élémentaires effectuées dans la vidéo requête. L'algorithme de recalage temporel *Dynamic Time Warping* (`DTW`) [Berndt and Clifford, 1994] est usuellement employé dans ces méthodes pour obtenir une invariance en terme de vitesse d'évolution des exemples. La figure 2.13 montre de façon intuitive le recalage entre deux séquences avec une variation de vitesse d'exécution de mouvement.

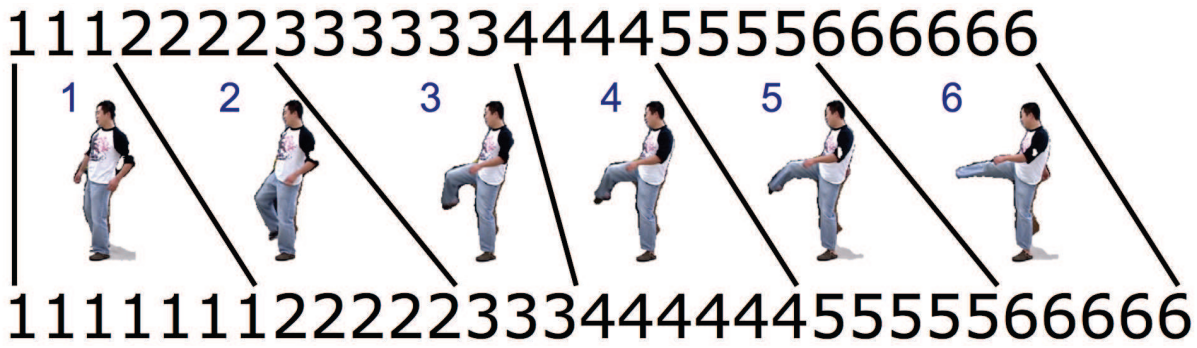

FIGURE 2.13 – Exemple de recalage entre deux séquences représentant l'action `Étirer la jambe`. Ces deux séquences sont exécutées avec une variation de vitesse non-linéaire entre elles (figure tirée de [Aggarwal and Ryoo, 2011]).

— [Gavrila and Davis, 1995] ont développé un modèle spatio-temporel de suivi de parties du corps du sujet. Le but est de recaler un squelette 3D sur le sujet à chaque image de la séquence. Plusieurs caméras sont utilisées pour obtenir une version 3D du squelette, composé de segments et de leurs jointures. Les écarts angulaires des jointures sont enregistrés au cours du temps et sont utilisés comme caractéristiques descriptives d'une action humaine. L'algorithme *Dynamic Time Warping* est utilisé pour comparer des séquences d'écart d'angles entre deux actions.

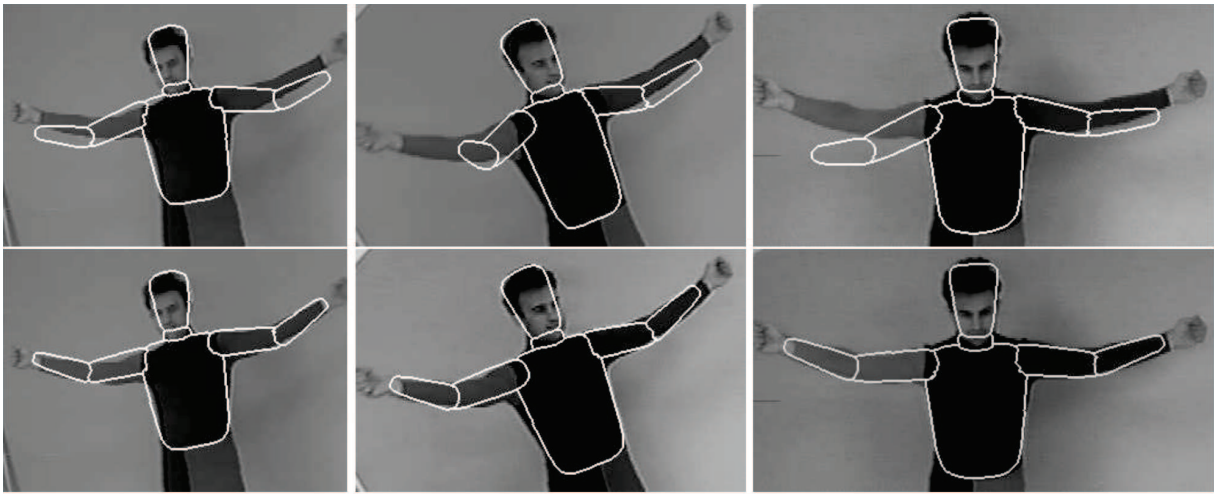

FIGURE 2.14 – Exemple d'appariement de squelette 3D. La tête, le torse, les bras et avant bras sont les éléments dont la position est estimée au cours du temps. Ces déplacements constituent par la suite la caractéristique descriptive de l'action étudiée.(figure tirée de [Gavrila and Davis, 1995]).

— [Efros et al., 2003] caractérisent le mouvement d'un sujet centré dans une séquence vidéo à l'aide du flot optique. Les composantes verticales et horizontales du champ vectoriel $\mathbf{F}$ associé sont séparées en parties positives $\mathbf{F}^+$ et négatives $\mathbf{F}^-$ conduisant à quatre composantes descriptives qui sont par la suite lissées. Ces composantes décrivent chaque image d'une vidéo séquence au cours du temps. La comparaison entre deux séquences vidéos se fait en calculant la similarité entre toutes les combinaisons possibles de paires d'images. On obtient une matrice de similarité décrivant la correlation entre deux séquences vidéos. La reconnaissance est réalisée en détectant des motifs particuliers sur la diagonale de la matrice de similarité. Des actions élémentaires issues de sports tels que le `Football` ainsi que le `Tennis` sont reconnues avec cette méthode.

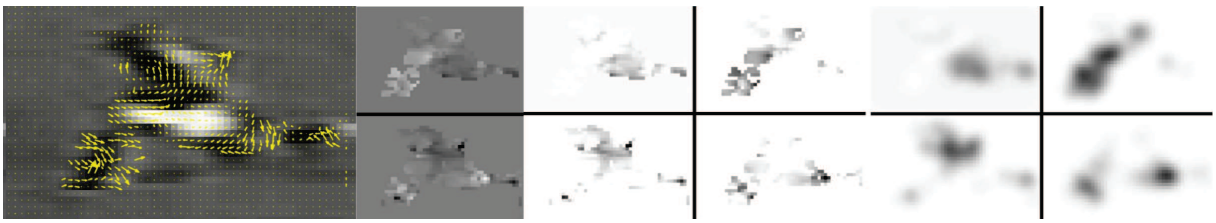

FIGURE 2.15 – De gauche à droite : Flot optique $\mathbf{F}_{x,y}$, composantes $\mathbf{F}_x$, $\mathbf{F}_y$, séparation $\mathbf{F}_x^+$, $\mathbf{F}_x^-$, $\mathbf{F}_y^+$, $\mathbf{F}_y^-$, composantes lissées par filtre Gaussien $\mathbf{Fb}_x^+$, $\mathbf{Fb}_x^-$, $\mathbf{Fb}_y^+$, $\mathbf{Fb}_y^-$ (figure tirée de [Efros et al., 2003]).

Méthodes basées sur les états probabilistes Les méthodes basées états probabilistes font partie de l'autre catégorie des méthodes séquentielles. Elles représentent les actions en construisant un modèle entrainé pour générer des séquences d'éléments descriptifs correspondant à ces actions. Les méthodes basées états probabilistes sont des approches séquentielles qui représentent une action comme un ensemble d'états au sens probabiliste. Un modèle statistique est entrainé de façon à générer une séquence d'états correspondant à l'action apprise. Les modèles de Markov caché `HMM` (*"Hidden Markov Model"*) ainsi que les réseaux Bayésien dynamiques `DBN`, initialement employés pour la reconnaissance vocale, sont largement utilisés dans ce cadre. Ils permettent de caractériser les états successifs d'un sujet dans une vidéo. Une activité est représentée comme un ensemble d'états cachés. Chaque image est caractérisée par un état. À l'image suivante, un système de transition quantifie la probabilité de passage d'un état à un autre. Ces probabilités de transitions et d'observations sont apprises durant l'entrainement du modèle. La figure 2.16 illustre intuitivement le fonctionnement du modèle `HMM`. Le modèle apprend les probabilités de transitions a_{ij} d'un état w_i à un état w_j, ainsi que les probabilités b_{ik} d'observer la pose numérotée k à l'état w_i .

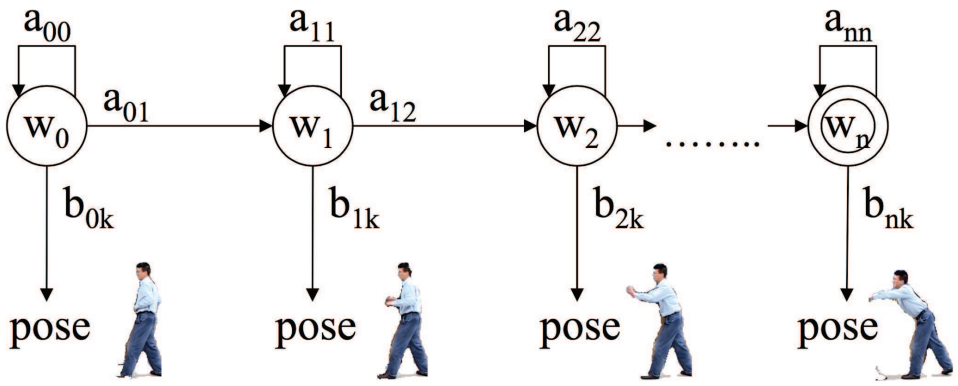

FIGURE 2.16 – Exemple d'un modèle `HMM` pour l'action `étirer le bras`. Chaque image correspond à une pose k dont la probabilité d'apparition b_{ik} est la plus forte suivant l'état w_i considéré (figure tirée de [Aggarwal and Ryoo, 2011]).

— [Yamato et al., 1992] sont parmi les premiers à avoir appliqué un modèle `HMM` pour la reconnaissance d'actions. Une suppression de fond est d'abord effectuée sur les séquences vidéos. Les images de chaque séquence sont ensuite quantifiées spatialement en plusieurs blocs et décrites à l'aide de vecteurs comptabilisant le nombre de pixels associés au sujet contenu dans chacun des blocs. Ces vecteurs sont traités comme des séquences d'observations. Chaque action est apprise en construisant un modèle `HMM` sur les probabilités des observations et de transitions d'état obtenues à partir de séquences vidéos labellisées. La figure 2.17 montre un exemple d'une action de `tennis` générée par cette méthode. Chaque image correspond à un symbole de la séquence. L'action est représentée comme un enchainement ordonné de symboles.

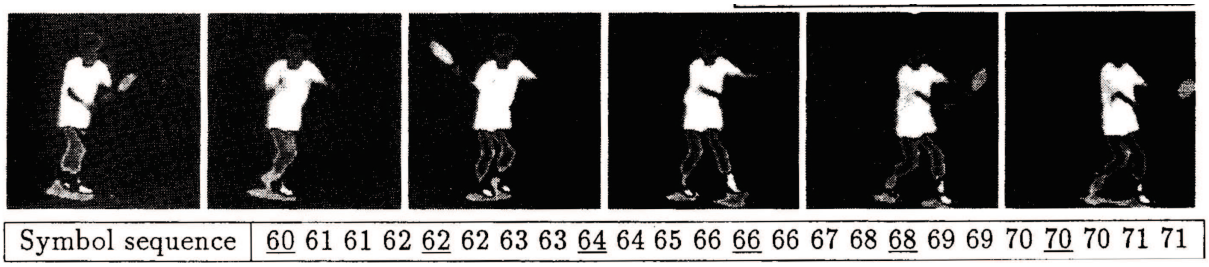

FIGURE 2.17 – Exemple d'une action de `tennis` traitée par la méthode de J.Yamato *et al.*. Les images sont obtenues après une suppression de fond sur la séquence vidéo. L'action est représentée par une séquence de symboles représentant chaque sous-évènement qui compose l'action (figure tirée de [Yamato et al., 1992]).

— [Wilson and Bobick, 2002] appliquent un modèle `HMM` pour la reconnaissance de gestes. Un geste est représenté par une trajectoire 2D décrivant le changement de position de la main. Chaque courbe est décomposée en séquence interprétable par la suite en séquence d'état. Un modèle `HMM` est alors utilisé pour apprendre les probabilités d'observations et les transitions entre les états.

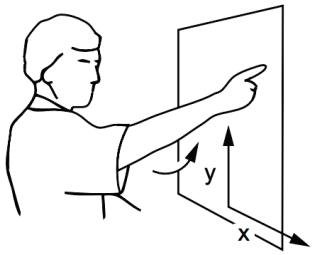

FIGURE 2.18 – La méthode de Wilson *et al.* permet de reconnaitre des gestes dans une vidéo en estimant les coordonnées (x, y) de la main au cours du temps. La méthode est employée dans le cadre d'interactions homme-machine dans des environnements contrôlés. (figure tirée de [Wilson and Bobick, 2002]).

Discussion **Les approches par volume spatio-temporel** caractérisent globalement la totalité des mouvements du sujet dans une vidéo et permettent une reconnaissance très précise des actions élémentaires. Cependant l'inconvénient de ces méthodes est leur difficulté à analyser des actions exécutées dans des scènes où le sujet ne peut être segmenté. En effet, l'ensemble des méthodes de la littérature utilisent des pré-traitements tels que la suppression de fond. Ces pré-traitements rendent ces approches sensibles aux bruits, aux occultations partielles, aux mouvements de caméra et d'autres éléments produisant des mouvements de perturbations. Ces méthodes sont donc applicables avec de fortes contraintes d'acquisitions et donc dans des scènes très spécifiques. L'étude des actions élémentaires à l'aide d'approches par volume spatio-temporel reste donc dans un cadre fortement contraint d'exécution de ces actions.

Les approches séquentielles considèrent les relations temporelles entre les caractéristiques décrivant une action élémentaire. L'aspect séquentiel fournit une description détaillée de l'enchaînement des actions et permet la détection d'actions élémentaires spécifiques. Les approches utilisant des modèles `HMM` apportent une représentation probabiliste des actions. Ces modèles établissent la topologie des actions et facilitent également l'intégration d'informations *a priori*. Néanmoins les méthodes `HMM` nécessitent une grande quantité de données d'apprentissage à mesure que les actions étudiées se complexifient. La complexification des transitions ainsi que les probabilités de génération d'observations constituent la principale cause de ce coût. De plus, les méthodes basées exemples et basées états probabilistes utilisent des éléments descriptifs sensibles aux variations visuelles.

Les limitations des méthodes globales sont clairement définies. L'appariement de squelettes, le suivi de sujet ainsi que la suppression de fond, utilisées dans la plupart des méthodes séquentielles sont des traitements qui maintiennent ces méthodes dans des cadres avec de fortes contraintes d'acquisitions. Les approches locales, présentées dans la partie suivante, tentent d'apporter une solution aux problèmes rencontrées par les méthodes globales en proposant une représentation différente des actions contenues dans des séquences vidéos.

2.2.2 Approches locales

Les approches locales considèrent la représentation des actions humaines d'un point de vue différent de celui des approches globales. Elles supposent que quelques régions spatio-temporelles pertinentes d'une séquence vidéo, caractérisant des mouvements locaux, sont suffisantes pour représenter une action. En effet, les volumes spatio-temporels utilisés dans les approches globales sont des objets 3D rigides. Si une méthode est capable d'extraire des éléments appropriés caractérisant chaque volume 3D d'actions, la reconnaissance d'action peut être interprétée comme un problème d'appariement entre deux objets. Une action n'est plus représentée comme un ensemble structuré en espace et en temps mais comme une collection d'éléments significatifs au sens de l'action. Les méthodes locales s'abstraient donc de contraintes spatio-temporelles dans une séquence vidéo et sont robustes aux occultations partielles, aux mouvements de caméras ainsi qu'à certains changements de points de vue.
La pertinence des approches locales dépend essentiellement des éléments significatifs extraits dans ces séquences. Elles suivent généralement un processus commun :

1. **Détection d'éléments d'intérêt** : Les éléments significatifs des actions effectuées dans une séquence vidéo sont appelés par la suite "éléments d'intérêt". Il existe différentes méthodes pour détecter ces éléments d'intérêt qui varient en fonction de la caractéristique jugée intéressante (coin spatio-temporel, *blob* 3D, etc.) [Laptev, 2005, Dollar et al., 2005, Willems et al., 2008].

2. **Description des éléments détectés** L'utilisation de descripteurs locaux permet la caractérisation des éléments d'intérêt détectés. Ces descripteurs ont pour but de décrire dans le voisinage d'un point d'intérêt une caractéristique particulière (orientation du mouvement, texture, gradient, etc). Ces descripteurs sont construits de façon à être robustes à différentes variations rencontrées lors de la captation (rotation, translation, illumination, etc). Les séquences étant représentées par une collection d'éléments d'intérêt, ces invariances permettent d'assurer l'unicité de la description d'un point d'intérêt entre une séquence originale et une séquence ayant subi une transformation quelconque.
[Laptev et al., 2008, Scovanner et al., 2007].

3. **Encodage des descripteurs** : L'encodage vectoriel des descripteurs permet de simplifier la représentation d'une séquence vidéo. Les descripteurs sont d'abord extraits d'un ensemble d'apprentissage de vidéos, regroupés, puis encodés selon différentes méthodes. On peut citer parmi ces méthodes l'approche du *Bag of Visual Word* [Peng et al., 2014, Lazebnik et al., 2006], *Fisher vector* [Oneata et al., 2013], *Vector of Locally Aggregated Descriptor* [Delhumeau et al., 2013], *Global Alignement Kernel* [Cuturi, 2011], ou encore *BossaNova* [Avila et al., 2013]. Ces approches fournissent une représentation compacte de l'ensemble de départ. Ils sont ensuite utilisés pour caractériser les descripteurs contenus dans une séquence vidéo. Cette étape fournit une représentation commune à chaque vidéo qui sert à mesurer leurs similarité.

4. **Classification** : L'étape de classification permet de distinguer différentes actions humaines contenues dans des vidéos en fonction de la similarité établie entre ces dernières.

Les principales différences entre les méthodes par approches locales sont les types d'éléments d'intérêt détectés, leurs descripteurs, ainsi que la façon dont ils sont combinés dans l'étape de classification. Dans la suite, nous passons en revue les méthodes de détection de points d'intérêt spatio-temporels les plus classiques de la littérature. Nous en profitons également pour rappeler les descripteurs utilisés dans chacune de ces méthodes pour obtenir une vision globale des couples détecteur/descripteur communément utilisés par les approches locales.

2.2.2.1 Points d'intérêt spatio-temporels

Les premières approches d'extraction d'éléments d'intérêt dans des videos sont basées sur des méthodes issues du domaine de la reconnaissance d'image. Ces méthodes utilisent des détecteurs de points d'intérêt 2D dans des images afin de les représenter comme des collections d'éléments d'intérêt. Ces dernières étendent, dans le domaine temporel, des détecteurs à l'origine appliqués spatialement. Un détecteur est associé à une fonction dont l'espace de départ est généralement une séquence d'images en niveaux de gris. Typiquement, les points d'intérêt spatio-temporels correspondent aux maxima locaux de cette fonction.

— [Laptev, 2005] a été parmi les premiers à proposer un détecteur de points d'intérêt spatio-temporels appelé `STIP`. Ce détecteur est une extension temporelle du détecteur 2D de [Harris and Stephens, 1988]. Il détecte les points spatio-temporels dont le voisinage possède une forte variation en espace et en temps, ce qui correspond à des *coins* spatio-temporels. Les éléments en mouvement dont la direction change brutalement sont détectés. Ces points d'intérêt sont décrits à l'aide de dérivées successives de l'image dans le voisinage des points d'intérêt [Koenderink and van Doorn, 1987]. L'utilisation d'un classifieur `SVM` permet de reconnaître différentes actions élémentaires, notamment sur la base de données `KTH`. La figure illustre les points d'intérêt extraits avec la méthode `STIP`.

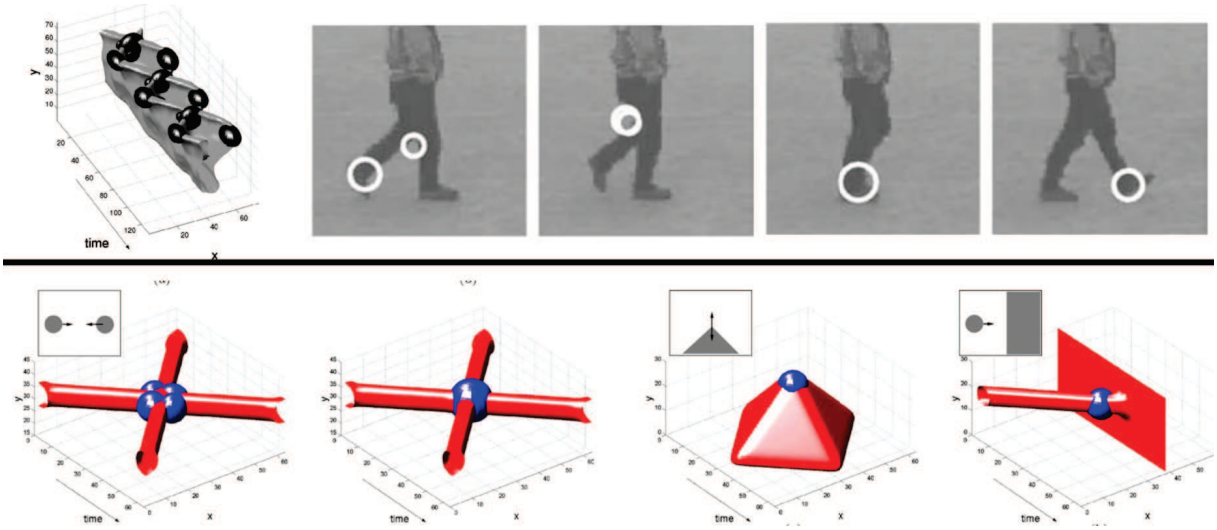

FIGURE 2.19 – Exemple de point d'intérêt spatio-temporel détectés lors du mouvement de marche (1ère ligne) (ex : extrémité du pied, genou, etc.). Les points sont détectés dans les zones de fortes déformations du volume généré par la réponse du détecteur (2ème ligne) (figure tirée de [Laptev, 2005]).

— [Dollar et al., 2005] proposent le détecteur et descripteur `cuboïd` pour la reconnaissance d'actions dans des vidéos. Le détecteur cuboïd consite en un filtrage Gaussien 2D dans le domaine spatial, ainsi que l'utilisation d'une paire d'ondelettes de Gabor 1D dans le domaine temporel. Les plus fortes réponses associées à ce détecteur correspondent aux mouvements périodiques significatifs présents dans la séquence vidéo. Le descripteur cuboïd consiste en une concaténation des informations de gradient et d'orientation de mouvements dans le voisinage des points d'intérêt. Cette méthode se révèle efficace pour la caractérisation de points d'intérêt autres que les coins spatiaux-temporels. La méthode des cuboïds produit de bons résultats sur des bases de données d'expressions faciales [Dollar et al., 2005]. La figure 2.20 présente quelques domaines d'application de la méthode `cuboïd`.

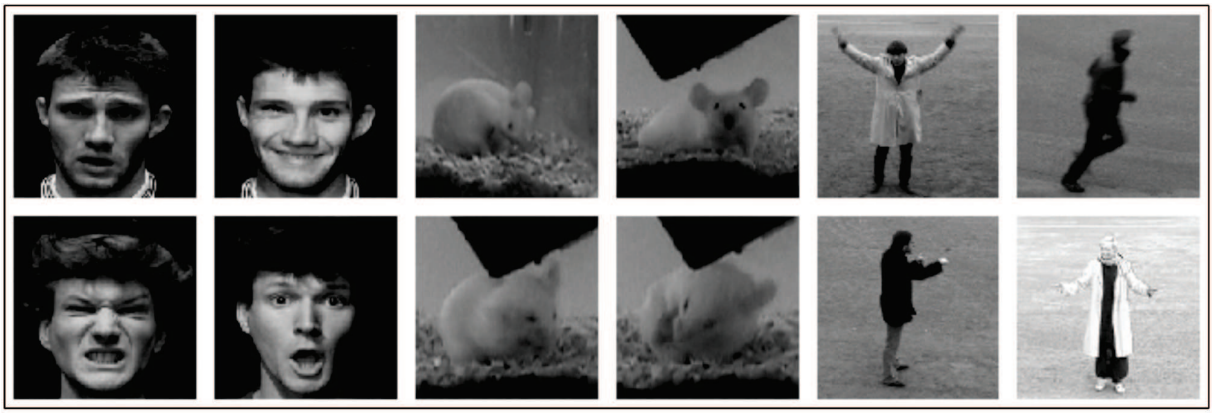

FIGURE 2.20 – Cadre d'application de la méthode des `cuboïd` : expressions faciales, comportement animal, actions humaines (figure tirée de [Dollar et al., 2005]).

— [Willems et al., 2008] étendent le détecteur 2D SURF [Bay et al., 2008] qui détecte des régions spatio-temporelles saillantes en utilisant le déterminant de la matrice Hessienne 3D. Les éléments d'intérêt détectés ont donc la forme de *blobs* spatio-temporels. Son efficacité en termes de temps de calcul est due à l'utilisation de vidéos intégrales [Ke et al., 2005] par analogie aux images intégrales. La figure 2.21 montre des points d'intérêt détectés par SURF sur des vidéos issues de la base de données KTH.

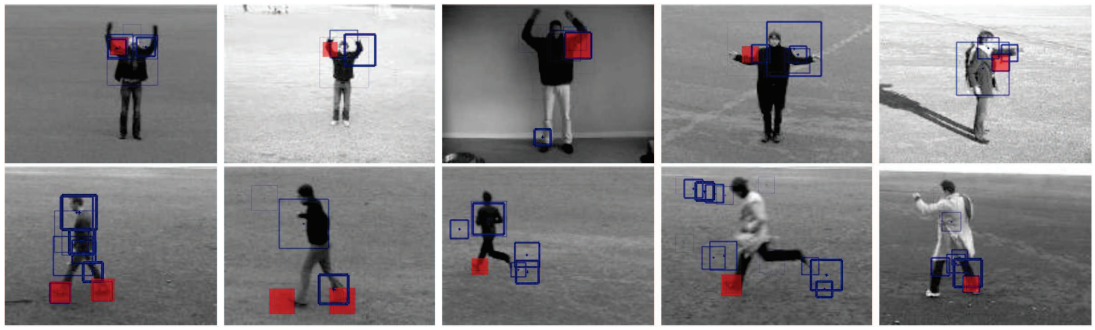

FIGURE 2.21 – Points d'intérêt détectés par la méthode SURF sur la base de données KTH Dataset (figure tirée de [Willems et al., 2008]).

	HOG3D	HOG/HOF	HOG	HOF	Cuboid	Esurf
Harris 3D	89%	**91,8%**	80,9%	**92,1%**	-	-
Cuboids	**90%**	88,7%	82,3%	88,2%	89,1%	-
Esurf	84,6%	88,7%	77,7%	88,6%	-	81,4%
Dense	85,3%	86,1%	79%	88%	-	-

TABLE 2.2 – Taux de reconnaissance obtenus sur la base de données `KTH` avec différentes combinaisons de détecteurs et descripteurs de la littérature (résultats issus de [Wang et al., 2009]).

	HOG3D	HOG/HOF	HOG	HOF	Cuboid	Esurf
Harris 3D	79,7%	78,1%	71,4%	75,4%	-	-
Cuboids	**82,9%**	77,7%	72,7%	76,7%	76,6%	-
Esurf	79%	79,3%	66%	75,3%	-	77,3%
Dense	**85,6%**	81,6%	77,4%	**82,6%**	-	-

TABLE 2.3 – Taux de reconnaissance obtenus sur la base de données `Hollywood` avec différentes combinaisons de détecteurs et descripteurs de la littérature (résultats issus de [Wang et al., 2009]).

Discussion Les méthodes de détections de points d'intérêt dans des vidéos ont permis d'explorer des problématiques plus complexes en termes d'analyse de mouvements, notamment grâce à leur robustesse à différentes variations visuelles. Les vidéos caractérisées par ces méthodes contiennent plus d'informations visuelles que celles étudiées avec les approches globales. Néanmoins, les évaluations menées sur ces différents détecteurs montrent le manque de performance de ces derniers à reconnaitre des actions humaines exécutées dans des environnements plus réalistes [Wang et al., 2009]. Les tableaux 2.2 et 2.3 montrent les taux de reconnaissance obtenus avec les détecteurs et descripteurs de points d'intérêt de la littérature sur la base `KTH` et `Hollywood` (ces résultats sont issus des travaux de [Wang et al., 2009]). Ces séquences vidéos réalistes comportent de nombreuses variations visuelles comme cela a été montré en début de chapitre. Le problème des méthodes d'extraction de points d'intérêt est avant tout la faible quantité de points générés, bien que ces derniers décrivent bien les mouvements exécutés dans les séquences. Les résultats du tableau 2.2 montrent également que dans les cas réalistes, l'extraction de points à l'aide de grilles denses surpasse les méthodes de détection d'éléments d'intérêt en termes de taux de reconnaissance [Wang et al., 2009].

2.2.2.2 Extraction dense de points

— [Laptev et al., 2008] proposent une amélioration de leur première méthode de reconnaissance `STIP`. Une approche multi-échelles d'extraction de points d'intérêt spatio-temporels est privilégiée par rapport à la méthode initiale de sélection d'échelle qui permettait d'extraire des points à des échelles spécifiques. Les points d'intérêt sont ici décrits à l'aide d'informations de gradient (descripteur `HOG`) et d'orientation de mouvement (descripteur `HOF`). Cette extension de leur méthode permet d'extraire un plus grand nombre de caractéristiques et se révèle intéressante dans le cadre de reconnaissance de mouvements présents dans des vidéos réalistes comme celles de la base de données `Hollywood Dataset`.

FIGURE 2.22 – Exemple d'actions contenues dans des extraits de films et reconnues par la méthode de Laptev *et al.* (figure tirée de [Laptev et al., 2008]).

— [Reddy and Shah, 2013] étudient les actions humaines dans des vidéos avec de faibles contraintes d'acquisition, notamment sur la base de données `UCF-50`. Leur méthode permet d'extraire des éléments d'intérêt spatio-temporels avec le détecteur cuboïd mais également de caractériser les éléments contextuels ainsi que les éléments en mouvement en seuillant la norme du flot optique estimé sur les séquences étudiées. Ces éléments, contextuels et de mouvement, fournissent une caractérisation dense des vidéos, et sont décrits à l'aide du descripteur colorimétrique `C-SIFT` [Van de Sande et al., 2010], et du descripteur `3D-SIFT` [Scovanner et al., 2007]. Les éléments contextuels et les éléments en mouvement sont par la suite fusionnés dans l'étape de classification. La figure 2.23 illustre un exemple de seuillage sur la norme du flot optique d'une séquence afin d'en distinguer les éléments contextuels (rouge) et en mouvement (vert). Cette figure montre également le gain en termes de reconnaissance lorsque le nombre de caractéristiques mises en jeu augmente.

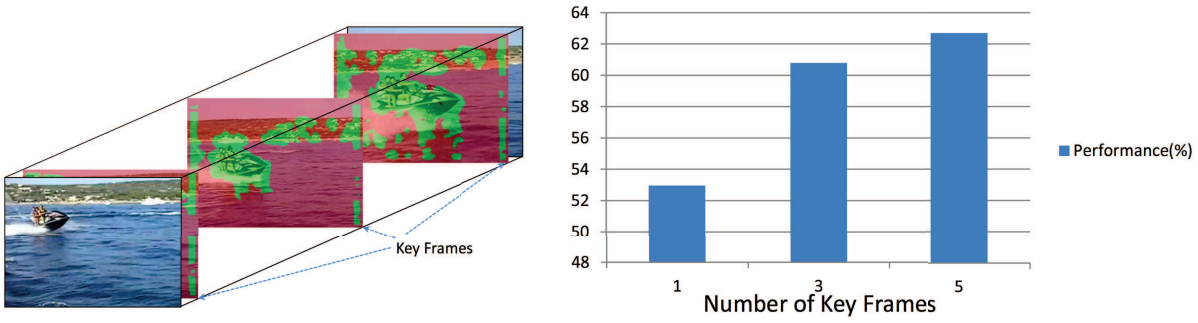

FIGURE 2.23 – Les éléments contextuels sont sélectionnés à partir d'un seuil appliqué sur le flot optique. Le diagramme de droite montre l'amélioration des performances de reconnaissance à mesure que le nombre d'éléments descriptifs augmente (figure tirée de [Reddy and Shah, 2013]).

— [Marszalek et al., 2009] proposent de caractériser des vidéos réalistes de façon dense. Deux types de détecteurs sont utilisés. Le détecteur de points d'intérêt spatio-temporels `STIP` pour les mouvements, et le détecteur Harris 2D [Harris and Stephens, 1988] pour les éléments contextuels. Cette méthode permet de reconnaître des actions dans des scènes complexes. Les éléments d'intérêt sont par la suite fusionnés de façon à prendre en compte l'importance de l'information contextuelle. La figure 2.24 montre un exemple de détection d'évènements temporels importants pour la caractérisation du mouvement avec le détecteur `STIP` et un exemple de détection d'éléments contextuels à l'aide du détecteur de Harris 2D.

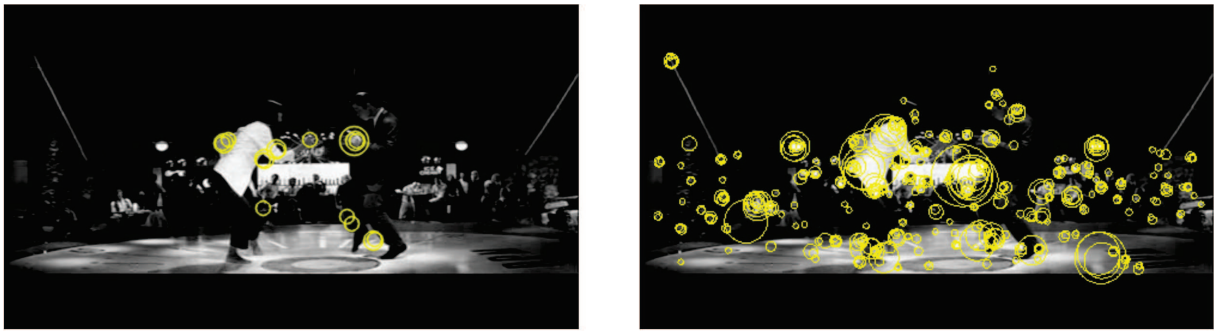

FIGURE 2.24 – À Gauche : éléments d'intérêt détectés par la méthode `STIP` pour caractériser le mouvement. À droite : éléments contextuels de la scène détectés par le détecteur 2D de Harris (figure tirée de [Marszalek et al., 2009]).

Discussion Les méthodes d'extraction denses ont montré leur efficacité quant à la reconnaissance d'actions dans des cadres génériques et réalistes. Fort de ces travaux, d'autres auteurs ont tenté d'améliorer la caractérisation des mouvements des vidéos en utilisant des approches autres que les points spatio-temporels, telles que les trajectoires de mouvements. La partie suivante montre comment ces dernières sont estimées puis caractérisées dans le processus de reconnaissance d'actions.

2.2.2.3 Trajectoires d'intérêt

Le but des méthodes basées sur des trajectoires d'intérêt est d'aller au-delà de la notion de point, trop localisé dans l'espace et dans le temps. Une trajectoire représente les différents mouvements effectués par un sujet au cours du temps et synthétise mieux les déformations temporelles de certaines régions dans une séquence vidéo.
La partie suivante présente une sélection de méthodes de la littérature proposant un modèle de trajectoire comme élément d'intérêt spatio-temporel.

— [Ullah and Laptev, 2012] caractérisent des actions élémentaires en appariant des trajectoires associées aux mouvements de différentes parties du corps. La figure 2.25 montre un exemple d'appariement de trajectoires associées à différents mouvements. Ces trajectoires sont extraites sur des bases de données de vidéos synthétiques pour la phase d'apprentissage. Les descripteurs d'orientation de mouvement `HOF` et de variation de gradient `HOG` sont utilisés pour caractériser ces trajectoires. Les mouvements appris sont par la suite retrouvés dans des vidéos réalistes d'actions humaines, notamment des actions sportives.

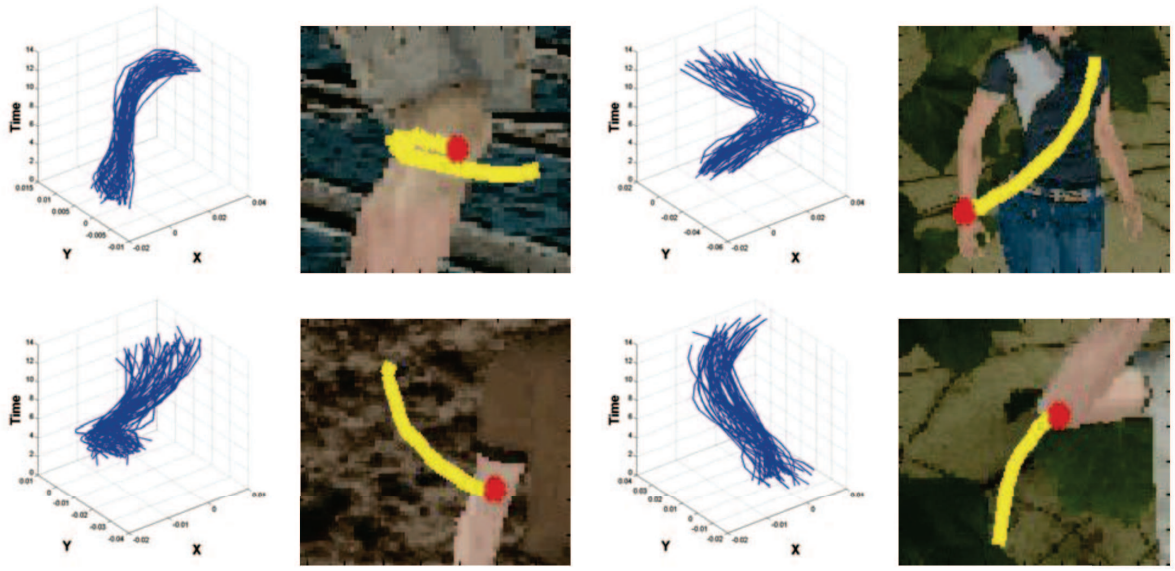

FIGURE 2.25 – Illustration de groupe de trajectoires de différentes articulations du corps (coude et poignet). Les trajectoires sont regroupées en bleu en fonction des mouvements qu'elles représentent (figure tirée de [Ullah and Laptev, 2012]).

— [Raptis and Soatto, 2010] proposent le descripteur `Tracklet` qui encode l'information d'orientation de mouvement le long de trajectoires de points de saillance. La taille des descripteurs dépend donc de la taille des trajectoires. L'algorithme `DTW` est utilisé pour comparer deux descripteurs de taille différente. On remarque sur la figure 2.26 les trajectoires extraites à partir de cette méthode. Les `Tracklet` sont regroupées en fonction des mouvements qu'elles représentent. La couleur de chaque trajectoire correspond au groupe auquel elle appartient.

FIGURE 2.26 – Exemple de trajectoire estimée par la méthode de M.Raptis *et al.*. La couleur des trajectoires dépend du groupe auquel appartient leur descripteur `Tracklet` (figure tirée de [Raptis and Soatto, 2010]).

— [Wang et al., 2011] utilisent une méthode d'extraction de points dense et estiment la position de ces points à différents intervalles de temps. L'utilisation de trajectoires permet de capturer beaucoup plus d'informations temporelles. Un exemple de trajectoires denses est illustré par la figure 2.27. Cette méthode a été améliorée en proposant une détection de personne dans les séquences vidéos ainsi qu'une approche de compensation de caméra en estimant les paramètres de l'homographie liant deux images consécutives [Wang and Schmid, 2013]. Cette approche produit des résultats parmi les meilleurs sur des bases de données de vidéos réalistes telles que `UCF-11 Dataset`, `UCF-50 Dataset` ou encore `Hollywood 2 Dataset` [Wang and Schmid, 2013].

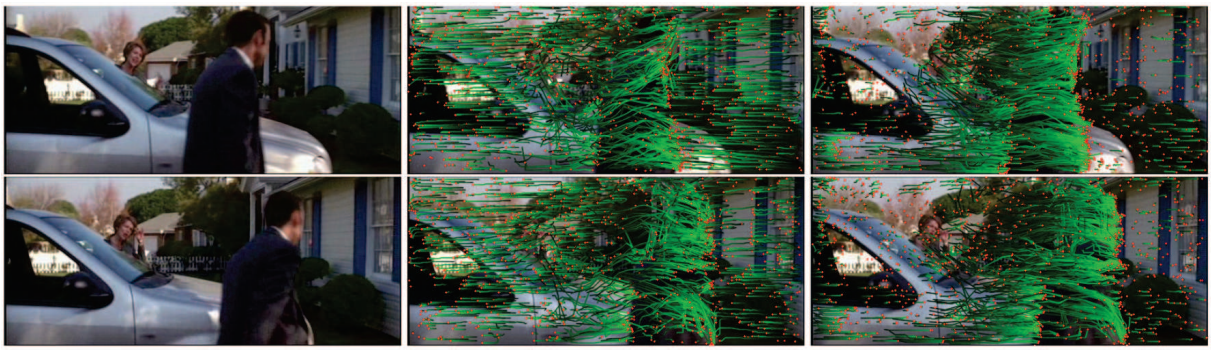

FIGURE 2.27 – 1ère colonne : deux images successives d'une séquence vidéo. 2ème colonne : trajectoires denses extraites à partir de la séquence vidéo. 3ème colonne : amélioration de l'estimation des trajectoires par compensation du mouvement de caméra (figure tirée de [Wang and Schmid, 2013]).

— [Vrigkas et al., 2014] extraient des trajectoires de mouvements à l'aide du flot optique associé à une séquence vidéo. Les actions sont représentées par un modèle de mélange de Gaussiennes (`GMM`) en appariant les trajectoires de toutes les séquences à l'aide de l'algorithme `k-mean`. Les trajectoires étant de tailles variables, un algorithme est utilisé pour détecter les paires similaires entre deux segments de courbes. Les plus longs segments communs entre deux courbes (`LCSS`) sont conservés. Cet algorithme a l'avantage d'être robuste au bruit et permet d'établir la similarité entre deux courbes de tailles différentes. Cette méthode fournit les meilleurs résultats sur des bases de données connues de la littérature (`Weizmann Dataset`, `KTH Dataset`, `UCF-11 Dataset`).

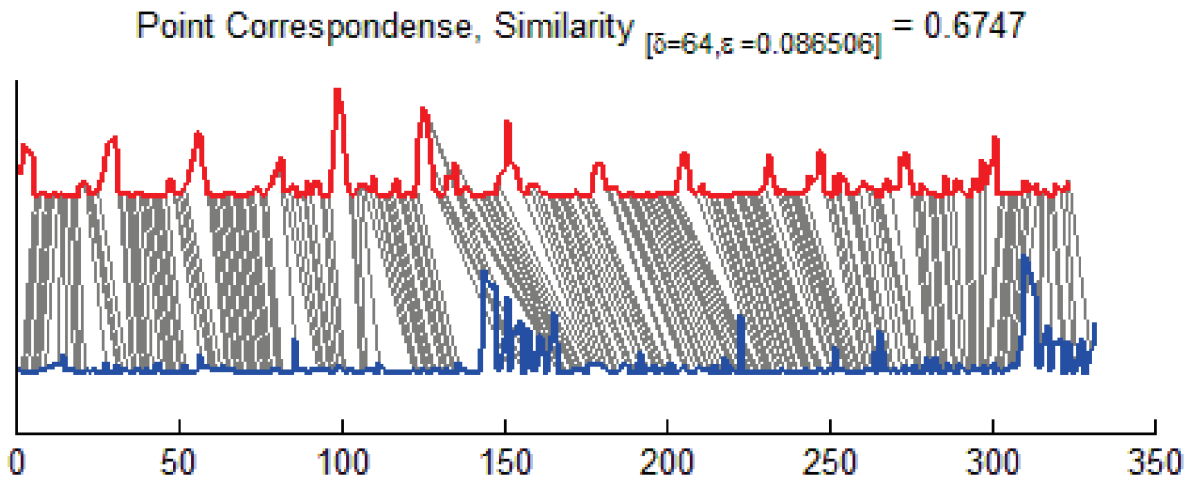

FIGURE 2.28 – Exemple d'appariement entre deux trajectoires de tailles différentes (bleue et rouge). L'algorithme `LCSS` détecte la plus longue séquence commune entre ces deux trajectoires (figure tirée de [Vrigkas et al., 2014]).

Discussion Les extractions denses de points et de trajectoires d'intérêt permettent d'atteindre les meilleurs taux de reconnaissance sur des bases d'actions à la fois complexes et très réalistes, ce que ne permettent pas les méthodes de détection de points d'intérêt "épars" ou encore les approches globales. Cependant, les méthodes basées sur des extractions denses souffrent du même inconvénient. L'échantillonnage dense de points conduit à des temps de calcul très élevés et une quantité importante de données en mémoire. Malgré leurs performances, l'intérêt des méthodes basées sur des extractions denses peut être remis en question. En effet, comme cela a été évoqué dans le Chapitre 1, on assiste à une augmentation exponentielle des données vidéos de part l'utilisation massive de téléphones mobiles, des applications temps-réel liées à la vidéo ainsi qu'un accès de plus en plus démocratisé à des hauts débits de connexion internet. Ce qui pose donc un problème en termes de temps de calcul.

2.2.2.4 Méthodes de réduction de caractéristiques

Certain auteurs tentent de contourner les problématiques liées aux approches dense en proposant des méthodes réduisant le nombre d'éléments à mettre en oeuvre lors du processus de reconnaissance.

— [Shi et al., 2013], montrent que l'utilisation d'un nombre fixe d'éléments issus d'un ensemble dense permet d'obtenir des résultats quasi-similaires à ceux des méthodes denses. À partir d'une extraction dense de points, ces derniers sont sélectionnés aléatoirement toutes les 160 images. Un total de 10000 éléments est conservé. Cela permet de garder plus de points issus des échelles les plus fines et de contrôler leur nombre. Cette méthode présente des taux de reconnaissance élevés sur la base `HMDB-51`, mais reste moins performante sur d'autres bases de grande taille comme `UCF-50 Dataset`.

— [Murthy and Goecke, 2013] développent une méthode pour optimiser la quantité de trajectoires dense générées par l'approche de [Wang et al., 2011]. Les auteurs apparient les trajectoires similaires afin de les fusionner en de nouvelles séquences de points. Ces fusions de trajectoires sont appelées `ordered trajectories`. La figure 2.29 illustre des trajectoires denses "réordonnées" par la méthode de Murthy *et al.*. On constate que le nombre de trajectoires ordonnées est moindre et que ces dernières sont plus localisées sur les éléments pertinents (parties du corps du joueur de basket). Les résultats observés avec cette méthode montrent qu'avec moitié moins de trajectoire et avec les mêmes paramètres, on obtient de meilleurs taux de bonne reconnaissance que l'approche classique des trajectoires denses sur la base `UCF-50`.

FIGURE 2.29 – Exemples de trajectoires denses "réordonnées" par la méthode de R.Murthy *et al.*. Comparativement à l'image de gauche, les trajectoires conservées sont celles qui correspondent au mieux aux mouvements du joueurs (figure tirée de [Murthy and Goecke, 2013]).

2.2.3 Conclusion

Nous avons vu un ensemble de méthodes de la littérature appliquées à la reconnaissance du mouvement humain dans des vidéos. Leur construction ainsi que leur performance varient en fonction des problématiques abordées mais également en fonction des avancées dans les domaines qui leur sont liés (flux internet vidéo en constante expansion, démocratisation des moyens de captation, augmentation des besoins d'application de surveillance, etc.). Nous faisons ici un point sur les avantages et les inconvénients de chacune de ces catégories de méthodes.

Approches globales Les approches globales sont les premières à avoir été appliquées au domaine de la reconnaissance des mouvements humains. Elles permettent de distinguer des mouvements très spécifiques grâce à des méthodes d'analyse fine des silhouettes ou l'analyse des mouvements des parties du corps d'un sujet. Cependant, cette finesse d'analyse n'est possible que dans des cas restreints. En effet, les fortes contraintes d'acquisition nécéssaires ainsi que les pré-traitements tels que la suppression de fond ou le suivi de sujet rendent ces méthodes très peu généralisables. Les mouvements humains reconnus sont généralement effectués dans des contextes très spécifiques (gestuelle de la main, actions liées à un contexte, etc.).

Approches locales Les approches locales permettent de traiter des actions humaines exécutées dans des situations plus complexes. Ces dernières utilisent une représentation des séquences vidéos robuste à différentes variations visuelles survenant lors de la captation et permettent donc une meilleure généralisation des actions. Cependant les approches de détection de points d'intérêt restent limitées dans le cas des vidéos réalistes.

Les approches d'extraction de points denses ou encore les méthodes basées trajectoires d'intérêt ont permit de pallier ce problème en apportant plus d'information dans le processus de reconnaissance. Cependant, cette quantité de données reste un problème en termes de complexité, mais aussi en stockage mémoire. Certaines approches tentent de répondre à cette problématique en réduisant le nombres d'éléments mis en oeuvre lors de la classification. Bien que ces méthodes de réduction d'éléments conduisent à une amélioration sur le taux de reconnaissance, le nombre d'éléments générés est toujours très élevé par rapport à des approches de détection de points épars, (10 à 20 fois plus élevé en moyenne [Wang et al., 2011, Murthy and Goecke, 2013]) et le temps de calcul reste également important. D'autant que ces méthodes ne font pas l'économie d'une extraction dense d'éléments et n'évitent donc pas le stockage de ces données en mémoire.

De plus, les trajectoires d'intérêt utilisées par ces méthodes, bien qu'elles apportent une information plus pertinente, ne sont pas totalement exploitées. En effet, entre une trajectoire et un cuboïd de même longueur temporelle, le gain moyen, en comparant sur plusieurs bases de données est de 2.6% [Wang et al., 2013]. On constate que sur la majorité des méthodes locales, les informations captées le long des trajectoires ou dans un voxel de même longueur sont les mêmes et prennent peu en compte les caractéristiques de ces trajectoires ([Raptis and Soatto, 2010, Wang et al., 2011, Murthy and Goecke, 2013, Shi et al., 2013]).

En résumé, les méthodes éparses offrent de meilleurs temps de calcul et une complexité moindre mais restent peu efficaces dans des situations réalistes. En parallèle, les approches denses ont montré leur performance sur des vidéos génériques mais deviennent extrêmement

coûteuses en temps de calcul du fait du nombre d'éléments à traiter. Elles sont dans ce cas peu efficientes notamment dans le cadre de reconnaissance d'actions en temps-réel.

Conclusion du chapitre Dans ce chapitre, l'état de l'art des méthodes de reconnaissance d'actions élémentaires a été présenté. Les différents types d'approches ont été explicitées de façon à mettre en évidence leurs avantages ainsi que leurs inconvénients. Dans le chapitre suivant, nous présentons notre méthode de reconnaissance d'actions élémentaires dans des vidéos Cette méthode rentre dans la catégorie des approches locales de détection d'éléments d'intérêts. Nous faisons le choix de signer les actions par la notion de mouvement. Le flot optique, communément utilisé pour estimer le mouvement dans des séquences d'image y est présenté. Nous verrons comment le champ vectoriel qui en résulte est utilisé pour analyser avec précision différentes caractéristiques spatio-temporelles à différentes échelles de mouvement. L'évaluation de cette approche dans le chapitre 4 montre comment l'information intrinsèque des mouvements analysés permet d'obtenir des taux de reconnaissance parmi les plus élevés de la littérature. Cette performance est attestée à la fois sur des bases de données de vidéos théoriques mais également de vidéos génériques, tout en conservant une complexité relativement faible.

Chapitre 3

Reconnaissance d'actions élémentaires

Chapitre 3

Reconnaissance d'actions élémentaires

Sommaire

Introduction du chapitre Dans la première partie de ce chapitre, nous introduisons nos travaux sur la détection de points d'intérêt pertinents pour la caractérisation d'actions humaines. La deuxième partie introduit notre approche de reconnaissance d'actions élémentaires. Le processus global de cette méthode, illustré par la figure 3.1, y est présenté et détaillé.

Figure 3.1 – Processus global de notre méthode de reconnaissance d'actions élémentaires.

3.1 Éléments d'intérêt spatio-temporels

Comme l'a montré le chapitre 1, les actions élémentaires humaines dans des vidéos sont des éléments difficiles à identifier et à interpréter. Les méthodes de la littérature permettant de détecter et de caractériser les informations pertinentes au sens du mouvement humain sont souvent peu efficaces dans des contextes très génériques. En effet les approches éparses (`STIP` [Laptev, 2005], `cuboid` [Dollar et al., 2005], `SURF` [Willems et al., 2008], etc.) caractérisent des évènements temporels liés aux mouvements humains, et sont efficaces sur les bases de données avec contraintes d'acquisition. Cependant, elles sont peu pertinentes sur des vidéos plus riches en informations visuelles, où les approches denses, en revanche, réalisent de meilleurs résultats [Wang et al., 2009]. Les approches denses captent beaucoup plus d'informations que les approches dites "éparses" dans des vidéos génériques. La contre-partie des approches denses est le nombre de données à traiter, qui lui, est beaucoup plus conséquent, entrainant des temps de calcul beaucoup plus importants. De plus, on constate que l'utilisation d'un très grand nombre de données ne leurs permettent pas d'être significativement plus efficaces que les approches éparses sur des bases de données avec contraintes d'acquisitions [Wang et al., 2009].

Pour pallier à ce problème, le but d'une méthode efficace de reconnaissance d'actions élémentaires devrait être de notre point de vue :

— D'utiliser un détecteur d'éléments d'intérêt permettant d'**obtenir un nombre raisonnable de données à traiter**.
— De capter l'information la plus cohérente par rapport aux mouvements humains tout en étant **robuste à certaines transformations visuelles survenant dans des vidéos avec peu de contraintes d'acquisition**.

Nous présentons dans cette section des approches de détection de points d'intérêt spatio-temporels de types "épars" qui traitent cette problématique. Cette partie correspond à la première étape de notre processus de reconnaissance d'actions élémentaires comme l'illustre la figure 3.2.

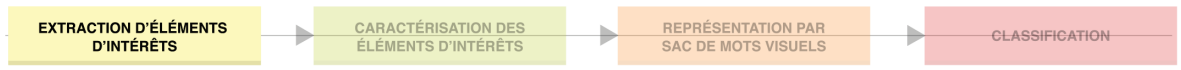

FIGURE 3.2 – L'extraction d'éléments d'intérêt correspond à la première étape de notre processus.

3.1.1 Détecteur de points d'intérêt orientés tenseur de structure

Les travaux de [Wang et al., 2009] sur l'évaluation des différents détecteurs de points d'intérêt de la littérature montrent que la méthode `STIP` de Laptev *et al.* reste la plus pertinente sur les bases de données de vidéos avec contraintes d'acquisition. Les travaux de [Schmid et al., 2000] ont également montré l'avantage du détecteur de la méthode `STIP` en terme de robustesse et de répétabilité face à d'autres détecteurs de points d'intérêt de la littérature. Cette méthode reste une référence pour la caractérisation de points d'intérêt spatio-temporels.
Néanmoins, on constate qu'elle ne permet pas de détecter tous les évènements temporels pouvant présenter un intérêt au sens du mouvement. Les points de la méthode `STIP` sont des coins spatiaux avec une forte variation de mouvement dans le temps, sur une courte période. Hors, certains évènements visuels intéressants pour caractériser une action, possèdent une forte variation temporelle mais ne sont pas nécessairement des coins spatiaux. Ces derniers ne sont pas détectés par la méthode `STIP`.

Notre but est de construire un détecteur de points d'intérêt mieux adapté à tous types d'évènements temporels. La partie qui suit montre comment on obtient un détecteur d'évènements temporels plus général en reformulant l'écriture du détecteur `STIP`.

Prenons deux exemples illustrant les inconvénients cités plus haut (voir figures 3.3 et 3.4).Ces deux cas illustrent de façon théorique, des évènements qui ne sont pas d'intérêt au sens de la méthode `STIP`. Le point de l'exemple 1 ne présente aucune structure angulaire forte assimilable à un coin mais présente un évènement temporel intéressant lors du changement brusque d'échelle. Le segment de l'exemple 2 possède un gradient élevé uniquement dans la direction horizontale et ne possède donc pas de coins spatiaux mais peut présenter un intérêt en terme de mouvement lors des changements brusques de déplacement.

FIGURE 3.3 – Exemple 1 : Point dont l'échelle augmente puis diminue au cours du temps jusqu'à disparaitre (voir vidéo).

FIGURE 3.4 – Exemple 2 : Segment qui se déplace aléatoirement au cours du temps (voir vidéo).

Lorsque l'on reprend l'écriture du détecteur Harris 3D[Laptev, 2005] :
$H_{3D} = \lambda_1\lambda_2\lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3$.
avec λ_1, λ_2 et λ_3 les valeurs propres du tenseur de structure associé à un pixel quelconque (x, y, t)
On constate que les deux valeurs propres λ_1 et λ_2 sont nulles pour l'exemple 1 et la valeur propre λ_1 est nulle pour l'exemple 2.
Dans les deux cas, même lorsque l'on observe un évènement temporel important (λ_3 élevé) le détecteur de la méthode STIP donnera une réponse faible (voir figure 3.5 et figure 3.6).

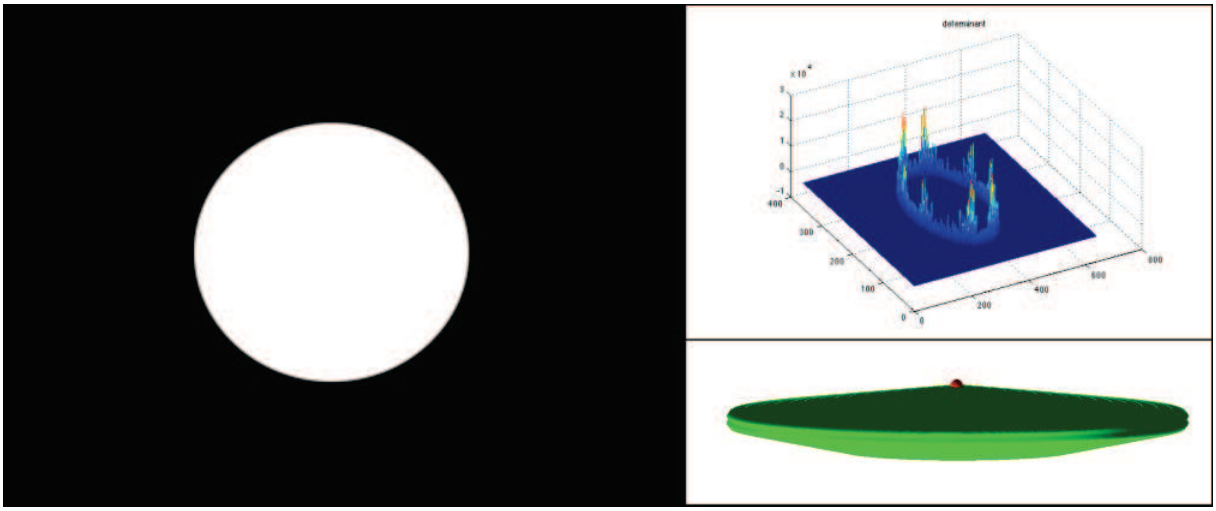

FIGURE 3.5 – Résultat du détecteur Harris 3D sur l'exemple 1 à l'image 19 de la séquence. Un point est détecté lorsque le disque disparait (échelle zéro), en bas à droite.

Considérons une image de niveaux de gris dont l'intensité est exprimée par la fonction I. La paramétrisation φ de I est telle que :

$$\varphi : \mathbb{R}^2 \longrightarrow \mathbb{R}^3 \tag{3.1}$$
$$(x, y) \rightarrow (x, y, I(x, y)) \tag{3.2}$$

On définit à partir de cette paramétrisation la Jacobienne J de I :

$$J = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ I_x & I_y \end{pmatrix}$$

avec $\frac{\partial I}{\partial x}$ noté I_x et $\frac{\partial I}{\partial y}$ noté I_y
En munissant $\mathbb{R}^3$ de la métrique g telle que :

$$g = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \mu \end{pmatrix}$$

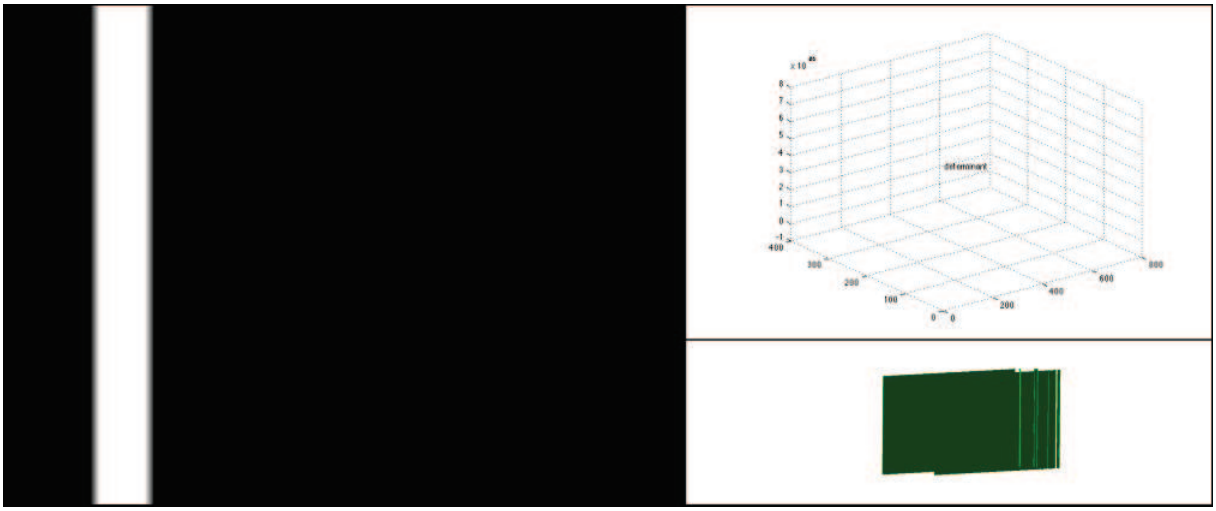

FIGURE 3.6 – Résultat avec du détecteur Harris 3D sur l'exemple 2 à l'image 6 . Aucun point n'est détecté durant la séquence.

On obtient alors :

$$\varphi^* g = \begin{pmatrix} 1 + \mu I_x^2 & \mu I_x I_y \\ \mu I_x I_y & 1 + \mu I_y^2 \end{pmatrix} \tag{3.3}$$

$$\tag{3.4}$$

avec $\varphi^* g = J^T g J$ qui est un tenseur de structure sur $\varphi(\mathbb{R}^2)$.

Dans [Rousseau et al., 2010], la paramétrisation suivant la métrique g ci-dessus est utilisée afin d'obtenir les gradients d'images couleurs selon une orientation particulière. La finalité de cette approche est de détecter des contours dans des images couleurs bruitées comme l'illustre la figure 3.7.

FIGURE 3.7 – Exemple de gradient obtenu sur une image bruitée. Le gradient est orienté dans la direction de la composante principale de l'image. (image tirée de [Rousseau et al., 2010])

Nous introduisons ci-dessous le principe de cette approche dans le cadre des images couleurs. Dans le cas où I est une image couleur dans l'espace `RGB`, on a la paramétrisation :

$$\varphi : \mathbb{R}^2 \longrightarrow \mathbb{R}^5 \tag{3.5}$$

$$(x, y) \rightarrow (x, y, I_1(x, y), I_2(x, y), I_3(x, y)) \tag{3.6}$$

Avec la Jacobienne J de I telle que :

$$J = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ I_1x & I_1y \\ I_2x & I_2y \\ I_3x & I_3y \end{pmatrix}$$

En munissant $\mathbb{R}^3$ de la métrique g telle que :

$$g = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

On obtient la matrice $\varphi^* g$ telle que :

$$\varphi^* g = \begin{pmatrix} 1 + I_1x^2 & I_1xI_1y \\ I_1xI_1y & 1 + I_1y^2 \end{pmatrix} \tag{3.7}$$

Dans ce cas, on s'intéresse uniquement à la composante R de l'image I.

Si la métrique g de $\mathbb{R}^5$ est telle que :

$$g = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

On a :

$$\varphi^* g = \begin{pmatrix} 1 + \sum_{i=1}^{3} I_i x^2 & \sum_{i=1}^{3} I_i x I_i y \\ \sum_{i=1}^{3} I_i x I_i y & 1 + \sum_{i=1}^{3} I_i y^2 \end{pmatrix} \tag{3.8}$$

Toutes les composantes sont ici prises en compte.

De manière générale, on peut choisir une direction privilégiée sur laquelle se projeter. Soit le vecteur unitaire $\mathbf{u} = (a, b, c)^T$. On considère la forme quadratique $Q = uu^T$ telle que la matrice associée à Q soit :

$$Q = \begin{pmatrix} a^2 & ab & ac \\ ab & b^2 & bc \\ ac & bc & c^2 \end{pmatrix}$$

On définit la métrique g associée $\mathbb{R}^5$ telle que :

$$g = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \bigoplus \begin{pmatrix} a^2 & ab & ac \\ ab & b^2 & bc \\ ac & bc & c^2 \end{pmatrix}$$

En prenant un vecteur $\mathbf{u}$ dans la direction de la première composante principale de l'image couleur, le résultat obtenu permet de filtrer le bruit présent dans l'image. En effet, le bruit est contenu dans la direction orthogonale à la composante principale (voir figure 3.7).

Dans le cas de séquences d'images en niveaux de gris I, nous proposons la paramétrisation suivante :

$$\varphi : \mathbb{R}^3 \longrightarrow \mathbb{R}^6 \tag{3.9}$$

$$(x, y, t) \rightarrow (x, y, t, I(x, y_0, t_0), I(x_0, y, t_0), I(x_0, y_0, t)) \tag{3.10}$$

$I(x, y_0, t_0)$ correspond à la séquence I avec les composantes y et t fixées. On ne considère donc que les lignes de l'image une à une. Il en est respectivement de même pour $I(x_0, y, t_0)$ et $I(x_0, y_0, t)$.

La Jacobienne associée à cette paramétrisation est donc :

$$J = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ I_x & 0 & 0 \\ 0 & I_y & 0 \\ 0 & 0 & I_t \end{pmatrix}$$

En utilisant la méthode ci-dessus pour la construction de métrique et en considérant le vecteur unitaire $\mathbf{u} = (a, b, c)^T$ on obtient g tel que :

$$g = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \bigoplus \begin{pmatrix} a^2 & ab & ac \\ ab & b^2 & bc \\ ac & bc & c^2 \end{pmatrix}$$

et

$$\varphi^* g = \begin{pmatrix} 1 + a^2 I_x^2 & abI_x I_y & acI_x I_t \\ abI_x I_y & 1 + b^2 I_y^2 & bcI_y I_t \\ acI_x I_t & bcI_y I_t & 1 + c^2 I_t^2 \end{pmatrix} \tag{3.11}$$

En jouant sur la valeur de a,b et c, on donne de l'importance aux composantes spatiales ou à la composante temporelle. On remarque que la formule associée au détecteur `STIP` est un cas particulier avec un vecteur $\mathbf{u} = \frac{1}{\sqrt{3}}(1,1,1)^T$:

$$g = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \bigoplus \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

$$I_1 = \varphi^* g = \begin{pmatrix} 1 + I_x^2 & I_x I_y & I_x I_t \\ I_x I_y & 1 + I_y^2 & I_y I_t \\ I_x I_t & I_y I_t & 1 + I_t^2 \end{pmatrix} \tag{3.12}$$

On a donc généralisé l'écriture du détecteur `STIP` de façon à obtenir une plus grande liberté quant à l'importance donnée aux composantes spatiales ou temporelles.

Dans le cas des deux exemples présentés précédemment, on utilise cette approche pour détecter des évènements temporels autres que des coins spatiaux avec de forts mouvements temporels. Soit le vecteur $\mathbf{u} = \frac{1}{\sqrt{7}}(1,1,5)^T$. La métrique g qui en résulte est :

$$g = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \bigoplus \frac{1}{\sqrt{7}} \begin{pmatrix} 1 & 1 & 5 \\ 1 & 1 & 5 \\ 5 & 5 & 25 \end{pmatrix}$$

On obtient donc :

$$I_u = \varphi^* g = \frac{1}{\sqrt{7}} \begin{pmatrix} 1 + I_x^2 & I_x I_y & 5I_x I_t \\ I_x I_y & 1 + I_y^2 & 5I_y I_t \\ 5I_x I_t & 5I_y I_t & 1 + 25I_t^2 \end{pmatrix} \tag{3.13}$$

$$\tag{3.14}$$

I_u étant le tenseur de structure généralisé suivant le vecteur u.

Les résultats obtenus en utilisant la méthode de détection de points d'intérêt spatio-temporels de Laptev avec la matrice I_u sont visibles sur la figure 3.8 et 3.9. On constate qu'en utilisant la méthode de détection de points d'intérêt spatio-temporels de Laptev et en accordant plus d'importance à la composante temporelle ($\mathbf{u} = \frac{1}{\sqrt{7}}(1,1,5)^T$) on détecte des points cohérents par rapport aux phénomènes présents dans les deux exemples. En effet, sur la figure 3.8 ,une série de

points apparait sur le contour du disque au moment où ce dernier commence à rétrécir (image du haut). Sur la figure 3.9 une série de points apparait tout le long des contours lorsque la ligne change de direction (image du bas).

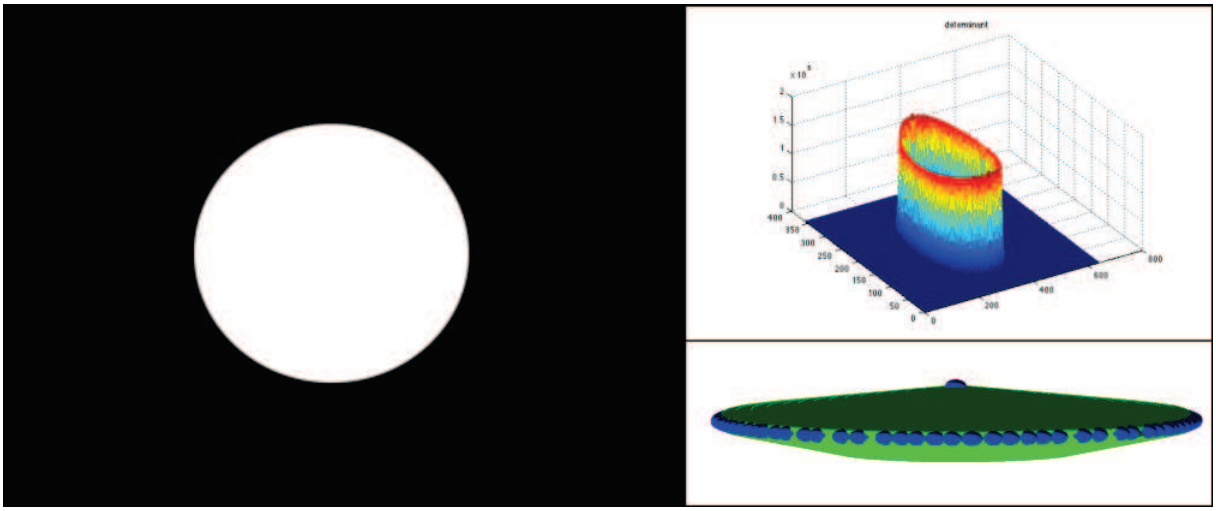

FIGURE 3.8 – Résultat avec le tenseur l_u. Des points sont détectés lors du changement d'échelle (image 19 de la séquence).

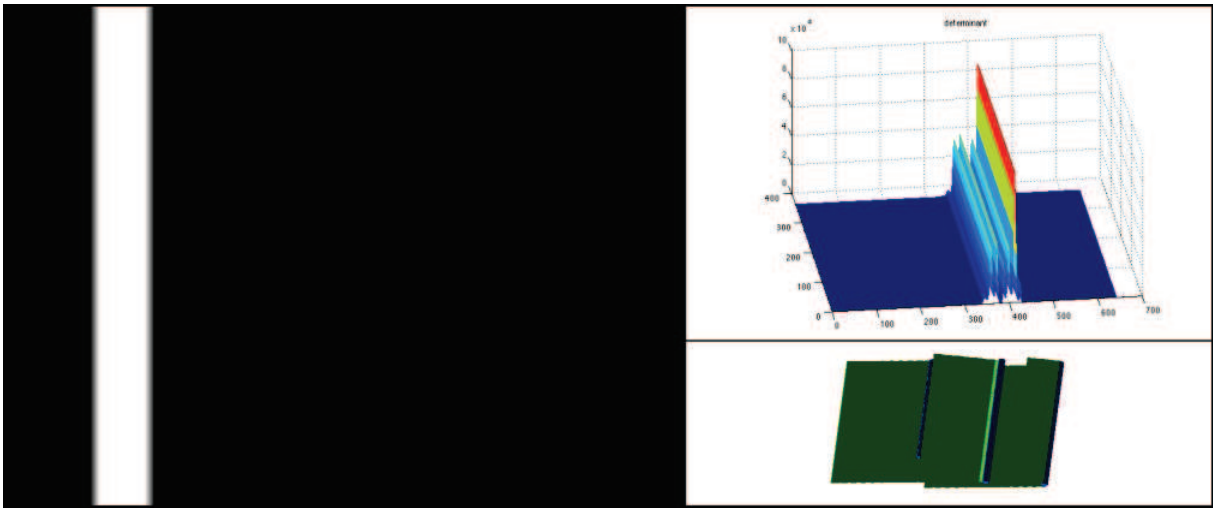

FIGURE 3.9 – Résultat avec le tenseur l_u. Des points sont détectés le long du contour lors du changement brusque de direction.

Dans ce cadre, on peut prendre en compte deux critères pour optimiser la détection de points d'intérêt :

— l'apprentissage de métrique sur les séquences vidéos.
— la maximisation du critère de répétabilité des points d'intérêt détectés en fonction de la direction du vecteur.

Apprentissage de métrique En considérant deux ensembles S et D de paires d'éléments que l'on considère comme similaires et dissimilaires, l'apprentissage de métrique a pour but de minimiser la distance entre les paires similaires et maximiser la distance entre les paires d'éléments dissimilaires [Bellet et al., 2013]. Dans notre cas, l'ensemble S correspond aux paires d'éléments d'intérêt et l'ensemble D correspond aux paires entre les éléments d'intérêt et les éléments négligeables pour la reconnaissance d'actions (pixels du fond par exemple).
Cette méthode n'est pas la plus probante pour notre problème pour deux raisons :

— La métrique g n'est pas utilisée comme une distance dans notre méthode, rien ne garantit que l'apprentissage de métrique corresponde à ce que l'on recherche.
— Deuxièmement, l'ensemble S de paires d'éléments d'intérêt est précisément ce que l'on cherche à déterminer avec notre détecteur. On ne peut fournir dans ce cadre une vérité terrain de points d'intérêt spatio-temporels.

Maximiser le critère de répétabilité La répétabilité est un critère très utilisé pour évaluer la robustesse d'un détecteur de point d'intérêt. Le but est d'évaluer la proportion de points d'intérêt conservés après différentes transformations appliquées à la vidéo (translation, échelle, rotation, ajout de bruit, changement de *frame rate*, etc) [Schmid et al., 2000]. Cependant, cette méthode même si elle permet de détecter les points d'intérêt les plus robustes, tend à réduire le nombre de points d'intérêt utilisés pour la reconnaissance d'activité. Hors, on sait que les taux de bonnes reconnaissances sont aussi liés au nombre de points d'intérêt mis en jeu [Wang et al., 2009, Laptev et al., 2008, Reddy and Shah, 2013].

Discussion Ce détecteur que nous avons proposé présente un certain nombre d'avantage, notamment la possibilité de généraliser le détecteur de la méthode `STIP` et de détecter des évènements temporels autres que des *coins* spatio-temporels. Cependant, pour les raisons évoquées ci-dessus, il ne sera pas utilisé pour la suite de nos travaux. Dans la partie qui suit, nous introduisons notre deuxième approche d'extraction d'éléments d'intérêt. Ces éléments spatio-temporels sont basés sur l'étude du mouvement apparent dans les vidéos séquences. Nous verrons comment ces éléments sont détectés et leur pertinence quant aux actions élémentaires présentes dans des vidéos.

3.1.2 Points critiques du flot optique

L'approche d'extraction d'éléments d'intérêt que nous présentons ci-dessous se base sur l'estimation du mouvement apparent dans une séquence d'images. Le flot optique est l'outil couramment utilisé dans le domaine de la vision par ordinateur pour estimer le mouvement apparent. Dans un premier temps nous introduisons le principe du flot optique, ainsi que les principales méthodes utilisées pour l'estimer.

3.1.2.1 Flot optique

Le flot optique d'une séquence vidéo correspond aux mouvements apparents des pixels au cours du temps. Il est caractérisé par un champ vectoriel $\mathbf{F}_t$ à deux composantes (déplacement horizontal u_t, déplacement vertical v_t) tel que : $\mathbf{F}_t = (u_t, v_t)$.
L'objectif est d'estimer le vecteur de déplacement $\mathbf{v}$ de chaque pixel entre deux images consécutives prises dans un intervalle de temps Δt. Cette estimation est faite en supposant le conservation de la luminance entre les deux images. Soit $I(x, y, t)$ la fonction exprimant l'intensité en niveau de gris d'une image au point (x, y, t). On cherche donc à estimer Δx, Δy,Δt tels que :

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \tag{3.15}$$

En supposant que les mouvements des pixels entre deux images consécutives sont petits, on développe $I(x, y, t)$ en série de Taylor :

$$I(x + \Delta x, y + \Delta y, t + \Delta t) = I(x, y, t) + \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta yt + \frac{\partial I}{\partial t}\Delta t \tag{3.16}$$

L'équation 3.15 devient :

$$\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t = 0 \tag{3.17}$$

En divisant par Δt on a :

$$\frac{\partial I}{\partial_x}v_x + \frac{\partial I}{\partial y}v_y + \frac{\partial I}{\partial t} = 0 \tag{3.18}$$

Avec v_x et v_y composantes horizontales et verticales du vecteur vitesse $\mathbf{v}$. En posant $I_x = \frac{\partial I}{\partial x}$, $I_y = \frac{\partial I}{\partial y}$ et $I_t = \frac{\partial I}{\partial t}$ on a finalement :

$$I_x v_x + I_y v_y = -I_t \tag{3.19}$$

Ce système, à deux inconnues (v_x, v_y) et une équation, est sous-déterminé. Cette équation traduit un problème communément appelé "problème de l'ouverture" dans le domaine de la vision par ordinateur. La figure 3.10 illustre ce problème. Quand le mouvement apparent est observé dans le voisinage restreint d'un contour, on ne peut estimer le vecteur de déplacement $\mathbf{v}$ de ce contour. Seule la composante orientée dans le sens du gradient peut être évaluée. Le flot apparent qui en résulte est appelé le *flot normal*.

Pour déterminer le mouvement apparent entre deux images consécutives, il est nécessaire de faire quelques suppositions complémentaires. Il existe plusieurs méthodes dans la littérature pour déterminer le flot optique à partir de cette équation. Les deux exemples explicités plus bas sont les méthodes les plus classiquement utilisées.

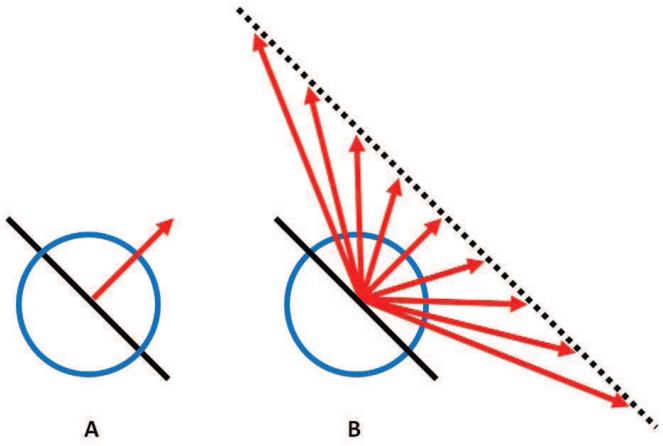

FIGURE 3.10 – Illustration du problème de l'ouverture. La direction du mouvement d'un contour (segment noir) ne peut être déterminée quand ce mouvement est observé dans un voisinage restreint (cercle bleu). Le mouvement apparent est orienté dans le sens du gradient, que le contour se déplace horizontalement ou verticalement (figure A).

Horn & Schunk La méthode de Horn & Schunk [Horn and Schunck, 1981] est une approche globale qui vise à minimiser la fonctionnelle ci-dessous :

$$E = \int\int [(I_x v_x + I_y v_y + I_t)^2 + \alpha^2(\|\nabla v_x\|^2 + \|\nabla v_x\|)^2] dxdy \tag{3.20}$$

Avec α une constante de régularisation. Cette fonctionnelle est minimisée en utilisant la méthode d'Euler-Lagrange telle que :

$$\frac{\partial L}{\partial u} - \frac{\partial}{\partial x}\frac{\partial L}{\partial u_x} - \frac{\partial}{\partial y}\frac{\partial L}{\partial u_y} = 0 \tag{3.21}$$

$$\frac{\partial L}{\partial v} - \frac{\partial}{\partial x}\frac{\partial L}{\partial v_x} - \frac{\partial}{\partial y}\frac{\partial L}{\partial v_y} = 0 \tag{3.22}$$

Avec L la fonction intégrante de E. On obtient :

$$I_x(I_x u + I_y v + I_t) - \alpha^2 \Delta u = 0 \tag{3.23}$$

$$I_y(I_x u + I_y v + I_t) - \alpha^2 \Delta v = 0 \tag{3.24}$$

Avec Δ l'opérateur Laplacien. En remplacant le Lapacien de u et v par son approximation en différences finies on peut écrire $\Delta u = \bar{u}(x, y) - u(x, y)$. Avec $\bar{u}$ la moyenne de u calculée dans un voisinage de (x, y). On a donc :

$$(I_x^2 + \alpha^2)u + I_x + I_y v = -\alpha^2 \bar{u} - I_x I_t \tag{3.25}$$

$$I_x + I_y u + (I_x^2 + \alpha^2)v = -\alpha^2 \bar{v} - I_y I_t \tag{3.26}$$

Ce système peut être résolu classiquement de façon itérative par la méthode Gauss-Seidel telle que :

$$u^{k+1} = \bar{u}^k - \frac{I_x(I_x \bar{u}^k + I_y \bar{v}^k + I_t)}{\alpha^2 + I_x^2 + I_y^2} \tag{3.27}$$

$$v^{k+1} = \bar{v}^k - \frac{I_y(I_x \bar{u}^k + I_y \bar{v}^k + I_t)}{\alpha^2 + I_x^2 + I_y^2} \tag{3.28}$$

Avec u^{k+1} et v^{k+1} l'estimation des composantes de $\mathbf{v}$ à l'itération k. L'avantage de la méthode de Horn & Schunk est qu'elle produit un champ vectoriel dense en interpolant le mouvement des pixels dans les zones où l'information de gradient est manquante, ce qui est un plus par rapport aux approches locales que nous allons décrire à présent.

Lucas & Kanade L'approche de Lucas & Kanade [Lucas and Kanade, 1981] suppose qu'entre deux images consécutives, le déplacement des pixels est petit et localement homogène. Pour un pixel p dont le vecteur de déplacement est $\mathbf{v}$, les pixels appartenant à un voisinage V_p suffisamment petit de p possèdent le même vecteur de déplacement $\mathbf{v}$. Il en résulte que :

$$\begin{cases} I_x(p_1)v_x + I_y(p_1)v_y = -I_t(p_1) \\ I_x(p_2)v_x + I_y(p_2)v_y = -I_t(p_2) \\ \quad .. \quad = \quad .. \\ \quad .. \quad = \quad .. \\ I_x(p_n)v_x + I_y(p_n)v_y = -I_t(p_n) \end{cases}$$

Avec $p_i \in V_p \quad i = (1, .., n)$

Le problème précédent devient donc sur-déterminé et peut s'écrire matriciellement comme $A\mathbf{v} = b$ avec :

$$A = \begin{bmatrix} I_x(p_1) & I_y(p_1) \\ I_x(p_2) & I_y(p_2) \\ . & . \\ . & . \\ . & . \\ I_x(p_n) & I_y(p_n) \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} v_x \\ v_y \end{bmatrix} \quad et \quad b = \begin{bmatrix} I_t(p_1) \\ I_t(p_2) \\ . \\ . \\ . \\ I_t(p_n) \end{bmatrix}$$

On a donc :

$$A^\top A\mathbf{v} = A^\top b$$

On obtient finalement :

$$\mathbf{v} = (A^\top A)^{-1} A^\top b$$

L'utilisation de la méthode des moindres carrés suppose que l'erreur dans le voisinage V_p suit une distribution normale de moyenne nulle. Cette supposition peut amener des erreurs d'estimation, notamment lorsque les mouvements de pixels entre deux images consécutives sont grands.

Dans notre approche, le flot optique utilisé est une variante de celui proposée par Horn & Schunk [Sun et al., 2010a]. Le chapitre 4 présente les avantages de cette version par rapport aux approches classiques que nous venons de discuter.

Le flot optique est une méthode très employée dans le domaine de la reconnaissance d'actions humaines. Il est utilisé en tant que descripteur pour décrire le mouvement apparent

dans le voisinage de points d'intérêt préalablement détectés (HOF [Laptev et al., 2008], MBH [Wang et al., 2011], etc.). Dans l'approche présentée ci-dessous, le flot optique est utilisé comme un outil permettant la détection d'éléments d'intérêt dans des vidéos. Ces éléments d'intérêt correspondent aux points décrivant une forte déformation du champ vectoriel associé au flot optique. Nous verrons comment sont définis ces points ainsi que leur intérêt quant à la caractérisation d'évènements temporels.

3.1.2.2 Points critiques d'un champ vectoriel

Caractérisation d'un point critique Un concept important dans l'étude de champ vectoriel est celui des points critiques. Un champ vectoriel $\mathbf{F}$ peut être interprété comme un système d'équations différentielles dans $\mathbb{R}^n$ tel que :

$$\begin{cases} \frac{\partial \mathrm{x}^1}{\partial t} & = \mathbf{F}^1(\mathrm{x}^1, ..., \mathrm{x}^n) \\ \frac{\partial \mathrm{x}^2}{\partial t} & = \mathbf{F}^2(\mathrm{x}^1, ..., \mathrm{x}^n) \\ ... & = ... \\ ... & = ... \\ \frac{\partial \mathrm{x}^n}{\partial t} & = \mathbf{F}^n(\mathrm{x}^1, ..., \mathrm{x}^n) \\ \mathrm{x}(t=0) & = \mathrm{x}_0 \end{cases}$$

Une solution $(\mathrm{x}^1(t), \mathrm{x}^2(t), ..., \mathrm{x}^n(t))$ de ce système peut être considérée comme une représentation paramétrique d'une courbe dans $\mathbb{R}^n$. Le champ vectoriel $\mathbf{F}(p)$ donne le vecteur tangent à la courbe au point p. Un point critique $a \in \mathbb{R}^n$ de ce système d'équations différentielles est défini tel que :

$$\frac{\partial \mathrm{x}^i}{\partial t}(a) = 0 \quad i = (1, ..., n)$$

et donc :

$$\mathbf{F}(a) = 0$$

Au voisinage d'un point critique a, le champ vectoriel peut être approché linéairement suivant une approximation de Taylor :

$$\mathbf{F}(a+h) = \mathbf{F}(a) + \mathbf{J}_a(h) + O(\|h\|^2)$$

Avec $\mathbf{J}_a$ la Jacobienne de $\mathbf{F}$ au point a. En fonction des valeurs propres de $\mathbf{J}_a$ on peut distinguer différents types de points critiques.

Classification des points critiques Un point critique a est dit hyperbolique si toutes les valeurs propres de la Jacobienne $\mathbf{J}_a$ possèdent une partie réelle $\Re$ non nulle. L'une des propriétés des points critiques hyperboliques concerne leur stabilité structurelle. En effet, ajouter une légère perturbation au champ vectoriel $\mathbf{F}$ ne change pas la topologie des lignes de flux autour d'un point critique. En fonction du signe des valeurs propres de leur Jacobienne, les points critiques hyperboliques peuvent être classés en différents types :

1. **Deux valeurs propres réelles α_1 et α_2**
   — α_1, α_2 positifs : *Noeud répulsif*
   — α_1, α_2 négatifs : *Noeud attractif*
   — $\alpha_1\alpha_2 < 0$: *Point selle*
2. **Deux valeurs propres complexes conjuguées α_1 et α_2**
   — $\Re(\alpha_1)$, $\Re(\alpha_2)$ positives : *Spirale répulsive*
   — $\Re(\alpha_1)$, $\Re(\alpha_2)$ negatives : *Spirale attractive*

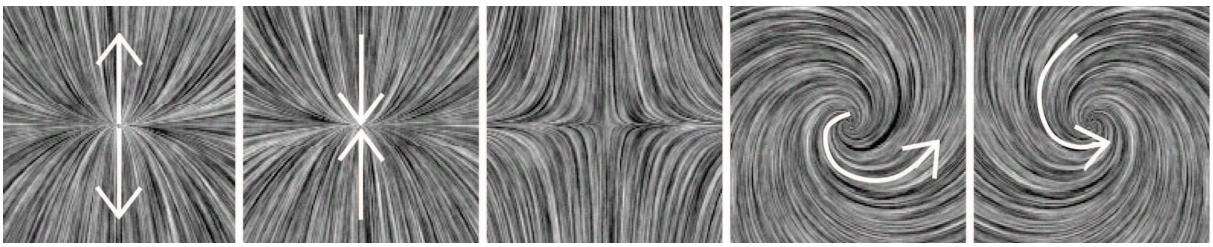

FIGURE 3.11 – Illustration de points critiques d'un champ vectoriel. De gauche à droite : Noeud répulsif, Noeud attractif, Point selle, Spirale répulsive, Spirale attractive.

La figure 3.11 illustre les différents types de points critiques hyperboliques d'un champ vectoriel. On constate que ces points caractérisent de façon pertinente sa déformation. Dans le cas où ce champ vectoriel est issu du mouvement apparent entre deux images consécutives, les points critiques hyperboliques caractérisent les régions spatio-temporelles du flot optique qui contiennent de fortes variations de mouvement. Les caractéristiques physiques de déformation du champ vectoriel au voisinage de ces points est propre aux mouvements d'extension, de repliement, de torsion fréquemment provoqués par les mouvements humains dans des vidéos. En effet, les mouvements générés par les éléments du fond ou par les mouvements de caméras sont généralement des mouvements de translation ou d'échelle. De part leur nature, les points critiques hyperboliques apportent donc une information essentielle sur les mouvements pertinents compris dans le flot optique ainsi que sur la nature physique de ces mouvements.

Pour estimer ces points critiques, nous séparons le flot optique en deux composantes : la divergence et le rotationnel. Nous verrons dans quelle mesure ces deux composantes permettent un compromis entre caractérisation pertinente des points critiques et temps de calcul.

Rotationnel et divergence Soit un champ de vecteur $\mathbf{F}_t = (u_t, v_t)$ avec u_t et v_t ses composantes horizontales et verticales, le rotationel Rot et la divergence Div de $\mathbf{F}$ sont définis comme suit :

$$Rot(\mathbf{F}_t) = \nabla \wedge \mathbf{F}_t = \frac{\partial v_t}{\partial x} - \frac{\partial u_t}{\partial y} \tag{3.29}$$

$$Div(\mathbf{F}_t) = \nabla \cdot \mathbf{F}_t = \frac{\partial u_t}{\partial x} + \frac{\partial v_t}{\partial y} \tag{3.30}$$

Le rotationnel et la divergence sont communément utilisés en physique pour l'étude de champs vectoriels liés à des mouvements fluides. L'importance de cette décomposition est liée au théorème de `Helmholtz` [Bhatia et al., 2013] qui énonce qu'un champ vectoriel $\mathbf{F}$ peut être recomposé à partir de sa divergence $Div(\mathbf{F})$ et de son rotationnel $Rot(\mathbf{F})$ si ces deux composantes sont connues en tout point.

Ces deux composantes sont pertinentes pour notre étude car elles caractérisent deux déformations physiques du flot optique au cours du temps :

— Le rotationel donne une information sur la manière dont un champ de vecteur peut « tourner » localement.

— La divergence mesure à quel degré un point du champ de vecteur est une source ou un puits.

Les définitions précédentes montrent que $Rot(\mathbf{F}_t)$ et $Div(\mathbf{F}_t)$ caractérisent l'allure des lignes de champ au voisinage d'un point p quelconque de ce champ. On remarque également que les singularités obtenues en prenant les extrema de ces deux composantes coïncident avec certains points critiques du flot optique, notamment les spirales et les noeuds, attractifs et répulsifs. Ces points critiques, comme vu précédemment, sont porteurs d'informations liés à de potentiels mouvements d'intérêt dans une séquence vidéo.

Caractériser ces points critiques à partir de la divergence et du rotationnel de $\mathbf{F}$ présente également un intérêt d'un point de vue calculatoire. En effet, ces composantes sont estimées à partir d'opérateurs linéaires différentiels du premier ordre (voir les équations 3.29 et 3.29).

Points critiques du flot optique : éléments d'intérêt spatio-temporels Nous utilisons les notions et outils précédemment cités pour la détection de points critiques du flot optique dans le cadre de la caractérisation d'actions élémentaires humaines dans des vidéos.

Les figures 3.13, 3.14, 3.15 montrent l'extraction de points critiques du flot optique issus de vidéos d'actions humaines élémentaires illustrées sur la figure 3.12.

FIGURE 3.12 – Vidéos issues de `UCF-50 Dataset`, `UCF-11 Dataset`, `KTH Dataset`

Sur chaque figure, de gauche à droite, on peut distinguer : 1) une image de la séquence à l'instant t, 2) la divergence du flot optique à l'instant t, 3) le rotationnel du flot optique à l'instant t, 4) les points critiques obtenus à l'instant t. Sur ces figures, on constate que les composantes de divergence et de rotationnel ne caractérisent pas les mêmes évènements, ce qui démontre leur complémentarité et l'importance de traiter ces deux caractéristiques du flot optique.

La vidéo illustrée par la figure 3.13 présente plusieurs coureurs cyclistes filmés à l'aide d'une caméra embarquée dans une voiture qui les suit. Le fond est en mouvement et les cyclistes ont une position stable par rapport à la caméra. On constate néanmoins que les points critiques obtenus sont majoritairement situés sur les coureurs cyclistes, notamment sur leurs roues ainsi que leurs pédales. Peu de points critiques sont obtenus au niveau du fond en mouvement. On remarque que les lignes de flux du flot optique associées au fond sont globalement homogènes et orientées dans le même sens. Cette configuration est typiquement liée à de forts mouvements de translation, générés par des mouvements de caméra ou par des éléments du fond. Ces perturbations, même fortes, ne génèrent pas de points critiques du flot optique car elles ne présentent pas de déformations complexes du champ vectoriel. On caractérise donc ici uniquement les éléments d'intérêt liés aux mouvements humains.

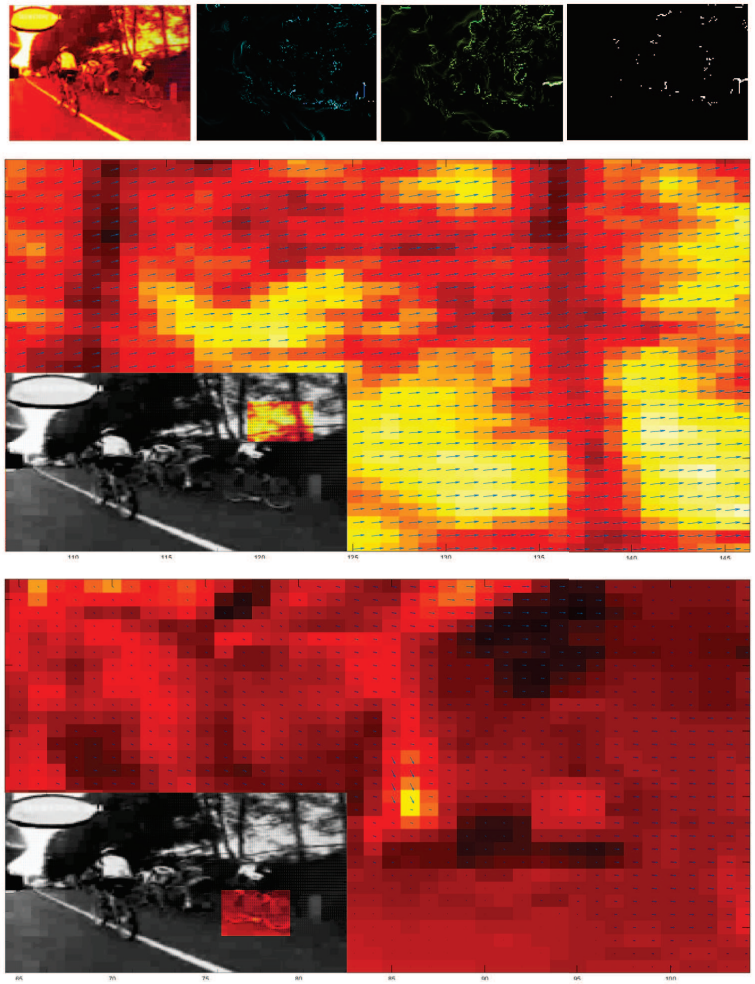

FIGURE 3.13 – Divergence, rotationnel et points critiques d'une vidéo issue de `UCF-50 Dataset`. (voir vidéo).

La vidéo associée à la figure 3.14 contient un enfant jouant à la balançoire. Le seul mouvement présent dans cette séquence est celui de la balançoire. Les points critiques détectés correspondent bien aux mouvements observés sur la séquence. On constate de fortes déformations du flot optique au niveau des bras ainsi que de la corde de la balançoire, caractérisant ainsi le mouvement de balancier.

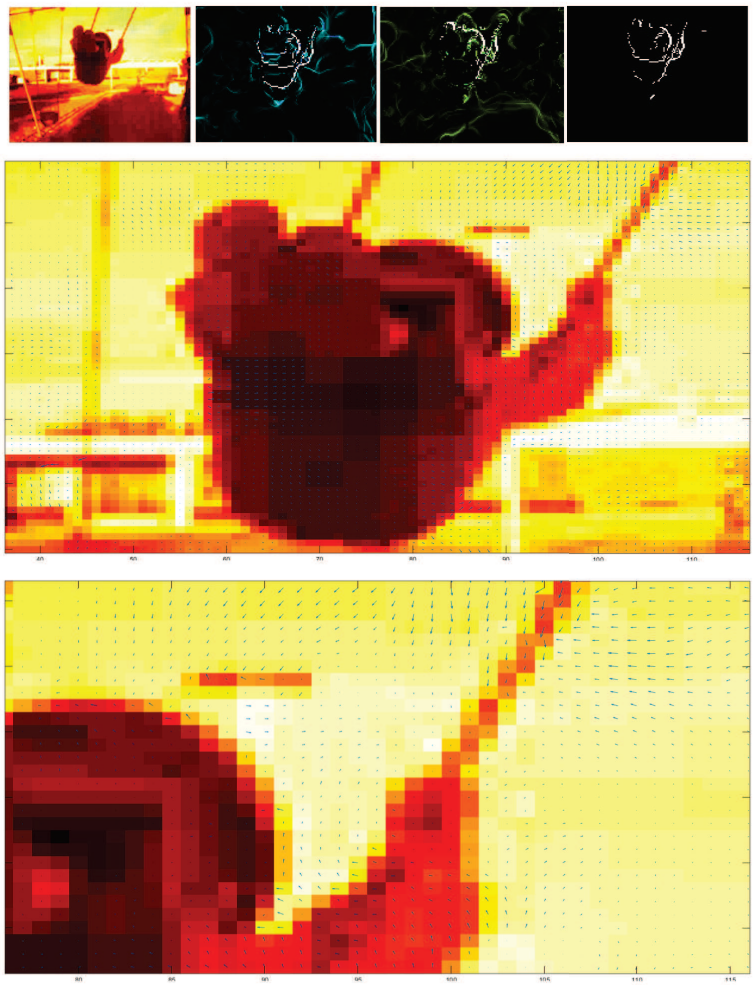

FIGURE 3.14 – Divergence, rotationnel et points critiques d'une vidéo issue de `UCF-11 Dataset`. (voir vidéo).

La figure 3.15 présente les points critiques détectés sur un sujet effectuant des mouvements techniques de boxe. Ces points critiques sont détectés tout le long du bras, au moment du repli de ce dernier. On remarque que les lignes de flux du champ vectoriel s'opposent au niveau des contours du sujet, ce qui crée une forte divergence dans ces régions. Les points critiques caractérisent naturellement ces situations.

FIGURE 3.15 – Divergence, rotationnel et points critiques d'une vidéo issue de `KTH Dataset`. (voir vidéo).

Les points critiques fournissent une caractérisation des mouvements humains dans des vidéos en s'appuyant sur des déformations complexes du champ vectoriel associé au flot optique. On constate que ces déformations sont relatives à des évènements temporels pertinents au sens du mouvement humain. L'étude des méthodes de reconnaissance d'actions humaines faite au chapitre 2 a montré l'importance d'éléments d'intérêt plus généraux tels que les trajectoires. Dans la partie suivante, nous verrons comment les points critiques estimés sont utilisés pour apporter une information temporelle plus riche.

3.1.3 Trajectoires de mouvements multi-échelles

Comme cela a été vu dans le Chapitre 2, la notion de trajectoires d'intérêt est fréquemment utilisée dans le domaine de la reconnaissance d'actions humaines. En effet, les trajectoires fournissent une information temporelle plus pertinente que les seuls points d'intérêt. Dans le cadre de la reconnaissance d'actions humaines, nous introduisons dans cette partie la notion de trajectoire de points critiques. Les points critiques estimés sont les points de départ de ces trajectoires. Elles décrivent donc les différentes positions de ces points au cours du temps et permettent d'avoir une meilleure caractérisation des évènements d'intérêt dans les vidéos.

3.1.3.1 Estimation des trajectoires de points critiques

Les points critiques sont suivis dans la séquence vidéo par un recalage de leur position en utilisant les composantes du flot optique situé dans le voisinage de ces points. Soit un flot optique $\mathbf{F}_t = (u_t, v_t)$. La position d'un point $P_t = (x_t, y_t)$ à l'image t est estimée à l'image suivante $t+1$ par le point $P_{t+1} = (x_{t+1}, y_{t+1})$ tel que :

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + Med_{F_t}(V_{(x_t,y_t)}) \tag{3.31}$$

Avec Med_F, un filtre médian appliqué spatialement au flot $\mathbf{F}_t$ en $V_{(x_t,y_t)}$ qui est un voisinage du point P_t. Le filtrage médian dans le voisinage des points suivis permet de maintenir la précision des contours situés dans le voisinage lors du suivi. La figure 3.16 illustre l'intérêt du filtrage médian pour le suivi de points situés sur des contours contrairement à une interpolation bilinéaire qui elle, mélange les informations du fond et des éléments d'intérêt et réduit donc la précision des contours.

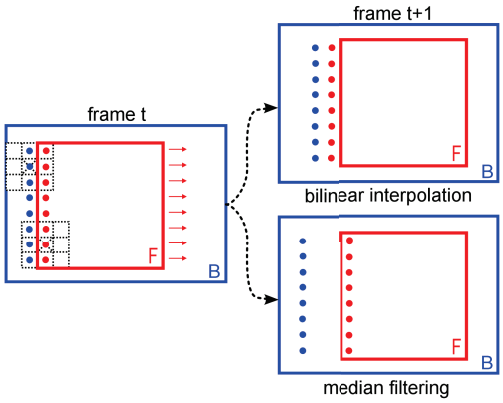

FIGURE 3.16 – Les points bleus appartiennent au fond, ceux en rouge correspondent à un élément en mouvement. Le suivi de points suivant un filtrage médian permet de conserver les informations de contours et permet donc une meilleure estimation de la position d'un objet à l'image suivante d'une séquence. (image tirée de [Wang et al., 2011]).

Cette approche d'estimation de trajectoires est utilisée par H.Wang *et al.* [Wang et al., 2011]. Elle a montré son efficacité par rapport à d'autres approches d'estimation de trajectoires telles

que les trajectoires de points d'intérêt SIFT [Sun et al., 2010b] ou encore les trajectoires issues de la méthode de [Lucas and Kanade, 1981]. Chaque point critique de la séquence est associé à sa trajectoire. La figure 3.17 illustre des trajectoires de points critiques estimées sur différentes séquences d'actions élémentaires. On distingue bien la cohérence entre les trajectoires obtenues et les mouvements présents dans ces séquences. En plus des points critiques qui fournissent une indication sur les régions spatio-temporelles comprenant de fortes déformations dues aux mouvements, les trajectoires de ces points critiques apportent une information complémentaire et plus riche sur les actions effectuées dans les séquences vidéos.

FIGURE 3.17 – Exemples de trajectoires de points critiques (voir vidéo).

3.1.3.2 Approche multi-échelle

Nous avons vu comment les points critiques permettent de caractériser la déformation du champ vectoriel associé au flot optique. Cependant, cette déformation peut être provoquée par différents types de mouvements. On peut obtenir des déformations dues à des mouvements rapides et localisés dans le temps ou des déformations lentes associées à des mouvements plus longs dans le temps. Les mouvements effectués par un sujet possèdent généralement plusieurs échelles spatio-temporelles caractéristiques. Les trajectoires issues de ce mouvement sont donc elles aussi associées à une ou plusieurs fréquences caractéristiques. Par exemple, une action de `course` peut être caractérisée par un mouvement lent de translation, mais aussi par des mouvements rapides de bras et de jambes, localisés dans le temps. Les trajectoires issues de ces différentes fréquences de mouvements sont aussi associées à différentes fréquences caractéristiques.

Afin d'analyser les différentes fréquences de mouvement pour les trajectoires extraites, une décomposition spatio-temporelle des séquences vidéos est réalisée. Une subdivision dyadique

spatio-temporelle est effectuée sur les séquences. Les sous-séquences obtenues sont filtrées par un noyau Gaussien spatio-temporel `G` afin de supprimer les éventuelles hautes fréquences. Le flot optique est ensuite estimé sur chacune de ces séquences. Chaque sous-séquence correspond à une échelle de subdivision.

— **La subdivision dyadique spatiale** permet d'estimer des points critiques de différentes échelles spatiales.
— **La subdivision dyadique temporelle** permet d'obtenir des trajectoires de même longueur mais pour des fréquences temporelles différentes. Ce point est détaillé par la suite.

Cette approche nous permet d'observer des mouvements de différentes fréquences et d'isoler les trajectoires associées à ces différentes fréquences. Les premières échelles dyadiques révèlent les mouvements les plus rapides : trajectoires courtes et de hautes fréquences, tandis que les échelles les plus basses mettent en évidence les mouvements les plus lents : trajectoires plus longues et de basses fréquences. Ceci permet une meilleure analyse et une meilleure caractérisation des trajectoires et des mouvements présents dans la séquence vidéo. La figure 3.18 présente notre processus de subdivision dyadique appliqué au séquence vidéo.

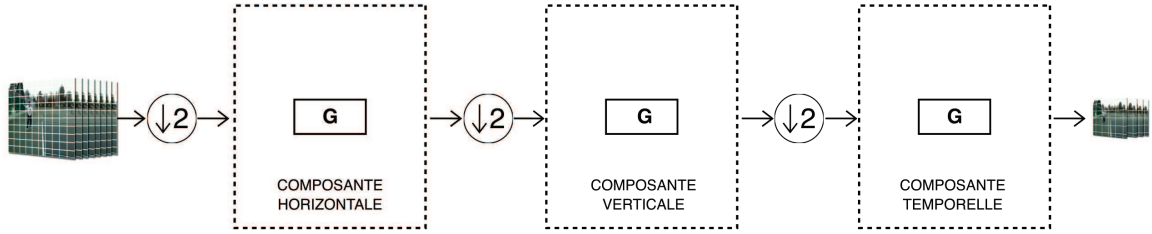

FIGURE 3.18 – Principe de notre approche multi-échelle. Une subdivision dyadique suivie d'un filtrage Gaussien `G` dans chaque composante sont effectués.

À chaque échelle de subdivision, des points critiques sont extraits. Les trajectoires de ces points critiques sont ensuite estimées. Les trajectoires obtenues sont ce que nous appelons par la suite des `trajectoires multi-échelles`. De part la subdivision dyadique, toutes les trajectoires obtenues sont de même longueur mais correspondent à différentes fréquences de mouvement. La figure 3.19 illustre des trajectoires multi-échelles de points critiques estimées sur différentes séquences vidéos d'actions élémentaires humaines. Les trajectoires rouges correspondent aux mouvements courts et rapides, les trajectoires bleues aux mouvements lents et de basses fréquences, les trajectoires vertes sont issues d'une échelle intermédiaire. On voit sur ces exemples que l'approche multi-échelle spatio-temporelle proposée permet une analyse robuste des différents mouvements en fonction des échelles de fréquence axuquelles ils sont exécutés. Par exemple, sur la deuxième illustration de la ligne du bas, on observe les trajectoires rouges au niveau des pieds du joueur et de la balle. Les trajectoires vertes correspondent au mouvement de la jambe, et les bleues correspondent au mouvement global du corps.

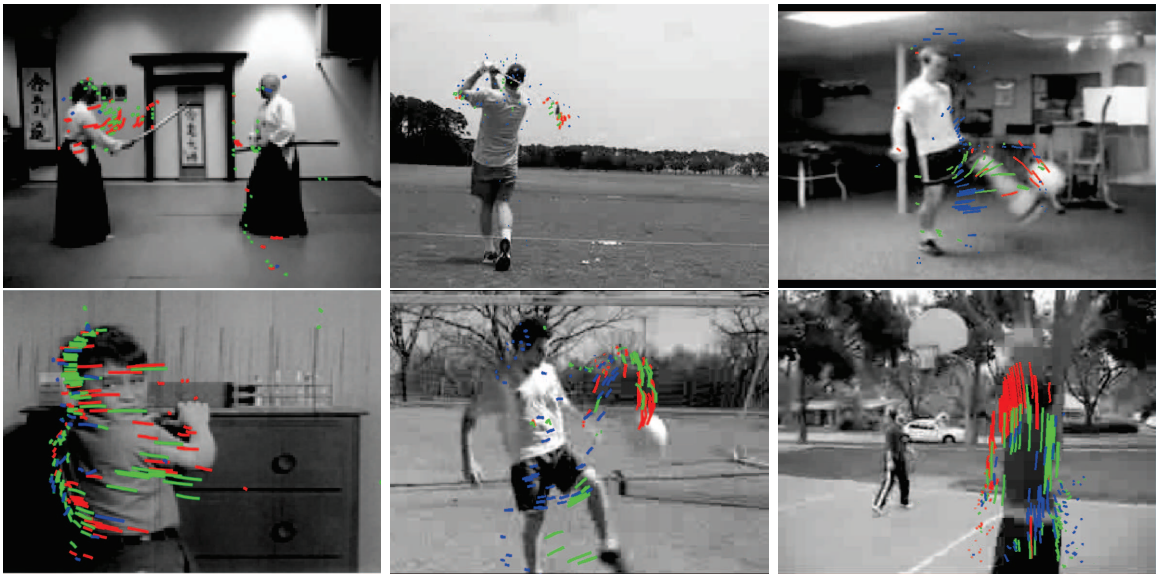

FIGURE 3.19 – Exemples de trajectoires multi-échelles de points critiques (voir vidéo).

Discussion Nous avons vu dans cette section la pertinence des points critiques en tant que points d'intérêt spatio-temporels. Ces derniers caractérisent les évènements temporels contenus dans les séquences vidéos en fonction des déformations du flot optique associées à ces séquences. Les trajectoires multi-échelles de ces points critiques apportent une description plus large des actions, en prenant en compte différentes fréquences de mouvement. Dans la section suivante, nous verrons quels descripteurs sont utilisés dans notre processus de reconnaissance d'actions pour tirer profit des informations obtenues.

3.2 Caractérisation du mouvement

Dans la section précédente, nous avons vu les différents éléments d'intérêt développés pour la reconnaissance d'actions élémentaires, notamment les points critiques du flot optique ainsi que les trajectoires multi-échelles. Dans cette section, nous verrons la deuxième étape du processus de reconnaissance illustrée par la figure 3.20. Elle consiste à caractériser localement les éléments d'intérêt afin d'en extraire des informations pertinentes et propres aux mouvements humains. La première sous-section traite de notre approche de compensation de mouvements de caméra dans le cadre de la reconnaissance d'actions élémentaires dans des vidéos génériques. Les autres sous-sections traitent des descripteurs utilisés pour tirer profit des informations apportées par nos éléments d'intérêt.

FIGURE 3.20 – La description d'éléments d'intérêt permet de capter les informations caractéristiques des mouvements humains.

3.2.1 Compensation du mouvement de caméra

Nous avons vu précédemment l'utilisation de trajectoires multi-échelles pour effectuer une analyse robuste des mouvements humains dans des séquences vidéos. La principale difficulté dans l'estimation des trajectoires est de garder une faible erreur d'estimation de position au cours du temps. Dans le cas de vidéos génériques, cela se révèle plus complexe du fait de nombreux mouvements de caméra ou de changements de point de vue. Ces différentes contraintes influent directement sur la qualité de l'estimation et donc sur la pertinence des trajectoires des mouvements étudiés.

Nous présentons ci-dessous quelques méthodes de reconnaissance d'actions humaines de la littérature proposant des approches de compensation de mouvements de caméra.

3.2.1.1 Méthodes de compensation de mouvements de caméra

La correction du mouvement de caméra est une problématique toujours active, particulièrement du fait de la mise à disposition récente de bases de données de vidéos réalistes. Ce mouvement ne peut pas toujours être totalement compensé. En effet, certains mouvements de caméra mènent à des changements de perspectives trop radicaux. Parmi les différentes stratégies de la littérature pour traiter cette problématique, nous donnons en exemple les travaux suivants :

— [Wang and Schmid, 2013] proposent une méthode d'estimation de trajectoires en détectant les humains présents dans une séquence vidéo. La méthode part du principe que deux images consécutives de la séquence sont liées par une homographie. Les paramètres de ces homographies sont estimés pour calculer la correspondance entre deux images

consécutives. Cette estimation se fait par une correspondance de points d'intérêt locaux de types SURF. Ce choix est motivé par la robustesse des points SURF au flou de bougé, généralement produit par les mouvements de caméra. Cette démarche permet d'obtenir un gain notable sur des bases de données d'actions humaines de référence. Les taux de reconnaissance de cette méthode sont présentés dans chapitre le 4. En contre partie, cette approche augmente significativement la complexité en termes de temps de calcul de par l'utilisation d'un processus *ad hoc* de détection automatique d'humains dans des images via un classifieur entrainé sur la base PASCAL VOC07 [Mottaghi et al., 2014] et l'utilisation de l'algorithme RANSAC [Fischler and Bolles, 1981] pour l'estimation des paramètres de l'homographie entre deux images consécutives.

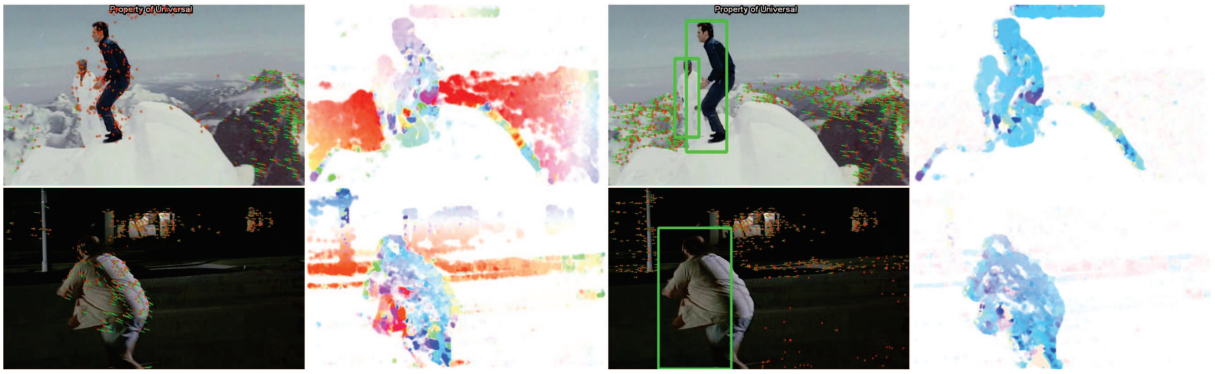

FIGURE 3.21 – Illustration de la méthode de compensation de mouvement de caméra proposée par Wang *et al.*. La détection de personne dans une séquence permet de mieux localiser les mouvements. (figure tirée de [Wang and Schmid, 2013]).

— [Jain et al., 2013] arguent que le mouvement d'une séquence vidéo peut être séparé en deux composantes : le mouvement dominant, dû à la caméra, et le mouvement résiduel, relatif aux actions présentes dans la séquence. Le mouvement dominant est extrait en utilisant l'estimation du flot affine entre deux images consécutives. La compensation est obtenue en soustrayant le flot affine au flot optique. Cette approche présente de bons résultats mais suppose l'estimation consécutive de deux champs vectoriels : le flot optique et le flot affine. La figure 3.22 illustre l'application de cette méthode sur deux images d'une séquence vidéo. On constate la diminution de l'amplitude des vecteurs liés aux pixels du fond et la conservation de ceux liés aux mouvements humains.

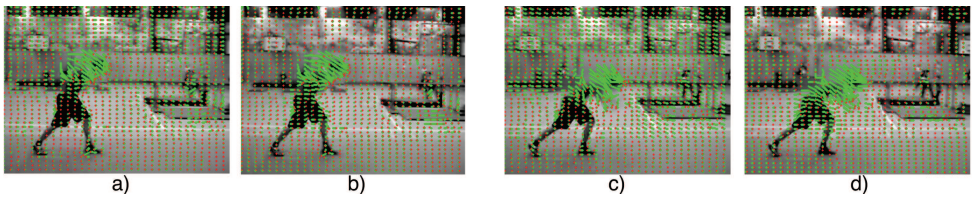

FIGURE 3.22 – Illustration de la méthode de compensation de mouvements de caméra proposée par Jain *et al.*. Les figures a) et c) représentent les vecteurs de déplacement des pixels d'une vidéo d'action humaine. Les figures b) et d) montrent l'application de la méthode de compensation sur le flot optique (figure tirée de [Jain et al., 2013]).

La méthode que nous proposons reprend le schéma global de la méthode de Jain *et al.* sans estimer le mouvement dominant par un champ vectoriel supplémentaire mais en se basant sur une estimation pyramidale du flot optique.

3.2.1.2 Approche de compensation proposée

L'objectif de cette approche proposée est de minimiser l'impact du mouvement de caméra avec un faible temps de calcul, tout en évitant des méthodes *ad-hoc* trop coûteuses. Le flot optique étant la base de notre approche, il est utilisé dans le but d'améliorer l'estimation des trajectoires, en compensant la plus grande partie des mouvements de caméra. La version pyramidale de l'estimation du flot optique, entre deux échelles est telle que :

$$\mathbf{F}_t^L = E_2(\mathbf{F}_t^{L+1}) + f([I_t^L + E_2(\mathbf{F}_t^{L+1})], I_{t+1}^L) \tag{3.32}$$

Avec L l'échelle de la pyramide, $\mathbf{F}_t^L = (u_t^L, v_t^L)$, E_2 un opérateur qui interpole une matrice de taille $k \times k$ en une matrice de taille $2k \times 2k$ et f le flot optique estimé entre deux images consécutives. La figure 3.23 illustre cette forme de représentation.

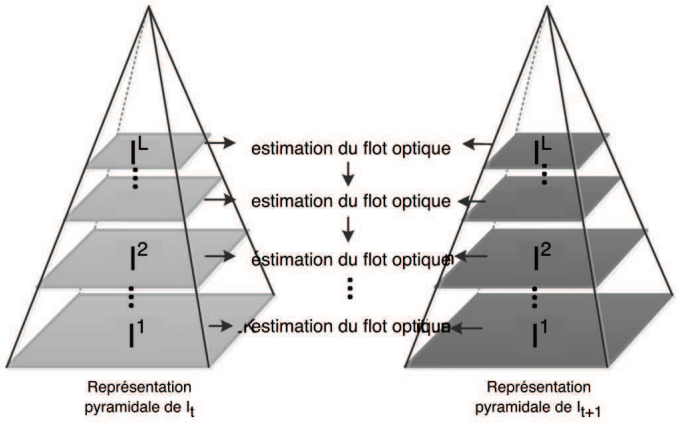

FIGURE 3.23 – Représentation de l'approche pyramidale de l'estimation du flot optique.

Le flot optique de la séquence est estimé et, dans le même temps, l'estimation obtenue à la dernière échelle de la pyramide est sauvegardée. À cette échelle, le champ vectoriel estimé correspond aux mouvements les plus larges et les plus présents globalement dans la séquence, souvent dûs aux mouvements de la caméra. Les mouvements rapides et plus localisés n'y sont pas inclus. Ce flot global est utilisé selon le même principe que celui du mouvement dominant dans [Jain et al., 2013]. Le mouvement de caméra est compensé directement durant le processus d'estimation du flot optique. Le temps de calcul reste donc le même. Nous obtenons donc :

$$\mathbf{F}_{comp}^0 = \mathbf{F}_{original}^0 - \mathbf{F}_{original}^N \tag{3.33}$$

Avec $\mathbf{F}_{original}^0$ l'estimation du flot optique de base entre deux images consécutives, $\mathbf{F}_{original}^N$ l'estimation du flot optique obtenu à la dernière échelle de la représentation pyramidale. Cette

composante représente le mouvement global de la caméra. La composante $\mathbf{F}^0_{comp}$ représente le flot optique entre les deux images consécutives avec la compensation du mouvement de la caméra. La figure 3.24 montre le résultat de notre approche sur une séquence vidéo comprenant un joueur de basket-ball.

La 1ère colonne présente la séquence qui contient un mouvement de translation de caméra. La 2ème colonne présente le flot optique original $\mathbf{F}^0_{original}$ de la séquence. La 3ème colonne comporte le flot issu de la dernière échelle $\mathbf{F}^N_{original}$ d'estimation du flot optique. La colonne 4 montre le résultat de la compensation $\mathbf{F}^0_{comp}$ sur la séquence originale. On remarque que les mouvements du joueur sont préservés tandis que le mouvement global de la caméra est compensé. En effet la direction ainsi que l'intensité de ce dernier son caractériser par la teinte bleu présente dans $\mathbf{F}^N_{original}$ et $\mathbf{F}^0_{original}$.

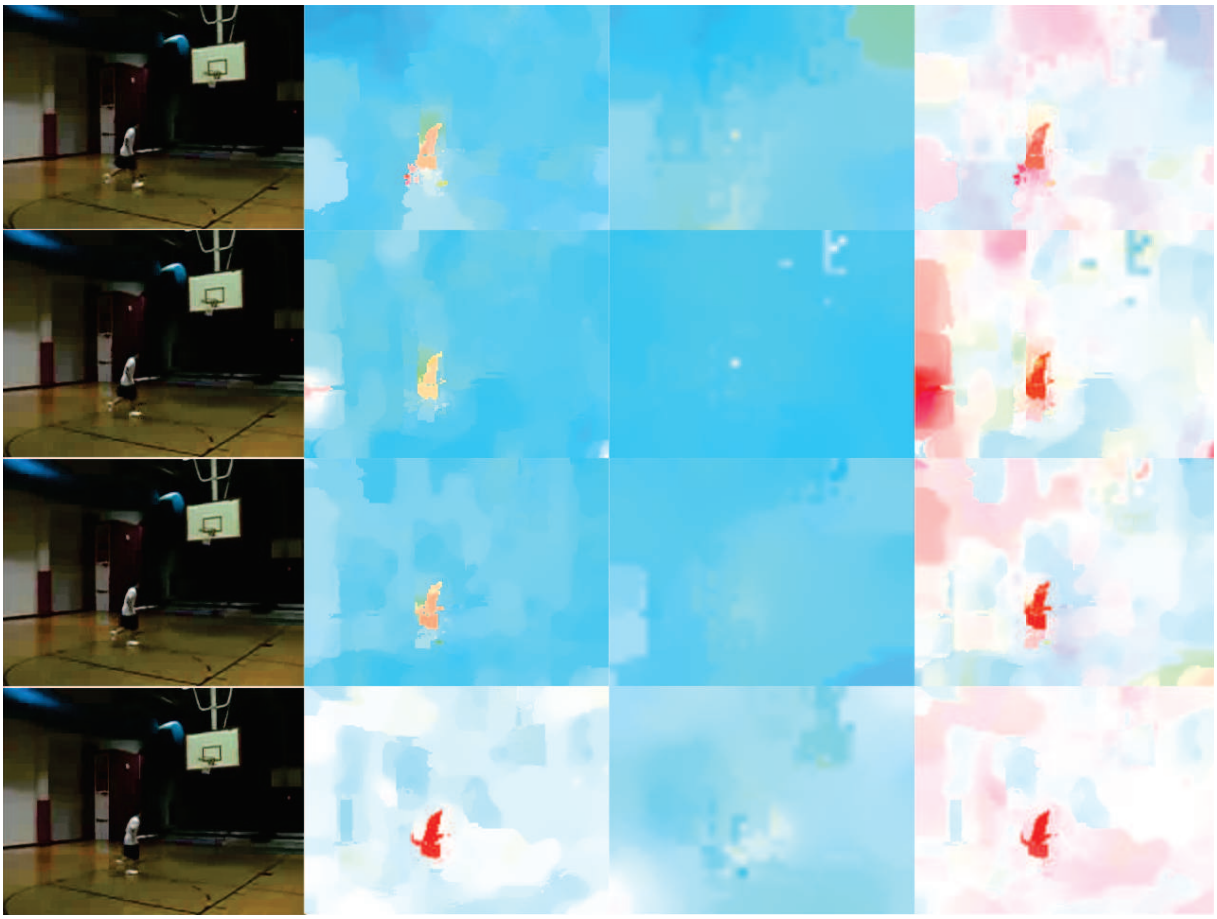

FIGURE 3.24 – 1ère colonne : quatre images consécutives avec un mouvement de caméra latéral sur les trois premières images, 2ème colonne : estimation du flot optique $\mathbf{F}^0_{original}$, 3ème colonne : estimation du mouvement global $\mathbf{F}^N_{original}$, 4ème colonne : compensation du mouvement de caméra. $\mathbf{F}^0_{comp}$ (voir vidéo).

L'avantage de notre approche est qu'elle permet une compensation du mouvement global de la caméra sans coût de calcul supplémentaire. La compensation se fait directement lors de l'estimation du flot optique, cela permet de proposer une approche avantageuse en termes de

taux de reconnaissance et en temps de calcul, ce qui est non négligeable sur des bases de données de grande taille ainsi que les applications en temps-réel.

3.2.2 Description fréquentielle des trajectoires de points critiques

Les trajectoires multi-échelles permettent de mettre en évidence différentes fréquences de mouvements dans les vidéos étudiées. La possibilité de caractériser ces fréquences est un avantage significatif pour la reconnaissance d'actions élémentaires. Dans ce cadre, il est naturel d'utiliser des outils de description fréquentiel afin de traiter l'information apportée par les trajectoires multi-échelles de façon adéquate.

La transformée de Fourier permet de caractériser les signaux non périodiques dans le domaine fréquentiel. Soit un signal discret $\mathbf{P} = [P_1, P_2, ..., P_t, ..., P_N]$. La transformée de Fourier discrète de $\mathbf{P}$ est définie comme suit :

$$X_k = \sum_{t=0}^{N-1} e^{\frac{-i2\pi kt}{N}} \cdot P_{t+1}, \quad k \in [\![0, N-1]\!] \tag{3.34}$$

avec le point $P_t = (x_t, y_t)$, N la longueur de la trajectoire et k la fréquence d'analyse. X_k est un nombre complexe représentant la phase et l'amplitude du signal projeté dans le domaine fréquentiel à la fréquence k. Le vecteur $\mathbf{X} = [X_0, X_1, ..., X_k, ..., X_{N-1}]$ contient les coefficients de la transformée de Fourier du signal $\mathbf{P}$. Les coefficients de Fourier décrivent donc le spectre fréquentiel du signal.

La transformée de Fourier et les coefficients qui en résultent sont très utilisés en traitement du signal, notamment dans la description de formes et de silhouettes [Zhang and Lu, 2002]. En effet, la caractérisation fréquentielle de signaux non-périodiques permet d'obtenir une représentation compacte de ces derniers. La forme globale des silhouettes étant représentée par les coefficients associés aux basses fréquences, tandis que les hautes fréquences caractérisent généralement les détails fins ainsi que le bruit.

Les trajectoires multi-échelles sont donc décrites à l'aide des coefficients de leur transformée de Fourier dans notre approche de reconnaissance d'actions élémentaires. Comme précédemment expliqué, le domaine fréquentiel fournit une représentation compacte et robuste des signaux mais aussi présente un avantage en termes d'invariance à différentes transformations.

Dans le cadre de la caractérisation de nos trajectoires, nous détaillons ci-dessous ces différentes invariances.

3.2.3 Invariance dans le domaine fréquentiel

Soit une trajectoire comportant N points séquentiels : $\mathbf{P} = [P_1, P_2, ..., P_t, ..., P_N]$ P_t étant un point quelconque de la trajectoire ayant comme position (x_t, y_t) à l'instant t. Dans la suite, on considère que la transformée de Fourier d'une trajectoire $\mathbf{P}$ est :
$\mathbf{X} = [X_0, X_1, ..., X_k, ..., X_{N-1}]$.

Invariance en translation L'invariance en translation est obtenue en soustrayant aux coordonnées (x_n, y_n) des points de la trajectoire $\mathbf{P}$ leur valeur moyenne $\tilde{x_n}$ et $\tilde{y_n}$ sur cette trajectoire :
$\tilde{x_n} = x_n - \sum_{t=1}^{N} \frac{x_t}{N}$ et $\tilde{y_n} = y_n - \sum_{t=1}^{N} \frac{y_t}{N}$

Invariance en rotation Afin d'obtenir une invariance en rotation, les trajectoires $\mathbf{P}_N$ sont traitées comme des vecteurs de nombres complexes et s'écrivent : $\mathbf{P}_i = [P_{i1}, P_{i2}, ..., P_{it}, ..., P_{iN}]$ $P_{it} = \tilde{x}_t + i\tilde{y}_t$ étant la représentation complexe du point P_t. Ainsi, pour une trajectoire $\mathbf{P}_{\theta i}$ représentant une rotation d'angle θ de la trajectoire initiale $\mathbf{P}_i$, la norme des coefficients de la transformée de Fourier de $\mathbf{P}_{\theta i}$ et celle des coefficients $\mathbf{P}_i$ sont égales. Il y a donc invariance par rapport à la rotation.

Invariance en échelle L'invariance par rapport à l'échelle est assurée par la normalisation de la transformée de Fourier en divisant ses coefficients par la première composante fréquentielle non nulle.
$\tilde{X_k} = \frac{X_k}{|X_0|}$, $k \in [\![0, N-1]\!]$

Finalement, le descripteur basé sur les coefficients de Fourier (`FCD`) est :

$$\texttt{FCD}_{[T_{iN}]} = [\left|\tilde{X_0}\right|, \left|\tilde{X_1}\right|, ..., \left|\tilde{X_k}\right|, ..., \left|\tilde{X_{N-1}}\right|], \quad k \in [\![0; N-1]\!] \tag{3.35}$$

Comme cela a été vu précédemment, la subdivision dyadique conduit à des trajectoires multi-échelles de même longueur N, le descripteur `FCD` est donc calculé sur les mêmes plages fréquentielles $\frac{k}{N}$ avec $k \in [\![0, N-1]\!]$.

La figure 3.25 illustre les différentes transformations précédemment citées sur une trajectoire quelconque. On constate que, pour chaque transformation géométrique, le spectre fréquentiel associé à la transformée de Fourier reste le même. Cela montre bien l'invariance naturelle à ces transformations, que l'on retrouve dans le domaine fréquentiel.

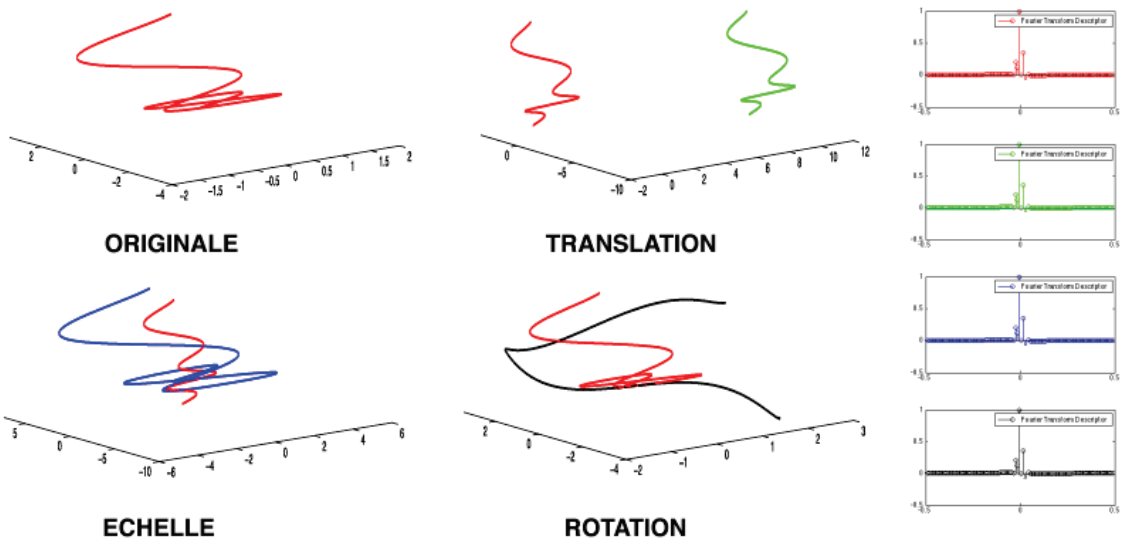

FIGURE 3.25 – Différentes transformations géométriques de la trajectoire originale (translation, échelle, rotation) aboutissant au même vecteur descripteur.

3.2.4 Lissage de trajectoire dans le domaine fréquentiel

Les trajectoires sont ensuite lissées en supprimant les coefficients de la transformée de Fourier correspondant aux très hautes fréquences, qui sont assimilées à du bruit ou des imprécisions de localisation. La figure 3.26 illustre cette étape. Ce traitement permet de rendre le descripteur robuste aux petites perturbations de mouvement.

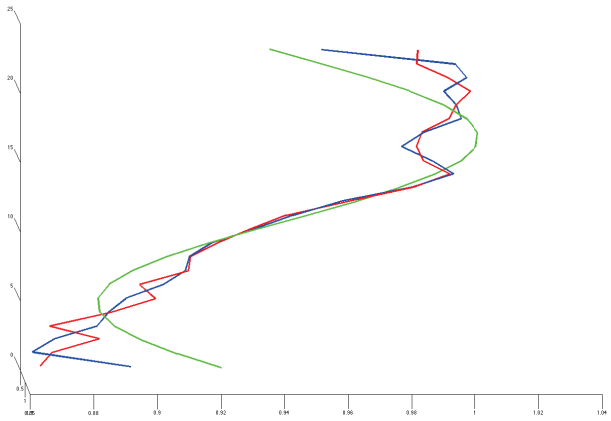

FIGURE 3.26 – Trajectoire originale (en rouge), trajectoire lissée en supprimant 50% des coefficients (en bleu), trajectoire lissée en supprimant 80% des coefficients (en vert).

3.2.5 Caractérisation des variations de formes et d'orientation du mouvement.

Les trajectoires multi-échelles sont caractérisées à partir de leurs coefficients de Fourier afin de prendre en compte l'information fréquentielle qu'elles apportent sur les mouvements étudiés. Les points critiques, qui fournissent également une information pertinente sur la localisation des évènements temporels d'intérêt, sont également caractérisés dans notre approche. Nous utilisons les descripteurs de contours `HOG` (Histogram of Orientation of 2D Gradient) et d'orientation de mouvement `HOF` (Histogram of Orientation of optical Flow) [Laptev et al., 2008] pour caractériser le voisinage spatio-temporel des points critiques détectés.

— Le descripteur `HOG` décrit l'évolution du gradient 2D dans le voisinage d'un point critique. Il permet d'encoder l'information de contour et de forme des mouvements présents dans la séquence.
— Le descripteur `HOF` caractérise l'information d'orientation du flot optique autour d'un point critique. Il décrit la variation locale de l'orientation du flot optique, qui est une information très caractéristique du mouvement.

Le descripteur `HOG` est basé sur les variations locales du gradient, le descripteur `HOF` est lié à l'estimation du flot otique et le descripteur `FCD` caractérise les différentes fréquences de mouvement présentes au cours du temps. Ces trois informations sont des éléments caractéristiques des mouvements humains dans des vidéos. Les expérimentations présentes dans le Chapitre 4 montrent la complémentarité de ces trois informations et leur pertinence dans le processus de classification. En effet, on constate, de part leur nature et leur construction, que la variation des contours, l'orientation du mouvement ainsi que l'information fréquentielle, sont des informations peu corrélées.

Nous avons vu dans cette section les différents descripteurs utilisés pour notre méthode de reconnaissance d'actions élémentaires. Ces descripteurs prennent en compte les différentes caractéristiques des éléments d'intérêt extraits. Nous utilisons un descripteur de contour et d'orientation du flot optique pour décrire les variations de gradients ainsi que l'orientation des mouvements dans le voisinage des points critiques. Les trajectoires multi-échelles, qui caractérisent les différentes fréquences de mouvement au cours du temps sont décrites à l'aide des coefficients de transformée de Fourier afin d'exploiter au mieux cette information fréquentielle. Dans la section suivante, nous verrons comment ces descripteurs sont utilisés et combinés dans le processus de reconnaissance d'actions élémentaires.

3.3 Étape de classification

Cette section développe l'étape de classification de notre méthode de reconnaissance d'actions. L'approche dite de "sac de mots visuels" [Lazebnik et al., 2006] est d'abord présentée. Cette approche permet de représenter de façon compacte les descripteurs de forme, d'orientation de mouvement et de fréquence d'une séquence vidéo. L'utilisation d'un classifieur de type `SVM` ainsi qu'une méthode de boosting pour la combinaison pertinente des informations sont ensuite détaillées.

3.3.1 Représentation par sac de mots visuels

FIGURE 3.27 – La représentation par sac de mots visuels permet de quantifier les descripteurs calculés à l'étape précédente.

Définition Afin d'évaluer les performances de notre méthode pour la reconnaissance, l'approche dite de "sac de mots" [Lazebnik et al., 2006] est utilisée. Cette approche est initialement issue du domaine de la recherche de documents. Elle permet de décrire un document à l'aide d'un dictionnaire de mots généralement plus compact. Un document est représenté par un histogramme d'occurence de mots de ce dictionnaire. Cette approche a également montré son efficacité dans le domaine de la recherche et de la classification d'images. Dans ce cadre, les images sont représentées par un dictionnaire de "mots visuels". Ces mots visuels sont constitués à partir d'un ensemble de caractéristiques extraites d'une base de données d'images. Ces caractéristiques peuvent être issues de descripteurs de formes, de couleurs, de textures, etc. Comme pour les documents, une image est ensuite représentée par un histogramme d'occurence des mots visuels contenus dans cette dernière. Nous utilisons cette approche pour représenter les vidéos d'actions humaines à partir des caractéristiques présentées dans la section précédente.
Cette représentation suit plusieurs étapes :

1. **Extraction de caractéristiques** Les points critiques et les trajectoires multi-échelles sont estimés sur un ensemble d'apprentissage de vidéos séquences, puis caractérisés à l'aide de descripteurs de forme, d'orientation de mouvements et de fréquence.
2. **Apprentissage du dictionnaire de mots visuels** Le dictionnaire est utilisé pour quantifier un ensemble de N descripteurs $(\mathbf{v}_1, ..., \mathbf{v}_N)$. La construction de ce dictionnaire se fait en partitionant l'ensemble des vecteurs descripteurs calculés sur la base de données d'apprentissage. Ce partitionnement se fait en utilisant un processus non-supervisé tel que l'algorithme des `K-moyennes`. Les centres $(\mathbf{V}_1, ..., \mathbf{V}_K)$ obtenus forment les "mots visuels" du dictionnaire.
3. **Quantification des caractéristiques à partir du dictionnaire**. Un vecteur descripteur $\mathbf{v}_n$ est ensuite associé à son "mot visuel" le plus proche $\mathbf{V}_k$ au sens de la distance euclidienne.
4. **Représentation des données par histogramme d'occurence de mots visuels du dictionnaire** Une séquence vidéo est finalement représentée par un histrogramme regroupant les occurences des mots visuels contenus dans cette séquence vidéo.

La figure 3.28 présente de façon intuitive le fonctionnement de la représentation par sac de mots visuels. La ligne du haut présente le processus d'apprentissage du sac de mots visuels. La ligne du bas montre comment une vidéo requête est caractérisée par les occurrences de mots visuels qui la composent.

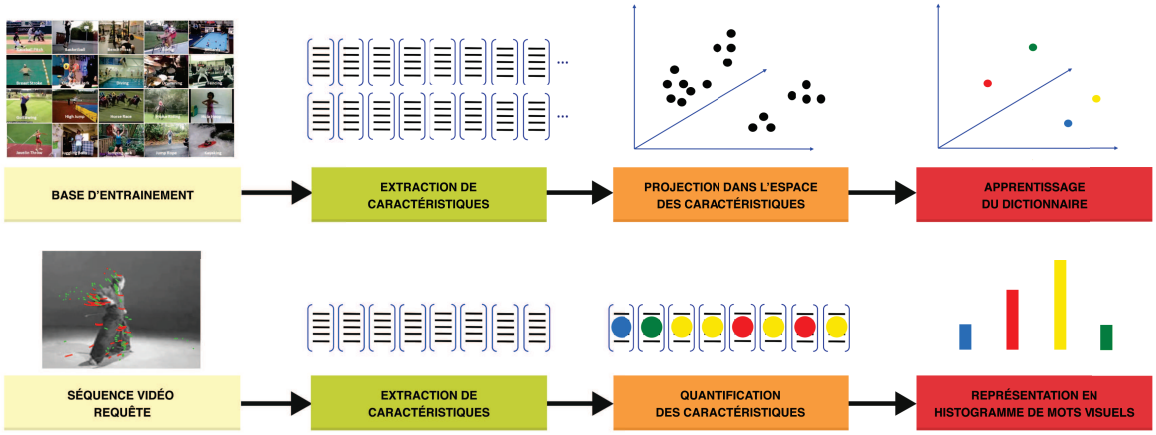

FIGURE 3.28 – Représentation intuitive de l'approche du sac de mots visuels.

Sac de mots multi-canaux L'approche du sac de mots visuels est pertinente lorsqu'on utilise des descripteurs locaux, ce qui est le cas de notre méthode. En effet, la représentation en histogramme de mots visuels permet de s'affranchir des contraintes spatiales et temporelles, ce qui est relativement efficace par rapport aux occultations partielles et certains changements de points de vue. Cependant la relation spatiale et temporelle entre les caractéristiques extraites donne également une information sur l'action observée. Pour récupérer une structure spatiale et temporelle lors de la représentation en histogramme, une version dite "multi-canaux" du sac de mots visuels est utilisée ([Laptev et al., 2008, Wang et al., 2011]). Cette extension du sac de mots visuels consiste à subdiviser une vidéo selon une certaine structure. Un histogramme de mots visuels est calculé sur chaque cellule de cette structure. L'histogramme global de la vidéo est la concaténation des histogrammes de chacune de ses cellules. La subdivision de la vidéo suivant une structure particulière est appelée un canal. La figure 3.29 montre quelques exemples de canaux ainsi que l'agencement de leurs cellules respectives.

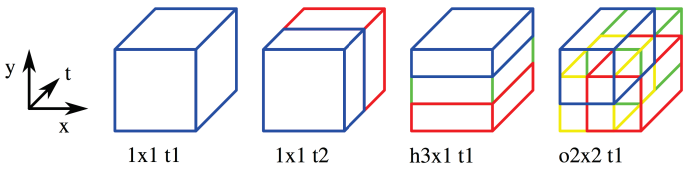

FIGURE 3.29 – Exemples de canaux spatio-temporels, avec leurs cellules respectives représentées par différentes couleurs.

La structure `1x1 t1` correspond à la représentation standard du sac de mot visuel. La structure `1x1 t2` est une subdivision temporelle en deux cellules, tandis que la structure `o2x2 t1` a une subdivision spatiale horizontale et verticale avec une zone de recouvrement centrale.
La version spatio-temporelle du sac de mots utilise différents canaux de ce type afin de combiner un plus grand nombre d'informations locales. Dans nos expérimentations nous utilisons les

canaux `1x1 t1` et `2x1 t1` qui ont montré de bon résultats en termes de taux de reconnaissance.

3.3.2 Classification supervisée par SVM

La représentation en sac de mots visuels permet d'obtenir pour chaque vidéo traitée une représentation standardisée et compacte de ces vidéos sous forme d'histogramme. L'étape de classification montre comment est construit un modèle discriminatif afin de distinguer des actions de différentes classes entres elles et ainsi attester de la pertinence des éléments d'intérêts et des caractéristiques utilisées pour la construction de ce modèle. Cette étape est la dernière de notre méthode de reconnaissance d'actions humaines comme le montre la figure 3.30.

FIGURE 3.30 – L'étape de classification permet d'évaluer la similarité entre deux vidéo séquences contenant une action en fonction de la pertinence des éléments d'intérêts utilisés.

Classifieur SVM Le modèle discriminatif employé pour l'étape de classification est un modèle de type `SVM` (*Support Vector Machine*) [Chang and Lin, 2011]. Soit un ensemble d'apprentissage composé des éléments $\{(\mathbf{x}_1, l_1), (\mathbf{x}_2, l_2), ..., (\mathbf{x}_p, l_p)\}$ avec $(\mathbf{x}_k, l_k) \in \mathbb{R}^N \times \{-1, 1\}$, l_k étant la classe de l'éléments $\mathbf{x}_k$. La classification est ramenée ici à un problème de discrimination linéairement séparable, où l'on cherche à trouver un hyperplan séparateur entre deux classes de données, associé à une fonction de décision linéaire. Cette fonction de décision h est telle que :

$$h(x) = \mathbf{w}^T\mathbf{x} + w_0 \tag{3.36}$$

Avec $\mathbf{w}$ un vecteur de poids. Il peut y avoir une infinité d'hyperplans satisfaisant la contrainte de séparation des deux classes de données. Le classifieur `SVM` cherche à trouver l'hyperplan séparateur dont la distance au plus proche exemple de deux classes est maximum. Cet hyperplan définit la `marge maximale` entre deux classes de données. Il vérifie donc les contraintes :

$$\arg\max_{w,w_0} \quad \min_k \{|| \, x - x_k \, ||: x \in \mathbb{R}^N, \mathbf{w}^T\mathbf{x} + w_0 = 0\} \tag{3.37}$$

$$l_k(\mathbf{w}^T\mathbf{x}_k + w_0) \geqslant 0 \tag{3.38}$$

L'intérêt de l'hyperplan à vaste marge est qu'il s'appuie sur les points les plus robustes aux variations de l'ensemble des points. La figure 3.31 montre deux exemples de plans séparateurs entre deux classes de données. Celui de gauche est choisi arbitrairement et celui de droite maximise la marge entre les deux classes.

FIGURE 3.31 – Exemple de plans séparateurs entre deux classes de données. L'exemple de droite montre un plan séparateur quelconque des deux classes de données. L'exemple de droite illustre un plan séparateur à vaste marge calculé à partir du modèle `SVM`.

Méthode à noyaux Dans le cas général, la frontière séparant les données d'apprentissage n'est pas linéaire. Dans ce cas, la prise en compte de cette non linéarité est effectuée par l'introduction de noyaux non linéaires. Un noyau correspond à une transformation non linéaire Ψ qui projette les données initiales du problème dans un espace de dimension supérieure. Dans ce nouvel espace il est plus probable que le problème initial soit linéairement séparable. L'hyperplan séparateur est donc défini comme :

$$h(x) = \mathbf{w}^T \Psi(\mathbf{x}) + w_0 \tag{3.39}$$

L'intérêt de la méthode des noyaux est que l'on peut remplacer la transformation Ψ par la fonction κ qui lui est associée telle que :

$$\kappa(x_i, x_j) = \Psi(x_i)^T \cdot \Psi(x_j) \tag{3.40}$$

κ définit donc un produit scalaire sur $\Psi(\mathbb{R}^N)$. Cela permet de ne pas effectuer de calcul dans un espace de dimension supérieur. En effet, la transformation Ψ n'a pas besoin d'être connue, seul sa fonction noyau κ est utilisée. κ est généralement représentée sous forme d'une matrice de Gram $\kappa_{ij} = \kappa(x_i, x_j)$.
La figure 3.32 montre l'intérêt de la méthode des noyaux dans un cas non linéairement séparable. On constate que la projection des données dans un espace plus grand suivant la transformation $\Psi : (x_1, x_2) \mapsto (x_1, x_1, x_1^2 + x_2^2)$ permet une séparation linéaire du problème.

En prenant en compte ces deux formulations du problème (Équations 3.38 et 3.39) nous utilisons deux types de noyaux pour la classification des vidéos d'actions humaines.

Noyau RBF multi-canaux Dans le cadre standard de classification, nous utilisons un classifieur muni d'un noyau Gaussien dit `RBF`. Pour prendre en compte les différents canaux du sac de mots visuels représentant une vidéo séquence, ce noyau κ_{RBF} est défini comme suit :

$$\kappa_{RBF}(x_i, x_j) = exp(-\sum_{c \in C} \frac{1}{A_c} D_{\chi^2}(H_i^c, H_j^c)) \tag{3.41}$$

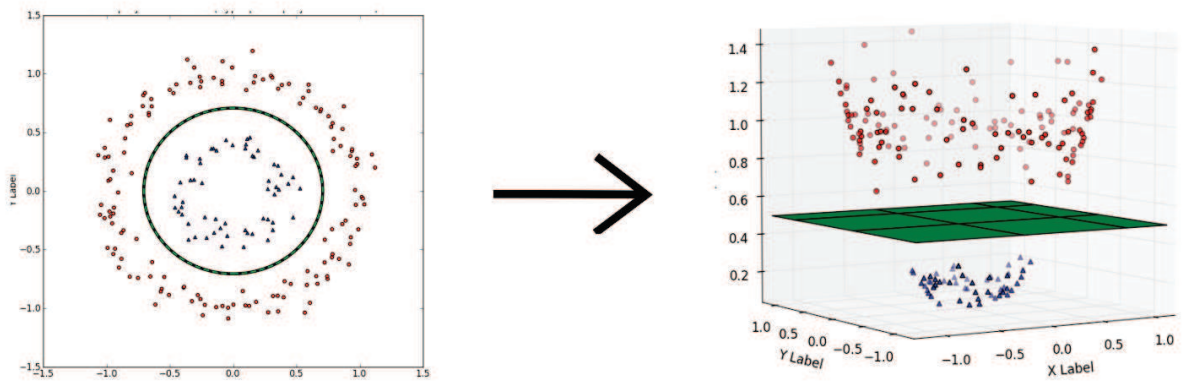

FIGURE 3.32 – La projection des éléments dans un espace plus grand suivant un noyau donné permet de trouver un hyperplan qui sépare linéairement ces éléments.

Où H_i^c et H_j^c sont respectivement les histogrammes des vidéos x_i et x_j relatifs au canal c comme défini plus haut. D_{χ^2} est la distance du χ^2 et A_c un coefficient de normalisation [Zhang et al., 2006]. L'ensemble C correspond à l'ensemble des canaux utilisés. Ce noyau agrège les histogrammes des différents canaux en fonction de leur pertinence. Il permet ainsi d'établir une distance entre des vidéos représentées par plusieurs histogrammes de différents canaux.

Noyau linéaire multi-canaux L'approche par sac de mots visuels assure une représentation parcimonieuse des séquences vidéos. Cependant, quand la dimension des histogrammes représentant une vidéo devient très grande, plonger les éléments d'apprentissage dans un espace de dimension plus grand n'améliore pas significativement les résultats. En effet, il est démontré que dans le cadre de classification de données creuses de très grandes dimensions ou d'un grand nombres d'instances, l'utilisation d'un noyau linéaire se révèle être plus efficiente qu'un noyau non-linéaire [Fan et al., 2008]. On utilise donc dans le cadre de classification de grandes bases de données ($instances > 3000$) un noyau linéaire κ_{Linear} tel que :

$$\kappa_{Linear}(x_i, x_j) = (H_i^C)^T H_j^C \tag{3.42}$$

Avec H_i^C et H_j^C qui sont respectivement les histogrammes résultants de la concaténation de tous les histogrammes de l'ensemble des canaux C. L' utilisation de plusieurs canaux est donc préservée dans ce cadre.

3.3.3 Fusion de caractéristiques par boosting

L'étape de classification suivant un modèle SVM permet de mettre en évidence l'efficacité des éléménts d'intérêts estimés. Cependant, pour prendre en compte de façon optimale la pertinence de chaque composante (forme, orientation du mouvement, fréquence) un algorithme Adaboost est utilisé [Hastie et al., 2009]. Pour se faire, un sac de mots visuels multi-canaux est appris pour chaque descripteur, puis, un classifieur SVM multi-classe est entrainé sur chaque sac de mots visuels. La figure 3.33 récapitule le processus mis en place pour une fusion de caractéristiques.

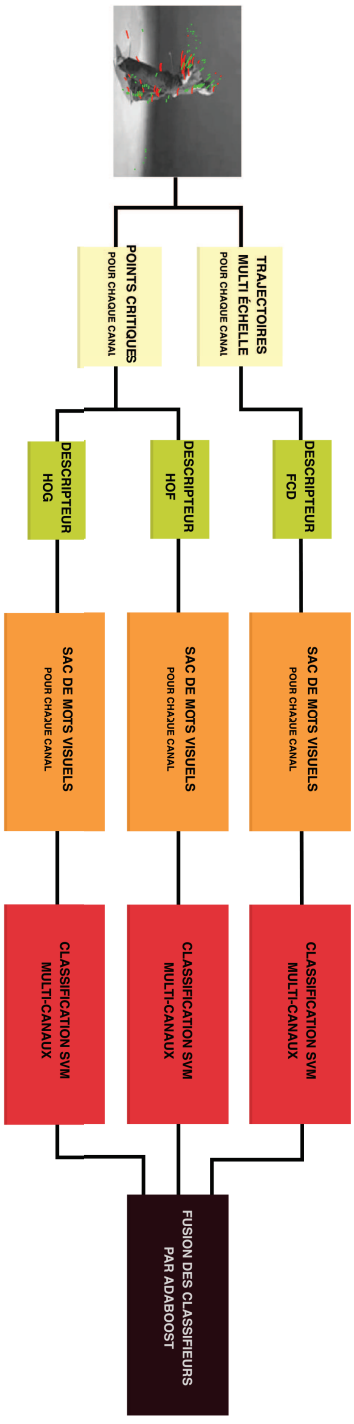

FIGURE 3.33 – Processus global de notre méthode de reconnaissance d'actions élémentaires.

L'algorithme Adaboost multi-classes est le suivant :

Algorithme 1 Fusion des caractéristiques par Adaboost multi-classe

Initialiser les poids des observations w_i tel que $w_i = 1/n$, $i = (1, .., n)$
pour m=1 **à** M **faire**
- Un classifieur $\mathcal{T}^{(m)}$ est entrainé sur chaque donnée pondérée par w_i
- Calcul de l'erreur pondérée

$$err^m = \sum_{i=1}^{n} w_i \cdot (c_i \neq \mathcal{T}^{(m)}(\mathbf{x}_i)) \backslash \sum_{i=1}^{n} w_i$$

- Calcul du poids associé au classifieur $\mathcal{T}^{(m)}$

$$\alpha^{(m)} = \log \frac{1 - err^m}{err^m} + \log(K - 1)$$

- Mise à jour des pondérations des données

$$w_i \leftarrow w_i \cdot exp(\alpha^{(m)} \cdot (c_i \neq \mathcal{T}^{(m)}(\mathbf{x}_i)))$$

- Renormalisation de w_i

fin pour
- Combinaison linéaire des classifieurs

$$C(\mathbf{x}) = \arg\max_k \sum_{m=1}^{n} \alpha^{(m)} \cdot \Pr\left(\mathcal{T}^{(m)}(\mathbf{x}) = k\right).$$

En considérant l'erreur pondérée sur chaque classifieur, l'algorithme Adaboost fournit les poids pour une combinaison linéaire efficace des classifieurs entrainés sur nos trois caractéristiques. Adaboost est dans notre approche utilisé durant le processus de classification `SVM`. Lors de la validation-croisée *Leave-One-Out* [Chang and Lin, 2011] l'erreur est calculée pour chaque observation, le poids associé à chaque classifieur est donc obtenu à la fin du processus de la validation-croisée. L'approche de boosting n'engendre donc pas de temps de calcul supplémentaire. Les pondérations sont ensuite stockées pour être réutilisées lors de la classification d'une vidéo requête.

Cette méthode permet à la fois d'obtenir une information sur la pertinence de chacune des trois composantes, et assure que ces dernières soit fusionnées de façon adéquate. La littérature a montré l'efficacité de ce procédé de fusion dans le domaine de la reconnaissance d'actions humaines [Peng et al., 2014]. Les résultats obtenus, et présentés dans le chapitre 4, montrent la complémentarité des caractéristiques mises en commun ainsi que le gain en termes de taux de bonne reconnaissance obtenu avec cette approche.

Conclusion du chapitre Ce chapitre a présenté des méthodes d'extraction d'éléments d'intérêts pertinentes au sens du mouvement. Nous avons vu par la suite le processus complet de la méthode de reconnaissance d'action que nous proposons. Cette dernière fait intervenir des caractéristiques propres aux mouvements humains, à savoir l'orientation du mouvement, la forme, ainsi que la fréquence. Nous avons vu comment ces caractéristiques sont combinées de façon à prendre en compte l'apport de chacune dans le processus de reconnaissance.

Dans le prochain chapitre nous présentons différentes évaluations menées sur des bases de données d'actions humaines de la littérature. Les résultats y sont présentés ainsi que l'influence des différents paramètres de notre méthode sur les taux de reconnaissance obtenus.

Chapitre 4

Évaluation : méthode de reconnaissance d'actions élémentaires

Chapitre 4

Évaluation : méthode de reconnaissance d'actions élémentaires

Sommaire

Dans le chapitre précédent, nous avons introduit notre méthode de reconnaissance d'actions élémentaires dans des vidéos. Dans ce chapitre, nous présentons les expérimentations menées afin d'évaluer les performances de cette méthode.

Dans la première partie, nous présentons les taux de bonne reconnaissance obtenus avec cette méthode sur des bases de données d'actions humaines, ainsi qu'une comparaison avec d'autres approches de la littérature.

Dans la seconde partie, nous évaluons l'influence des paramètres de notre méthode sur le taux de reconnaissance obtenu sur une sélection de bases de données. L'influence de l'approche multi-échelle, ainsi que celle de la méthode de compensation de mouvement de caméra sont également évaluées.

Dans la dernière partie, nous introduisons une méthode originale permettant d'évaluer la capacité de généralisation de notre approche par apprentissage-croisé de bases de données.

4.1 Expérimentations sur des bases de données vidéos

4.1.1 Introduction

Bases de données utilisées Dans cette section, nous évaluons la méthode précédemment présentée sur différentes bases de données de la littérature. Nous utilisons deux bases de données vidéos avec de fortes contraintes d'acquisitions (`KTH Dataset` et `Weizmann Dataset`) ainsi que deux bases de données composées de vidéos génériques (`UCF-11 Dataset` et `UCF-50 Dataset`). Pour chacune de ces bases nous présentons :

— les taux de reconnaissance pour chaque classe d'actions avec la combinaison des descripteurs `FCD, HOG, HOF`.

— le taux de reconnaissance par descripteurs (`FCD, HOG, HOF`), ainsi que le taux de reconnaissance après combinaison par Adaboost de ces descripteurs suivant une validation croisée *Leave-One-Out* à l'aide d'un classifieur SVM.

— une comparaison avec des taux de reconnaissance de la littérature sur la base considérée.

Flot optique utilisé Le flot optique utilisé sur ces trois premières bases de données est celui proposé par [Sun et al., 2010a]. Ce dernier se base sur le modèle de [Horn and Schunck, 1981] en apportant des améliorations à différentes étapes (pré-traitements des images, recalage, etc.). L'amélioration la plus notable étant l'utilisation de filtre médian, à chaque itération, durant le processus d'affinement du vecteur de déplacement $\mathbf{v}$.

La figure 4.1 montre le flot optique obtenu avec cette méthode comparativement à celles de [Lucas and Kanade, 1981] et [Horn and Schunck, 1981] sur une séquence issue de la base de données `Middlebury` . On remarque la précision de l'estimation au niveau des contours par rapport aux autres approches. Le résultat obtenu par cette méthode est visuellement le plus proche de la vérité terrain.

FIGURE 4.1 – a) vérité terrain b) flot optique [Lucas and Kanade, 1981] c) flot optique [Horn and Schunck, 1981] d) flot optique de Sun *et al.* [Sun et al., 2010a]. (utilisation du code fourni par Middlebury College [Baker et al., 2007]

Pour la base `UCF-50`, le flot optique `Deepflow` proposé par [Weinzaepfel et al., 2013]est utilisé. Cette méthode d'estimation du flot optique figure parmi les plus performantes sur des bases de données d'évaluation telles que la base `Sintel` [Wulff et al., 2012]. L'utilisation de ce flot optique sur cette base de données est justifiée par le nombre de vidéos à traiter (6680 vidéos). En effet, le flot optique `Deepflow` possède des temps d'exécution relativement faibles et est fourni avec une implémentation C++, ce qui permet d'obtenir des estimations de flot optique sur des séquences vidéos trois fois plus rapidement qu'avec la méthode de Sun *et al.*. Cependant, la méthode de Sun *et al.* est préférée pour la conservation des structures de l'images dans l'estimation du mouvement. La figure 4.2 illustre le résultat obtenu sur des exemples de la base `Sintel` avec la méthode de [Lucas and Kanade, 1981], [Sun et al., 2010a] et [Weinzaepfel et al., 2013].

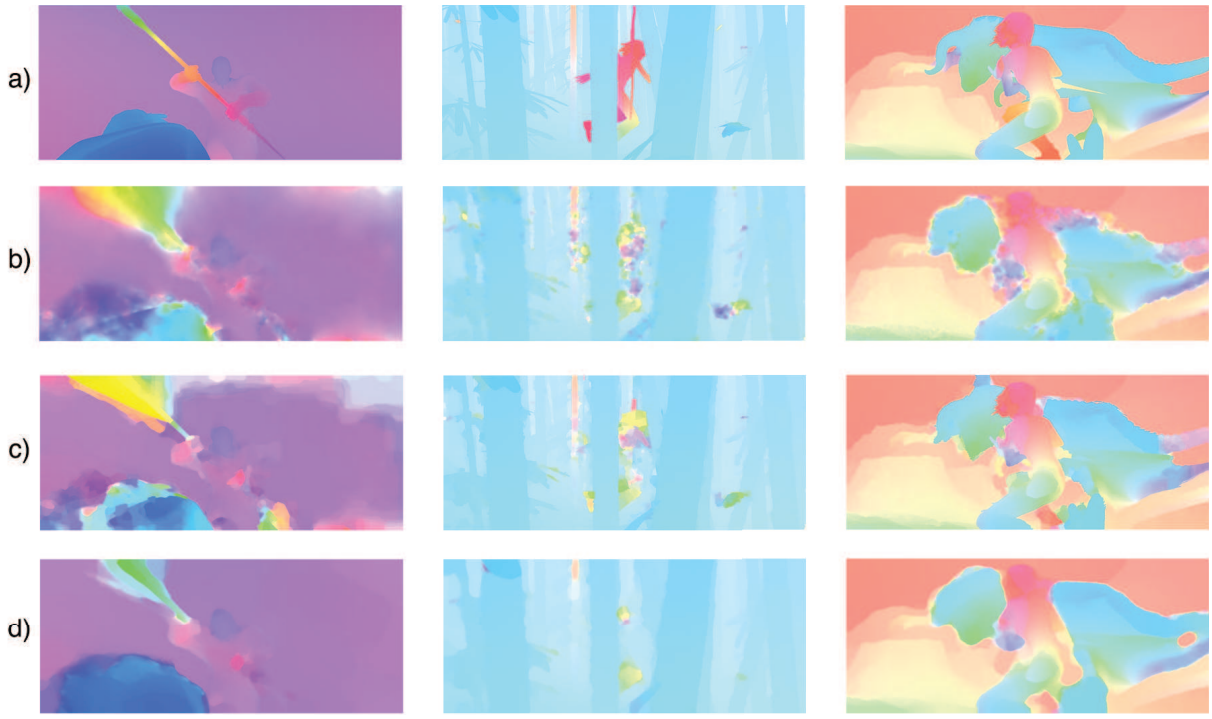

FIGURE 4.2 – Comparaison des flots optiques sur trois exemples de la bases de données Sintel a) vérité terrain b)flot optique [Horn and Schunck, 1981] c) flot optique de Sun *et al.* [Sun et al., 2010a]. d) flot optique de [Weinzaepfel et al., 2013].

4.1.2 Évaluation sur des bases de données de la littérature

4.1.2.1 Résultats sur la base de données KTH

Les résultats de classification obtenus avec notre méthode sur la base de données `KTH` sont présentés sur la figure 4.3 et les tableaux 4.1 et 4.2. Les résultats sont issus d'une validation croisée *Leave-One-Out* à l'aide d'un classifieur SVM.

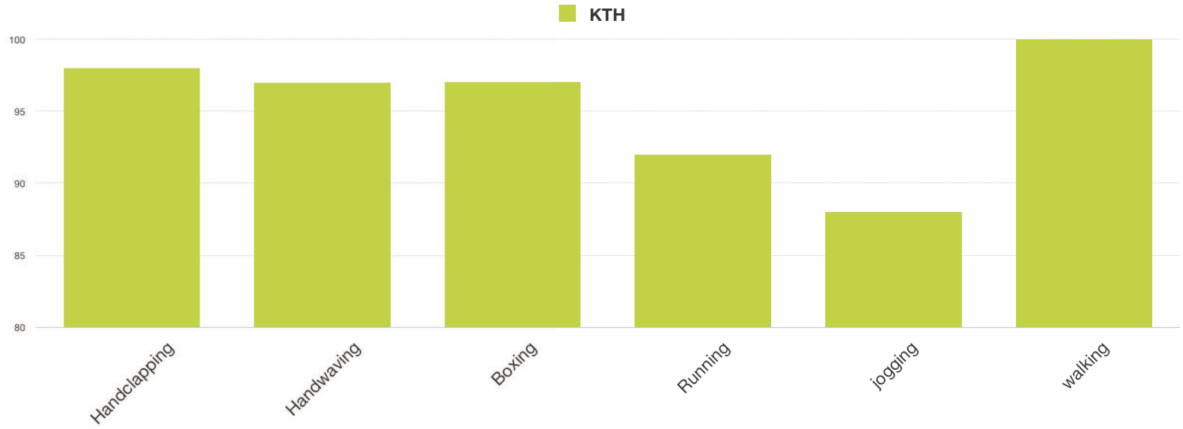

FIGURE 4.3 – Taux de reconnaissance selon les classes d'actions de `KTH Dataset` avec combinaison des descripteurs.

Descripteur	% de rec.
FCD	85,4%
HOG	91,9%
HOF	91,9%
Combiné	**95,3%**

TABLE 4.1 – Taux de reconnaissance par descripteur pour la base `KTH`.

Méthode	% de rec.
Dollar *et al.*	89,1%
Laptev *et al.*	92,1%
Wang *et al.*	94,2%
Vrigkas *et al.*	98,3%
Notre approche	**95,3%**

TABLE 4.2 – Taux de reconnaissance de la littérature pour `KTH Dataset`.

Les expérimentations montrent que le descripteur `FCD` seul donne des résultats satisfaisants en termes de reconnaissance. On observe 20,6% de confusion entre les classes `Boxing` et `Handshaking` avec le descripteur FCD, ce qui peut s'expliquer par la similarité de ses actions en terme de périodicité et donc de contenu fréquentiel. Cependant des classes comme `Running, Jogging` et `Walking` sont fortement discriminées par le descripteur `FCD`. On obtient seulement 4,6% de confusion entre ces trois classes. En effet, elles sont visuellement similaires

mais s'exécutent à différentes fréquences de mouvement. La combinaison des descripteurs via la méthode Adaboost permet d'obtenir, grace à la complémentarité des caractéristiques utilisées parmi les meilleurs taux de reconnaissance de la littérature comme l'indique le tableau 4.2.

4.1.2.2 Résultats sur la base de données Weizmann

Les résultats de classification obtenus avec notre méthode sur la base de données `Weizmann Dataset` sont présentés sur la figure 4.4 et dans les tableaux 4.3 et 4.4. Les résultats sont issus d'une validation croisée *Leave-One-Out* à l'aide d'un classifieur SVM.

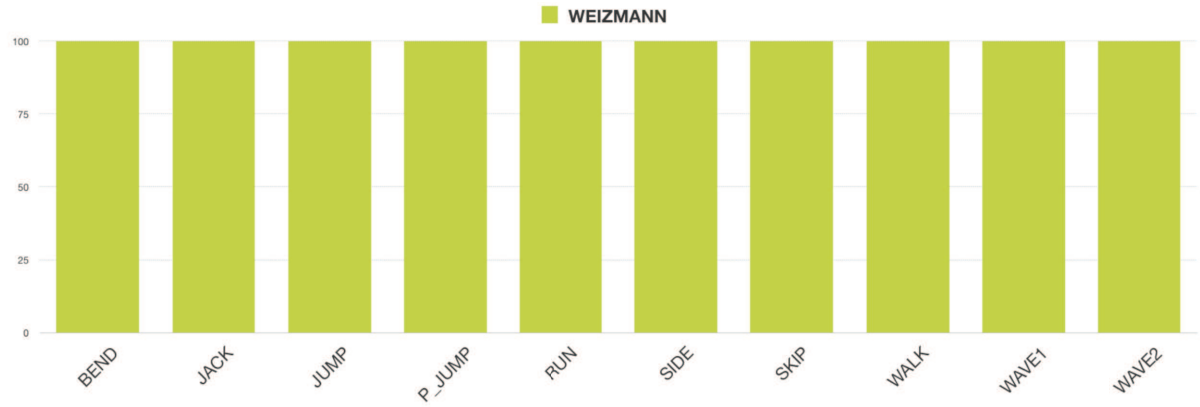

FIGURE 4.4 – Taux de reconnaissance selon les classes d'actions de la base `Weizmann Dataset`.

Descripteur	% de rec.
FCD	100%
HOG	95,5%
HOF	100%
Combiné	**100%**

TABLE 4.3 – Taux de reconnaissance par descripteur pour la base `Weizmann`.

Méthode	% de rec.
Gorelick *et al.*	97,8%
Blank *et al.*	99,6%
Vrigkas *et al.*	100%
Gorelick *et al.*	100%
Notre approche	**100**

TABLE 4.4 – Taux de reconnaissance de la littérature pour `Weizmann Dataset`.

La base `Weizmann` est très bien discriminée par notre approche. Les descripteurs `FCD` et `HOF` qui sont basés sur le mouvement apparent dans les vidéos obtiennent des taux de reconnaissance de 100%. Cette base présente un bon nombre d'actions visuellement similaires telles que `Skip`,`Jump`,`Side` et `Run`, ce qui peut expliquer la performance moindre du descripteur visuel `HOG`. Cependant après combinaison des descripteurs par la méthode Adaboost, le taux global obtenu est de 100%.

4.1.2.3 Résultats sur la base de données UCF-11

Les résultats de classification obtenus avec notre méthode sur la base de données `UCF-11 Dataset` sont présentés sur la figure 4.5 et les tableaux 4.5 et 4.6. Les résultats sont issus d'une validation croisée *Leave-One-Out* à l'aide d'un classifieur SVM.

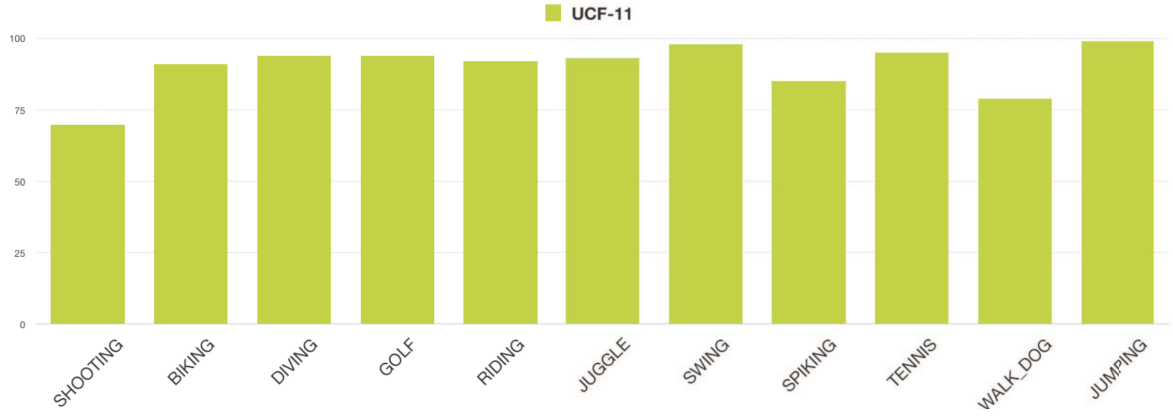

FIGURE 4.5 – Taux de reconnaissance selon les classes d'actions de la base `UCF-11 Dataset`.

Descripteur	% de rec.
FCD	74,0%
HOG	86,9%
HOF	66,4%
Combiné	**89,9%**

TABLE 4.5 – Taux de reconnaissance par descripteur pour la base `UCF-11`.

Méthode	% de rec
J. Liu *et al.*	71,2%
Wang *et al.*	85,4%
Reddy *et al.*	87,1%
Vrigkas *et al.*	95,1%
Notre approche	**89,9%**

TABLE 4.6 – Taux de reconnaissance de la littérature pour `UCF-11 Dataset`.

La base `UCF-11 Dataset` présente un ensemble de classes d'actions diverses et exécutées dans des contextes visuels très variables. La combinaison des descripteurs par Adaboost permet un gain de reconnaissance de 3,01%. Les erreurs des descripteurs ne sont pas commises sur les mêmes données selon la caractéristique qu'ils représentent. Le taux de reconnaissance obtenu sur cette base de données est l'un des meilleurs de la littérature (voir tableau 4.6).

4.1.2.4 Résultats sur la base de données UCF-50

Les résultats de classification obtenus avec notre méthode sur la base de données `UCF-11 Dataset` sont présentés sur la figure 4.6 et les tableaux 4.7 et 4.8. Les résultats sont issus d'une 25-validation croisée à l'aide d'un classifieur SVM comme le suggère les auteurs [Reddy and Shah, 2013].

Descripteur	% de rec.
FCD	53,5%
HOG	84,8%
HOF	73,8%
Combiné	**88,3%**

TABLE 4.7 – Taux de reconnaissance par descripteur pour la base `UCF-50`.

Méthode	% de rec.
Reddy *et al.*	76,9%
Todorovic *et al.*	81,0%
Murthy *et al.*	87,3%
Wang *et al.*	91,2%
Notre approche	**88,3%**

TABLE 4.8 – Taux de reconnaissance de la littérature pour `UCF-50 Dataset`.

La base de données `UCF-50 Dataset` contient un ensemble de 50 classes d'actions réparties en 5000 vidéos issues de Youtube. Notre méthode, basée sur une extraction de points épars, obtient un résultat proche des meilleures méthodes de la littérature, qui pour la plupart, sont basées sur des approches denses [Wang and Schmid, 2013, Murthy and Goecke, 2013, Peng et al., 2014]. Ce constat est détaillé plus précisément dans la section suivante. Le gain global en combinant les descripteurs est de 3,43%. Les résultats obtenus sur cette base montrent que la méthode permet de discriminer un nombre important d'actions à partir d'un grand nombre d'exemples.

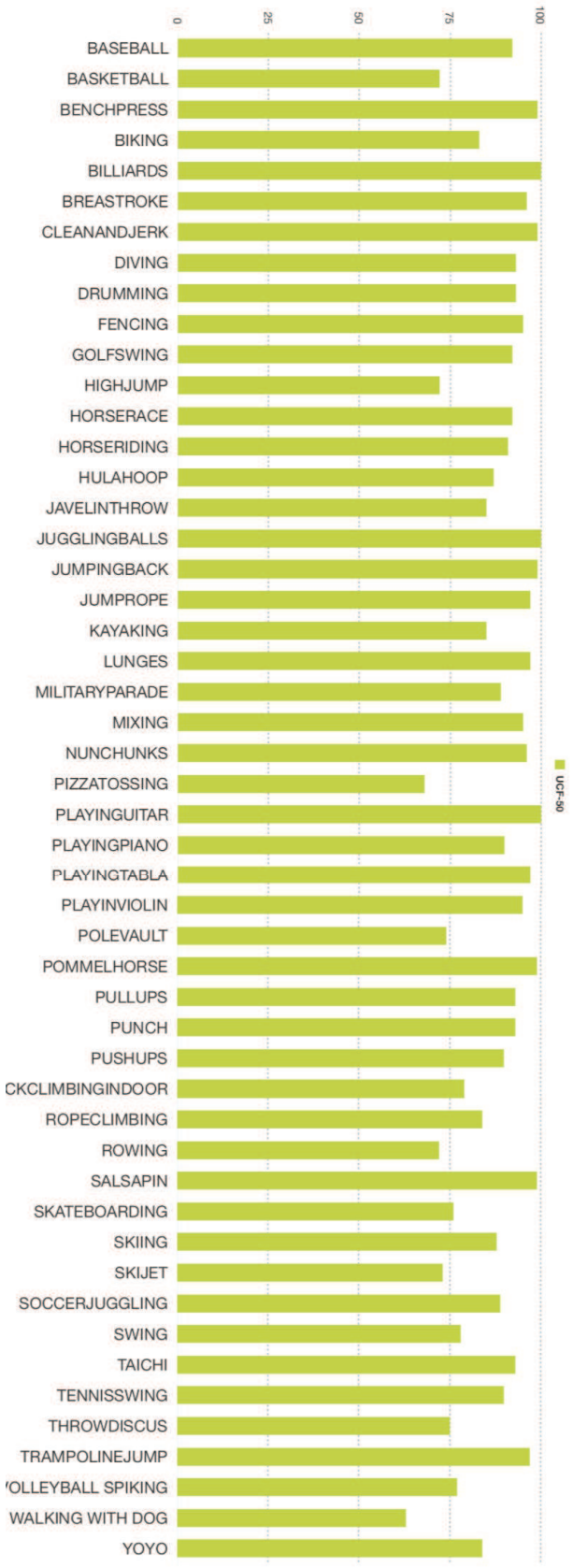

FIGURE 4.6 – Taux de reconnaissance selon les classes d'actions de la base `UCF-50 Dataset`.

4.1.3 Complexité

Temps de calcul de la méthode Cette méthode a été implémenté sous Matlab sur un serveur muni d'un processeur QuadCore à 3,1 Ghz et 24 GB de RAM.
Le flot optique est l'étape la plus coûteuse en temps de calcul (Figure 4.7). L'extraction des éléments d'intérêt et le calcul des trajectoires multi-échelles, qui constituent le coeur de notre méthode, sont les étapes les plus rapides. La méthode, étant une approche basée sur la détection de points d'intérêts, permet un temps de calcul moindre que celui des méthodes denses comme cela est montré dans la partie suivante. Nous obtenons un ratio de 2,03 sec par image pour le calcul du flot optique et un ratio de 1,.1 sec par video pour l'extraction des éléments d'intérêts et le calcul des descripteurs.

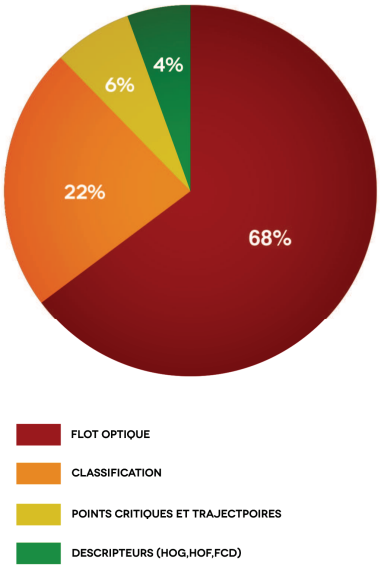

FIGURE 4.7 – Proportion du temps de calcul de chaque étape de la méthode.

Caractéristiques extraites par image Le nombre de caractéristiques (points d'intérêts, trajectoires, régions, etc.) extraites par image pour obtenir un taux de reconnaissance donné est un indicateur couramment utilisé dans la littérature. Cet indice permet d'effectuer une comparaison, en termes de complexité, entre les méthodes de reconnaissance d'actions.

Le tableau 4.9 présente le ratio nombre de caractéristiques/image pour différentes méthodes de la littérature dense (Wang *et al.*, Raptis *et al.*), éparse (Laptev *et al.*, Willem *et al.*, Dollar *et al.*), ainsi que pour la nôtre, pour la base de données `KTH`. On constate que le nombre de caractéristiques par image obtenu par notre méthode est proche des méthodes de détection de points d'intérêts épars. De même, pour ce nombre de caractéristiques par image donné, le taux de reconnaissance obtenu, est lui, proche de méthodes d'extraction dense de points. Ceci montre bien la pertinence de notre méthode, qui obtient de bons résultats avec une complexité moindre.

Méthode	Laptev *et al.*	Willem *et al.*	Dollar *et al.*	Wang *et al.*	Raptis *et al.*	**Notre approche**
caract./frame	31	19	44	225	78	**27**
%	92,1	83,3	89,1	94,2	83.3	**95,3**

TABLE 4.9 – Nombre moyen de caractéristiques par image et taux de bonne reconnaissance associé à la base `KTH Dataset` (données issues de [Wang et al., 2009])

Le tableau 4.10 montre le nombre de caractéristiques extraites par image par notre méthode sur la base `UCF-50` en comparaison de deux autres méthodes de la littérature [Wang et al., 2013] et [Shi et al., 2013].

Méthode	Wang *et al.*	Shi *et al.*	**Notre approche**
caract./image	230	65.3	**70,6**
%	91.2	83.3	**88,3**

TABLE 4.10 – Nombre moyen de caractéristiques par image et taux de bonne reconnaissance associé à la base `UCF-50 Dataset`

Shi *et al.* [Shi et al., 2013] proposent une sélection aléatoire de 10000 caractéristiques dans un ensemble extrait sur une grille dense. Wang *et al.* [Wang et al., 2013] produisent un nombre de caractéristiques par image proche de 230 pour le traitement de vidéos génériques. Cette méthode donne un taux de reconnaissance parmi les meilleurs de l'état de l'art mais génère un nombre de caractéristiques très élevé, en plus de l'utilisation de 8 échelles spatiales de trajectoires et de 30 canaux de partition spatio-temporelle. De plus, 15% du temps d'exécution de la méthode est réservé à l'écriture des données générées sur disque dur. [Murthy and Goecke, 2013] ont testé sur une vidéo le nombre de caractéristiques mis en œuvre comparé à la méthode de [Wang and Schmid, 2013] après l'étape de correspondance de trajectoires. Cette méthode met en oeuvre 1,85 fois moins de trajectoires que la méthode des trajectoires denses (11657 contre 21647 caractéristiques). On estime donc que la méthode de correspondance de trajectoires génère un total de 124,32 caractéristique par image sur la base UCF-50 avec un taux de 87,3%.

Pour la base UCF-50, notre méthode met en oeuvre 10800 points par vidéo et donc un nombre de caractéristiques par image de 70,6 pour un taux de reconnaissance de 88,30 %. On constate ici que la différence de complexité des meilleures approches de la littérature avec la notre permet de relativiser l'écart de taux de reconnaissance obtenu sur cette base de données.

4.1.4 Discussion

Le gain obtenu avec la méthode de fusion tardive Adaboost illustre la complémentarité des caractéristiques choisies. On constate un gain moyen de 3,67 % sur l'ensemble des bases de données utilisées. La caractérisation de la fréquence des trajectoires de mouvement combinée à l'information d'orientation de mouvement (`HOF`) et l'information de variation de gradient (`HOG`) permet d'atteindre un taux de reconnaissance parmi les meilleurs de la littérature. Notons que seul le mouvement des points critiques est caractérisé, ce qui représente un avantage significatif en termes de complexité. En effet, les taux de reconnaissance obtenus sont proches des méthodes basées sur des stratégies denses avec beaucoup moins de caractéristiques mises en œuvre. Les points critiques décrivent de façon pertinente les actions présentes dans les séquences vidéos étudiées et la méthode de fusion montre son efficacité dans le processus de reconnaissance. Les taux de reconnaissance obtenus sur les différentes bases de données illustrent les performances de la méthode proposée dans différentes situations : reconnaissance d'actions dans des vidéos avec contraintes d'acquisition, reconnaissance d'actions dans des vidéos génériques, discrimination d'actions avec de fortes similarités visuelles et discrimination d'un grand nombre de classes d'actions.

4.2 Évaluation de l'influence des paramètres

Dans cette partie nous évaluons les différents paramètres de notre méthode ainsi que leur influence sur le taux de reconnaissance obtenu sur les bases de données de la littérature. La méthode utilise très peu de paramètres. Ces derniers sont :

— C_p : le nombre de points critiques extraits,

— N : la taille des trajectoires,

— s : le nombre d'échelles utilisées pour l'approche multi-échelles.

Les expériences menées sur l'ensemble des bases de données ont révélé que le paramètre C_p est celui qui influe le plus le taux de bonne reconnaissance. Ce paramètre est à mettre en corrélation avec l'approche multi-canaux du `BoW` qui multiplie le nombre de points critiques générés en fonction du nombre de cellules utilisées. Le nombre s d'échelles spatio-temporelles aide à l'amélioration du taux de reconnaissance sur les bases de données réalistes où l'information de fréquence est plus riche que sur les bases de données avec contraintes d'acquisitions.

La taille N des trajectoires a été fixée à 16 trames pour chaque base de données. En effet, la méthode d'estimation de trajectoires est issue de [Wang et al., 2011] où l'auteur étudie l'influence de la taille des trajectoires et met en évidence l'intérêt de conserver des trajectoires d'environ 15 trames. Comme nous utilisons une analyse dyadique, nous prenons la puissance de 2 la plus proche de cette longueur, à savoir 16.

La compensation du mouvement de caméra sur les bases de données génériques permet également d'observer une amélioration des résultats sur un grand nombre de classes d'actions. Ces points sont détaillés dans les sections qui suivent.

4.2.1 Variation du nombre de caractéristiques

Nombres de points critiques Nous évaluons ici l'influence du nombre de caractéristiques utilisées sur le taux de reconnaissance. L'évolution du taux de reconnaissance en fonction du nombre de points critiques générés sur `KTH Dataset` et `UCF-11 Dataset` est présentés sur la figure 4.8. Ce taux de reconnaissance est obtenu en utilisant la version standard du `BoW`, et seule la cellule `1x1t1` est utilisée afin d'observer l'influence du nombre de caractéristiques. Ces bases de données contiennent respectivement des vidéos avec de fortes contraintes d'acquisition et des vidéos génériques.

On constate que le nombre de points critiques nécéssaires pour atteindre un bon taux de reconnaissance se stabilise rapidement pour la base de données avec contraintes d'acquisition (300 points critiques). `UCF-11 Dataset` nécessite un plus grand nombre de points critiques générés avant d'obtenir un taux de reconnaissance stable. Dans les expérimentations menées, nous utilisons un total de 500 points critiques pour les bases de données avec contraintes d'acquisitions `KTH` et `Weizmann`. Pour la base de données `UCF-11` et `UCF-50` nous utilisons respectivement un total de 800 et 1200 points critiques par canal et par échelle de fréquence.

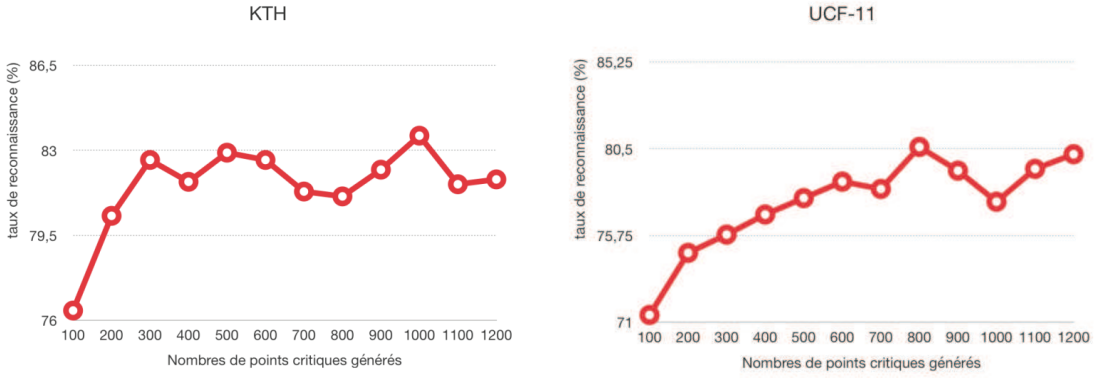

FIGURE 4.8 – Évolution du taux de reconnaissance en fonction du nombre de points critiques générés.

Taille du dictionnaire de mots visuels L'évolution de la taille du dictionnaire, constitué à partir des caractéristiques extraites, est évaluée sur les bases de données étudiées et présentée sur la figure 4.9. Le dictionnaire du `BoW` est construit de façon non-supervisée à l'aide de l'algorithme des `k-moyennes`. L'évolution du taux de reconnaissance permet d'observer la séparabilité des données dans l'espace des caractéristiques en fonction de la taille du dictionnaire du `BoW`. L'intérêt de ces caractéristiques en termes de représentation et de description des actions étudiées est donc souligner dans le cas où le taux de reconnaissance est élevé pour une taille de dictionnaire faible.

Les bases de données avec contraintes d'acquisition obtiennent de bons résultats avec des tailles de dictionnaire relativement faibles pour chaque descripteur. On constate qu'à partir d'un dictionnaire composé de 1000 mots visuels, le taux de reconnaissance évolue peu. C'est le nombre de mots visuels retenu par la suite pour les bases `KTH` et `Weizmann`. Le descripteur de trajectoire `FCD` permet d'obtenir un taux de reconnaissance de 100% sur `Weizmann Dataset`, avec une taille de dictionnaire assez faible, contrairement aux autres descripteurs. Cette base de données est la

FIGURE 4.9 – Évolution du taux de reconnaissance en fonction de la taille du dictionnaire pour les bases `KTH`, `Weizmann`, `UCF-11`, `UCF-50`.

seule à ne posséder aucun mouvement de caméra, un fond complètement uniforme et des acteurs quasi statiques. Il n'y a donc aucune forme de perturbation dans le processus d'estimation des trajectoires, ce qui les rend d'autant plus pertinentes pour caractériser les mouvements.

Les bases de données de vidéos génériques nécessitent un nombre de mots visuels plus important pour la formation des dictionnaires. Le taux de reconnaissance tend à se stabiliser à partir de 2000 mots visuels pour `UCF-11` et 5000 pour `UCF-50`. Ces taux sont respectivement choisis pour ces deux bases de données. En effet, ces bases comportent beaucoup plus d'informations visuelles et nécessitent de mettre en œuvre un plus grand nombre de caractéristiques afin d'être correctement caractérisées. Comme cela a été vu dans le Chapitre 3, afin d'obtenir un temps de calcul convenable, nous utilisons un noyau linéaire dans l'étape de classification pour les bases de données génériques.

4.2.2 Échelle de fréquence utilisée

L'approche multi-échelles de notre méthode est évaluée ci-dessous sur la base de données `UCF-11`. Pour les évaluations qui suivent, le nombre maximal d'échelles utilisées est de 3. Au delà de ce chiffre, les sous-séquences obtenues par subdivision dyadique sont de trop basse résolution et trop courtes temporellement pour représenter une information pertinente. Le tableau 4.11 montre les taux de reconnaissance obtenus pour chaque descripteur dans deux situations : l'approche standard qui correspond à l'utilisation d'une seule échelle de fréquence, et l'approche multi-échelles où trois échelles de fréquence sont utilisées.
On constate que le gain global avec l'approche multi-échelles est de 4,58%, et que le descripteur `FCD`, caractérisant les trajectoires de points critiques gagne 8,38%.

UCF 11	Approche standard ($s = 1$)	Approche multi-échelle ($s = 3$)
FCD	49,04 %	57,42 %
HOG	73,61 %	78,52 %
HOF	62,23 %	68,24 %
Combiné	**80,49** %	**85,07** %

TABLE 4.11 – Taux de reconnaissance de chaque descripteur en fonction du nombre d'échelles s de fréquence.

La figure 4.10 illustre l'évaluation de l'approche multi-échelles sur deux bases de données : `KTH` et `UCF-11` composées respectivement de vidéos avec contraintes d'acquisition et de vidéos génériques. La figure montre le taux de reconnaissance global en fonction de chaque échelle de fréquence ainsi que le taux de reconnaissance obtenu lorsque l'on cumule les échelles 1 et 2, puis 1, 2 et 3.

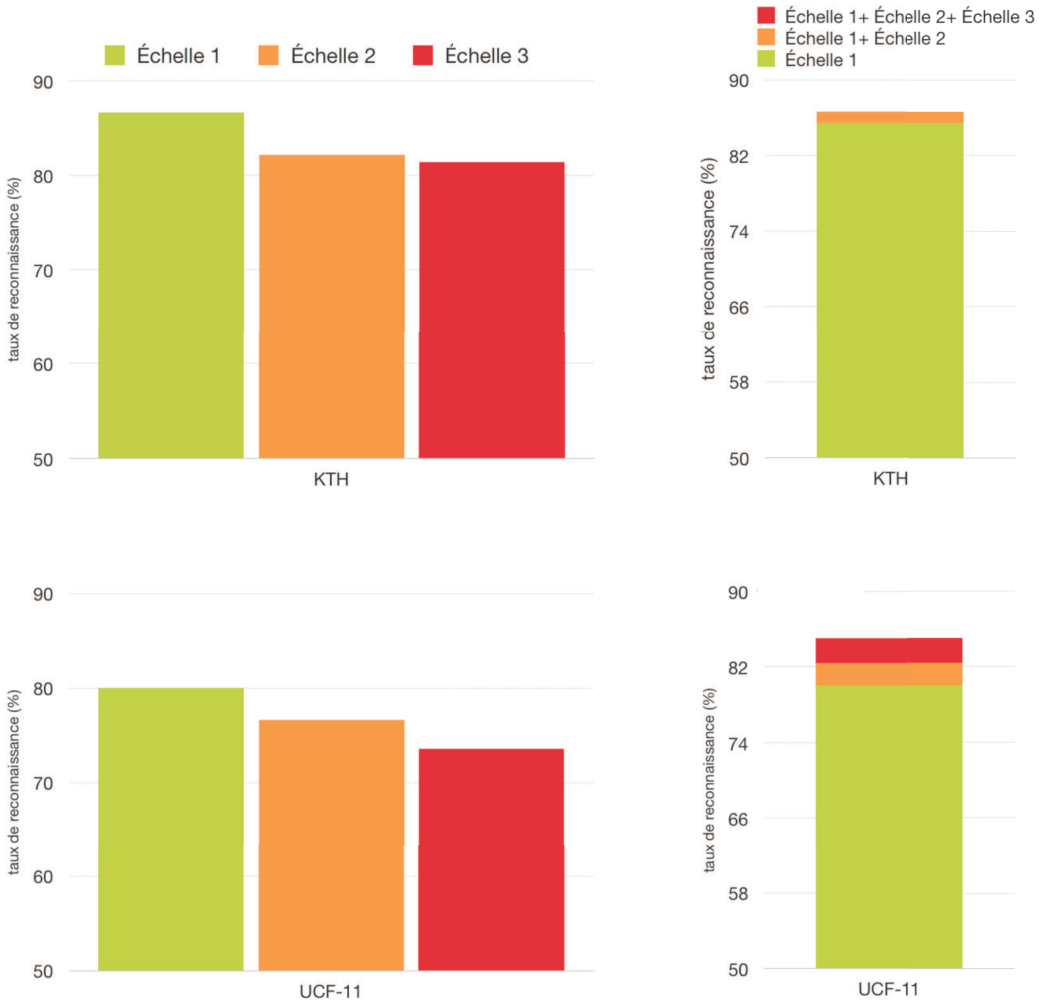

FIGURE 4.10 – À gauche : taux de reconnaissance obtenus en utilisant une échelle de fréquence fixée. À droite : taux de reconnaissance obtenus en cumulant les échelles de fréquences (1ère ligne : `KTH Dataset`, 2ème ligne : `UCF-11 Dataset`).

Pour chaque base de données, on remarque que l'utilisation de la première échelle seule ($s = 1$) permet une meilleure reconnaissance que les échelles 2 et 3 (80,04% pour `UCF-11 Dataset` et 86,64% pour `KTH Dataset`). Cela est cohérent car l'échelle 1 encode l'information des mouvements de hautes fréquences, qui contient le plus de détails, et favorise donc la distinction entre les différentes actions. On constate une différence entre les deux bases de données. Le cumul des échelles de fréquence pour la base de données `UCF-11` permet d'augmenter sensiblement le taux de reconnaissance (2,51% de gain moyen à chaque ajout d'échelle). Cependant pour la base `KTH`, le taux de reconnaissance en cumulant les échelles de fréquence reste quasiment le même qu'avec la première échelle seule (0,085% de gain moyen à chaque ajout d'échelle). Cette différence peut être expliquée par la nature des deux bases de données étudiées.

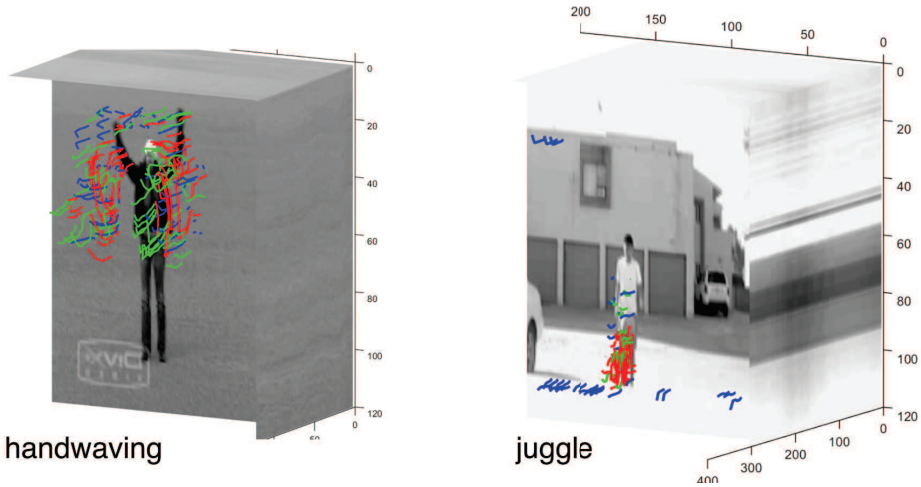

FIGURE 4.11 – Trajectoires multi-échelles sur une vidéo de la base `KTH` et de la base `UCF-11`. Les trajectoires des différentes échelles ont majoritairement la même localisation sur la vidéo de `KTH` alors qu'au contraire sur la vidéo de `UCF-11` elles sont plus dispersées spatialement. Cela est dû à la plus grande variété de mouvements, les différentes fréquences sur cette base de données.

En effet, la base `KTH Dataset` possède des vidéos où les acteurs sont immobiles et effectuent des mouvements répétitifs et peu naturels. L'information fréquentielle sur ces vidéos n'est pas très variée. Sur la figure 4.11, on constate que pour une vidéo représentant la classe `handwaving` (*Mouvement de la main*), les trajectoires correspondant aux différentes échelles de fréquence sont toutes localisées aux mêmes endroits et caractérisent donc la même information de mouvement.

En revanche sur les bases de données génériques, l'information fréquentielle est plus riche du fait des différents mouvements naturels effectués dans des contextes réalistes par les sujets. La figure 4.11 montre un exemple de l'action `Juggle` (*Jongle*) de la base de donnée `UCF-11`. Hormis certaines trajectoires non pertinentes, on observe que les trajectoires de la dernière échelle ($s = 3$) sont localisées sur le corps du sujet au niveau du torse, tandis que les trajectoires de la première échelle ($s = 1$) sont localisées au niveau des pieds du joueur, les trajectoires de la deuxième échelle ($s = 2$), sont elles, au niveau des jambes. Les différentes trajectoires caractérisent des mouvements différents et apportent donc une information complémentaire lors du cumul des échelles.

4.2.3 Compensation du mouvement de caméra

L'influence de l'étape de compensation des mouvements de caméra, décrite dans le chapitre 3 est présentée dans le tableau 4.12. La compensation de mouvement a été utilisé sur la base de données `UCF-11`. Cette base de données est constituée de vidéos génériques capturées à l'aide de caméras non statiques et présente donc de nombreux mouvements parasites. Nous étudions ici l'impact de la compensation de mouvement sur le taux de bonne reconnaissance.

UCF 11	Taux sans compensation	Taux avec compensation
FCD	49,04 %	64,05 %
HOG	73,61 %	83,34 %
HOF	62,49 %	74,06 %
Combiné	**80,49** %	**87,89** %

TABLE 4.12 – Comparaison des résultats obtenus sur chaque descripteur avec et sans compensation de mouvement de caméra.

Le gain obtenu sur le descripteur `FCD` est de 15,01 %, ce qui illustre l'importance de la compensation de mouvements de caméra dans le processus d'estimation des trajectoires. L'augmentation des performances de la méthode quant aux descripteurs `HOG` et `HOF` montre que le flot optique après compensation permet une meilleure caractérisation des mouvements. En effet, les point critiques sont mieux localisés et l'information locale autour de ces derniers est moins perturbée et donc plus pertinente.

La figure 4.12 illustre le gain obtenu sur chaque classe d'action de la base `UCF-11` en utilisant notre méthode de compensation de mouvements de caméra. Les classes possédant les plus forts gain en terme de reconnaissance sont : `Spiking`,`Walk with a dog`,`Tennis` et `Diving` (respectivement 15, 11 et 9% de gain).

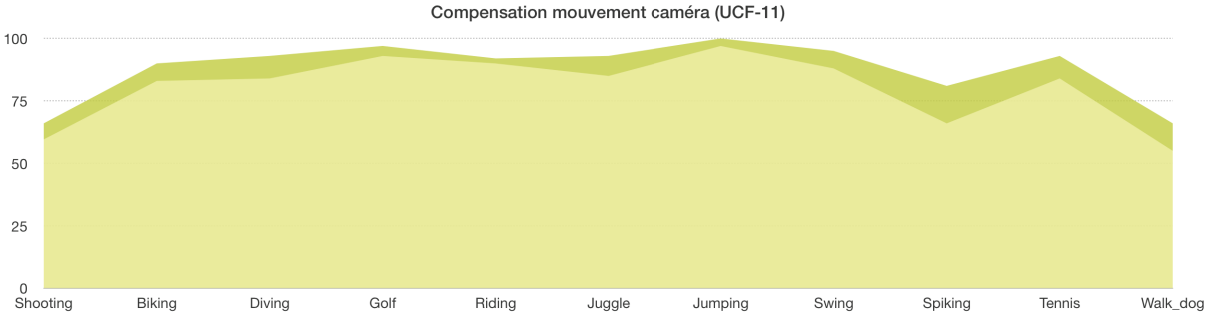

FIGURE 4.12 – Taux de reconnaissance avec la compensation de mouvement (vert foncé) et sans compensation (vert clair). On constate un gain significatif pour chaque classe d'actions de la base `UCF-11 Dataset`.

Dans la meilleure des configurations ($s = 3$), le gain global de reconnaissance avec la compensation de caméra est de 3,01% sur la base de données `UCF-11`. Nous atteignons ainsi un taux de reconnaissance de 89,99 %, comme présenté dans la section précédente. Ce taux est l'un des meilleurs de la littérature sur cette base de données.

4.3 Évaluation de la généricité de la méthode

Dans cette section, une approche originale est présentée afin d'évaluer la généricité de notre méthode. Le but est d'évaluer à quel point la méthode caractérise les actions dans des vidéos, indépendamment de la base de données utilisée.

4.3.1 Biais visuels des bases de données

Cette expérimentation est basée sur les travaux de [Torralba and Efros, 2011]. L'objectif initial de ces travaux est de mettre en lumière les biais visuels contenus dans la plupart des bases de données de référence de vidéos. Cette problématique est très importante dans le domaine de la reconnaissance de formes mais est souvent négligée dans la littérature. En effet, les bases de données sont construites afin de représenter de la façon la plus variée possible le monde réel. Les auteurs mettent en avant les différentes causes de ces biais visuels.

— **Biais de sélection** Les sources d'où proviennent les données induisent un biais visuel appelé le biais de sélection. Les bases de données tendent généralement à sélectionner un certain type d'images ou à utiliser certains mots clés pour illustrer une classe donnée. Les scènes de nature, les scènes urbaines, certains animaux ou objets sont généralement collectés à partir de mêmes sources et tendent à être similaires dans les bases de données. La figure 4.13 montre comment est illustrée la classe d'image `car` dans les bases de données `Caltech-101` [Fei-Fei et al., 2007] et `SUN` [Xiao et al., 2014]. On constate que la base de données `SUN` (fond bleu) a tendance à représenter la classe `car` avec des prises de vue de voitures en biais, tandis que la base `Caltech-101` (fond vert) l'illustre plutôt avec des des prises de vue de côté. Ces différentes prises de vue génèrent un biais de sélection et peuvent influencer les décisions d'un classifieur, en fonction de la base avec laquelle il a été entrainé.

FIGURE 4.13 – Exemples de la classe `car` pour la base `SUN` (fond bleu) pour la base `Caltech-101` (fond vert). On constate la différence de représentation pour ces deux bases de données d'images.

— **Biais de captation** Le biais de captation correspond aux contraintes d'acquisitions et aux habitudes de captation pour la prise de vue de certains sujets ou objets. Ce biais visuel peut être également vu comme une forme de biais "social". En effet, les photos sont prises de façon à ce que les objets, monuments, où lieux photographiés soient reconnus de tous. On constate donc que la grande majorité des images présentes sur internet ne fournissent pas un échantillonnage aléatoire des points de vue recouvrant un objet ou une scène mais plutôt une représentation formatée par nos habitudes sociales. La figure 4.14 illustre ce biais de captation. On observe les six premières images proposées par `Google Images` en tapant la requête *"Street image"* (deux premières lignes) et la requête *"Tour Eiffel"* (deux dernières lignes). Pour chacune d'entre elles, les images obtenues représentent la requête selon un même angle de vue. Un classifieur peut être influencé, en termes de reconnaissance, par ce biais de captation.

FIGURE 4.14 – Six premières requête de `Google Images` pour les mots clés *"Street image"* et *"Tour Eiffel"*. Les photos obtenues sont globalement prises avec le même angle de vue.

— **Biais de l'ensemble négatif** L'ensemble négatif est ce qu'une base de donnée représente comme étant "le reste du monde" en comparaison au monde clos défini par l'ensemble des données de cette base. Une base de données très peu représentative du monde extérieur ou avec une pondération non-uniforme des données entre les classes biaise les décisions d'un classifieur. L'ensemble négatif définit dans l'espace des caractéristiques les frontières de décisions du classifieur en fonction de la base de données sur laquelle il a été entrainé. La figure 4.15 montre l'ensemble des classes d'action de la base de données `Weizmann`. Chaque action est effectuée dans le même contexte. On constate que l'ensemble visuel formé par cette base est très pauvre. Le fond est uniforme, l'angle de vue est le même, les acteurs sont visuellement similaires. Le "reste du monde" est mal représenté dans cette base de données.

FIGURE 4.15 – Illustration des classes de la base `Weizmann`. Les actions sont toutes effectuées dans un contexte qui varie peu et sont donc représentées dans un monde "clos".

Jason Salavon, dans sa production artistique *"100 special moments"* utilise une centaine de photos issues d'internet sur quatre thématiques différentes, et les superpose entre-elles. La figure 4.16 montre les différents résultats obtenus. Les thématiques sont `Jeunes mariés`, `Enfants sur le père noel`, `Jeune joueur de base-ball`, `Étudiant diplômé`. Le résultat permet d'observer une image "moyenne" représentant une forme qui illustre chacun de ces thèmes. Ces travaux artistiques révèlent bien le biais de captation, présent sur internet, expliqué plus haut.

FIGURE 4.16 – *"100 special moments"* de Jason Salavon. Les images moyennes obtenues par empliment successif illustrent clairement la thématique à laquelle elles sont associées (`Jeunes mariés`, `Enfants sur le père noel`, `Jeune joueur de base-ball`, `Étudiant diplômé`).

Nous avons reproduit le même procédé que Jason Salavon sur un ensemble de vidéos pour les actions `Diving, Golf, Clean and jerk, HorseRace`. Une image est aléatoirement choisie sur un ensemble de 120 vidéos pour chacune de ces actions. La figure 4.17 illustre le résultat obtenu. On distingue, comme pour les productions plus haut, une image moyenne qui correspond visuellement à l'action effectuée ou le contexte global où elle est exécutée. Cet échantillon montre une tendance, dans les bases de données de la littérature, à developper une forme de biais de captation pour certaines classes d'actions.

FIGURE 4.17 – "Images moyennes" obtenues sur les vidéos des actions `Diving, Golf, Clean and jerk, HorseRace`. Pour chaque action, on obtient une image significativement représentative de ces dernières. On retrouve visuellement une pause stéréotypée du joueur de golf, du culturiste, ainsi que la forme globale d'une piscine et d'une scène de course de chevaux.

L'objectif des travaux qui suivent est de répondre à cette question : " Comment un classifieur entrainé sur une base de données généralise l'information quand il est testé sur d'autres bases de données, comparativement aux performances obtenues quand il est testé sur la base de données d'origine " [Torralba and Efros, 2011].

4.3.2 Expérimentations

Dans le contexte de la reconnaissance d'actions humaines, cette méthologie est employée afin d'évaluer la généricité de l'approche présentée dans le Chapitre 3. Le but est d'estimer à quel degré la caractérisation des actions humaines par cette méthode est robuste aux biais visuels contenus dans les bases de données.

Pour cette expérimentation, nous utilisons quatre bases de données très utilisées dans la littérature : `KTH`, `Weizmann`, `UCF-11`, `UCF-50` et `HMDB-51`. Nous utilisons les classes d'actions `Walk` et `Wave`, qui sont des actions communes à toutes les bases de données choisies (notons que pour `UCF-11 Dataset`, l'action `Wave` est représentée par la classe `golf` qui est l'action la plus proche dans cette base représentant un mouvement de la main).

Notre méthode est entrainée avec 200 exemples positifs et 200 exemples négatifs pour chaque base de données. Un sur-échantillonnage est effectué sur les bases composées d'un nombre de données insufisant. L'ensemble de test, pour chaque base , est composé de 100 exemples positifs de cette base et 100 exemples négatifs issus de l'ensemble des bases de données. Cette configuration des exemples négatifs de l'ensemble de test permet de diminuer l'effet du biais de l'ensemble. L'objectif de cette évaluation est d'observer la différence de taux de reconnaissance quand le classifieur est entrainé sur une base de donnée, puis testé sur d'autres bases.

4.3.3 Résultats d'apprentissage croisé

Les tableaux 4.13 et 4.14 présentent les résultats obtenus. Les lignes correspondent au classifieur entrainé sur une base de données et testé sur les autres. Les colonnes correspondent donc aux résultats obtenus quand le classifieur est testé sur une base de données et entrainé sur les autres.

Comme l'a montré [Torralba and Efros, 2011], les meilleurs résultats sont obtenus quand on entraine et teste le classifieur sur la même base de donnée (94,7 % en moyenne pour `Walk` et 95,1 % pour `Wave`) ce qui conforte l'idée que les bases sont biaisées. `Weizmann Dataset` et `UCF-11 Dataset` sont les bases les moins performantes en termes de généralisation avec respectivement 39,75 % and 35,35 % de baisse de reconnaissance en moyenne pour les deux actions étudiées. Les fortes contraintes d'acquisition et le peu d'exemples dans la base `Weizmann` peuvent expliquer la difficulté à obtenir un bon taux de généralisation. Kuehne *et al.* [Kuehne et al., 2011] ont montré que les vidéos génériques issues de Youtube contiennent un biais visuel bas niveau dû aux habitudes de captation des utilisateurs. Cela peut en partie expliquer le faible taux de généralisation obtenu avec `UCF-11` où l'on obtient 42% de baisse pour la classe d'action `Walk`. En revanche, avec `HMDB-51` qui contient des vidéos issues de différentes sources, on obtient 15 % de baisse pour la classe d'action `Walk`.

action	train/test	KTH	Weizmann	UCF-11	HMDB51	Base d'entrainement	Autres bases	% perdu
walk	KTH	**97%**	96%	56%	62%	97%	71,3%	26,4%
	Weizmann	66%	**100%**	51%	55%	100%	57,3%	42,7%
	UCF-11	54%	50%	**95%**	61,5%	95%	55,1%	42%
	HMDB51	79%	79,5%	62,5%	**87%**	87%	73,6%	15%
	Autres bases	66,3%	75,1%	56,5%	59,5%	94,7%	64,3%	32,1%

TABLE 4.13 – Apprentissage croisé pour la classe `Walk` quand le classifieur est entrainé sur une base de données (ligne) et testé sur les autres (colonnes).

action	train/test	KTH	Weizmann	UCF-11	HMDB51	Base d'entrainement	Autres bases	% perdu
wave	KTH	**99,5%**	73,5%	60%	50%	99,5%	61,1%	38,5%
	Weizmann	65%	**100%**	73,5%	51%	100%	63,1%	36,8%
	UCF-11	58,5%	85,5%	**94,5%**	58%	94,5%	67,3%	28,7%
	HMDB51	50,5%	75%	69,5%	**86,5%**	86,5%	65%	24%
	Autres bases	58%	78%	67,6%	53%	95,1%	64,1%	32%

TABLE 4.14 – Apprentissage croisé pour la classe `Wave` quand le classifieur est entrainé sur une base de données (ligne) et testé sur les autres (colonnes).

Le manque de comparaison avec d'autres approches similaires ne nous permet pas de conclure définitivement sur la robustesse de la méthode par rapport aux biais visuels des bases de données. Cependant, en observant les résultats obtenus, on constate que la méthode tend à caractériser de façon générique les actions observées. En effet, on obtient un taux de bonne reconnaissance moyen de 64,2 % lorsque le classifieur est entrainé sur une base de données et testé sur les autres. Les actions `Walk` et `Wave` sont donc globalement bien généralisées par notre méthode.

4.3.4 Bases de données hybride

Les bases de données sélectionnées représentent différents aspects des actions `Walk` et `Wave`. `KTH Dataset` et `Weizmann Dataset` contiennent des vidéos où les actions sont réalisées par des acteurs, de façon non naturelle. Dans les bases de données `UCF-11 Dataset` et `HMDB51 Dataset`, les actions sont exécutées dans différentes situations et différents contextes. Cela amène une variabilité visuelle ainsi que du bruit. Ces éléments apportent une représentation réaliste des actions `Walk` et `Wave`.
Ces deux types de bases de données, avec contraintes et génériques, contiennent des informations complémentaires quant aux actions élémentaires. La figure 4.18 illustre, pour l'action `Walk` des exemples issus de ces deux types de base de données (`KTH Dataset`, `Weizmann` et `UCF-11`). La nature et la variabilité des informations visuelles délivrée par ces bases diffèrent. On peut observer dans le tableau 4.13 que les bases `KTH` et `HMDB-51`, respectivement une base avec contraintes et une base générique, obtiennent de bonnes performances en généralisation.

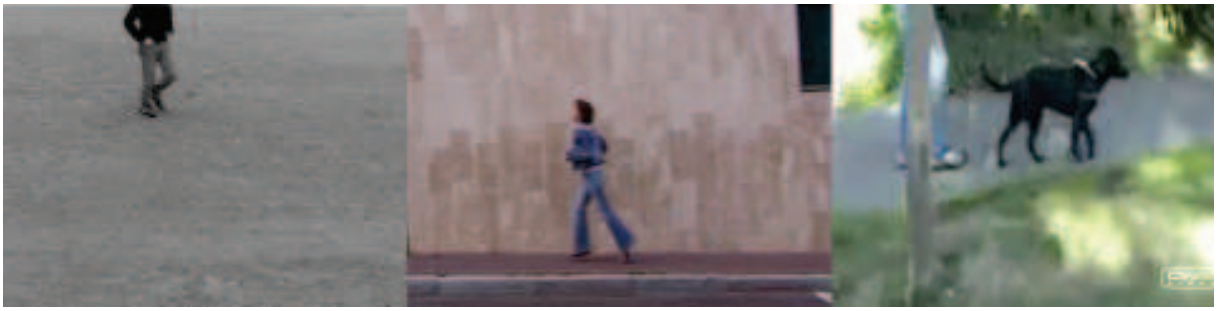

FIGURE 4.18 – Représentation de la classe d'action `Walk` sous différentes bases de données (`KTH Dataset`, `Weizmann Dataset`, `UCF-11 Dataset`). On remarque, les différents points de vue et contexte fourni par chaque base de données. (voir vidéo).

En partant de ces observations, nous tentons d'améliorer le processus de généralisation précédent en utilisant un mélange pondéré de bases de données dans la phase d'entrainement. Nous utilisons le pourcentage perdu de chaque base de données comme une pondération pour construire une nouvelle base de données qui donne plus d'importance aux vidéos issues de base de données possédant un bon taux de généralisation. Cette nouvelle base de données contient 200 exemples positifs et 200 exemples négatifs composés de chaque base de données proportionnellement à leur poids obtenus en normalisant les taux de perte.

Les tableaux 4.15 et 4.16 montrent les résultats obtenus avec cette base de données hybride. Le taux moyen obtenu en testant sur toutes les autres bases de données est clairement plus élevé comparativement à ceux obtenu dans les tableaux 4.13 et 4.14 (83,6% pour `Walk` et 83,3% pour `Wave`). La perte est juste de 6,9% en moyenne, ce qui est moitié moins que la perte sur `HMDB51 Dataset` qui obtient le meilleur taux de généralisation (15 % for `Walk` et 24% for `Wave`). Cette nouvelle base de données, qui est un mélange des précédentes, permet une représentation robuste des actions `Walk` et `Wave`.

L'évaluation de biais de base de données ainsi que la généricité des méthodes de reconnais-

action	train/test	base mélange	KTH	Weizmann	UCF-11	HMDB51	autres bases	% drop
walk	base mélange	**90%**	85%	93%	74.5%	82%	83,6%	7,1%

TABLE 4.15 – Évaluation croisée pour l'action `Walk` quand le classifieur est entrainé sur la base de données mélange et testé sur les autres bases (colonnes).

action	train/test	mixture dataset	KTH	Weizmann	UCF-11	HMDB51	autres bases	% drop
wave	mixture dataset	**89,5%**	81%	93,5%	88%	71%	89,5%	6,8%

TABLE 4.16 – Évaluation croisée pour l'action `Wave` quand le classifieur est entrainé sur la base de données mélange et testé sur les autres bases (colonnes).

sance est assez récente dans la littérature et seuls quelques articles traitent de cette problématique ([Khosla et al., 2012, Sultani and Saleemi, 2014]). La construction de bases de données mélange issues de différentes bases de données, en fonction de leur taux de généralisation est un travail préliminaire mais apporte quelques directions quant à la construction de classifieur robuste à différentes représentations d'actions humaines comme nous le verrons dans le chapitre 5.

Conclusion du chapitre Dans ce chapitre, nous avons mis en avant les performances de notre méthode quant à la caractérisation d'actions humaines élémentaires dans des bases de données de la littérature. Après une présentation des résultats obtenus sur ces bases de données nous avons présenté l'intérêt en terme de complexité de notre approche face à d'autres méthodes de la littérature. L'influence de différents paramètres de notre approche ont également été évalué. La dernière partie de ce chapitre a montré la capacité de notre approche à généraliser la représentation des actions à l'aide d'une méthode d'apprentissage croisé sur différentes bases de données.

Le chapitre suivant introduit la deuxième partie de nos travaux sur la reconnaissance du mouvement humain, notamment les actions complexes ou activités humaines. Nous verrons à la fois la différence entre la notion d'actions élémentaires et activités humaines mais également comment les travaux présentés jusqu'ici s'intègrent dans un processus de caractérisation et de reconnaissance d'activités humaines.

Chapitre 5

Reconnaissance d'activités humaines

Chapitre 5

Reconnaissance d'activités humaines

Sommaire

Introduction du chapitre Dans la première partie de ce chapitre, nous revenons sur la différence sémantique entre une activité et une action élémentaire. Nous verrons comment les méthodes de la littérature se sont employées à résoudre les problématiques propres à la caractérisation d'activités humaines.

Nous exposerons dans un second temps la méthode que nous proposons pour la caractérisation et la reconnaissance d'activités humaines dans des vidéos sans contraintes d'acquisitions.

5.1 Introduction

5.1.1 Différences sémantiques entre actions élémentaires et activités

Une activité est considérée comme le niveau d'interprétation sémantique supérieur par rapport aux actions élémentaires présentées au chapitre 3. En effet, les activités humaines mettent généralement en œuvre plusieurs actions élémentaires, ou sous-évènements, pour former un ensemble plus complexe, dont l'interprétation varie généralement en fonction du contexte où elles sont effectuées. Communément, les activités représentent :

— les comportements humains du quotidien (manger, nettoyer, se promener, etc).
— les actions sportives (basket-ball, course à pied, tennis, etc).
— les interactions humaines (rentrer dans une pièce, discuter, se battre).
— les enchaînements d'actions dans un cadre contrôlé (jeu de cartes, préparation de recettes, etc).

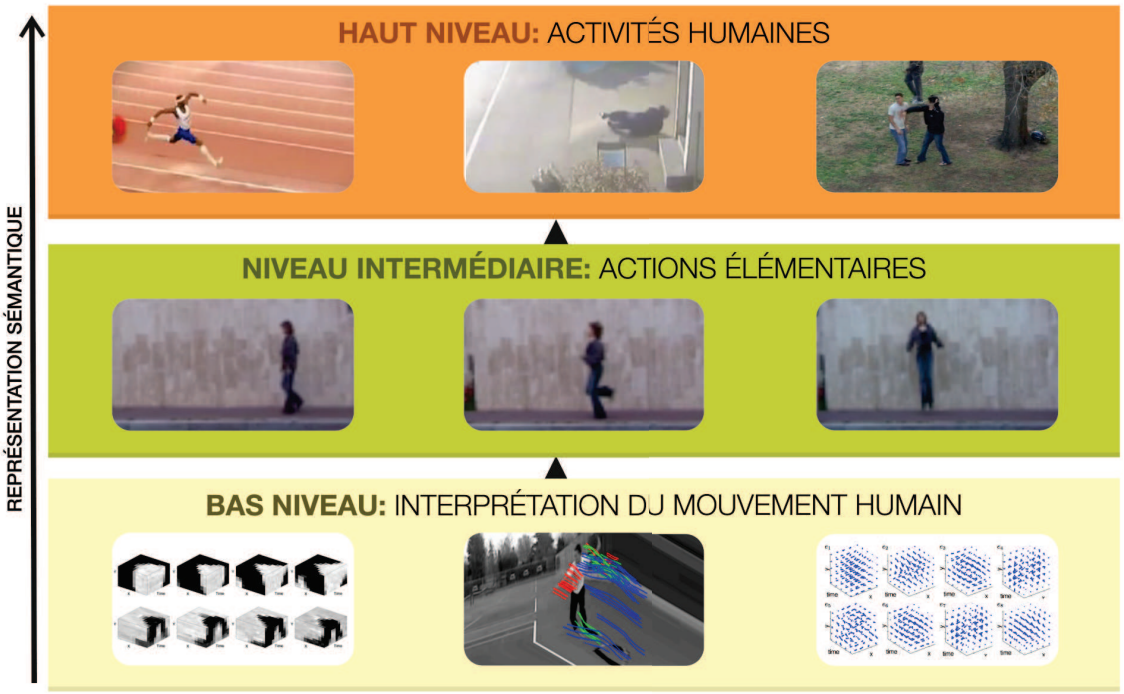

FIGURE 5.1 – Hiérarchie de représentation sémantique. Bas niveau : détection d'éléments d'intérêt liés aux mouvements humains. Niveau intermédiaire : reconnaissance des actions humaines élémentaires. Haut niveau : caractérisation d'activités humaines.

La caractérisation des activités humaines est une thématique qui suscite un intérêt grandissant. L'étude des activités humaines est un élément essentiel pour un grand nombre d'applications telles que la vidéo-surveillance, l'indexation automatique de vidéos, les systèmes d'interactions homme-machine, les systèmes de diagnostic de patients, dans le domaine de la santé, notamment pour les personnes âgées et les enfants en bas âge. Pour être utiles à l'utilisateur, ces différentes applications nécessitent d'interpréter les mouvements humains à un plus haut niveau de compréhension sémantique que la caractérisation d'actions élémentaires. Cependant, la reconnaissance d'activité n'est pas une tâche simple et regroupe de nombreuses problématiques. En effet, ces activités humaines nécessitent un temps d'observation beaucoup plus long que les simples actions élémentaires. Cela requière la gestion de variations visuelles durant ce temps d'observation, notamment celui du changement de point de vue au cours du temps, la variation de l'illumination, les occultations partielles, la prise en compte de la séquentialité des actions élémentaires ainsi que les variations de vitesse d'exécution des activités. Elles requièrent donc des méthodes qui prennent en compte cet aspect temporel de façon robuste.

On retrouve dans la littérature différentes méthodes tentant de répondre aux problématiques engendrées par la reconnaissance d'activités humaines. Nous faisons le choix de présenter un ensemble de méthodes réparties en deux catégories : les méthodes supervisées, qui utilisent des données d'entrainement labellisées pour l'apprentissage d'activités, et les méthodes non-supervisées, qui s'appuient sur la découverte automatique d'éléments caractérisant une activité humaine. La figure 5.2 reprend la hiérarchie de ces approches de reconnaissance.
Dans la suite, nous donnons une liste non-exhaustive des méthodes supervisées, puis des méthodes non-supervisées de la littérature. Nous verrons la façon dont ces dernières interprètent la notion d'activité ainsi que l'avantage et les inconvénients de chacune.

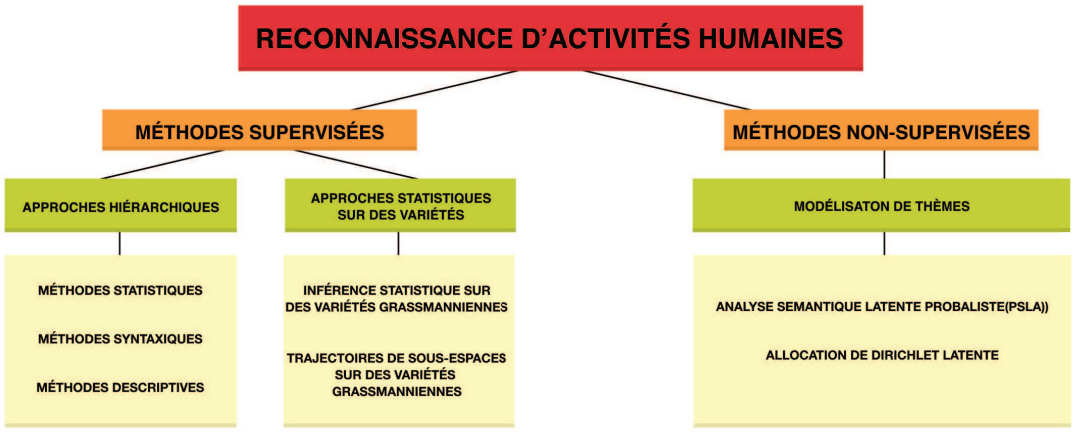

FIGURE 5.2 – Typologie des méthodes de la littérature pour la reconnaissance d'activités humaines.

5.2 Caractérisation d'activités humaines : un bref état de l'art

5.2.1 Méthodes supervisées

Nous reprenons ici un ensemble de méthodes supervisées de la littérature pour la reconnaissance d'activités humaines. On y retrouve deux grands types d'approches. D'une part, les approches hiérarchiques, qui se basent sur la détection et la hiérarchisation de sous-évènements qui composent une activités. D'autre part, les approches statistiques sur des variétés Riemannienne, qui analysent la représentation géométrique des activités sur des espaces particuliers.

5.2.1.1 Approches hiérarchiques

Les approches hiérarchiques permettent de détecter et reconnaître des sous-évènements, aussi appelés *actions atomiques* ou *actions élémentaires*, afin de les représenter de façon hiérarchique et décrire une action plus complexe d'un point de vue sémantique. Ces sous-évènements sont reconnus dans les séquences vidéos à partir d'approches à un seul niveau de représentation (présentées dans le chapitre 2). Les activités sont par la suite caractérisées par les relations structurelles inhérentes à ces sous-évènements.

L'avantage de l'aspect hiérarchique de ces approches est qu'il offre un cadre de reconnaissance compréhensible sémantiquement et calculatoirement efficace. Ces approches permettent de reconnaitre des activités avec des structures plus complexes (activités de groupe, interactions humain/objets, etc.). La représentation des activités en organisation de sous-évènements sémantiquement interprétables permet l'intégration d'information *a priori* sur ces activités. Elles nécessitent également moins de données d'entrainements que les méthodes à un seul niveau de représentation (notamment les méthodes séquentielles `HMMs` [Rabiner, 1989]) pour obtenir des résultats équivalents. En effet, l'aspect hiérarchique gère la redondance de sous-évènements présents dans les activités, tandis qu'un modèle basé `HMMs` doit retenir un plus large nombre de transitions et de probabilités d'observations [Oliver et al., 2002].

On distingue trois types de méthodes hiérarchiques utilisées pour la reconnaissance d'activités humaines : les méthodes statistiques, syntaxiques et descriptives.

Méthodes statistiques Les méthodes statistiques se servent de modèles probabilistes tels que les réseaux bayésiens dynamiques (`DBNs`) [Murphy, 2002], de réseaux de Petri [Worgan et al., 2011] ou encore les modèles de Markov cachés hiérarchiques (`HHMM`) [Karaman et al., 2014]. Ces modèles sont organisés en plusieurs couches de niveaux hiérarchiques (généralement deux). Le premier niveau est utilisé pour reconnaitre les sous-évènements composant les activités dans des vidéos. Le second niveau prend les résultats du premier et les interprète comme une structure séquentielle d'observations composant une activité.

— Olivier *et al.* [Oliver et al., 2002] ont proposé le modèle de Markov caché à deux niveaux hiérarchiques (`LHMMs`). Leur méthode permet de reconnaître deux niveaux de complexité (sous-évènements et activité) avec un modèle probabiliste de type `HMMs`. Par sa nature, ce modèle impose que les sous-évènements d'une activité soient exécutés de façon strictement séquentielle. Des activités effectuées dans des salles de conférences telles que `une personne donnant une présentation` et `conversation face à face` sont reconnues grâce à des sous-évènements tels que `une personne est présente`, `plusieurs personnes sont présentes`. Chaque niveau du `HMM` est entrainé séparément avec des données labellisées. La figure 5.3 illustre de façon intuitive le fonctionnement du modèle `LHMMs`. Chaque sous-évènement est appris par des modèles `HMMs` : `étirer la main`, `main tendue` afin de reconnaitre les transitions et observations d'une activité plus complexe `Coup de poing`.

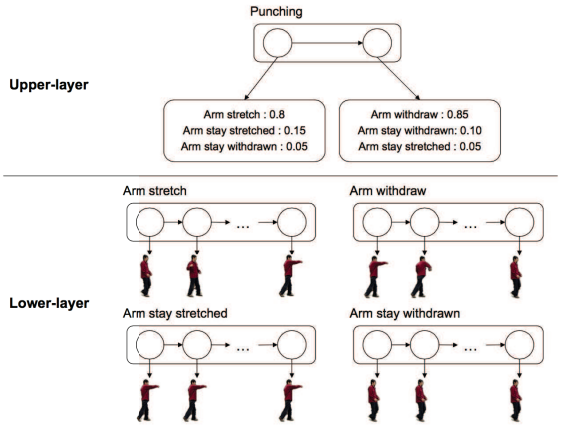

FIGURE 5.3 – Illustration du modèle `LHMMs` de Olivier *et al.*. Une action complexe `Coup de poing` est apprise à partir de modèle `HMMs` caractérisant les sous-évènements `Main tendue`, `Main étirée` (figure tirée de [Aggarwal and Ryoo, 2011]).

Méthodes syntaxiques Les méthodes syntaxiques représentent une activité comme une chaine de caractères où chaque caractère correspond à un sous-évènement. Elles s'appuient sur des outils développés dans le domaine des languages de programmation pour interpreter une activité comme un ensemble de règles permettant de générer des chaines de caractères. Les modèles de grammaire non contextuelle `CFG` (" [Nijholt, 1980], ou grammaire non contextuelle stochastique (`sCFG`) [Lari and Young, 1990] sont utilisés dans ce cadre. La production de règles issues de modèles du type `CFG` conduit logiquement à une description hiérarchique des activités.

— Ivanov *et al.* [Ivanov and Bobick, 2000] ont developpé une méthode syntaxique utilisant un modèle *HMM* pour l'apprentissage de sous-évènements dans des vidéos. Les résultats obtenus avec cette couche d'apprentissage sont convertis en une séquence d'actions simples, le niveau supérieur basé sur un modèle de type *CFG* traite ces séquences comme une chaîne de caractères en utilisant des techniques d'analyse lexicale. Une activité est finalement représentée comme une chaîne stochastique de règles. La figure 5.4 montre des exemples de gestes étudiés par cette approche lexicale. La méthode interprète les mouvements de la main et reconnait le dessin d'un carré.

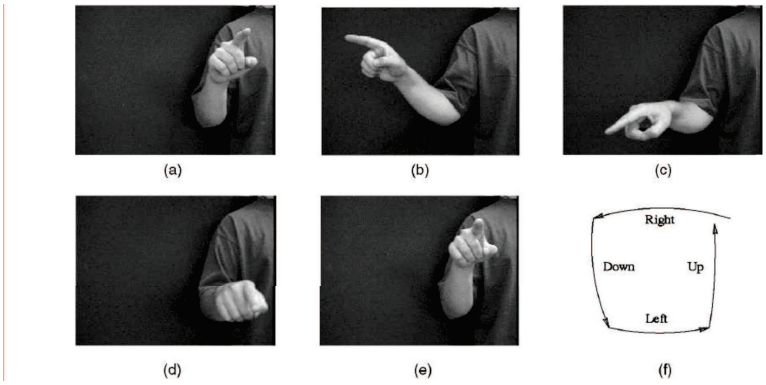

FIGURE 5.4 – Exemple de l'application de la méthode de Bobick *et al.* [Ivanov and Bobick, 2000]. Le sujet effectue à l'écran les gestes de la main formant un carré. L'action est décomposée en structure grammaticale simple : `Right, Down, Left, Up` (figure tirée de [Ivanov and Bobick, 2000]).

Méthodes descriptives Les approches hiérarchiques basées description décrivent explicitement la structure spatio-temporelle des activités humaines. Les sous-évènements d'une activité sont décrits suivant leurs relations temporelles et spatiales. Une activité est représentée comme l'occurence de ses sous-évènements dans le temps. Les prédicats temporels de Allen *et al.* [Allen and Ferguson, 1994] sont largement repris dans ces méthodes pour la représentation structurelle des activités. Ces prédicats permettent de spécifier les relations (séquentielles, co-occurrences, combinaisons) entre les sous-évènements ainsi que l'intervalle de temps entre les occurrences de ces sous-évènements. Les intervalles de temps des méthodes descriptives intègrent de l'information *a priori* permettant de gérer les cooccurrences de sous-évènements. Les modèles `CFG` sont utilisés dans ces méthodes pour formaliser la structure des sous-évènements obtenus. Ces modèles `CFG` ne sont pas utilisés comme dans les méthodes syntaxiques. Ces dernières s'en servent directement pour la reconnaissance de sous-évènements, les méthodes descriptives les utilisent pour exploiter la syntaxe formelle que fournit ces modèles. La reconnaissance d'une activité se fait en cherchant les sous-évènements qui satisfont les relations spécifiées dans sa représentation formelle. Un algorithme permettant de résoudre ce problème de satisfaction de contraintes est ensuite mis en oeuvre pour la reconnaissance.

— Ryoo *et al.* [Ryoo and Aggarwal, 2009] présentent une méthode de reconnaissance d'activités descriptive avec plusieurs niveaux de représentation. Un niveau d'extraction de silhouette corporelle, de reconnaissance de gestes et un niveau pour la caractérisation sémantique de l'activité. Les activités sont ensuite reconnues en effectuant des correspondances hiérarchiques entre les différents niveaux de représentation de l'activité. La figure 5.5 illustre la méthode et montre comment chaque partie du corps est décrite par une succession de texte.

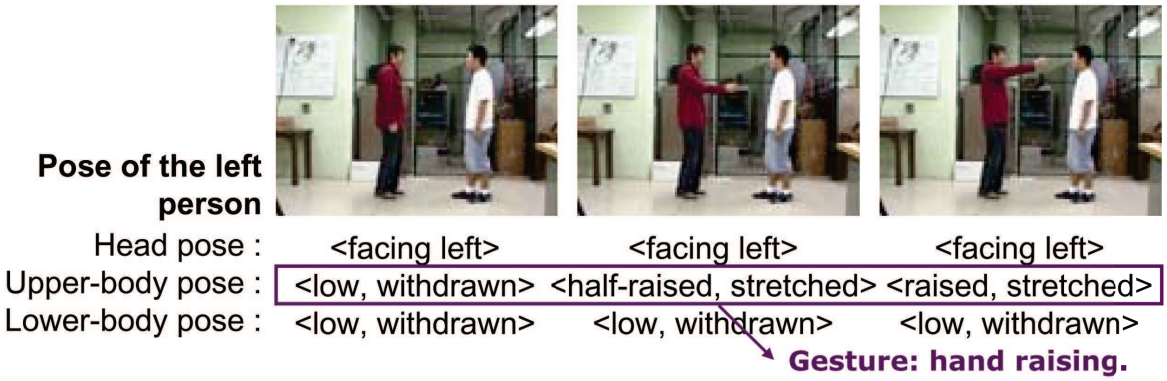

FIGURE 5.5 – Exemple de la reconnaissance du geste `Main levée` avec la méthode développée par Ryoo *et al.* Le niveau de représentation de geste permet de caractériser chaque image de la séquence. On constate que l'état des différentes parties du corps (tête, partie haute, partie basse) est décrit sémantiquement (figure tirée de [Ryoo and Aggarwal, 2009]).

Discussion La grande limitation des méthodes hiérarchiques, notamment les méthodes statistiques et syntaxiques, reste leur incapacité à représenter deux sous-évènements exécutés en même temps dans une séquence vidéo. L'aspect séquentiel strict de ces approches fait qu'elles rencontrent des difficultés à modéliser un sous-évènement A qui est exécuté, débute ou fini en même temps qu'un sous-évènement B.
Les méthodes descriptives remédient à ce problème en supposant des contraintes d'acquisitions. En effet, les méthodes de la littérature utilisent des caractéristiques issues d'approches globales pour leur premier niveau d'apprentissage (acquisition de mouvements de squelette, reconnaissance gestuelle, etc.). De plus, les méthodes descriptives compensent mal les mauvaises détections de sous-évènements dans les niveaux de représentation supérieurs. Ces éléments en font udes approches très sensibles au bruit [Hongeng et al., 2004, Siskind, 2011].

5.2.1.2 Modèles statistiques sur des variétés

Variétés de Stiefel et de Grassmann Les approches décrites par la suite se placent dans le cadre d'utilisation de variétés Riemanniennes pour caractériser une activité, notamment les variétés de Grassmann et de Stiefel. Nous profitons de cette section pour donner quelques définitions qui seront utilisées dans la suite du manuscrit.

— **Une variété de Stiefel** est l'ensemble formé par tous les d-repères orthonormés de $\mathbb{R}^n$ notée $\mathscr{S}_{n,d}$. Les éléments de cette variété sont des matrices de taille $n \times d$ dont les colonnes sont des vecteurs unitaires. $\mathscr{S}_{n,d}$ peut également être considéré comme un groupe quotient de $SO(n)$, groupe spécial orthogonal de $\mathbb{R}^n$ et donc comme une variété Riemannienne lorsqu'il est munit de la métrique g telle que :
$g(X,Y) = tr(Y^t X), \quad$ avec $X \in SO(n)$ et $Y \in SO(n)$.

En effet, en considérant les sous-groupes de rotation $SO(n-d)$ et l'application ϕ_s : $SO(n-d) \rightarrow SO(n)$ telle que :

$$\phi_s(V) = \begin{bmatrix} I_d & 0 \\ 0 & V \end{bmatrix} \quad \in SO(n) \tag{5.1}$$

$\phi_s(SO(n-d))$ est l'application qui effectue une rotation dans $SO(n)$ et qui laisse invariant les d premier éléments. Une relation d'équivalence peut être définie entre deux éléments O_1 et O_2 de $SO(n)$ telle que :

$$O_1 = O_2 \phi_s(V)$$

$O_1 \sim O_2$ si et seulement si leur d premières colonnes sont identiques. La classe d'équivalence qui en découle est :

$$[O]_s = \{O\phi_s(V) \mid V \in SO(n-d)\}$$

$\mathscr{S}_{n,d}$ est vu comme l'ensemble formé par ce type de classes d'équivalences. Il est donc noté $SO(n)/SO(n-d)$.

— **Une variété Grassmannienne** correspond à l'ensemble des sous-espaces vectoriels de dimension d de $\mathbb{R}^n$ généralement noté $\mathscr{G}_{n,d}$. Une variété Grassmannienne peut être considérée comme un sous-groupe quotient de $SO(n)$ et donc comme une variété Riemannienne lorsqu'elle est muni de la métrique g citée plus haut.
En considérant les sous-groupes produits de rotation $SO(d) \times SO(n-d)$ et l'application $\phi_g : SO(d) \times SO(n-d) \rightarrow SO(n)$ tel que :

$$\phi_g(V_1, V_2) = \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix} \quad \in SO(n) \tag{5.2}$$

Une relation d'équivalence est établi sur $SO(n)$ entre deux éléments O_1 et O_2 si :

$$O_1 = O_2 \phi_g(V_1, V_2)$$

, avec $V_1 \in SO(d)$ et $V_2 \in SO(n-d)$.
$O_1 \sim O_2$ si et seulement si les d premières colonnes de O_1 sont une rotation des d premières colonnes de O_2 et les $(n-d)$ dernières colonnes de O_1 sont une rotation des $(n-d)$ dernières colonnes de O_2. La classe d'équivalence qui en découle est :

$$[O]_g = \{O\phi_g(V_1, V_2) \mid V_1 \in SO(d), V_2 \in SO(n-d)\}$$

$\mathscr{G}_{n,d}$ correspond à l'ensemble formé par ce type de classe d'équivalence. Il est noté $SO(n)/SO(d) \times SO(n-d)$.

Application exponentielle Un outil important dans l'étude des variétés Riemanniennes est l'application exponentielle. Elle permet de paramétriser localement une variété Riemannienne M autour d'un point $p \in M$ à partir de l'espace tangent $T_p(M)$ de ce point. L'application exponentielle projette un vecteur $\mathbf{v}$ de l'espace tangent $T_p(M)$ sur la variété M en partant du point p, dans la direction du vecteur $\mathbf{v}$ avec une vitesse constante $\mid \mathbf{v} \mid$ pour une unité de temps donnée.

Soit M une variété Riemannienne, p un point de M, $T_p(M)$ l'espace tangent à M au point p et $\mathbf{v}$ un point appartenant à $T_p(M)$. L'application exponentielle, au point p, $Exp_p : T_p(M) \rightarrow M$ est telle que :

$$Exp_p(\mathbf{v}) = \alpha_{\mathbf{v}}(1)$$

avec $\alpha_{\mathbf{v}}$ l'unique géodésique satisfaisant $\alpha_{\mathbf{v}}(0) = p$ et $\alpha'_{\mathbf{v}}(0) = \mathbf{v}$.

De même, pour reprojeter dans l'espace tangent $T_p(M)$ d'un point p un élément quelconque W de la variété M on utilise l'application inverse $iExp_p : M \rightarrow T_p(M)$ telle que :

$$iExp_p(\alpha_{\mathbf{w}}(1)) = \mathbf{w}$$

avec $\mathbf{w} \in T_p(M)$, $\alpha_{\mathbf{w}}(1) = W$, $\alpha_{\mathbf{w}}$ l'unique courbe satisfaisant $\alpha_{\mathbf{w}}(0) = p$ et $\alpha'_{\mathbf{w}}(0) = \mathbf{w}$.

La figure a) 5.6 montre un exemple d'espace tangent sur une surface Riemannienne. Les points P_1 et P_2 et leur espace tangents T_{P1} et T_{P2} sont représentés ainsi que les géodésiques le long de la surface de la variété, dans la direction des vecteurs $\mathbf{V}_1$ et $\mathbf{V}_2$.

La figure b) 5.6 montre un exemple de géodésique obtenue par l'application exponentielle. $\mathbf{V}_1$ et $\mathbf{V}_2$ appartiennent à T_{P_1}, l'espace tangent au point P_1. $Exp_P(\mathbf{V}_1)$ et $Exp_P(\mathbf{V}_2)$ sont les projections des points $\mathbf{V}_1$ et $\mathbf{V}_2$ sur la variété par l'application exponentielle.

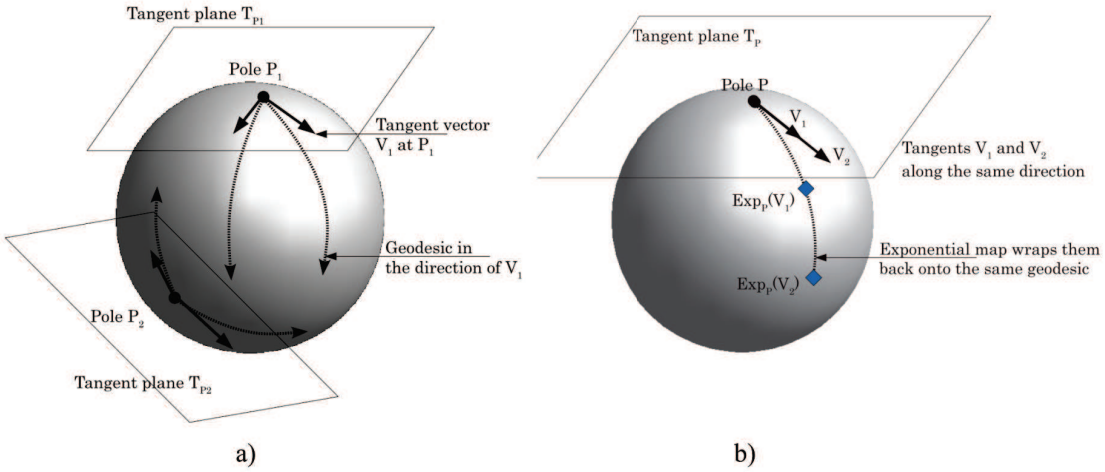

FIGURE 5.6 – Illustration de plans tangents aux points P_1 et P_2 d'une variété Riemannienne. Les vecteurs $\mathbf{V}_1$ et $\mathbf{V}_2$ appartenants aux plans tangents T_P sont projetés sur la variété le long de la géodésique passant par le point P dans la direction des vecteurs $\mathbf{V}_1$ et $\mathbf{V}_2$ (figure tirée de [Turaga et al., 2011])

Les variétés de Grassmann et de Stiefel ont été utilisées dans la littérature pour différents problèmes de vision par ordinateur (textures dynamiques [Doretto et al., 2003], reconnaissance faciale [Aggarwal et al., 2004], modélisation d'activités humaines [Veeraraghavan et al., 2005], etc). Le but de ces méthodes est de reformuler des problèmes de reconnaissance d'activités comme des problèmes d'inférence statistique sur des variétés Riemanniennes.

Les méthodes statistiques sur des variétés représentent les activités comme des évolutions de poses, représentées par des séries temporelles. Des modèles dynamiques tels que les modèles autoregressifs à moyenne mobile (`ARMA`) [Van Overschee and De Moor, 1991, Doretto et al., 2003] permettent, à partir des paramètres de ces modèles, de caractériser ces séries temporelles par des sous-espaces vectoriels de dimensions finies ([Turaga et al., 2011, Turaga and Chellappa, 2009]). Ces sous-espaces vectoriels sont des points appartenant à des variétés Grassmanniennes.

— Turaga *et al.* [Turaga et al., 2011] ont proposé une méthode de discrimination d'activité à l'aide de variétés Grassmanniennes. Les activités sont représentées par des silhouettes au cours du temps, caractérisées par des séries temporelles. Ces séries temporelles sont par la suite modélisées à l'aide d'un processus dynamique linéaire (`LV-LDS`) introduit par Turaga *et al.* permettant une meilleure représentation de l'évolution des silhouettes des sujets étudiés notamment dans le cas de changement abrupt de poses comme le montre la figure 5.7. Ces séries temporelles sont représentées par des modèles `ARMA` dont les paramètres sont projetés sur une variété Grassmannienne. L'objectif de la méthode est d'apprendre la distribution de probabilité liée aux paramètres des modèles `ARMA` sur la variété en utilisant le barycentre de Karcher [Krakowski and Manton, 2007]. Les paramètres d'une famille de densité de probabilité associés à une activité particulière sont estimés à l'aide de cette méthode. Ces estimations sont réalisées sur les espaces tangents aux données et sont ramenées sur la variété à l'aide de l'application exponentielle. Cette approche permet de caractériser des vidéos pour la reconnaissance faciale, ainsi que des activités humaines capturées selon plusieurs points de vue.

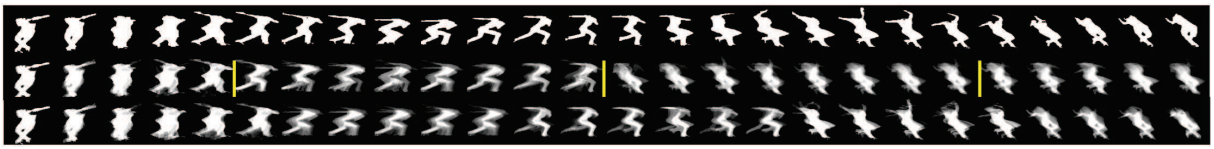

FIGURE 5.7 – Illustration de la représentation des silouhettes par la méthode de Turaga *et al.* La première ligne correspond aux silhouettes d'un joueur de golf au cours du temps. La deuxième ligne représente les silhouettes du joueur généré par un modèle `ARMA` avec les transitions entre les changements de poses représentées par des lignes verticales jaunes. La troisième ligne correspond aux silhouettes du joueurs générés par le modèle `LV-LDS` utilisé par l'approche de Turaga *et al.* [Turaga et al., 2011]. Contrairement au modèle `ARMA`, qui suppose une évolution linéaire des paramètres représentant la série temporelle associés aux silhouettes, cette modélisation permet une transition plus douce entre les différentes poses composant une activité (figure tirée de [Turaga and Chellappa, 2009]).

— Weixin *et al.* [Li et al., 2013] ont représenté une vidéo en séquence d'attributs dynamiques. Ces attributs sont obtenus à partir de modèles génératifs appelé système binaire dynamique `DBS`. Ces attributs sont agencés en sacs de mots appelés *Bag of Word for Attributes Dynamics*. Les données sont ensuite considérées comme des trajectoires des paramètres du modèle estimé sur une variété Grassmannienne. La figure 5.8 montre un exemple de caractérisation temporelle d'attribut dans une vidéo. Les courbes obtenues sont ensuite projetées sur une variété.

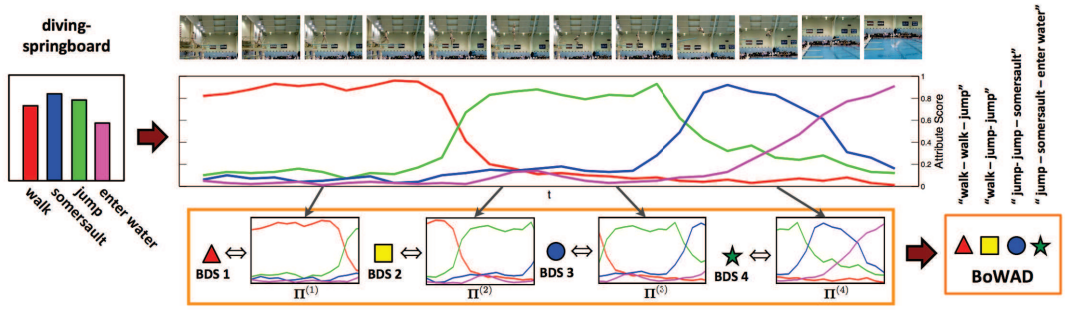

FIGURE 5.8 – Caractérisation d'attribut au cours du temps avec la méthode de Weixin *et al.*. La méthode caractérise l'évolution des proportions des attributs représentant l'activité `diving` au cours du temps (figure tirée de [Li et al., 2013]).

Discussion Les méthodes basées sur des variétés Riemanniennes obtiennent de bons résultats de reconnaissance sur des bases d'action complexes [Li et al., 2013, Turaga and Chellappa, 2009, Shirazi et al., 2012]. La prise en compte des activités d'un de vue géométrique permet une bonne généralisation des activités étudiées. Cependant, on constate que pour la construction des séries temporelles caractérisées par des modèles `ARMA`, est basée sur une estimation de poses pour chaque image d'une séquence vidéo. Globalement, ces méthodes sont utilisées sur des bases de données avec de fortes contraintes d'acquisitions, notamment des vidéos d'actions prises sous plusieurs angles de vue avec des caméras statiques.

5.2.2 Méthodes non-supervisées

Le principe des approches non-supervisées est de découvrir des activités ou sous-évènements d'activités non connues *a priori* dans des séquences vidéos. Dans ce cadre, les méthodes de modélisation de thèmes (*topic modeling*), issues du traitement automatique des documents sont employées dans le domaine de la reconnaissance d'activités ([Blei et al., 2003, Hofmann, 1999]). Nous explicitons, ci-dessous, l'espace dans lequel les méthodes de modélisation de thèmes projettent les activités, puis, nous énumérons par la suite un ensemble de méthodes de modélisation de thème de la littérature.

Variété et métrique de Fisher Dans les approches de modélisation de thèmes, les activités humaines sont représentées sous forme de vecteurs probabilités de sous-évènements. L'espace dans lequel ces probabilités sont définies est appelé une variété statistique. Une variété statistique est une variété Riemannienne dont les éléments sont des distributions de probabilités. La métrique Riemannienne associée à ces variétés est la métrique de Fisher. Cette métrique peut être interprétée comme un développement à l'ordre 2 de la divergence de Kullback-Leibler entre deux distributions de probabilités.
Soit $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_L)$ un point d'une variété statistique M de coordonnées et une distribution de probabilité $p(x \mid \boldsymbol{\theta})$ issue de la variable aléatoire X, la métrique de Fisher est définie par le tenseur métrique g tel que :

$$g_{jk}(\theta) = \int_{\mathbb{R}} \frac{\partial \log p(x \mid \boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \log p(x \mid \boldsymbol{\theta})}{\partial \theta_k} p(x \mid \boldsymbol{\theta}) dx$$

avec j et k les indices de la matrice associée au tenseur métrique g, $\boldsymbol{\theta}$ les coordonnées d'un point de la variété statistique, $\partial\theta_j$ et $\partial\theta_k$ les dérivés partielles en j et k.
Cette variété est utilisée dans le domaine de la recherche de documents et notamment dans la découverte automatique de thèmes. Comme vu précédemment, l'espace associé aux paramètres des lois multinomiales étant une variété statistique, les documents sont généralement représentés, dans ce domaine, par des distributions de probabilités de thèmes suivant une loi multinomiale. Les documents sont donc caractérisés par des proportions de thèmes, définie par des points sur cette variété statistique. La représentation géométrique de cette variété Riemannienne est un simplexe dont la dimension dépend du nombre de thèmes. La métrique de Fisher caractérise ces distributions en donnant plus d'importance aux points proches des sommets, synonyme de thèmes "purs", comme l'illustre la figure 5.9.

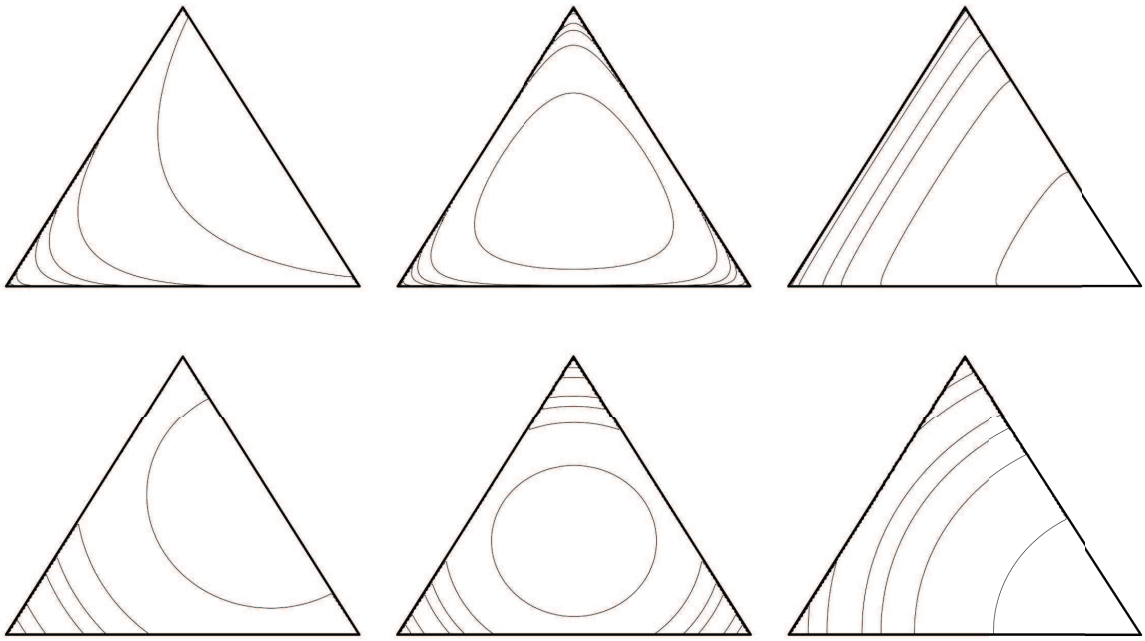

FIGURE 5.9 – Exemples de tracés dont les points sont à égale distance du centre de l'arrête de droite (colonne de gauche), le centre du simplexe (colonne du milieu) et le coté droit du simplexe (colonne de droite). La ligne du haut présente ces tracés en utilisant la métrique d'information de Fisher, celle du bas, avec la métrique Euclidienne usuelle. On constate que la métrique de Fisher donne plus d'importance aux points proches des sommets du simplexe et privilégie donc les thèmes "purs"[Lafferty and Lebanon, 2005].

5.2.2.1 Modélisation de thèmes

Les approches de modélisations de thèmes sont initialement issues du domaine de la fouille de données et de l'indexation automatique de documents textuels. Ces approches utilisent des modèles génératifs probabilistes afin d'apprendre automatiquement la distribution statistique des mots contenus dans les documents d'un corpus [Blei et al., 2003, Tavenard et al., 2013, Emonet et al., 2014]. Le but est de retrouver les thèmes sous-jacents de ces documents ainsi que leurs distributions. Les termes suivant sont classiquement employés dans ce domaine :

— Un *mot* w est l'unité de base d'un document. Il est défini comme faisant partie d'un dictionnaire indexé par $\{1, ..., V\}$.

— Un *thème* z est une variable latente issue d'une distribution sur les mots w, tel que $z \in Z = \{z_1, ..., z_K\}$. Les thèmes sont les éléments que l'on cherche à déterminer.

— Un *document* est une séquence de N mots noté par $\mathbf{d} = (w_1, w_2, ..., w_N)$ avec w_n, le $n^{\text{ième}}$ mot du document.

— Un *corpus* est une collection de M documents noté $D = \{\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_M\}$.

Il existe différents types de modèles génératifs utilisés dans la fouille de données. On dis-

tingue :

Le modèle Unigramme : [Manning et al., 2008] : représente un document $\mathbf{d}$ comme un ensemble de N mots w générés indépendamment les uns des autres. La probabilité d'un document $\mathbf{d}$ est la suivante :

$$p(\mathbf{d}) = \prod_{n=1}^{N} p(w_n)$$

Le mélange d'unigramme : introduit par Nigam *et al.* [Nigam et al., 2000] Pour un document donné, un thème z est d'abord choisi aléatoirement. Chaque mot de ce document est ensuite généré selon la probabilité conditionnelle $p(w_n \mid z)$. La probabilité d'un document $\mathbf{d}$ est :

$$p(\mathbf{d}) = \sum_{k=1}^{K} p(z_k) \prod_{n=1}^{N} p(w_n \mid z_k)$$

avec $p(w_n \mid z_k)$ la probabilité d'avoir un mot w_n sachant le thème z_k, K le nombre de thèmes et N le nombre de mots du dictionnaire.

L'analyse sémantique latente probabiliste (PSLA) : contrairement aux précédentes approches, Hofmann *et al.* [Hofmann, 1999] considèrent qu'un document peut être composé de plusieurs thèmes. Si $p(z \mid \mathbf{d})$ est la proportion d'un thème z pour le document $\mathbf{d}$, la probabilité conjointe d'un document $\mathbf{d}$ et du mot w_n est :

$$p(\mathbf{d}, w_n) = p(\mathbf{d}) \sum_{k=1}^{K} p(w_n \mid z_k) p(z_k \mid \mathbf{d})$$

Avec $p(z_k \mid \mathbf{d})$ la probabilité que le thème z_k soit dans le document $\mathbf{d}$ et $p(w_n \mid z_k)$ la probabilité d'avoir un mot w_n sachant le topic z_k.

Allocation de Dirichlet latente (LDA) : l'allocation de Dirichlet latente, proposée par Blei *et al.* [Blei et al., 2003], permet une representation plus naturelle de la distribution des thèmes compris dans un document. Contrairement aux modèles `PSLA`, les documents sont représentés comme un mélange de thèmes, où chaque thème est caractérisé par une distribution multinomiale sur les mots du dictionnaire. Les paramètres mis en jeu dans l'algorithme `LDA` sont :

— α, le paramètre d'une distribution de Dirichlet.
— $\theta = (\theta_{\mathbf{1}}, ..., \theta_{\mathbf{n}})$, les paramètres d'une loi multinomiale, qui est générée par une distribution de Dirichlet.
— β, une matrice de taille $K \times V$ qui représente la distribution des mots du dictionnaire selon un thème particulier ($\beta_{ij} = p(w_j = 1 \mid z_j = 1)$).

L'algorithme LDA suit le processus génératif suivant :
— Générer θ à partir d'une distribution de Dirichlet de paramètre (α)
— Pour chaque N mot visuel w_n d'un document $\mathbf{d}$:
1. Choisir un thème z_n à partir d'une loi multinomiale de paramètre θ
2. Choisir un mot issu du thème z_n en fonction de la probabilité $p(w_n \mid z_n, \beta)$

La distribution marginale d'un document est :

$$p(\mathbf{d} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{k=1}^{K} p(z_k \mid \theta) p(w_n \mid z_k, \beta) \right) d\theta$$

α et β sont générés une fois par corpus de documents. L'apprentissage du LDA passe par la recherche du couple (α, β) qui maximise la log-vraisemblance l telle que :

$$l(\alpha, \beta) = \sum_{\mathbf{d}=1}^{M} logP(\mathbf{w_d} \mid \alpha, \beta)$$

Cet algorithme connait un large succès dans la reconnaissance de scènes, d'objets dans des images, et d'activités humaines. Contrairement aux approches hiérarchiques supervisées qui ont du mal à traiter la cooccurrence de sous-évènements de part leur aspect strictement séquentiel, la découverte automatique de thèmes permet de caractériser un nombre Kdéfini de sous-évènements effectués en même temps dans une vidéo. En effet, les activités sont représentées par les probabilités des sous-évènements qui la composent. La figure 5.10 illustre le principe du `LDA`. Il existe un certain nombre de thèmes (topics) composés d'une collection particulière de mots (colonne gauche). Pour chaque document, on génère une distribution de thèmes (colonne de droite), puis, pour chaque mot de ce document, un thème est choisi suivant cette distribution (points colorés). Un mot correspondant à ce thème est ensuite sélectionné et intégré au document.

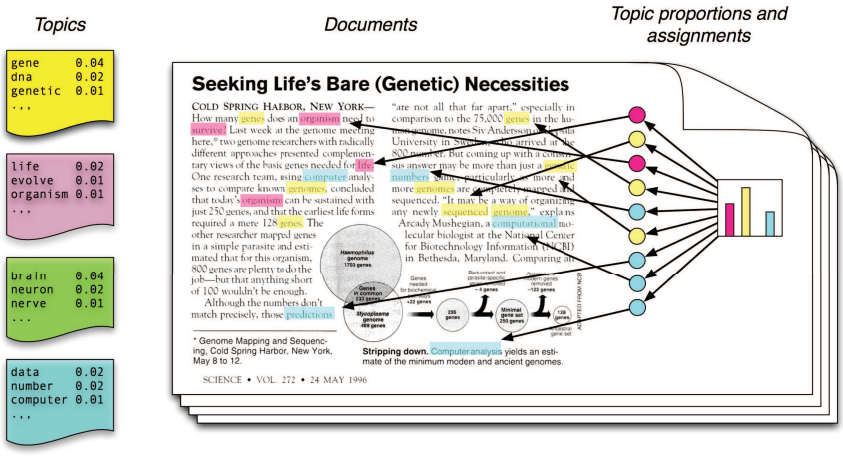

FIGURE 5.10 – Illustration du principe de l'algorithme LDA. On fait ressortir du document des mots associés au thème de la génétique (jaune), biologie (violet), neurologie (vert) et informatique (cyan). L'algorithme représente le document comme une proportion de mots associés à ces différents thèmes. (figure tirée de [Blei et al., 2003].)

— Niebles *et al.* [Niebles et al., 2008] ont utilisé l'algorithme `PSLA` et `LDA` pour apprendre des actions humaines dans des vidéos de façon non-supervisée. Les auteurs utilisent le détecteur et descripteur cuboïd [Dollar et al., 2005] pour construire leur sac de mot (`BoW`). Pour la caractérisation d'activités, plus longues dans le temps, la méthode est utilisée sur plusieurs sous-séquences temporelles de la vidéo. Les thèmes sont donc découverts au cours du temps afin de caractériser une video comme un enchainement temporel de thèmes. La figure 5.11 illustre différents "patchs" spatio-temporels détectés sur des vidéos de la base `KTH Dataset`. La couleur est attribuée en fonction de la classe d'action la plus probable auquelle appartient ce patch. On constate que les activités dans les vidéos sont donc décomposées en proportion de sous-évènements élémentaires.

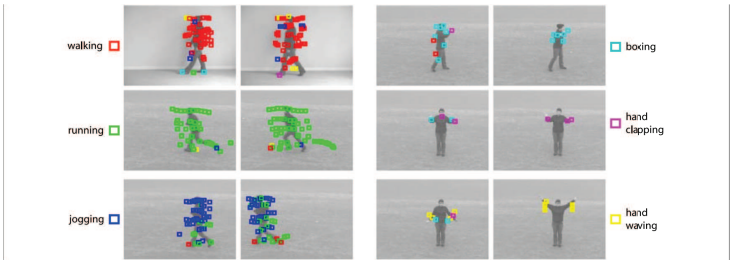

FIGURE 5.11 – Des Cuboïd sont extraits par la méthode de J.C Niebles *et al.* sur la base `KTH Dataset`. Les sous-évènements sont automatiquement découverts par la méthode et chaque cuboïd est coloré en fonction de la classe d'appartenance la plus probable. Cela permet d'attribuer des probabilités d'apparition de sous-évènements dans chaque vidéo en fonction de la proportion de "patchs" de différentes classes détectés (figure tirée de [Niebles et al., 2008]).

— Y.Wang *et al.* [Wang et al., 2007] ont développé la méthode Semi-Latent Dirichlet Allocation (`S-LDA`), une version semi-supervisée du *LDA*. La grande différence entre le `S-LDA` et la version originale du `LDA` est que la variable z_n, correspondant aux thèmes du modèle, est connue durant le processus d'apprentissage du modèle. L'avantage de cette méthode est qu'elle permet d'intégrer l'information provenant des classes d'actions directement dans le modèle durant l'apprentissage. Il y a donc, par construction, une correspondance entre les thèmes générés par le `S-LDA` et les classes d'actions connues. La figure 5.12 illustre le modèle graphique du `S-LDA`. On constate que la variable z représentant les thèmes, est observée, contrairement à l'approche initiale du `LDA`.

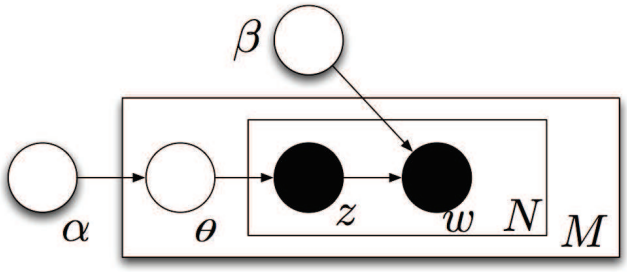

FIGURE 5.12 – Représentation graphique du `Semi-Latent Dirichlet Allocation` lors de la phase d'entrainement de l'algorithme. La pastille noir associée à la variable z montre que cette dernière est observée, contrairement à l'algorithme `LDA`. (figure tirée de [Wang et al., 2007]).

Discussion En dépit de leurs avantages, les méthodes de modélisation de thèmes rencontrent certains problèmes dans le cadre de la reconnaissance d'activités humaines. D'une part, il est difficile de déterminer de façon optimale le nombre K de thèmes à choisir pour l'apprentissage des modèles génératifs. Deuxièmement, il est difficile de décrire sémantiquement les thèmes découverts par le LDA. Il n'y a aucune garantie que les thèmes générés correspondent aux classes d'actions des séquences vidéos. Cela est d'autant plus critique lorsqu'on trouve un nombre de thèmes K optimal différent du nombre de classes d'actions prédéfini.
La construction du "sac de mots" des méthodes de la littérature s'appuie sur des descripteurs globaux ou des descripteurs qui supposent peu de mouvements de caméra ([Niebles et al., 2008, Wang et al., 2007], ce qui est rarement le cas pour des vidéos réalistes. D'autres méthodes [Wang et al., 2007] utilisent des pré-traitements coûteux en termes de temps de calcul (suivi de cible, stabilisation de scène, etc) ce qui les rend difficilement applicables dans des contextes d'utilisations en temps-réel.
Enfin, ces méthodes restent limitées en terme de résultats pour la caractérisation de sous-évènements. Sur des bases de données d'actions élémentaires telles que `KTH Dataset` ou Weizmann, elles présentent des résultats très peu élevés par rapport à d'autres méthodes à un seul niveau de représentation [Niebles et al., 2008, Wang et al., 2007, Tavenard et al., 2013]. Blei *et al.* dans [Blei et al., 2003] insistent sur les limitations des modèles probabilistes quand il s'agit de classification de données. On constate effectivement qu'un classifieur SVM entrainé sur une représentation en mot visuels est plus efficace qu'un classifieur SVM entrainé sur une représentation en thèmes, qui est cependant plus compacte.

5.2.3 Conclusion

Ce point sur les méthodes de l'état de l'art nous permet de dégager les éléments essentiels des méthodes de reconnaissance d'activités humaines pertinentes :

— **Être capable de caractériser la structure intrinsèque des activités**
L'organisation en sous-évènements au cours du temps est l'un des principaux attributs d'une activité. Cette caractéristique est largement exploitée dans la littérature, que cette structure soit représentée d'un point de vue statistique, linguistique informatique, sémantique ou géométrique.

— **Être capable de généraliser la représentation des activités**
Les activités humaines sont généralement effectuées sur de longues durées. Entre deux exemples vidéos représentant une même activité, la structure temporelle de ces vidéos peut parfois varier (légère redondance de sous-évènements, durées d'exécution différentes, etc). Généraliser la représentation structurelle des activités revient à être le plus robuste possible aux petites variations de structure.

— **Être applicable dans des cas variés d'activités humaines**
Beaucoup de méthodes de la littérature s'intègrent dans des contextes d'applications très précis (salle de conférence, restaurant, parking,...). Ces méthodes sont optimisées pour un environnement donné et sont difficilement généralisables dans d'autres contextes.

— **Être adaptée aux vidéos avec peu de contraintes d'acquisitions**
La capacité à représenter et généraliser des activités humaines variées implique que ces activités doivent être reconnues et caractérisées dans divers contextes d'acquisition. Le besoin de traiter et analyser ces vidéos, avec peu de contraintes d'acquisition, est de plus en plus important de part l'augmentation de données vidéos issues de téléphone portable.

— **Être robuste pour la reconnaissance de sous-évènements**
Reconnaitre des activités dans des vidéos génériques suppose que les sous-évènements qui les composent soient caractérisés de façon robuste. Cela comprend l'invariance à différentes variations visuelles et géométriques et la capacité à reconnaître un même sous-évènement dans différents contextes d'acquisitions.

— **Être efficace en temps de calcul**
La reconnaissance d'activités humaines trouve son utilisation directe dans beaucoup d'applications en temps-réel (vidéo-surveillance, aide aux personnes agées,...). Ce type d'applications impose donc un temps de calcul réduit, notamment pour la caractérisation des sous-évènements qui composent ces activités.

La section suivante présente l'approche que nous proposons dans le cadre de la reconnaissance d'activités humaines. Cette approche propose une représentation originale et efficace de la structure des activités. Cette représentation permet une généralisation naturelle des activités humaines. Pour rester dans un cadre de reconnaissance d'activités dans des contextes variés et divers, cette approche s'appuie sur l'utilisation de notre méthode de reconnaissance d'actions humaines élémentaires présentée au Chapitre 3.

5.3 Décomposition d'activités en séquences d'actions élémentaires

Dans la précédente section, nous avons vu que la structure temporelle est un élément primordial pour l'analyse des activités humaines dans des vidéos. La structure temporelle d'une activité est essentiellement composée d'une succession temporelle de sous-évènements, suivant un ordre propre à l'activité étudiée. Cette structure s'articule entre les sous-évènements comme un "chemin" unique ou faiblement variable, caractéristique d'une activité. Dans cette section, nous décrivons comment nous caractérisons ce chemin temporel afin de le placer dans un cadre où la généralisation des activités humaines se fait de façon intuitive. En effet, nous verrons comment cette structure, extraite selon un formalisme stochastique, sera par la suite caractérisée dans un cadre déterministe.

5.3.1 Contributions

La méthode présentée dans ce chapitre repose sur la structure d'une activité et sur l'estimation, au cours du temps, de la proportion des actions élémentaires qui la compose. Ces actions élémentaires sont caractérisées avec notre méthode discriminative de reconnaissance présentée au chapitre 3. Elle permet d'obtenir un premier niveau de représentation des activités, invariant à diverses transformations géométriques, robuste aux mouvements de caméra et donc pertinent quant à la caractérisation de vidéos avec peu de contraintes d'acquisition.

L'une des particularités de cette approche est que l'entrainement du classifieur de notre premier niveau de représentation se fait avec un mélange de bases de données d'actions élémentaires. L'intérêt d'utiliser un mélange de base de données pour l'aspect générique de la reconnaissance des actions élémentaires est montré dans cette section.

Une autre originalité de cette méthode repose sur la façon de caractériser la structure inhérente aux activités. Elles sont projetées en tant que courbes dans un espace géométrique sémantique défini par les actions élémentaires. Le "chemin" représentant cette structure devient donc une trajectoire dans cet espace, défini par les proportions d'actions élémentaires au cours du temps. Ces activités sont par la suite analysées et discriminées en utilisant la géométrie de ces trajectoires. Nous verrons que cette étape permet une généralisation très naturelle et pertinente des activités étudiées. Cette caractérisation est effectuée en respectant les propriétés géométriques de l'espace sémantique utilisé.

5.3.2 Apprentissage des actions élémentaires sur un mélange de bases de données

5.3.2.1 Méthode de reconnaissance d'actions élémentaires

La section précédente a montré que le principal inconvénient des approches de reconnaissance d'activités est leur application restreinte à des environnements contrôlés (méthodes hiérarchiques) ou leur hypothèse de forte contraintes de capture des vidéos (méthodes de modélisation de thèmes). La caractérisation des actions élémentaires de vidéos issues de différentes sources d'acquisition dans des contextes variés est un enjeux du domaine. Pour tenter d'y répondre, nous utilisons notre approche de reconnaissance présentée dans le chapitre 3. Cette approche est utilisée pour deux raisons :

1. elle a montré sa robustesse sur différents types de base de données, avec et sans contraintes de captation notamment de part la possibilité de caractériser des mouvements à différentes échelles comme le montre la figure 5.13. Elle assure donc une reconnaissance fiable des actions élémentaires dans des vidéos,
2. le chapitre 4 a permis de démontrer la capacité de cette méthode à généraliser la représentation des actions humaines à travers des expérimentations sur des mélanges de base de données.

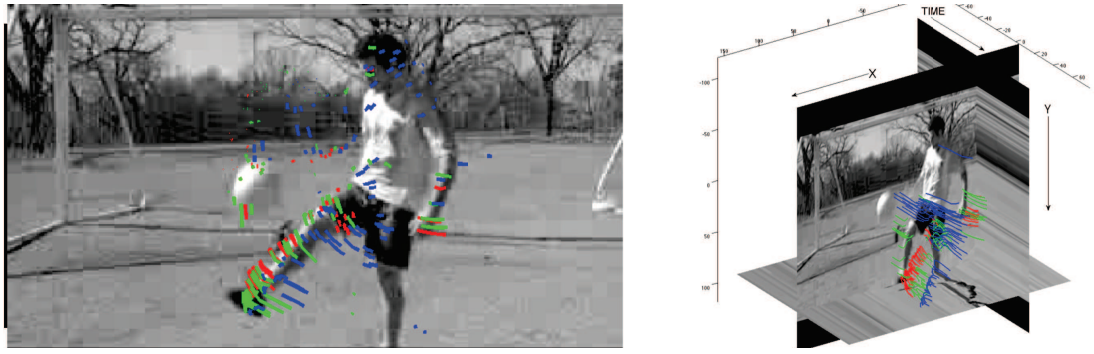

FIGURE 5.13 – Exemple de trajectoires multi-échelles extraites avec notre méthode de reconnaissance d'actions élémentaires sur une vidéo de jongles de football de la base `UCF-11` (voir vidéo).

Cette méthode est donc pertinente pour fournir un premier niveau de représentation sémantique robuste à différentes variations. L'originalité de cette étape réside dans l'entrainement de ce premier niveau avec un mélange de bases de données, contrairement à ce qui se fait généralement dans la littérature. Le concept de mélange de bases de données exposé au chapitre 4 est ici repris pour apporter une représentation des actions élémentaires la plus générique possible.

5.3.2.2 Apprentissage par mélange de bases de données

L'estimation des actions élémentaires étant effectuée sur des vidéos requêtes issues de différents contextes, le classifieur de notre méthode doit donc être entrainé avec différents types de représentation. Et ce, d'autant que la plupart des bases utilisées pour la reconnaissance possèdent d'important biais visuels. On peut pointer différentes causes menant à ces biais visuels : le biais de sélection (source des données), le biais d'acquisition (contraintes d'acquisitions, habitudes de capture), le biais de l'ensemble négatif d'apprentissage (représentation de l'ensemble du reste du monde) (voir chapitre 4). L'entrainement du classifieur avec un mélange de bases de données permet une plus grande robustesse pour la reconnaissance d'actions élémentaires dans des vidéos issues de différentes sources.

Construction de la base hybride Pour constituer ce mélange de bases de données (base hybride), nous utilisons deux bases de vidéos avec contraintes d'acquisitions (`KTH Dataset`, `Weizmann Dataset`) et deux bases de vidéos génériques (`UCF-11 Dataset`, `UCF-50 Dataset`). Ces bases de données fournissent deux types de représentation d'actions élémentaires.

Les bases `KTH` et `Weizmann`, possèdent des vidéos où les actions élémentaires sont exécutées dans un environnement contraint (caméra statique, fond uniforme, etc). La plupart des mouvements sont joués de manière répétitive et peu naturelle (par exemple l'action `boxe` dans `KTH Dataset` ou `handwave` dans `Weizmann Dataset`).
Dans `UCF-11` et `UCF-50`, les actions élémentaires sont capturées dans des situations de la vie quotidienne. Les mouvements de caméra, changement de points de vue et occultations partielles y sont très présents.
Ces deux catégories de bases de données, avec et sans contraintes d'acquisition, contiennent des informations complémentaires. L'une permet d'obtenir une représentation précise des mouvements correspondant aux actions élémentaires. L'autre présente ces actions dans des contextes d'exécution naturels et permet de renforcer la robustesse du classifieur en apportant de la variabilité visuelle. La figure 5.14 montre un exemple des vidéos utilisées sur chacune des bases de données.

Actions élémentaires sélectionnées Pour la construction du premier niveau de représentation, le classifieur est entrainé avec les actions élémentaires `Jump` (saut), `Run` (course), `Walk` (marche), et `Handwave` (mouvement de la main). Ces actions élémentaires font partie de nombreuses activités humaines du quotidien et se retrouvent aussi dans des actions complexes tels que les activités sportives. Le classifieur est entrainé sur ces actions afin d'obtenir une base de représentation à la fois décorrélée et la plus générale possible pour constituer cette base hybride. Un ensemble de 32 vidéos par classe est sélectionné.

FIGURE 5.14 – Actions élémentaires `Walk`, `Run`, `Jump`, et `Handwave` issues des bases de données `UCF-11 Dataset`, `UCF-50 Dataset`, `Weizmann Dataset` et `KTH Dataset`.

Apprentissage des actions élémentaires Le Tableau 5.1 montre les résultats obtenus après une validation-croisée de type *Leave-One-Out* avec notre classifieur entrainé sur ce mélange de bases de données. Le taux de reconnaissance est de 96.87%. Les rares confusions apparaissent entre les classes sémantiquement proches (`Run`, `Walk`). L'information apportée par chaque descripteur est combinée à travers le processus de fusion tardive Adaboost [Hastie et al., 2009] introduit dans le chapitre 3. Le taux de reconnaissance par descripteur et les pondérations obtenues avec l'algorithme Adaboost sont présentés dans le Tableau 5.2. Le poids correspondant au descripteur `HOG` est le plus faible parmi ces trois descripteurs. En effet, mélanger des vidéos de différentes bases de données augmente la variabilité visuelle et affaibli la pertinence de l'information issue du gradient (`HOG`), tandis que l'information liée aux variations de mouvement reste relativement stable.

action	`Saut`	`Marche`	`Course`	`Mvt. de la main`
`Saut`	90,62%	0%	9,37%	0%
`Marche`	0%	100%	0%	0%
`Course`	0%	3,12 %	96,87%	0%
`Mvt. de la main`	0%	0%	0%	100%

TABLE 5.1 – Matrice de confusion après une validation-croisée de notre méthode de reconnaissance sur les actions élémentaires.

Descripteurs	FCD	HOF	HOG
Taux de rec.	94,53%	95,31%	53,12%
Poids Adaboost	1,82	2,91	0,42

TABLE 5.2 – Taux de reconaissance par descripteur et poids obtenus après fusion tardive avec Adaboost.

5.3.3 Fenêtre d'observation des actions élémentaires

Pour calculer la proportion d'actions élémentaires contenue dans une séquence vidéo, les décisions du classifieur SVM de notre méthode de reconnaissance d'actions élémentaires sont transformées en probabilités *a posteriori* suivant l'approche proposée [Wu et al., 2004]. On évalue donc les probabilités de chaque action élémentaire estimées dans une image.
Pour intégrer une structure temporelle à cette estimation, ces probabilités sont estimées dans une fenêtre temporelle glissante. Une image t de la séquence est donc caractérisée par la proportion d'actions élémentaires contenue dans la fenêtre temporelle de taille $[t-N; t+N]$ avec $N \in [6, 10]$. On suppose donc ici que les actions élémentaires sont effectuées sur de courtes périodes temporelles. Cette supposition s'appuie sur les travaux de [Schindler and Van Gool, 2008] qui démontrent que sur des bases de vidéos d'actions élémentaires, très peu d'images sont nécessaires pour atteindre un taux correct de bonne reconnaissance.

La Figure 5.15 présente l'application de notre approche sur une vidéo issue de la base `Weizmann Dataset`. L'action exécutée est `Jack`, une action composée à la fois d'un saut alterné et d'un mouvement des bras du haut vers le bas. Les courbes du graphe représentent l'évolution de la proportion des actions élémentaires au cours du temps. La courbe cyan correspond à l'action `Mouvement de la main`, la courbe bleue à l'action `Saut`, la courbe verte à l'action `Marche` et la courbe rouge à l'action `Course`.
On constate que l'on retrouve sur ce graphe à la fois la périodicité du mouvement exécuté sur la séquence et l'alternance entre une proportion forte des actions `Mouvement de la main` et `Saut`, qui caractérisent l'action `Jack`.

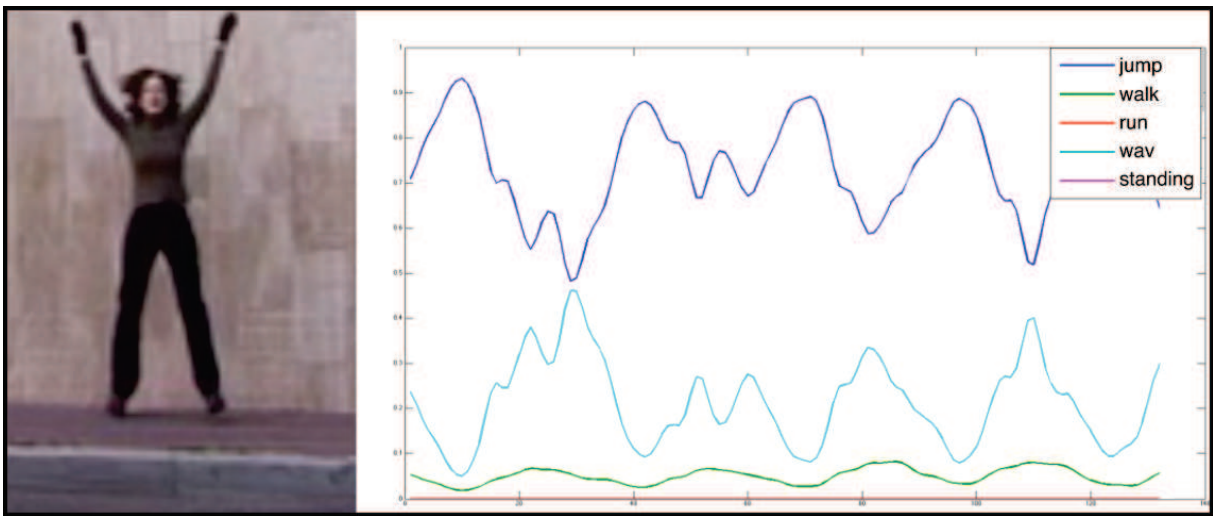

FIGURE 5.15 – L'action complexe `Jack` de `Weizmann Dataset` et sa représentation en séquence de probabilités d'actions élémentaires. La périodicité du mouvement effectué dans la vidéo se retrouve bien à travers l'évolution des courbes de probabilités (voir vidéo)

La Figure 5.16 illustre l'analyse des probabilités de ces mêmes actions élémentaires sur une séquence vidéo représentant l'action complexe `Tir au panier` issue de `UCF-11 Dataset`. L'image avec la barre verte pointe le moment où le joueur lève ses mains et se prépare au tir. Ce mouvement est caractérisé sur le graphe par une forte proportion de l'action élémentaire `Mouvement de la main` au même instant. L'image avec la barre bleue correspond au moment où le joueur effectue un saut. A l'instant correspondant sur le graphe, on distingue une forte probabilité de l'action élémentaire `Saut`.

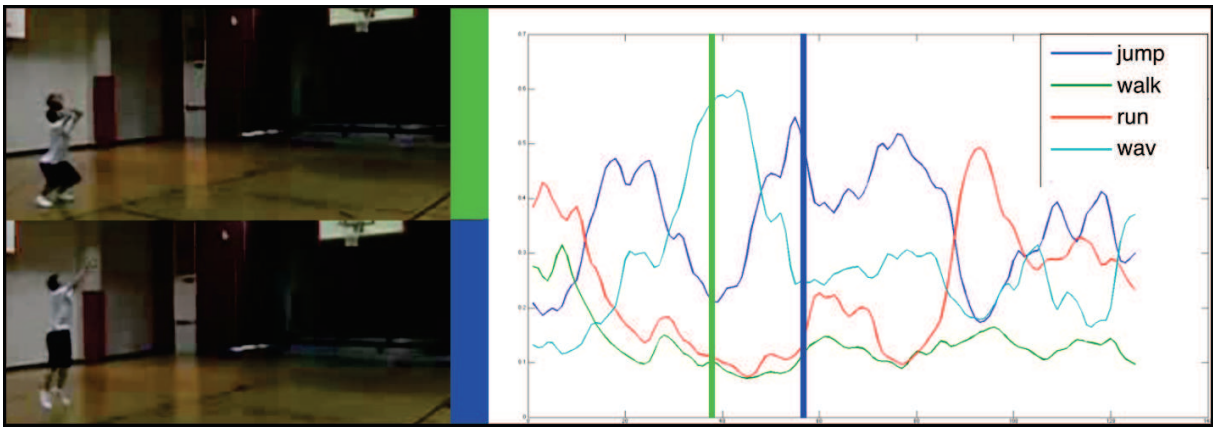

FIGURE 5.16 – L'action complexe `Basketball` de `UCF-11 Dataset` et sa représentation en séquence de probabilités d'actions élémentaires. On constate la correspondance entre les mouvements effectués sur la vidéo et l'estimation au cours du temps des probabilités d'actions élémentaires (voir vidéo).

Ces exemples montrent que notre méthode permet une représentation pertinente des actions élémentaires. En effet, ces actions sont apprises à partir d'un mélange de données issues de bases différentes, et nous constatons que ces actions sont retrouvées, sur de courtes périodes, tout au long de vidéos génériques. L'intérêt de cette approche est la correspondance naturelle entre les actions exécutées dans la vidéo et les probabilités des actions élémentaires apprises par notre modèle.

5.3.4 Caractérisation de l'absence d'action

L'un des points fort des méthodes supervisées de reconnaissance d'activités est leur capacité à gérer la cooccurrence des actions élémentaires. En effet, ces situations permettent de mettre en évidence des actions non entrainées par le classifieur qui sont alors représentées par des mélanges pertinents d'actions élémentaires connues. Utiliser des séquences de probabilités d'actions élémentaires permet de gérer les cas de cooccurrence d'actions élémentaires.

Dans d'autres cas, il peut arriver qu'aucun mouvement ne soit effectué dans la séquence. Le flot optique lié aux images de la séquence correspondante est quasi-nul. Dans ces cas de figure il est important d'adapter la réponse du classifieur. En effet, l'action la plus probable estimée par le classifieur ne correspond pas avec ce qui est visuellement représenté dans la séquence vidéo de part la normalisation des probabilités d'apparition des actions. De plus, une légère variation de mouvement entraine, dans ce cas, de grosses perturbations quant aux valeurs des probabilités

estimées. La Figure 5.17 illustre les problèmes que l'on rencontre dans ce cas. En effet quand le sujet n'effectue aucune action dans la séquence, le classifieur donne des résultats incohérents ou chaotiques du fait de l'absence de mouvement. L'utilisation de la classe `Standing` permet une meilleure représentation des actions élémentaires contenues dans la séquence, en évitant les problèmes de mauvaise estimation dûs à la normalisation des probabilités d'actions

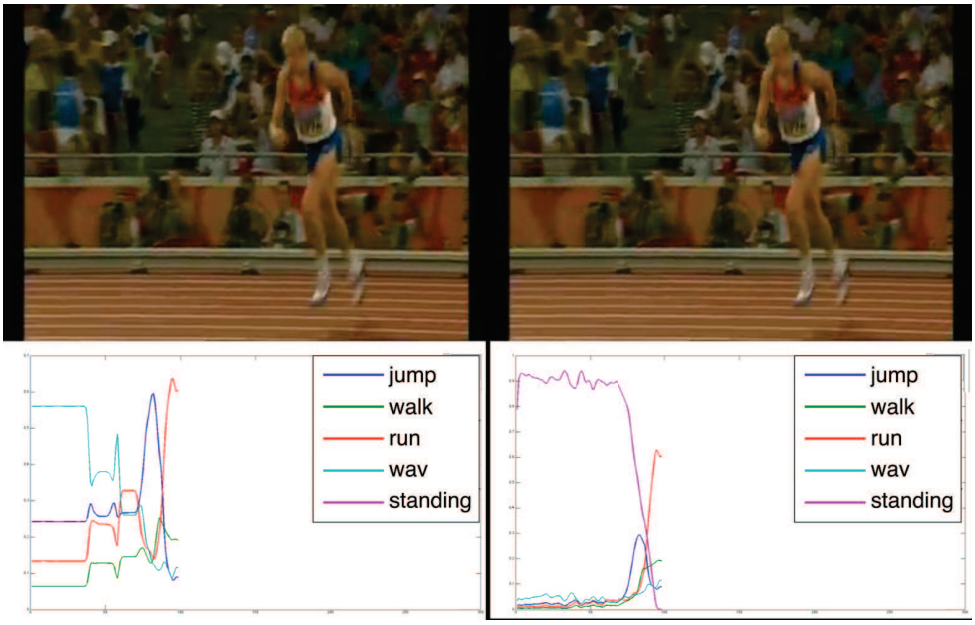

FIGURE 5.17 – Avant et après l'utilisation de classe `Standing` (respectivement image gauche et image droite). La méthode caractérise bien au début de la séquence l'absence d'action qui, initialement, génère des résultats incohérents (voir vidéo).

Pour évaluer l'absence ou la présence de mouvement, la quantité de mouvement contenue dans la fenêtre temporelle est estimée en se basant sur l'énergie du flot optique. Chaque image de la séquence est subdivisée horizontalement et verticalement et l'énergie moyenne du flot optique est calculée sur chacun de ces blocs. Au final, à chaque image k de la séquence est associée une valeur $coef_{standing}(t) \in [\![0,1]\!]$. Une valeur élevé de coefficient correspond à une forte présence d'action dans le voisinage d'une image k et inversement. Les probabilités d'actions en sortie du classifieur sont alors renormalisées. On introduit donc une classe, artificiellement générée, que l'on nomme `Standing`. Cette classe permet de determiner la présence ou non de mouvement dans la séquence à un instant t. Le vecteur de probabilités *a posteriori* d'actions élémentaires devient après renormalisation :

$$Prob_{estimates}(t) = [coef_{standing}(t) * (\lambda_1(t), ..., \lambda_k(t), ..., \lambda_L(t)), 1 - coef_{standing}(t)]$$

avec $\lambda_k(t)$ la probabilité de l'action élémentaire k au temps t.

La Figure 5.18 illustre l'analyse au cours du temps des probabilités des actions élémentaires ainsi que de la classe `Standing` sur une séquence vidéo comportant l'action complexe `Saut en hauteur` issue de la base de données `Olympic dataset`.

Le moment où l'athlète s'apprête à s'élancer est indiqué par une barre verticale magenta. Il y a très peu de mouvement à ce moment de la séquence, on constate bien que la classe majoritaire correspondante est la classe `Standing`. Le moment où l'athlète est en pleine course est indiqué par une barre verticale rouge. L'action élémentaire `Course` est majoritaire sur cette période. Enfin en bleu, on a l'instant où l'athlète effectue son saut par-dessus la barre. L'action `Saut` est celle qui a la plus forte probabilité à cet instant de la séquence.

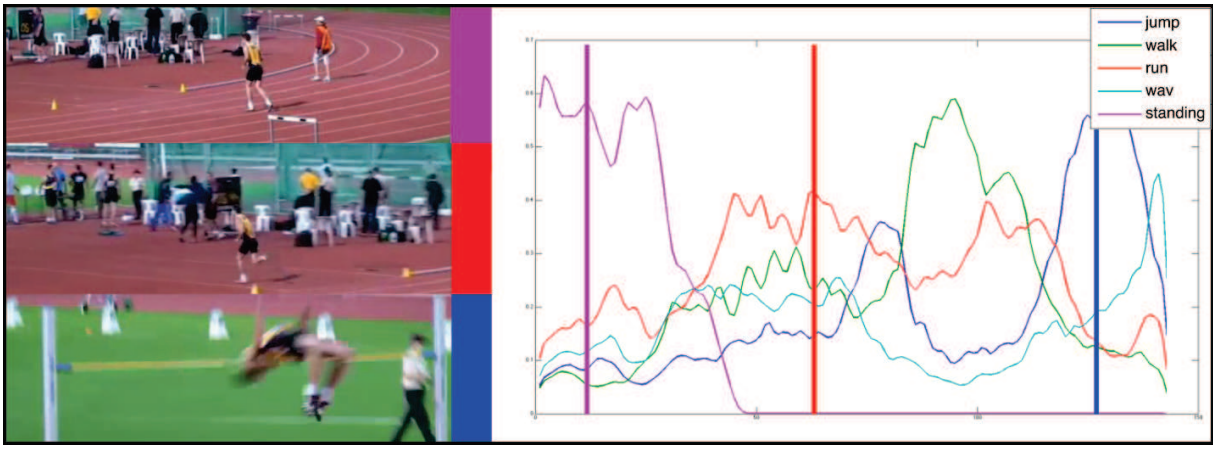

FIGURE 5.18 – L'action complexe `Saut en hauteur` et sa représentation en séquence de probabilités d'actions élémentaires. L'inactivité de l'athlète est bien caractérisée par la classe `Standing` (courbe magenta) en début de séquence (voir vidéo).

5.4 Caractérisation des trajectoires d'actions dans le simplexe

5.4.1 Trajectoires d'activités dans l'espace sémantique des actions élémentaires

Les séquences de probabilités *a posteriori* estimées dans la section précédente décrivent la proportion des actions élémentaires contenues dans une activité humaine. L'évolution de ces proportions au cours du temps décrit la structure propre des activités humaines comme un "chemin" s'articulant de façon continue et lisse entre les actions élémentaires. Pour analyser cette structure, nous projetons ces séquences dans un espace qui est défini par les actions élémentaires. Ce processus est résumé par la figure 5.19.

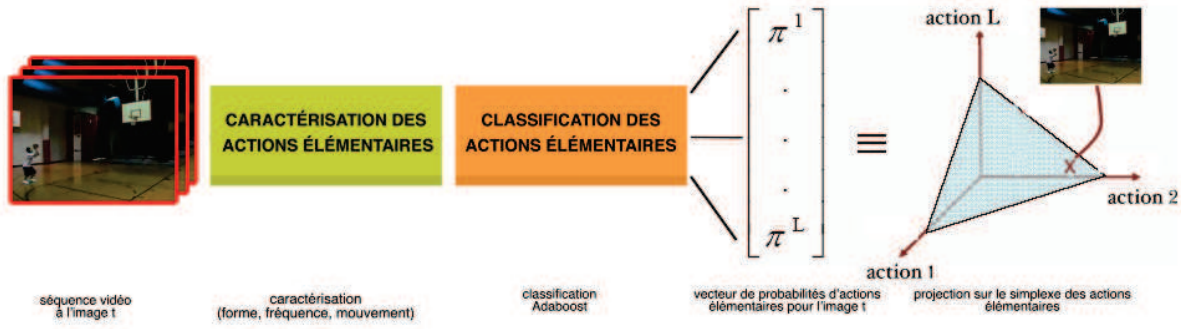

FIGURE 5.19 – Processus global de caractérisation d'activités.

5.4.1.1 Variétés statistiques

Simplexe sémantique La proportion d'actions élémentaires pour une image t est considérée comme une distribution de probabilités. Les activités évoluent donc dans l'espace des paramètres des lois multinomiales. Cet espace est représenté par un simplexe de taille L, qui est une variété statistique. La valeur de L est définie en fonction du nombre d'actions élémentaires apprises. Ces variétés ont été présentées dans la sous-section 5.2.2 par rapport aux approches de modélisation de thèmes. Dans notre cas, chaque sommet de cette variété statistique correspond à une probabilité $p = 1$ d'obtenir une action élémentaire donnée. Le simplexe $\mathcal{P}_L$ est défini comme suit :

$$\mathcal{P}_L = \{\boldsymbol{\theta} \in \mathbb{R}^{L+1} | \sum_{i=1}^{L+1} \theta_i = 1, \theta > 0\}, \quad \boldsymbol{\theta} = (\theta_1, ..., \theta_{L+1})$$

Isométrie entre simplexe et hyper-sphère positive Les distances géodésiques sont généralement difficiles à calculer et font appel à des méthodes d'optimisations coûteuses. Dans notre cas, la métrique de Fisher, associée au simplexe $\mathcal{P}_L$, correspond à la métrique Euclidienne à la surface de l'hypersphère positive $\mathcal{S}_L^+$. Cette transformation est obtenue par le difféomorphisme $\mathcal{F}$ tel que :

$$\mathcal{F} : \begin{cases} \mathcal{P}_L \rightarrow \mathcal{S}_L^+ \\ \boldsymbol{\theta} = (\theta_1, ..., \theta L + 1) \rightarrow \boldsymbol{\pi} = (2\sqrt{\theta_1}, ..., 2\sqrt{\theta_{L+1}}) \end{cases}$$

avec :

$$\mathcal{S}_L^+ = \{\boldsymbol{\pi} \in \mathbb{R}^{L+1} | \sum_{i=1}^{L+1} \pi_i^2 = 2, \pi > 0\},$$

La transformation $\mathcal{F} : \mathcal{P}_L \rightarrow \mathcal{S}_L^+$ étant une isométrie, la distance géodésique entre deux points (π_{k1}, π_{k2}) peut être calculée comme la plus petite courbe sur $\mathcal{S}_L^+$ qui connecte $(\mathbb{F}(\pi_{k1}), \mathbb{F}(\pi_{k2}))$ (voir Figure 5.20). Ces courbes sont les arcs de grands cercles sur $\mathcal{S}_L^+$ [Lafferty and Lebanon, 2005], leur longueur est donnée par la formule :

$$d_{\mathcal{S}_L^+}(\mathcal{F}(\theta_{k1}), \mathcal{F}(\theta_{k2})) = d_{\mathcal{S}_L^+}(\pi_{k1}, \pi_{k2}) = 2\cos^{-1}(\pi_{k1}\pi_{k2}^\top/4)$$

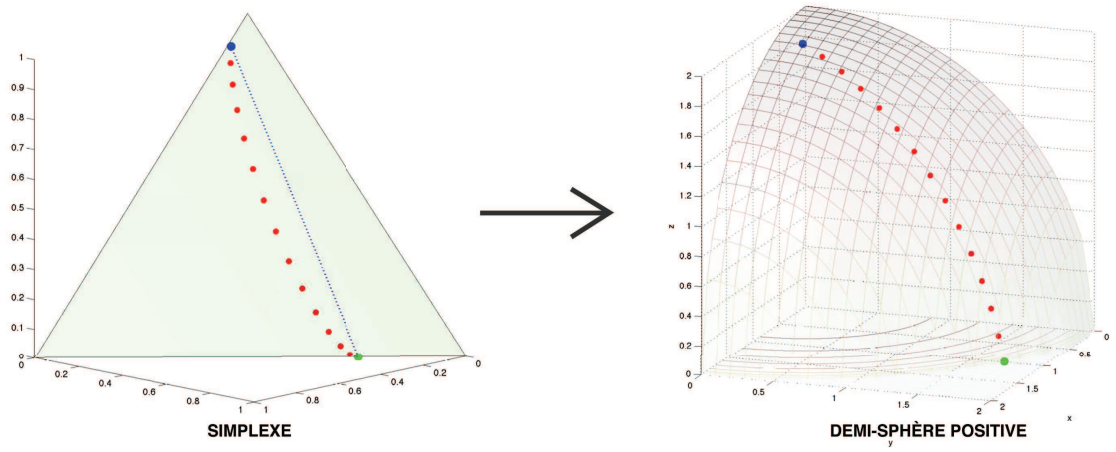

FIGURE 5.20 – Isométrie entre le simplexe $\mathcal{P}_L$ et la demi-sphère positive $\mathcal{S}_L^+$. On constate que la géodésique suivant la métrique de Fisher, entre deux points du simplexe correspond à l'arc de grand cercle reliant ces deux points sur l'hyper-sphère positive.

5.4.1.2 Trajectoires d'activités

Grâce à cette isométrie, l'étude des trajectoires d'activités sur $\mathcal{P}_L$ revient à étudier les trajectoires sur $\mathcal{S}_L^+$, plus interprétables et calculatoirement plus efficaces à extraire. Les trajectoires d'activités sur la demi-sphère positive $\mathcal{S}_L^+$ décrivent la structure propre aux activités, à la fois d'un point de vue géométrique mais aussi sémantique. Dans les figures qui suivent, la couleur des trajectoires sur l'hypersphère reprend la légende des couleurs des courbes d'actions élémentaires (voir figure 5.17).

La figure 5.21 représente la trajectoire de l'action complexe `Jack` sur la demi-sphère positive $\mathcal{S}_L^+$. On constate que cette trajectoire prend la forme d'une spirale, en s'étalant entre les pôles représentant les actions élémentaires `Jump` et `Handwave`. On retrouve ici le caractère périodique de cette action complexe, à la fois dans la forme de la trajectoire mais aussi son orientation sur le simplexe qui caractérise une redondance des actions élémentaires `Jump` et `Handwave`.

D'autres exemples de trajectoires sémantiques sont donnés sur la figure 5.22. On projete sur l'hyper-sphère l'activité sportive `Baseball`. Sur les deux exemples représentés, on constate que la courbe est interprétable d'un point de vue sémantique. L'action élémentaire dominante est le `Jump` de part la position du joueur. La trajectoire évolue vers le centre de l'hypersphère en se rapprochant de l'action élémentaire `Run` quand le joueur effectue un mouvement de balancier, puis vers l'action élémentaire `Wave` lors du lancer de la balle. On note également que les deux exemples présentent l'activité sportive `Baseball` avec deux points de vue différents, bien qu'il existe néanmoins une similitude entre les courbes et leur position, ce qui illustre la robustesse de notre approche pour traiter des vidéos sans contraintes d'acquisition.

5.5 Similarité de trajectoires d'activités humaines

5.5.1 Similarité par distance de Hausdorff entre trajectoires

L'approche proposée permet de rendre compte de la structure des activités sur la demi-sphère positive. Les trajectoires obtenues témoignent de cette structure. Il est donc nécessaire d'évaluer si cette représentation permet de discriminer des trajectoires de différentes activités. Afin d'évaluer la similarité entre les trajectoires des différentes activités, des expérimentations préliminaires ont été réalisées en utilisant la distance de Hausdorff.
Cette distance permet d'obtenir un indice de similarité entre deux ensembles fermés de points P et Q. Soit d une distance définie sur l'hypersphère $\mathcal{S}_L^+$, la distance de Hausdorf entre deux ensemble P et Q est telle que :

$$d_H(P,Q) = \max\{\sup_{p\in P} \inf_{q\in Q} d(p,q), \sup_{q\in Q} \inf_{p\in P} d(p,q)\}$$

La figure 5.23 illustre le calcul de la distance de Hausdorff entre deux ensembles. On considère, avec cette mesure, la plus grande des distances entre le point p de P le plus éloigné de Q et le point q de Q le plus éloigné de P.

Cette distance est un premier test de similarité afin d'évaluer les distances inter et intra-classes entre différentes trajectoires d'activités. Les distances moyennes entre les activités étudiées au sens de la distance de Hausdorff sont présentées dans le Tableau 6.1 du Chapitre 6.

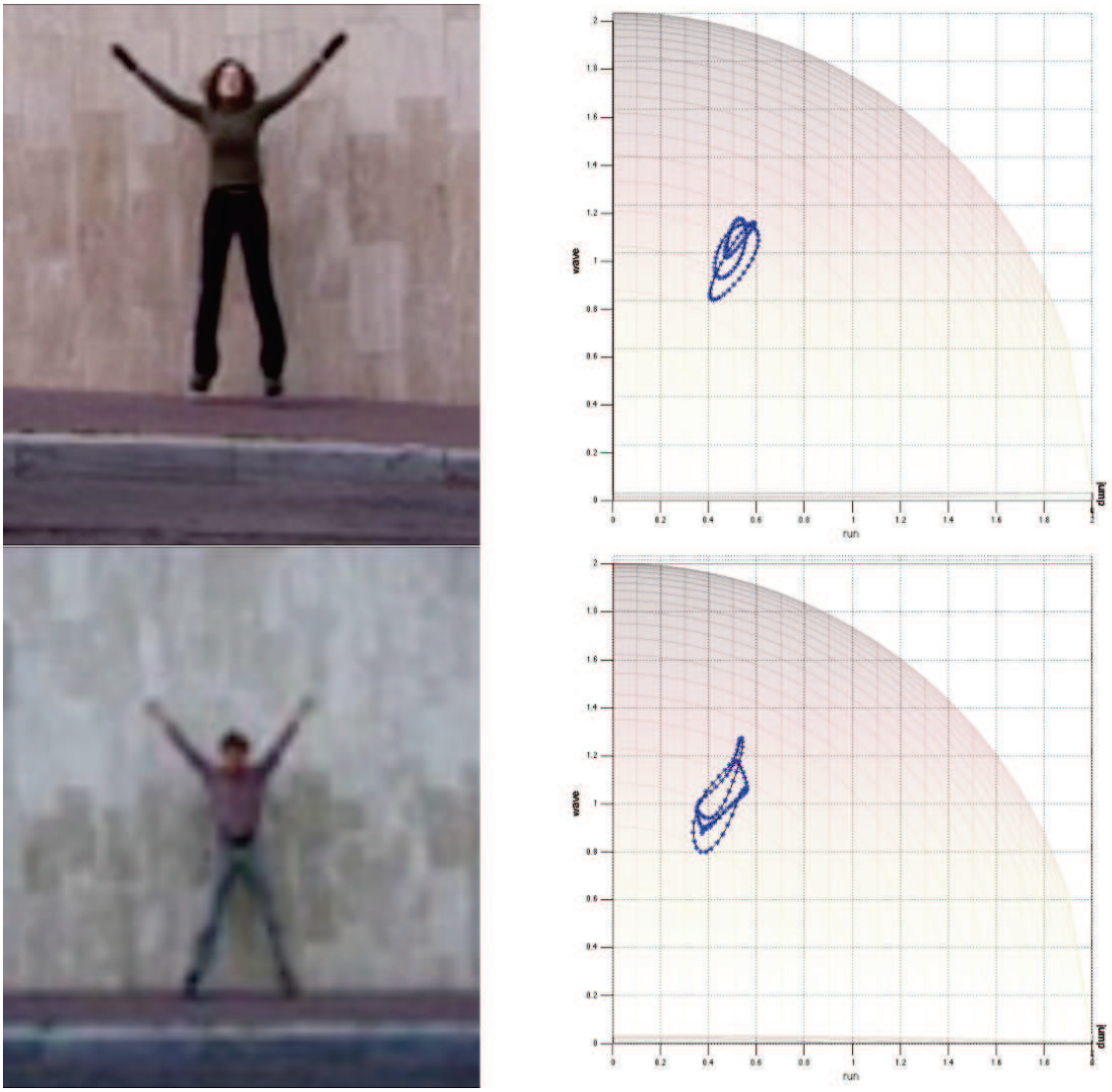

FIGURE 5.21 – Projection de l'action complexe `Jack` sur la demi-sphère positive. On constate que la redondance des actions `Jump` et `Wave` est caractérisée par une trajectoire en forme de spirale. Cette trajectoire illustre bien le caractère périodique de l'action `Jack` en évoluant entre les actions élémentaires `Jump` et `Wave` sur le simplexe (voir vidéo).

Les premières observations de l'utilisation de la distance de Hausdorff sont que les distances inter-classes sont plus élevées que les distances intra-classes. De plus, on constate que les trajectoires diffèrent entre elles, en terme de position mais aussi en terme de forme. Un autre élément qui tend à discriminer ces trajectoires est l'ordre dans lequel les actions sont enchainées au cours du temps, ce qui n'est pas pris en compte avec cette distance. En effet, on constate sur la figure que la distance de Hausdorff considère les trajectoires comme des nuages de points et ne considère donc ni l'ordre d'apparition, ni la forme de la trajectoire.

La distance de Hausdorff est un premier indicateur de similarité mais pour décrire la structure des trajectoires d'activités en prenant en compte les différentes variations de position sur la

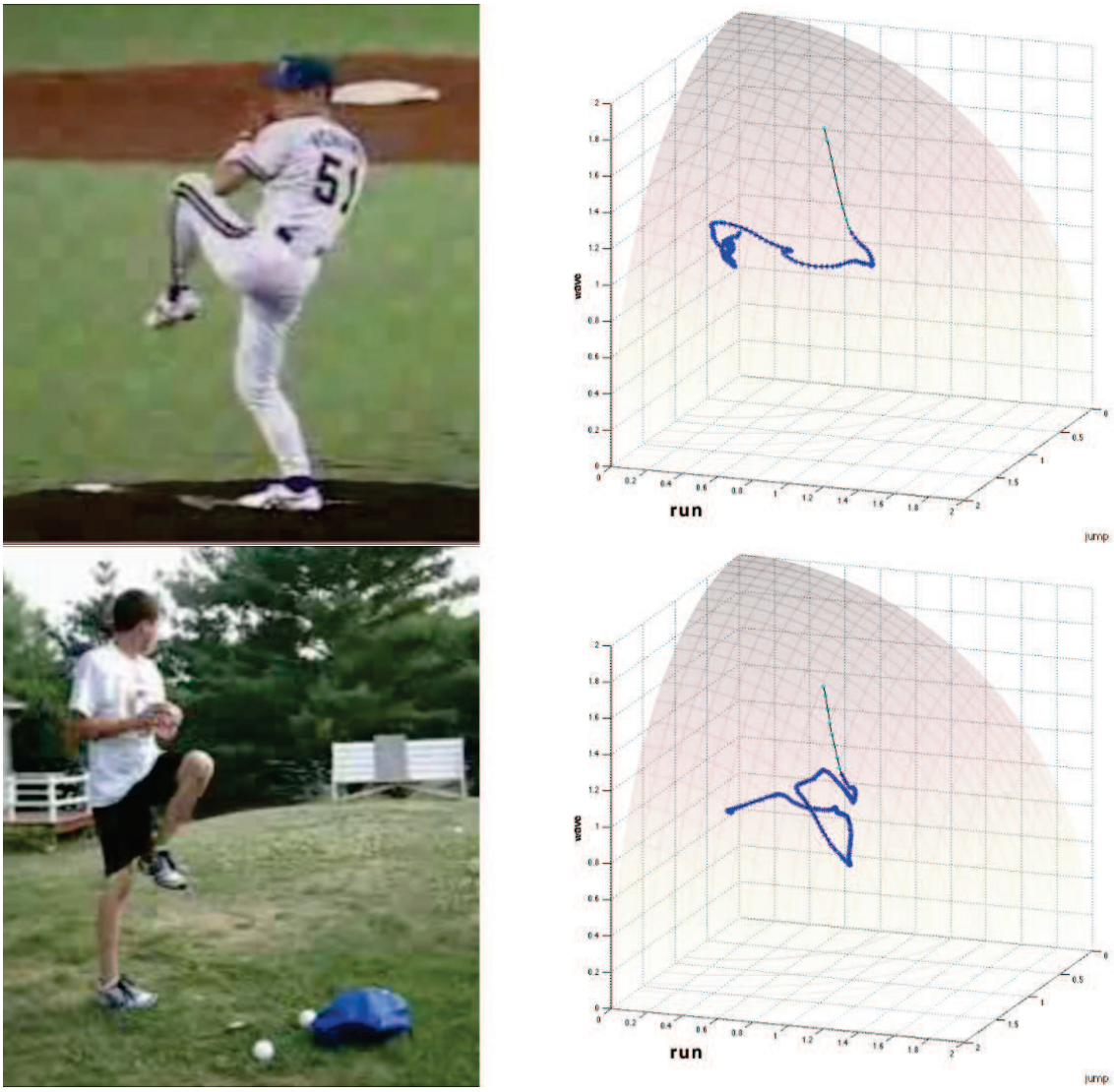

FIGURE 5.22 – Projection de l'activité sportive `Baseball`. On remarque que l'allure des deux courbes est similaire bien que les activités soit effectuées dans deux contextes différents, avec deux points de vue différents (voir vidéo).

demi-sphère, il nous faut construire un descripteur plus adapté.

5.5.2 Similarité par descripteur de Fourier sur une fonction cumulative de courbure

Avec l'approche proposée, les activités sont représentées sur la variété $\mathcal{S}_L^+$ par des trajectoires. Le but est de caractériser ces activités dans cet espace géométrique. La forme de ces trajectoires caractérise la structure intrinsèque des activités dans notre approche. Le fait de représenter cette structure d'un point de vue géométrique nous permet d'utiliser des outils propres à cet espace et ainsi généraliser la représentation des activités.

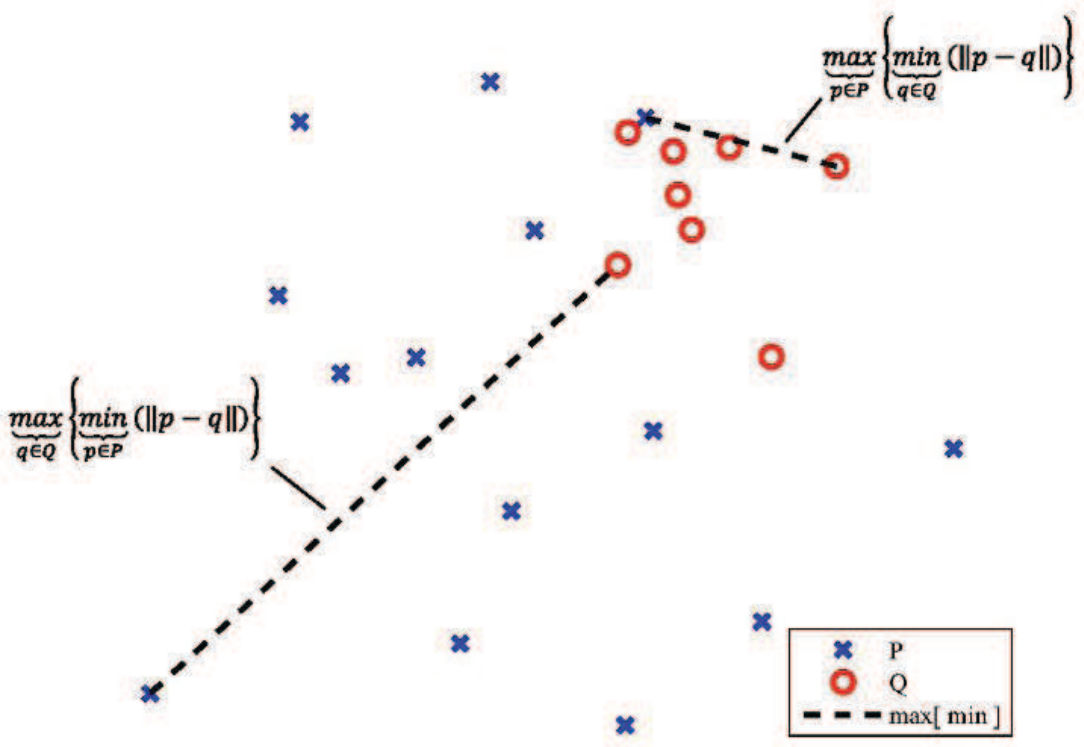

FIGURE 5.23 – Illustration de la distance de Hausdorff entre deux nuages de points P et Q.

5.5.2.1 Fréquence de transition entre actions élémentaires

En utilisant le difféomorphisme $\mathcal{F}$, nous obtenons des trajectoires qui se situent sur la surface de la demi-sphère positive $\mathcal{S}_L^+$. Les coordonnées cartésiennes des trajectoires (liées aux probabilités) sont converties en coordonnées sphériques. Elles sont définies par une coordonnée radiale r (dans notre cas $r = 2$) et L coordonnées angulaires $\phi_1, \phi_2, ..., \phi_L$ avec $\phi_L \in [0, 2\pi]$ et $\phi_k \in [0, \pi]$ pour $k = (1, ..., L-1)$. On décrit ainsi les trajectoires sur la sphère par l'évolution au cours du temps de leur coordonnées angulaires. Cette description permet d'obtenir, pour chaque activité, une courbe décrivant les variations de transition dans le temps entre chaque action élémentaire. On peut donc ainsi observer différentes formes de transitions entre les actions élémentaires : rapides, lentes, redondantes, etc. Décrire l'évolution angulaire au cours du temps des trajectoires sur $\mathcal{S}_L^+$ peut être fait dans le domaine fréquentiel.

En effet, les coefficients de transformée de la Fourier fournissent une description robuste de la fréquence de variation entre des actions élémentaires et donc, de façon plus générale, de la forme de la trajectoire. La forme des trajectoires est l'élément discriminant dans notre approche car les actions élémentaires effectuées correspondent à des positions particulières sur la sphère. D'autre part, l'ordre d'exécution des actions élémentaires dicte l'allure que prend la trajectoire sur le simplexe. Dans le domaine fréquentiel, les informations générales de la forme des tra-

jectoires sont incluses dans les coefficients correspondant aux basses fréquences. On peut donc "débruiter" ou éliminer les informations les moins pertinentes des trajectoires représentant une activité en seuillant les hautes fréquences. Considérer les variations angulaires permet d'assurer que les traitements effectués dans le domaine fréquentiel préservent les résultats obtenus dans le domaine de définition de $\mathcal{S}_L^+$. En effet, la composante radiale n'étant pas affectée, la somme des probabilités associées est égale à 1.

Les trajectoires d'activités sont des courbes ouvertes sur S_L^+. Les courbes reconstruites à partir des coefficients de Fourier associés aux basses fréquences, tendent à se refermer et à osciller dans le voisinage de leur extrémités lorsque les hautes fréquences sont supprimées.
Pour pallier ce problème nous utilisons une méthode proposée par U.Yoshinori *et al.* [Uesaka, 1984] permettant l'analyse fréquentielle de courbes ouvertes en préservant leur extrémités.

5.5.2.2 Caractérisation de courbes ouvertes

Une courbe ouverte C est constituée d'un ensemble de points $(x(t), y(t))$ tel que :

$$C = \{(x(t), y(t)) \mid t = (1, 2, ..., N)\}$$

Le fait d'utiliser une fenêtre temporelle dans notre méthode permet d'estimer les points $(x(t), y(t))$ à pas de temps d_t constant. Le but est d'interpoler la courbe avec des points estimés à pas de distance d_s tel que :

$$C = \{(x(s), y(s)) \mid s = (1, 2, ..., N)\} \text{ avec } d_s = N/S$$

S étant la longueur totale de la courbe.

La fonction de courbure K est définie comme la différence entre les segments consécutifs, de pas d_s, de C tel que :

$$K(s) = \varphi(s) - \varphi(s-1)$$
$$\text{avec } \varphi(s) = \arctan_2(y(s), y(s-1)) - \arctan_2(x(s), x(s-1))$$

L'utilisation de la fonction $\arctan_2$ permet de calculer les écarts d'angles dans l'intervalle $[-\pi, \pi]$, tel que :

$$-\pi \leqslant K(s) \leqslant \pi$$

Les écarts d'angles dans le sens horaire et anti-horaire sont donc respectivement de signe positif et négatif.

Finalement, la fonction de courbure totale ψ est définie comme l'intégrale de la fonction K. Ceci correspond à la somme totale de la courbure angulaire le long de la courbe ouverte.

$$\begin{cases} \psi(0) = K(0) \\ \psi(s) = \psi(s-1) + K(s) \mid s = (1, 2, ..., N-1) \end{cases}$$

La transformée de Fourier discrète de ψ est :

$$FD_k = \sum_{s=0}^{N-1} \psi(s) \cdot \exp(-i2\pi k)$$

ψ ne comporte que des valeurs réelles donc seule la moitié des FD_k est utilisée pour le descripteur de forme.

La figure 5.24 illustre la pertinence de cette représentation. On constate qu'en supprimant des coefficients de Fourier associés aux hautes fréquences, la trajectoire caractérisée par la transformée de Fourier sur les coordonnées sphériques n'a plus ses points de départ et d'arrivé à leurs positions initiales. Ce qui n'est pas le cas lorsque la courbe est caractérisée par la transformée de Fourier appliquée à la fonction de courbure totale de la trajectoire.

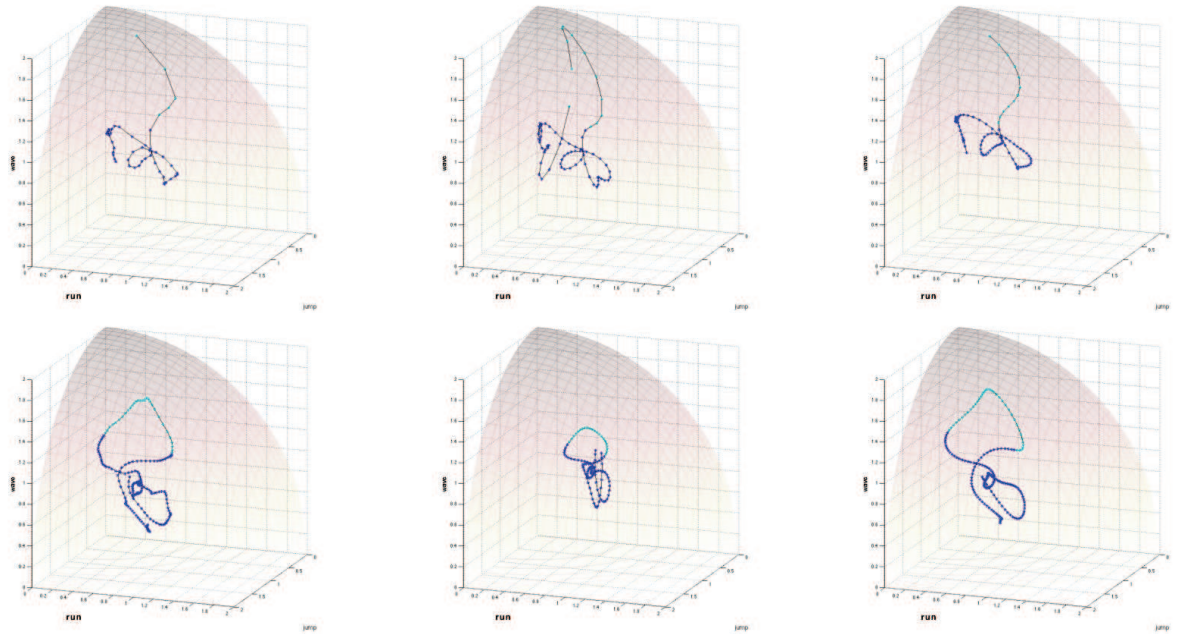

FIGURE 5.24 – Lissage de trajectoires sur $\mathcal{S}_L^+$. Gauche : trajectoires initiales. Centre : Trajectoires simplifiées reconstruites à partir du descripteur Fourier standard. Droite : Trajectoires simplifiées reconstruites à partir des coefficients de Fourier de la fonction cumulative de courbure.

5.5.2.3 Propriétés du descripteur

Dans le chapitre 3, section 3.2, nous avons présenté un exemple usuel de descripteur fréquentiel à partir des coefficients de transformée de Fourier. Les invariances en translation, rotation et échelle de ce descripteur de Fourier ont également été explicitées. Cependant, pour la caractérisation de trajectoire sur le simplexe, le descripteur de forme doit être exempt de différentes invariances géométriques.

En effet, la position sur le simplexe dépend des actions exécutées durant l'activité. Deux activités différentes peuvent avoir des formes de trajectoires similaires sans pour autant partager les mêmes activités élémentaires. Elles seront à deux positions différentes sur S_L^+. Le descripteur ne doit donc pas posséder d'invariance en translation, échelle ou rotation.
Pour répondre à ces contraintes, le descripteur de Fourier, pour une coordonnée angulaire ϕ_l est finalement défini comme :

$$F_{\phi_l} = \Re[FD_0, FD_1, ..., FD_{\frac{N-1}{2}}]$$

avec $\Re$: partie réelle.

Nous explicitons ci-dessous les différentes transformations effectuées pour prendre en compte les propriétés que l'on souhaite obtenir sur ce descripteur de forme :

Échelle Les coefficients de Fourier ne sont pas normalisés par la composante continue FD_0 pour s'abstenir de l'invariance en échelle.

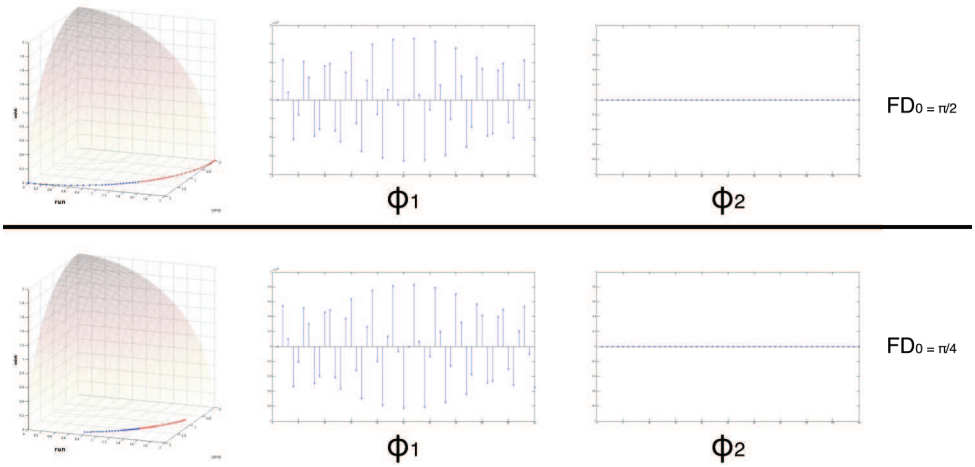

FIGURE 5.25 – Deux trajectoires avec la même forme mais ayant deux positions différentes sur le simplexe. La première trajectoire (ligne du haut) s'étend de l'action `Wave` à l'action `Run`. La seconde trajectoire (ligne du bas) s'étend de l'action `Wave` à l'action `Jump`. De part la concaténation des coefficients de Fourier de chaque coordonnée angulaire ϕ, les deux descripteurs sont différents (voir vidéo).

Translation Le descripteur global de la trajectoire est finalement la concaténation des descripteurs de chaque coordonnée angulaire de la trajectoire. Le fait de concaténer les descripteurs permet d'assurer la non-invariance en translation sur le simplexe comme l'illustre la figure 5.26.

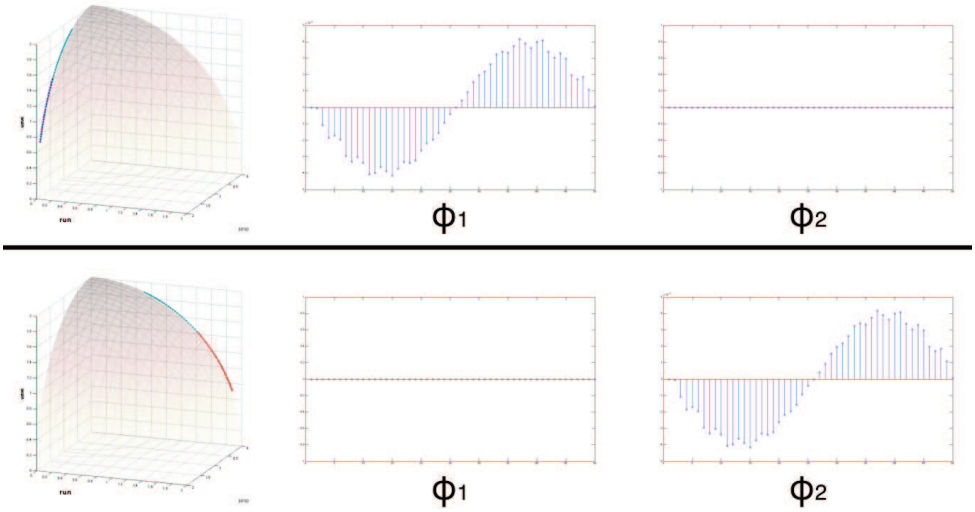

FIGURE 5.26 – Deux trajectoires avec la même forme mais ayant deux positions différentes sur le simplexe. La première trajectoire (ligne du haut) s'étend de l'action `Wave` à l'action `Run`. La seconde trajectoire (ligne du bas) s'étend de l'action `Wave` à l'action `Jump`. De part la concaténation des coefficients de Fourier de chaque coordonnée angulaire ϕ, les deux descripteurs sont différents (voir vidéo).

Sens d'éxecution La trajectoire n'a pas le même sens sémantique si elle commence à l'action `Wave` et se finit à l'action `Jump` ou inversement de l'action `Jump` à l'action `Wave`. La partie réelle du descripteur est conservée afin que le sens d'exécution de la trajectoire soit un élément discriminant. Deux trajectoires comme précédemment citées auront, pour chaque fréquence, un signe opposé.

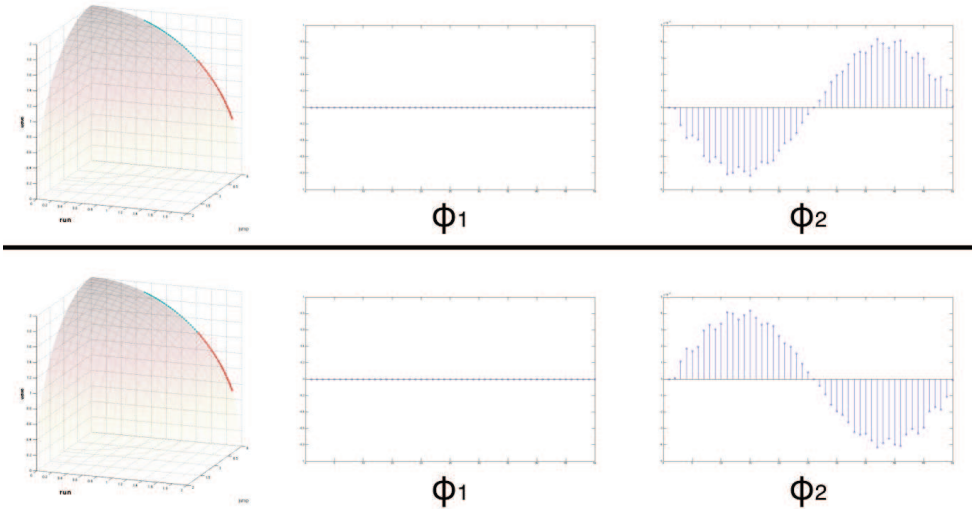

FIGURE 5.27 – Deux trajectoires identiques mais avec un ordre d'exécution des actions élémentaires sont inversé. La première trajectoire (ligne du haut) s'étend de l'action `Jump` à l'action `Run`. La seconde trajectoire (ligne du bas) s'étend de l'action `Run` à l'action `Jump`. On constate que les coefficients de Fourier associés aux coordonnées angulaires ϕ sont de signe opposé (voir vidéo).

Conclusion du chapitre Dans la première partie de ce chapitre, un bref état de l'art des méthodes de reconnaissance d'activités humaines a été réalisé. Nous avons montré les avantages et inconvénients de ces approches et mis en avant dans la deuxième partie du chapitre les contributions apportées par notre méthode, notamment le fait de pouvoir traiter des vidéos génériques. Représenter les activités comme des séquences de probabilités d'actions élémentaires permet une description riche et généralisable des activités. Les activités, formées par des séquences de probabilités d'actions élémentaires, sont représentées par des trajectoires sur la surface de cette sphère. Nous caractérisons ces activités à l'aide de descripteurs de Fourier calculés sur une fonction cumulative de courbure pour respecter au mieux la forme des trajectoires ainsi que la géométrie de l'espace où elles évoluent.

Dans le chapitre suivant, la méthode présentée ici ainsi que le descripteur utilisé seront évalués sur des vidéos d'activités sportives issues de différentes bases de données.

Chapitre 6

Évaluation : méthode de reconnaissance d'activités

Chapitre 6

Évaluation : méthode de reconnaissance d'activités

Sommaire

Introduction du chapitre Dans le chapitre précédent, nous avons présenté notre méthode de reconnaissance d'activités humaines dans des vidéos. Dans ce chapitre nous présentons les différentes expérimentations effectuées avec cette méthode.

La première partie du chapitre reprend l'ensemble des activités étudiées pour cette évaluation, ainsi que les résultats obtenus avec notre méthode.

Dans la deuxième partie, nous évaluons l'influence des paramètres de notre méthode, notamment l'influence du pourcentage de coefficients de Fourier conservés sur le taux de reconnaissance global. Nous évaluons également l'influence de la taille de la fenêtre d'observation ainsi que la base de données mélange utilisée pour l'apprentissage des classes d'actions élémentaires.

6.1 Expérimentations sur une base mixte à 3 activités

6.1.1 Introduction

Bases de données utilisées Les vidéos d'activités sportives sont les activités humaines les plus répandues dans les bases de données de la littérature. Nous évaluons l'approche développée dans le chapitre 5 sur ce type d'activités humaines. Cette évaluation de notre méthode est effectuée sur l'ensemble suivant :

— un `ensemble test` d'activités sportives, issues de différentes bases de données. Nous construisons cet ensemble à partir des bases de données `Olympic Sport`, `UCF-11` et `UCF-50`. Nous évaluons cette méthode sur trois classes d'activités sportives : `High Jump`, `Basket-Ball` et `Baseball`. Chaque classe d'activités comporte un total de 30 vidéos. La figure 6.1 illustre quelques exemples d'activités de cette base de données.

FIGURE 6.1 – Exemples d'activités utilisées. `High Jump`, `Basket-Ball` et `Baseball`.

Actions élémentaires utilisées Comme cela a été explicité dans le chapitre 5, les actions élémentaires considérées pour la construction de notre simplexe sémantique sont : `Walk`, `Run`, `Jump` et `Wave`. La classe `Standing`, présentée dans la section 5.3.4, constitue notre cinquième action. Les exemples de ces quatre actions sont issues de deux bases de données contrôlées (`KTH`, `Weizmann`) et deux bases de données génériques (`UCF-11`, `UCF-50`).

Échantillonnage de trajectoires Dans les travaux qui suivent, les trajectoires de chaque activité sont ré-échantillonnées de sorte qu'elles aient toutes le même nombre de points. Le but est d'établir une similarité entre les trajectoires durant le processus de classification.

6.1.2 Résultats de classification avec la distance de Hausdorff

La distance de Hausdorff, présentée dans le chapitre 5, section 5.5.1, est appliquée sur `l'ensemble test`. On calcule ici les distances moyennes de trajectoires par classe d'activités. Les résultats des distances moyennes sont exposés dans le Tableau 6.1. Cependant, comme cela a été souligné dans le chapitre précédent, la distance de Hausdorff ne tient compte que de la distance entre différents ensembles de points. La forme des trajectoires ainsi que l'ordre d'exécution des actions élémentaires ne sont pas considérés par cette distance, alors qu'ils sont discriminants de l'action effectuées.

action	High jump	Baseball	BasketBall
High jump	0.20	0.41	0.48
Baseball	-	0.25	0.36
Basketball	-	-	0.20

TABLE 6.1 – Distances de Hausdorff moyennes entre activités

On constate néanmoins que la distance moyenne entre les activités de même classe reste plus faible que les distances inter-classes. À partir de cette information fourni par la distance de Hausdorff, on peut supposer que la position sur le simplexe fournit une caractéristique discriminante des activités étudiées.

6.1.3 Résultats de classification avec la fonction cumulative de courbure

Le descripteur fréquentiel appliqué à la fonction cumulative de courbure (section 5.5.2) est employé pour caractériser chaque trajectoire d'activité. Les résultats qui suivent présentent la matrice de confusion obtenue après une validation croisée de type *Leave-One-Out* à l'aide d'un classifieur SVM muni d'un noyau RBF.

Résultats sur l'ensemble test Le tableau 6.2 montre les résultats obtenus avec la validation croisée *Leave-One-Out* sur chaque classe d'activité.

action	High jump	Baseball	BasketBall
High jump	100%	0	0
Baseball	0	100%	0
Basketball	10%	0	90%

TABLE 6.2 – Matrice de confusion obtenu quand un classifieur SVM est entrainé sur notre description de trajectoire. Le taux de reconnaissance est de 96,6%.

On obtient un taux de reconnaissance de 96.6%. Cela illustre la capacité du descripteur de Fourier à fournir une bonne caractérisation des trajectoires d'activité sur le simplexe sémantique. Les trajectoires de même activité partagent généralement la même forme et le même ordre d'enchainement des actions élémentaires comme l'illustre la figure 6.2. La description de trajectoires avec les coefficients de Fourier permet une description naturelle et robuste de leur forme dans le domaine fréquentiel, comme cela a été présenté dans le chapitre 3.

En comparaison, nous appliquons le détecteur STIP fourni par [Laptev et al., 2008] sur les vidéos d'activités utilisées dans notre experience. La matrice de confusion est présentée dans le Tableau 6.3. Le taux global de reconnaissance pour la méthode STIP est ici de 86%.

Le taux de reconnaissance obtenu avec une validation croisée *Leave-One-Out* met l'accent sur le caractère discriminatif de cette représentation. Nous avons souligné en début de chapitre l'importance de l'aspect temporel lors de la caractérisation d'actions complexes ou activités, qui nécessitent, toutes deux, un temps d'observation plus long. Les résultats obtenus montrent que la méthode proposée dans ce chapitre permet une meilleure représentation des activités plutôt que les points d'intérêts locaux de l'approche STIP sur cet ensemble de données. Les trajectoires d'activités sur le simplexe sémantique permettent la prise en compte de cette structure temporelle des actions élémentaires exécutées dans chaque classe d'activités.

action	High jump	Baseball	BasketBall
High jump	100%	0	0
Baseball	0	80%	20
Basketball	0%	20	80%

TABLE 6.3 – Matrice de confusion obtenu quand un classifieur SVM est entrainé sur les points d'intérêt STIP. Le taux de reconnaissance est de 86,6%.

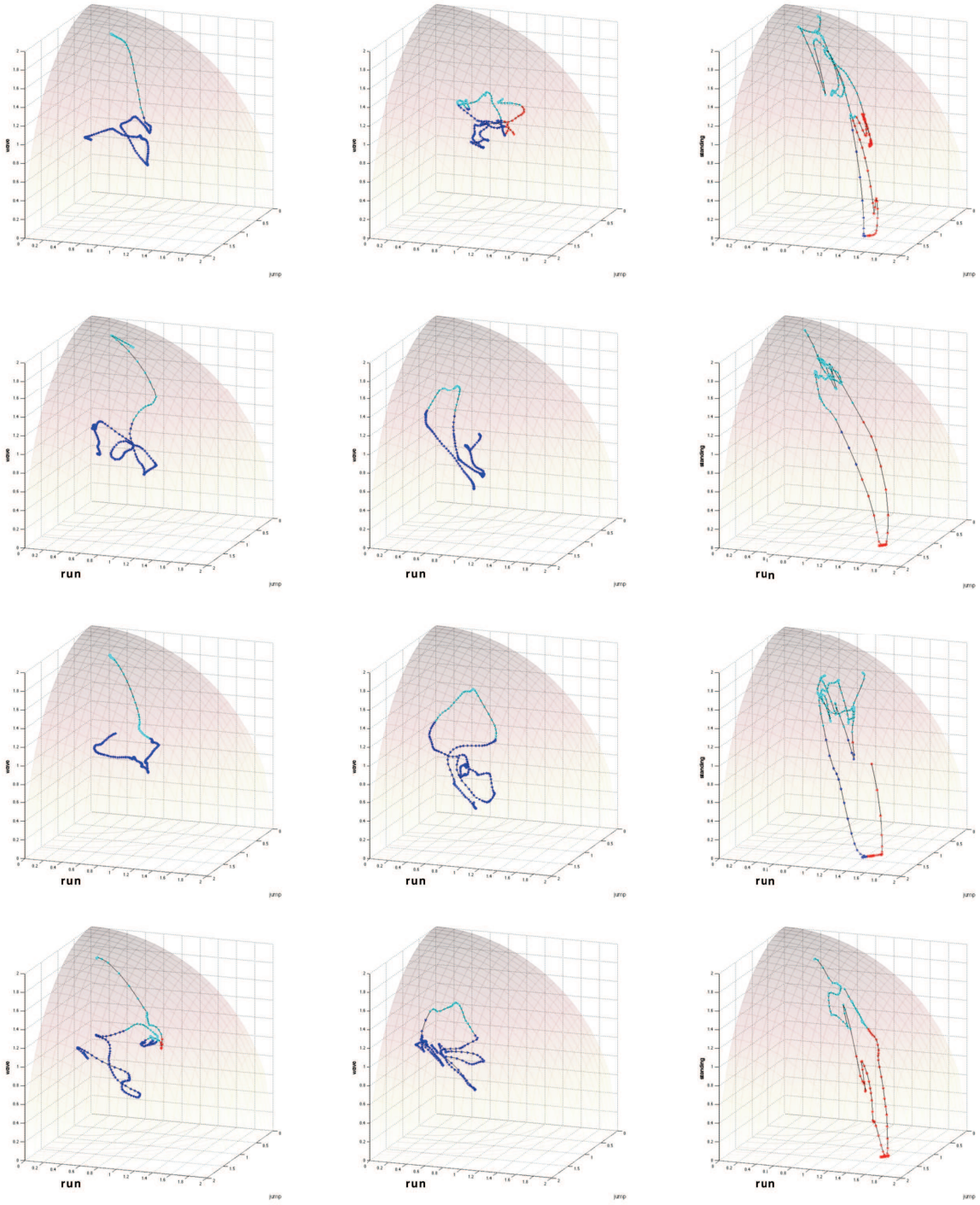

FIGURE 6.2 – Exemples de courbes d'activités obtenues par notre méthode. Les activités illustrées sont `Baseball`, `Basket-Ball` et `High Jump` (colonnes).

6.2 Évaluation de l'influence des paramètres

Dans cette partie nous évaluons les différents paramètres de notre méthode ainsi que leur influence sur les taux de reconnaissance obtenus. Les paramètres utilisés dans notre méthode sont :

— C_F : le pourcentage de coefficient de Fourier caractérisant les trajectoires,

— N : la taille de la fenêtre temporelle d'observation,

— P : la proportion de vidéos contraintes et génériques dans la base de données mélange.

6.2.1 Variation du pourcentage de coefficients de Fourier conservés

Nous évaluons ici l'influence du pourcentage de coefficients de Fourier C_F supprimés sur le taux de reconnaissance. L'évolution du taux de reconnaissance sur l'ensemble d'activités présentés précédemment, en fonction du pourcentage de coefficients est présentée sur la figure 6.3. Ce taux de reconnaissance est obtenu en utilisant un classifieur de type `SVM` muni d'un noyau `RBF` sur les descripteurs de trajectoires.

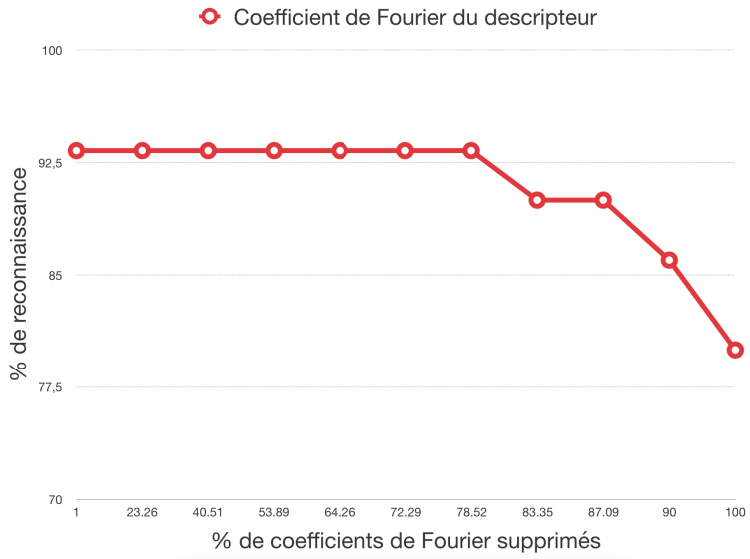

FIGURE 6.3 – Évolution du taux de reconnaissance en fonction du pourcentage C_F de coefficients de Fourier supprimés.

On constate que le taux de reconnaissance reste constant jusqu'à la suppression de plus de 78,52% des coefficients de Fourier. Passé ce pourcentage, on observe une diminution importante du taux de reconnaissance. Ce graphe montre bien que les coefficients de Fourier offrent une représentation compacte et pertinente des trajectoires. En effet, 20% de l'information fréquentielle reste discriminative pour la reconnaissance des activités.

Le constat précédent est vérifié visuellement avec les figures 6.4, 6.5 et 6.6. En effet, ces figures montrent respectivement des trajectoires d'activités `Baseball, Basket-ball` et `High`

`Jump` avec des valeurs de C_F qui croissent de façon logarithmique. La trajectoire notée "DC" est la composante continue, elle correspond à la géodésique sur la demi-sphère reliant le point de départ et le point d'arrivée de la trajectoire initiale. Comme le laissait penser la figure 6.3, on constate que globalement, ces courbes conservent la même allure jusqu'à environ 80% de coefficients supprimés. C'est cette proportion qui a été retenue dans notre étude.

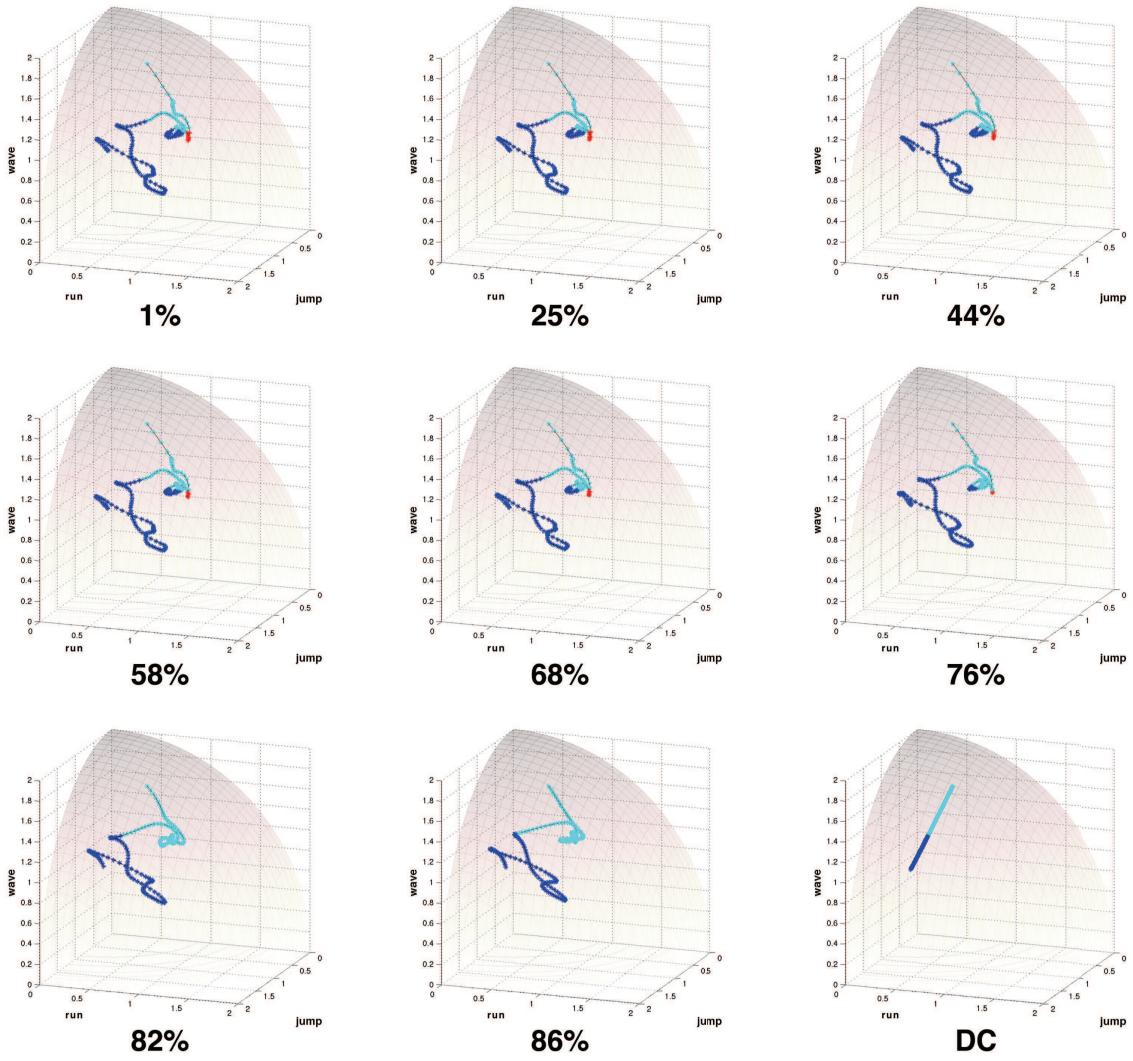

FIGURE 6.4 – Lissage des trajectoires d'actions complexes (`Base-ball`) en supprimant un pourcentage croissant de coefficients de la transformée de Fourier.

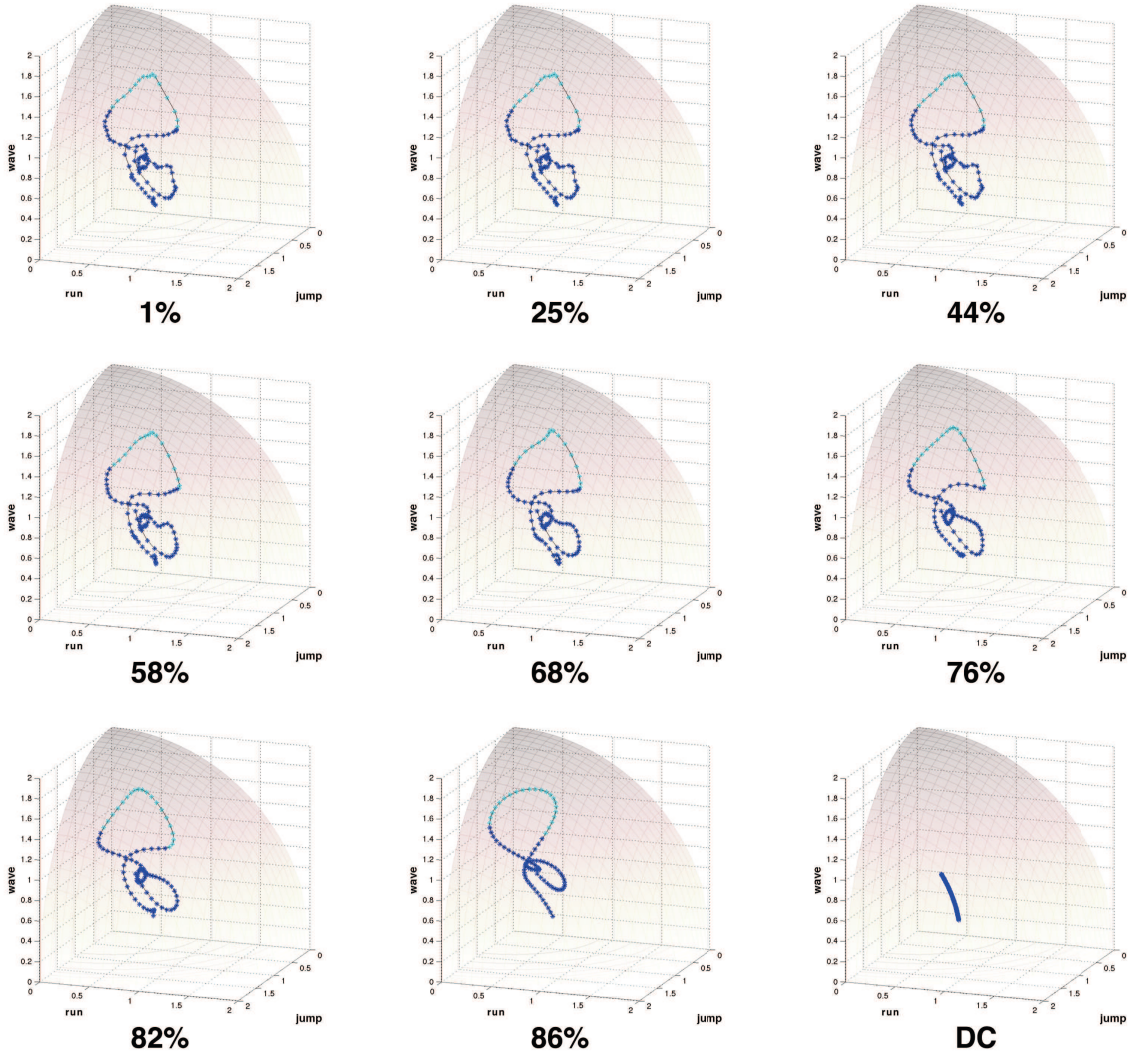

FIGURE 6.5 – Lissage des trajectoires d'actions complexes (`Basket-ball`) en supprimant un pourcentage croissant de coefficients de la transformée de Fourier.

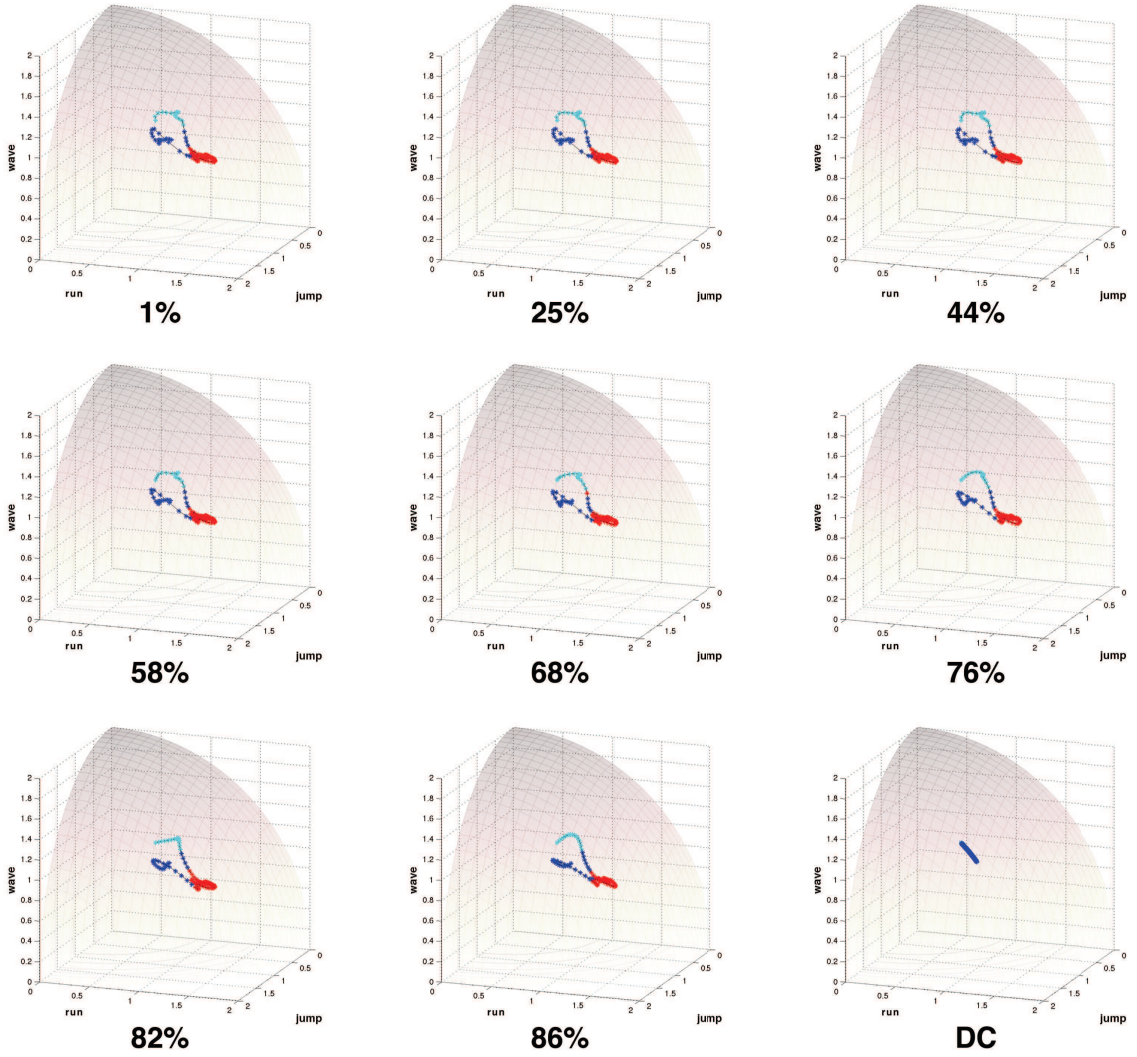

FIGURE 6.6 – Lissage des trajectoires d'actions complexes (`High Jump`) en supprimant un pourcentage croissant de coefficients de la transformée de Fourier.

6.2.2 Influence de la taille de la fenêtre d'observation

La taille N de la fenêtre d'observation $[t-N;t+N]$ (avec t une image quelconque d'une séquence vidéo) est également un paramètre influent dans notre méthode. Elle correspond à la taille de la fenêtre temporelle utilisée pour estimer la probabilité d'actions élémentaires dans une séquence vidéo. La détection d'actions élémentaires exécutées très rapidement peut être faussée dans le cas où la fenêtre d'observation est trop grande. La figure 6.7 montre l'évolution du taux de reconnaissance en fonction de la taille N de la fenêtre d'observation.

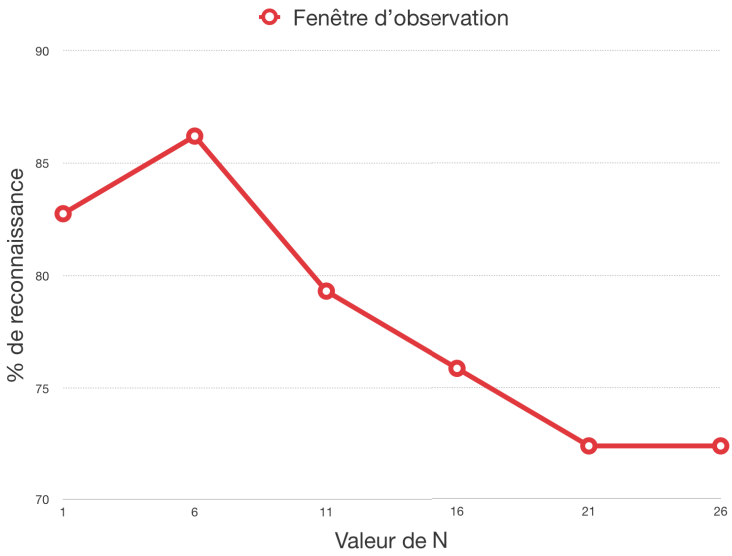

FIGURE 6.7 – Évolution du taux de reconnaissance sur l'ensemble test en fonction du paramètre N.

On remarque que le meilleur taux est obtenu avec $N = 6$, ce qui revient à observer les actions élémentaires sur une fenêtre de 12 frames. C'est le nombre qui a été retenu dans nos expérimentations.

Ce nombre est en adéquation avec les travaux de [Schindler and Van Gool, 2008] qui montrent dans leurs expérimentations que des actions élémentaires, notamment celles de KTH, peuvent être reconnues dans des fenêtre variant entre une à dix images.

La figure 6.8 illustre un exemple d'activité `High Jump` pour les valeurs $N = 1$ et $N = 26$. On constate que pour $N = 26$, certaines transitions entre actions élémentaires sont omis par la courbe, notamment le passage par l'action élémentaire `Jump` en rouge sur la première ligne, ou encore l'action `Wave` en cyan sur la deuxième ligne.

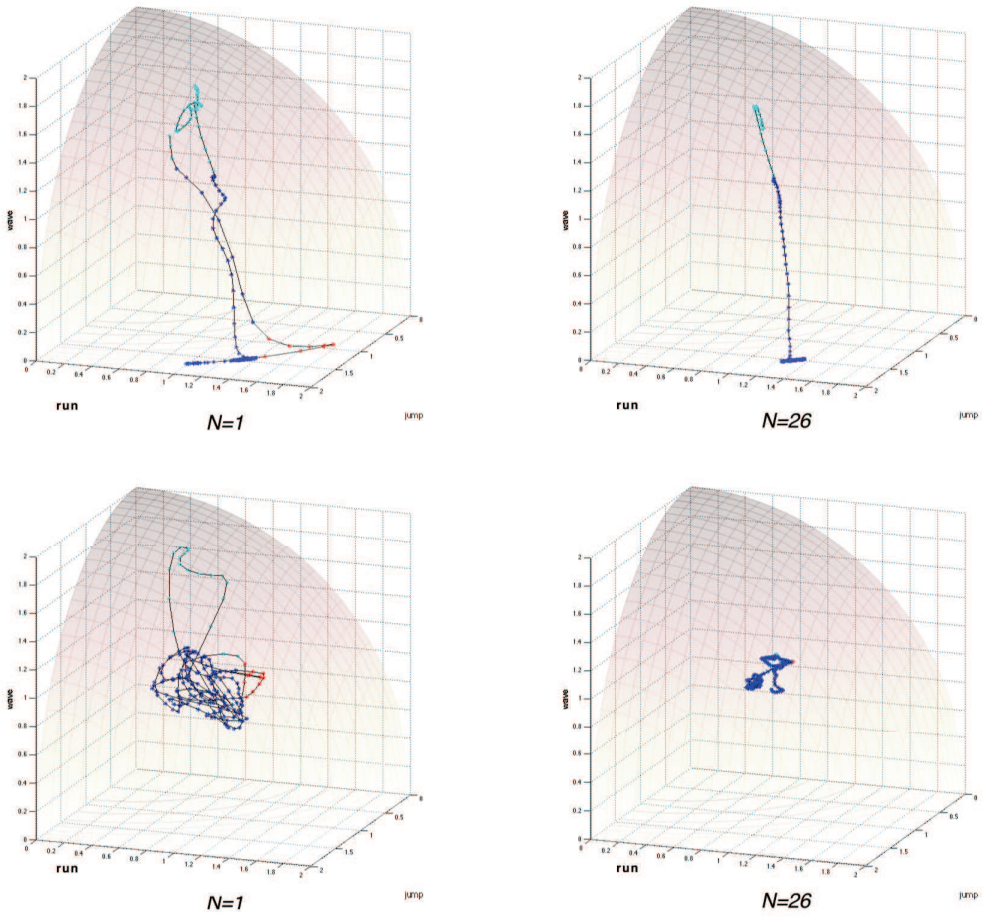

FIGURE 6.8 – Courbe d'une activité `High Jump` pour les valeurs $N = 1$ et $N = 26$. On constate que la courbe liée à $N = 26$ est moins précise sur les transitions d'actions élémentaires que la courbe liée à $N = 1$.

Cependant, l'ensemble des séquences que nous utilisons totalise en moyenne 141,58 images par vidéo. Les actions élémentaires considérées sont également exécutées sur de courtes durées. Dans le cadre de caractérisation d'activités beaucoup plus longues que des actions sportives, le nombre optimal pour la taille de la fenêtre élémentaire sera probablement amené à changer.

6.2.3 Proportion d'exemples génériques et d'exemples contrôlés

Nous évaluons ici l'influence de la base hybride d'apprentissage sur le taux de reconnaissance. Nous utilisons, pour l'apprentissage des actions élémentaires, un mélange de bases de données. Cette base hybride est composée de vidéos avec contraintes d'acquisition et de vidéos génériques. Pour évaluer cette influence, nous entrainons une base de données d'un total de 320 vidéos, en faisant varier le pourcentage de vidéos génériques qui la composent. Les vidéos contrôlées sont issues des bases `KTH` et `Weizmann` et les vidéos génériques de `UCF-11`, `UCF-50` et `HMDB-51`.

La figure 6.9 illustre la variation du taux de reconnaissance de la méthode sur l'ensemble test en fonction de la proportion de vidéos génériques contenues dans la base hybride d'apprentissage (paramètre P). On constate que les meilleurs taux sont obtenus lors des mélanges des deux types de bases de données (20% et 40%) de vidéos génériques.

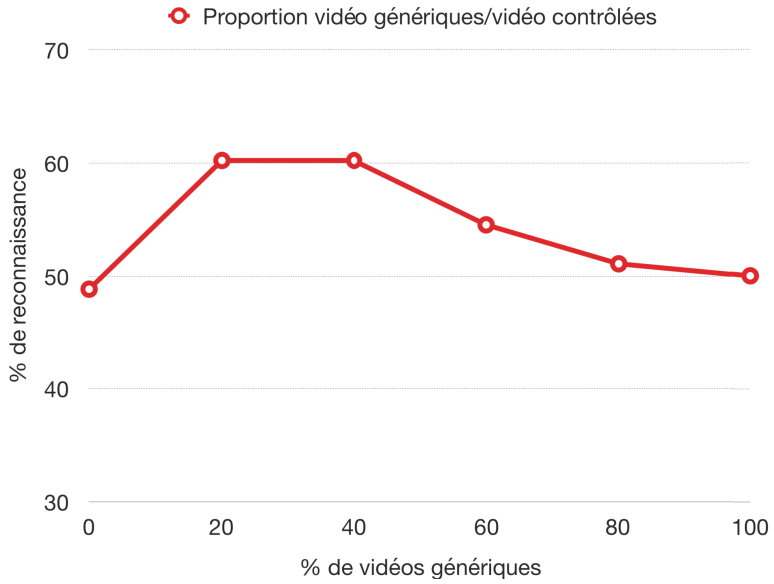

FIGURE 6.9 – Évolution du taux de reconnaissance sur l'ensemble test d'activités en fonction du paramètre P.

La figure 6.10 montre le taux de reconnaissance par classe d'activités. On remarque que les activités `High Jump` et `Baseball` sont relativement peu sensibles aux variations de la proportion de vidéos génériques, contrairement à l'activité `Basket-Ball` qui est en soi une activité plus variable et complexe.

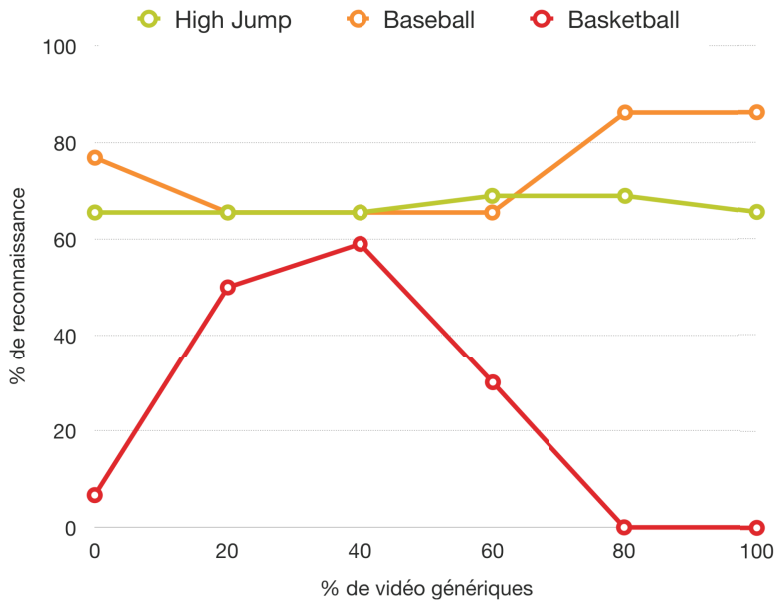

FIGURE 6.10 – Évolution du taux de reconnaissance sur les activités `High Jump`, `Baseball` et `Basketball` en fonction du paramètre P.

Cela peut s'expliquer par le fait que l'enchainement des actions élémentaires associées aux activités `High Jump` et `Baseball` est beaucoup moins fluctuant que celui de l'activité `Basketball`.

D'un point de vue qualitatif, la figure 6.11 illustre les courbes de probabilités d'actions élémentaires d'un exemple d'activités de `Basketball` dans le cas $P = 0$ et $P = 4$. On constate qu'en rajoutant des exemples génériques à la base de données hybride, l'action `Jump` est mieux détectée dans le cas $P = 4$ que pour $P = 0$. Ce qui montre bien l'intérêt de mélanger dans l'apprentissage différents types de bases de données.

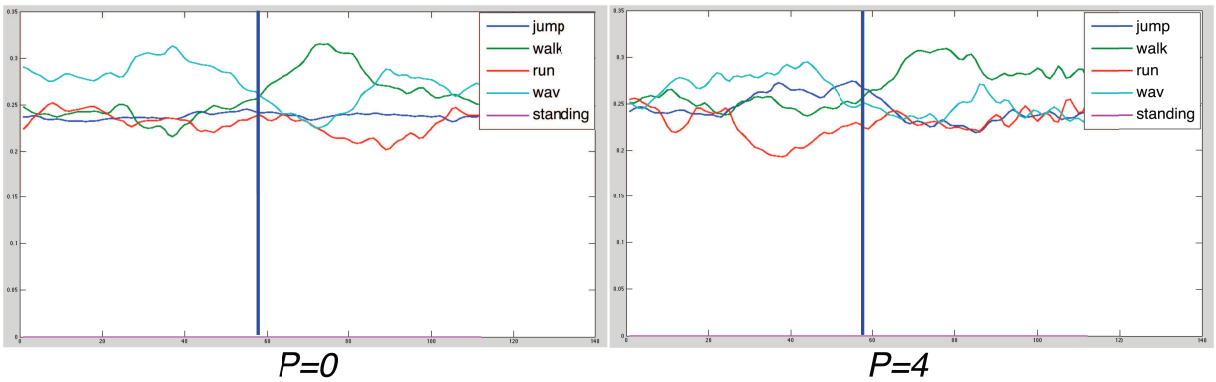

FIGURE 6.11 – Courbe de probabilité d'actions élémentaires pour une activité de `Basketball` pour $P = 0$ et $P = 4$.

6.3 Expérimentations sur une base complexe à 16 activités

6.3.1 Introduction

Olympic Sport dataset La méthode est également évaluée sur la base de données `Olympic Sport` [Niebles et al., 2010]. Cette base comporte un ensemble de 783 vidéos réparties en 16 classes. La figure 6.12 illustre l'ensemble des classes d'activités de cette base de données. Les activités sportives sont capturées dans des conditions génériques (mouvements de caméra, changements de point de vue, fonds non statiques, etc.). Les durées d'acquisition des vidéos varient fortement (durée minimale : 2sec, durée maximale : 23,72 sec). Toutes ces caractéristiques en font actuellement l'une des bases les plus complexes pour la reconnaissance d'activités humaines

FIGURE 6.12 – Les 16 classes d'activités de la base `Olympic Sport`.

6.3.2 Résultats de classification avec la fonction cumulative de courbure

Le descripteur fréquentiel appliqué à la fonction cumulative de courbure est employé pour caractériser chaque trajectoire d'activité. Les résultats qui suivent présentent la valeur MAP (*Mean Average Precision*) par classe ainsi que la matrice de confusion obtenue après une 10-validation croisée avec un classifieur SVM muni d'un noyau *RBF*. Il est à noté que ce critère n'est pas celui utilisé pour les bases de données précédentes (*Leave-One-Out*) mais il a été retenu afin de comparer nos résultats, présentés dans le tableau 6.4, à ceux de la littérature récente ([Li et al., 2013]).

	Laptev *et al.*	Tang *et al.*	Li *et al.*	**Notre méthode**
activité				
high-jump	52,4%	18,4%	83,9%	64,1%
long-jump	66,8%	81,8%	91,9%	40,5%
triple-jump	36,1%	16,1%	75,7%	84,3%
pole-vault	47,8%	84,9%	76,5%	76,7%
gym-vault	88,6%	85,7%	91,4%	70,7%
shot-put	56,2%	43,3%	79,4%	93,5%
snatch	41,8%	88,6%	73,4%	80,6%
clean-jerk	83,2%	78,2%	85,4%	60,0%
javelin throw	61,1%	79,5%	76,7%	72,1%
hammer throw	65,1%	70,5%	79,2%	41,1%
discus throw	37,4%	48,9%	66,9%	48,9%
diving-plat	91,5%	93,7%	82,0%	73,5%
diving-board	80,7%	79,3%	82,3%	95,5%
basket-layup	75,8%	85,5%	60,8%	64,1%
bowling	66,7%	64,3%	73,0%	41,7%
tennis serve	39,6%	49,6%	73,2%	62,2%
mean AP	62,0%	66,8%	78,2%	**66,9%**

TABLE 6.4 – Critère MAP pour trois travaux de la littérature ([Laptev et al., 2008, Tang et al., 2012, Li et al., 2013]) et notre méthode. Le taux de reconnaissance obtenu par notre approche est de 66,9%.

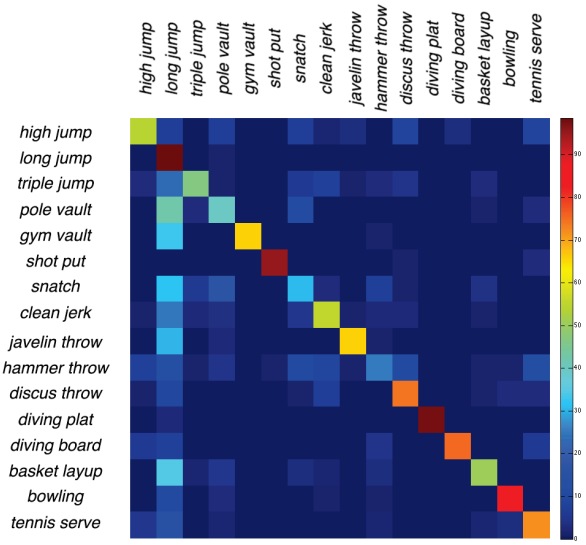

FIGURE 6.13 – Matrice de confusion obtenue sur les 16 classes d'activités.

Nous obtenons un taux de reconnaissance sur la base `Olympic Sport` proche des méthodes de l'état de l'art sur cette base [Laptev et al., 2008, Tang et al., 2012, Li et al., 2013]. Les trajectoires d'activités sur le simplexe sémantique permettent de prendre en compte la structure temporelle des actions élémentaires exécutées pour chaque classe d'activité. Cependant notre méthode est uniquement basée sur cet enchainement temporel et nous constatons qu'un certain nombre d'activités de cette base présentent une structure d'enchainement d'actions élémentaires similaires. Ces activités seront donc plus difficilement discriminées par notre approche.

C'est le cas des activités telles que : `bowling, discus throw, hammer throw, shot put` et `javelin throw` qui représentent toutes une action de type *lancé*. La figure 6.14 illustre l'enchainement des actions élémentaires de ces activités.

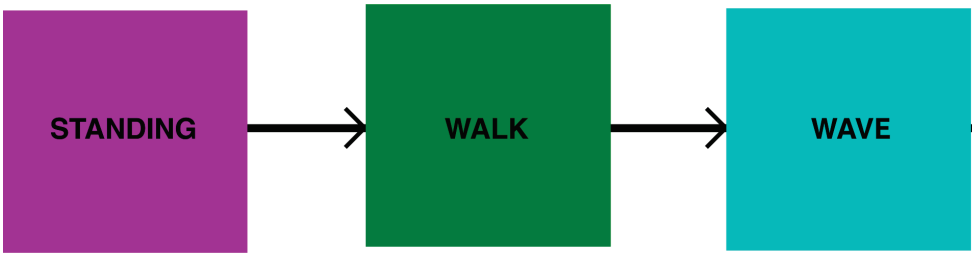

FIGURE 6.14 – Enchainement d'actions élémentaires commun aux activités de lancés.

Les figures 6.15, 6.16, 6.18, 6.17 illustrent la similarité en terme d'enchainement d'actions élémentaires de ces activités. La 6.19 montre la similarité d'enchainement d'actions élémentaires pour l'activité `hammer throw`, cependant pour cette dernière, l'action `walk` est mélangée à l'action `wave` lorsque l'athlète avance en faisant tourner son marteau.

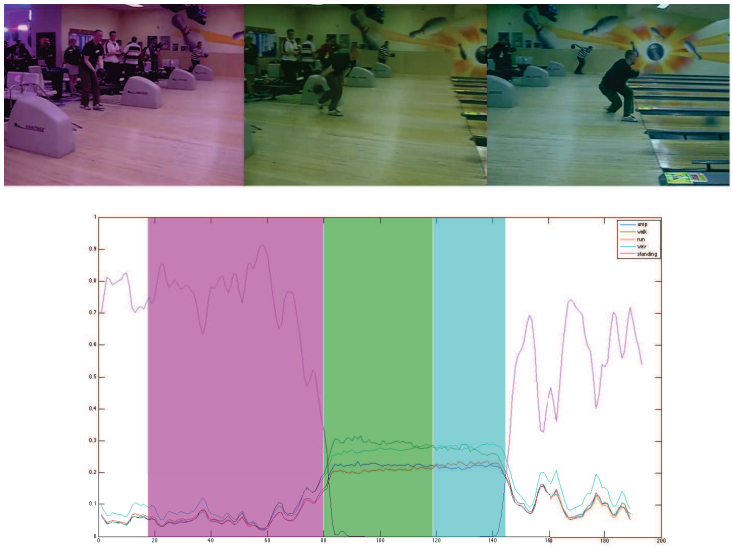

FIGURE 6.15 – Exemple de courbes de probabilités pour une activité `Bowling`.

Comme le montrent ces figures, les activités citées plus haut se décomposent généralement en une succession des actions élémentaires `Standing`, `Walk` et `Wave`.

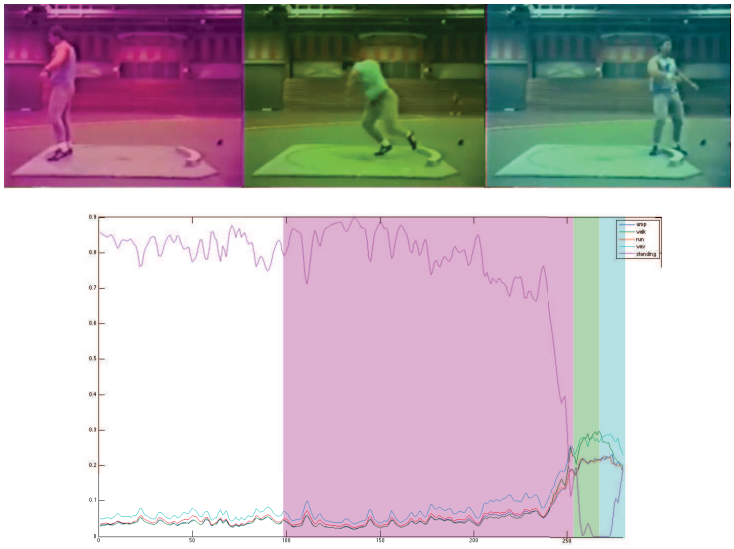

FIGURE 6.16 – Exemple de courbes de probabilités pour une activité `Shot put`.

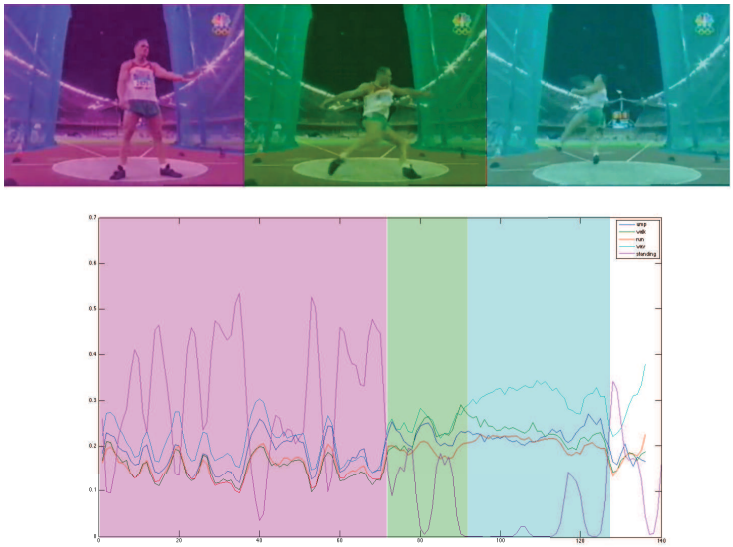

FIGURE 6.17 – Exemple de courbes de probabilités pour une activité `discus`.

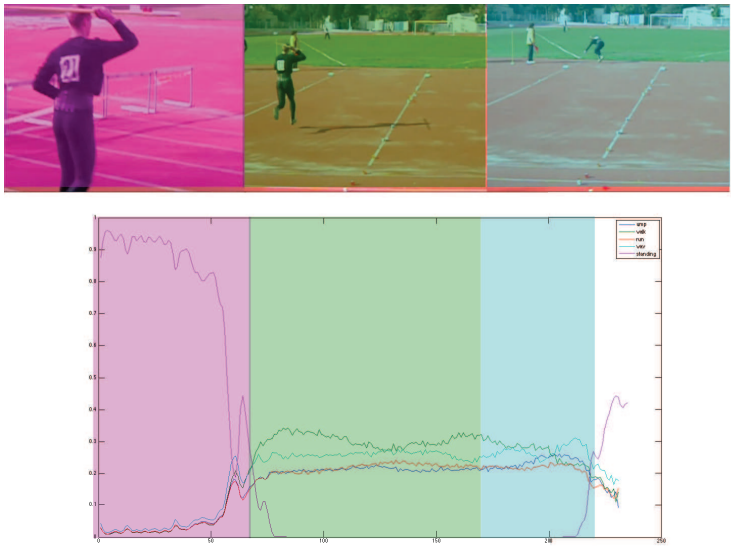

FIGURE 6.18 – Exemple de courbes de probabilités pour une activité `javelin`.

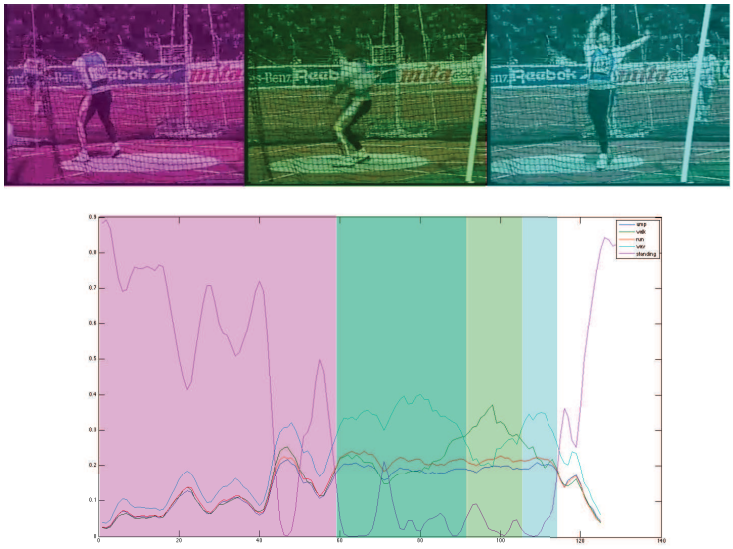

FIGURE 6.19 – Exemple de courbes de probabilités pour une activité `hammer`. On constate un mélange des actions `Wave` et `Walk` avant le lancé.

Regroupement des classes comportant un enchainement similaire d'actions élémentaires Les activités `bowling`, `discus throw`, `hammer throw`, `shot put` et `javelin throw` illustrent un même concept et sont composées du même enchainement d'actions élémentaires. Nous les avons donc regroupées dans une même classe `throw` et évaluons notre méthode sur ce nouvel ensemble composé de 12 classes. Les résultats du tableau 6.5 présentent le critère MAP *Mean Average Precision* par classe ainsi que la matrice de confusion obtenue après une 10-validation croisée.

activité	**Notre méthode**
high-jump	98,0%
long-jump	100%
triple-jump	100%
pole-vault	100%
gym-vault	92,2%
throw	84,7%
snatch	98,0%
clean-jerk	96,2%
diving-plat	95,2%
diving-board	94,9%
basket-layup	99,0%
tennis serve	100%
mean AP	96,5%

TABLE 6.5 – Précision moyenne obtenue quand un classifieur SVM est entrainé sur notre descripteur de trajectoires. Le taux de reconnaissance est de 96,5%.

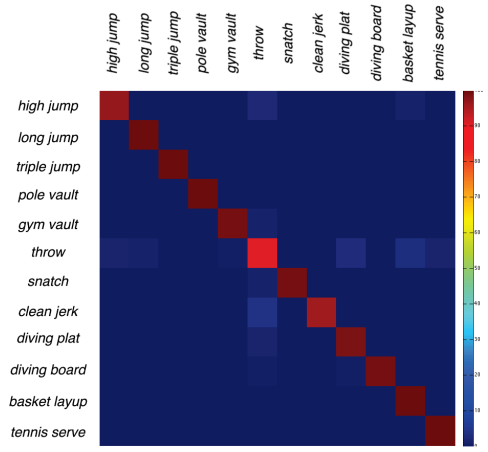

FIGURE 6.20 – Matrice de confusion obtenue sur les 12 classes d'activités.

On constate la très nette amélioration des résultats dans ce cas de figure. Les 12 classes considérées en regroupant les activités de *lancé* sont très bien discriminées par notre méthode et le taux global de classification atteint à présent 96,5%.

Même si des informations bas niveau sont utilisées à l'étape de caractérisation des actions élémentaires via le descripteur `HOG`, notre méthode considère les activités humaines comme un enchainement séquentiel d'actions élémentaires. Les activités sont donc caractérisées uniquement par la notion de dynamique temporelle. Cependant, dans des cas où certaines activités partagent une structure d'actions élémentaires similaires il est probablement intéressant d'introduire des informations supplémentaires, notamment des informations contextuelles (couleur du fond, texture, etc.), certainement aux prix d'un temps de calcul plus élevé.

Conclusion du chapitre Dans ce chapitre nous avons montré l'intérêt de caractériser les trajectoires d'activités sur le simplexe sémantique à l'aide de notre descripteur fréquentiel. L'influence de différents paramètres de notre méthode a également été évaluée, justifiant notamment notre choix de base de donnée hybride pour l'apprentissage, ainsi que celui de la taille de la fenêtre d'observation pour l'estimation des actions élémentaires.

Chapitre 7

Conclusion et perspectives

Chapitre 7

Conclusion et perspectives

Sommaire

7.1 Bilan

Tout au long de ce manuscrit, nous avons introduit un ensemble d'outils liés aux problèmes de la représentation des mouvements humains dans des vidéos. Les motivations liées à ces travaux sont la caractérisation et la reconnaissance automatique d'actions et d'activités humaines.

Nous avons évoqué dans le chapitre 1 l'intérêt grandissant pour cette thématique à travers divers éléments factuels et exemples. Cet intérêt amène la communauté en vision par ordinateur à répondre à différentes problématiques dues à la grande variabilité de la représentation, l'ambiguité du vocabulaire employé, la prise en compte du contexte visuel ainsi qu'au choix de la durée d'observation pour la détection et la caractérisation des actions.

Dans le chapitre 2, nous avons présenté une liste de bases de données de la littérature qui ont pour objectif d'illustrer, dans des vidéos tests, l'ensemble des problématiques précédemment citées. L'évolution de la complexité de ces bases est explicitée, que ce soit en termes de nombres de classes d'actions représentées, du nombre total de vidéos, mais également des caractéristiques visuelles illustrées. Dans ce même chapitre, un ensemble de méthodes de la littérature sont présentées. Ces dernières sont séparées en deux grandes catégories. D'une part, les approches globales, représentant les actions humaines comme un ensemble cohérent en temps et espace de caractéristiques propres à chaque méthode. D'autre part, les approches locales, qui se basent sur des éléments caractéristiques locaux, maximisant une fonction liée à la caractéristique jugée d'intérêt (gradient, périodicité, orientation du mouvement,etc.). Les limitations de chacune de ces approches sont détaillées. Les méthodes globales rencontrent des difficultés pour caractériser les actions dans des contextes avec peu de contraintes d'acquisition. Les méthodes locales d'extraction d'éléments d'intérêt épars se révèlent être peu efficaces pour caractériser des actions dans des vidéos génériques. Enfin, les méthodes locales utilisant une grille dense de points d'intérêt sont efficaces en terme de taux de reconnaissance mais posent des problèmes calculatoires et de stockage.

Notre méthode de reconnaissance d'actions élémentaires humaines est présentée dans le chapitre 3. Cette approche répond aux défis du chapitre 2 en caractérisant les actions humaines à l'aide de l'estimation du flot optique des séquences vidéos. Des points critiques sont extraits afin de décrire les zones de fortes déformations du flot optique, caractéristiques des mouvements humains dans des vidéos. Les points critiques, extraits à différentes échelles de fréquence ainsi que leur trajectoires permettent d'obtenir une description robuste des mouvements humains. Nous introduisons également dans ce chapitre notre approche de compensation de mouvement de caméra pour l'analyse de vidéos avec peu de contraintes d'acquisition.

Le chapitre 4 montre comment cette méthode, avec la fusion tardive de caractéristiques d'orientation de mouvements, de fréquences et de variations de gradient, obtient des taux de reconnaissance parmi les meilleurs de la littérature. Cette approche à l'avantage de ne caractériser que le mouvement des points critiques, ce qui est un avantage calculatoire. Les résultats obtenus montrent les qualités de notre méthode dans des contextes variés : reconnaissance d'actions dans des vidéos avec contraintes d'acquisition, reconnaissance d'actions dans des vidéos génériques, discrimination d'actions avec de fortes similarités visuelles, discrimination d'un grand nombre

de classes d'actions.
Le chapitre 4 illustre également la capacité de notre approche à généraliser l'information liée aux actions humaines. Cela est évalué sur des mélanges de base de données de la littérature, afin de montrer que notre méthode entrainée sur une base de données retrouve les mêmes actions avec une autre base de données, indépendamment du biais visuel contenu dans ces bases. Les résultats de cette étude montrent que le mélange de base de données, dans la phase d'apprentissage, améliore la caractérisation des actions humaines dans des vidéos génériques.

La conclusion du chapitre 4 sur le mélange des bases de données est le point de départ de notre approche de reconnaissance d'activités humaines présentée au chapitre 5. Nous partons du fait qu'une activité humaine est constituée d'un enchainement temporel spécifique d'actions élémentaires au cours du temps. Les probabilités d'occurence de ces actions élémentaires sont estimées, au cours du temps, par la méthode présentée au chapitre 3. Le classifieur de cette dernière est entrainé sur une base de donnée mélange, permettant une détection plus robuste des actions élémentaires dans des vidéos génériques, comme cela a été montré dans le chapitre 4. On obtient par cette approche une séquence de probabilités *a posteriori* d'actions élémentaires. Ces séquences de probabilités évoluent comme un ensemble ordonné de point sur une variété statistique, représentée par un simplexe dont la dimension dépend du nombre d'actions élémentaires considérées. Une activité est donc représentée *in fine* par une trajectoire sur ce simplexe. L'utilisation de notre approche discriminative pour détecter les actions élémentaires permet une correspondance intuitive entre les actions apprises et les actions effectuées dans des vidéos.
Le fait de caractériser les actions élémentaires comme des séquences de probabilités permet, contrairement à d'autres méthodes de la littérature, de gérer la cooccurrence de ces actions. les trajectoires obtenues sur le simplexe sémantique sont décrites à l'aide d'un descripteur fréquentiel afin de prendre en compte leur forme dans le processus de discrimination. Le descripteur retenu tient compte de la géométrie de l'espace où évoluent les trajectoires et permet de distinguer différentes activités telles que des actions sportives. La robustesse de cette approche, l'influence de la fenêtre temporelle d'observation ainsi que la pertinence du descripteur construit sont illustrés dans le chapitre 6.

À l'issue de ces travaux, nous avons développé une chaine complète de traitement permettant la reconnaissance d'activités humaines dans des vidéos, l'information utilisée est extraite au niveau pixel. La figure 7.1 montre l'articulation de nos différentes contributions dans un processus de reconnaissance d'activités. Les points critiques et leurs trajectoires multi-échelles permettent un premier niveau d'interprétation du mouvement humain. Le processus de reconnaissance développé en utilisant l'information fréquentielle combinée à d'autres caractéristiques pertinentes fournit une caractérisation robuste des actions élémentaires. Les probabilités d'occurrences de ces dernières sont projetées dans le simplexe sémantique en tant que trajectoires caractéristiques des différentes activités humaines.

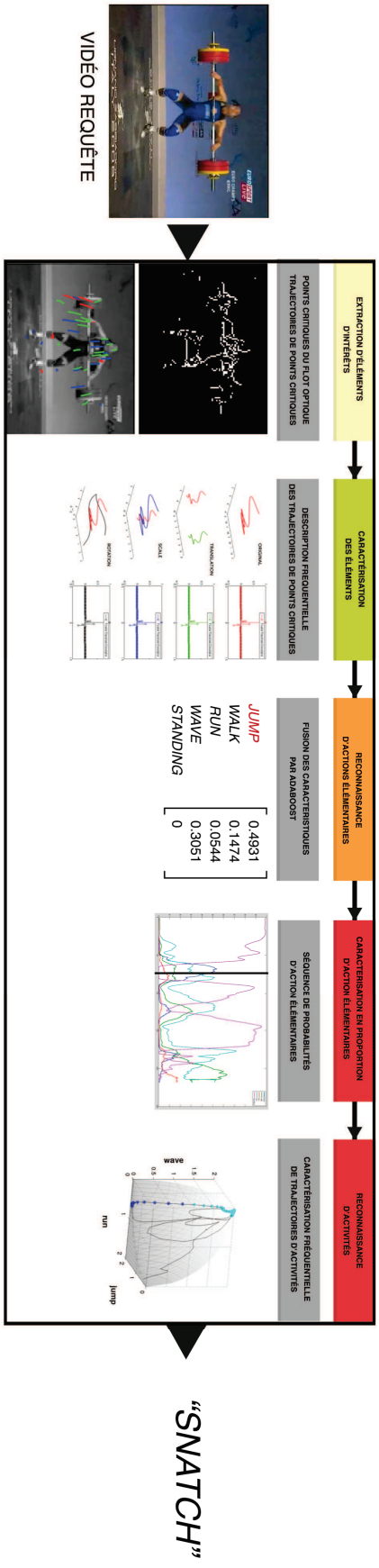

FIGURE 7.1 – Récapitulatif des différentes contributions apportées par notre méthode dans un processus de reconnaissance d'activités humaines (voir vidéo).

7.2 Perspectives du travail

7.2.1 Pistes de recherche ouvertes par la méthode

Les différents travaux développés ont permi d'entrevoir de nouvelles pistes de recherche. Ces pistes sont à la fois des perspectives d'amélioration de notre méthode, mais également des perspectives d'application de cette méthode à d'autres domaines d'étude.

Renforcement des singularités du flot optique L'estimation du flot optique est une étape importante pour notre méthode. Les points critiques du champ vectoriel associé au flot optique sont les points d'intérêts de notre méthode. Obtenir une bonne estimation du mouvement apparent est une étape coûteuse. Il serait intéressant de se focaliser sur les singularités utilisées dans pour notre méthode, notamment les points critiques qui précisément fournissent une caractérisation efficace des mouvements effectués dans les vidéos. Le développement de méthodes d'estimation du flot optique permettant de renforcer l'estimation juste au voisinage des singularités et leur suivi permettrait d'améliorer la robustesse des points critiques sans coût de calcul excessif sur l'ensemble de la vidéo.

Utilisation d'approche générative pour la représentation des caractéristiques Les vecteur de Fisher ont montré leur efficacité dans la classification d'images [Krapac et al., 2011]. D'autres méthodes de reconnaissance d'actions humaines utilisent également les vecteurs de Fisher dans leur processus de classification [Wang and Schmid, 2013]. Cette méthode permet d'apprendre la distribution statistique des vecteurs de caractéristiques extraits, en remplacement de l'approche par sac de mots visuels. L'utilisation de vecteurs de Fisher ou de vecteurs `VLAD` (*"Vector of Locally Aggregated Descriptors"*) est une piste d'amélioration de notre méthode, notamment en terme de taux de reconnaissance.

Étude des textures dynamiques Notre méthode de reconnaissance d'actions élémentaires se base sur le mouvement contenu dans une vidéo. Ce mouvement est décrit à l'aide d'informations fréquentielles, de formes, et d'orientation de mouvement. Les textures dynamiques sont définies comme des structures avec une quasi-périodicité spatiale et temporelle. L'application de notre méthode de reconnaissance d'actions élémentaires sur des textures dynamiques est une piste envisageable. En effet, l'utilisation des points critiques du flot optique est pertinente dans ce cadre, de sorte que ces derniers caractérisent les zones de forts repliement du flot optique, largement présentes dans les vidéos de textures dynamiques. Ces données peuvent également être décrites efficacement à l'aide du descripteur fréquentiel `FCD` et le descripteur de forme `HOG`. La figure 7.2 illustre des trajectoires multi-échelles extraites à l'aide de notre méthode de reconnaissance d'actions élémentaires sur des vidéos de textures dynamiques de la base de données `DynTex` [Péteri et al., 2010]. Les trajectoires multi-échelles caractérisent bien les perturbations de mouvements contenues dans les vidéos à différentes échelles, même si il pourrait être intéressant d'introduire dans l'estimation du flot optique des méthodes dédiées aux fluides incompressibles à divergence nulle [Corpetti et al., 2002].

FIGURE 7.2 – Exemples de trajectoires muti-échelles extraites sur des vidéos de textures dynamiques de la base de données `DynTex`. Les textures dynamiques illustrées sont : `Flag`, `Traffic` et `Calm water` (voir vidéo).

Évolution sémantique de phénomènes physiques Notre approche de caractérisation de courbes sur le simplexe peut être appliquée à d'autres champs d'étude que les activités humaines. Toutes données variables au cours du temps et pouvant être caractérisée par un ensemble d'états discret peut être caractérisée par cette approche. Elle est d'ailleurs l'objet d'une collaboration entre les laboratoires `MIA`, `LASIE` de L'université de La Rochelle et le laboratoire `IRISA/INRIA` de Rennes autour du projet GdR-ISIS multidisciplinaire TAbaSCO (*"TrAjectoires Semantiques pour l'étude de la COrrosion"*) pour l'étude de la fracturation par hydrogène de matériaux. L'évolution de cette fracture, représentée par différents états sémantiques, est analysée comme une trajectoire dans un simplexe sémantique propre aux états de dégradation considérés du matériau.

7.2.2 Perspectives générales de la reconnaissance d'actions humaines

La reconnaissance d'actions humaines dans des vidéos est un sujet complexe et peut être abordée de différentes façons comme l'a montré le chapitre 2. Les méthodes développées à ce jour proposent des approches très variées, mais tendent à suivre un schéma qui a fait ses preuves depuis plusieurs années : extraction de caractéristiques, descripteurs de caractéristiques, mise en commun des descripteurs (*"pooling"*) et classification des données. La distinction entre les méthodes se fait de plus en plus sur les outils utilisés dans chacune des briques de ce processus.

La méthode de reconnaissance d'actions développée est l'une des plus performantes parmi les méthodes utilisant l'approche des sac de mot visuels pour la mise en commun des descripteurs. Bien que cette méthode soit basée sur une approche d'extraction d'éléments d'intérêts, nous constatons que les performances de cette dernière augmente avec le nombre d'éléments mis en oeuvre dans le processus de reconnaissance. Ce fait est également pointé dans la littérature [Wang et al., 2009]. Cette problématique du nombre de caractéristiques utilisées pour caractériser avec pertinence les actions humaines exécutées dans une vidéo reste un inconvénient important dans le domaine. Les récentes bases de données génériques illustrent cette problématique. En effet, les points d'intérêt `STIP` extraits dans les bases de données `HMDB-51` et `UCF-101` sont disponibles en téléchargement. Cependant l'ensemble des points d'intérêts de ces bases sont plus

volumineux en espace mémoire que les vidéos brutes. On compte 3,5Gb de données `STIP` contre 2Gb de vidéos pour la base `HMDB-51` et 25,96Gb de données `STIP` contre 6,93Gb de vidéos pour `UCF-101`. Le temps de calcul associé ainsi que l'espace mémoire alloué aux données traitées est un frein pour de nombreuses méthodes quant à une utilisation pratique, notamment en temps-réel.

De ce fait, un grand nombre de travaux récent sur la reconnaissance d'actions humaines dans des vidéos se focalisent d'avantage sur le processus de reconnaissance [Peng et al., 2014, Ji et al., 2014, Sapienza et al., 2014, Zhou et al., 2015]. Elles tentent d'optimiser les différentes parties du processus, notamment avec des stratégies d'échantillonnage pour l'extraction d'éléments d'intérêt, l'encodage de caractéristiques, ou encore la fusion de ces caractéristiques.

Ces dernières années, les travaux sur la reconnaissance d'actions humaines dans des vidéos s'orientent vers des approches d'apprentissage profond (*"deep learning"*). Les méthodes de deep learning ont fait avancer de façon significative un bon nombre de problèmes en vision par ordinateur. En effet, elles sont appliquées à des problèmes variés tels que la reconnaissance d'objets [Krizhevsky et al., 2012], la reconnaissance faciale [Fan et al., 2014], la classification d'images [Simonyan et al., 2013], etc. On retrouve aujourd'hui ces modèles appliqués à la reconnaissance d'actions humaines. Les caractéristiques issues de réseaux de neurones à convolution (`CNN`) permettent de traiter des bases de données plus grandes et d'atteindre des taux de reconnaissance sur ces dernières jusqu'alors inaccessibles aux méthodes basées "sac de mots". Les différentes architectures proposées dans la littérature témoignent des avancées du domaine [Simonyan and Zisserman, 2014, Wang et al., 2015, Hasan and Roy-Chowdhury, 2015, Nasrollahi et al., 2015].

Ces dernières années, les travaux sur la reconnaissance d'actions humaines dans des vidéos s'orientent vers des approches d'apprentissage profond (*"deep learning"*). Les méthodes de deep learning ont fait avancer de façon significative un bon nombre de problèmes en vision par ordinateur. En effet, elles sont appliquées à des problèmes variés tels que la reconnaissance d'objets [Krizhevsky et al., 2012], la reconnaissance faciale [Fan et al., 2014], la classification d'images [Simonyan et al., 2013], etc. On retrouve aujourd'hui ces modèles appliqués à la reconnaissance d'actions humaines. Les caractéristiques issues de réseaux de neurones à convolution (`CNN`) permettent de traiter des bases de données plus grandes et d'atteindre des taux de reconnaissance sur ces dernières jusqu'alors inaccessibles aux méthodes basées "sac de mots". Les différentes architectures proposées dans la littérature témoignent des avancées du domaine [Simonyan and Zisserman, 2014, Wang et al., 2015, Hasan and Roy-Chowdhury, 2015, Nasrollahi et al., 2015].

La reconnaissance des actions humaines dans des vidéos est encore une thématique jeune. Le champ d'application de ce domaine est vaste et induit de nombreuses perspectives. Les méthodes permettent actuellement d'interpréter automatiquement les actions humaines effectuées dans un flux vidéo. Elles répondent à des requêtes telles que : *"Y a-t-il quelqu'un en train de se battre dans cette video ?"*. La finalité de ces méthodes pourrait, en plus d'inférer des actions apprises, aider à la compréhension globale d'une scène. Ainsi, en étant capable d'analyser les interactions qui surviennent dans une scène, on pourrait prédire des évènements potentiels, avant coureurs

d'un danger et répondre par exemple à des requêtes telles que : *"Cette personne est t-elle en danger ?"* ou *"Cette personne a t-elle un comportement suspect ?"*.

Enfin, la compréhension de scènes complexes peut également être utile dans des cadres comme celui de l'assistance aux personnes âgées qui est un domaine en fort développement. Les chutes ou autres actions anormales peuvent être caractérisées par des méthodes de reconnaissance d'actions. À un plus haut niveau sémantique, il serait intéressant de caractériser de façon plus spécifique et plus précise les comportements de personnes âgées, notamment celles atteintes de maladies dégénératives telle que la maladie d'Alzheimer [Boujut et al., 2013]. Ces méthodes pourraient être en mesure de détecter des *comportements moteurs aberrants*[1], validés par des praticiens. Ces comportement anormaux sont, dans la plupart des cas, caractérisés par la répétition de même séquence d'actions. Dans le contexte actuel, ces applications ne cesseront de croître dans les années à venir, ce qui laisse une place de choix à cette thématique de recherche dans le domaine de la vision par ordinateur.

1. Sous ce terme, on regroupe les comportements d'errances, de vérification, d'une activité répétitive ou excessive, de déambulation sans but apparent ou dans un but inapproprié (Source http://www.francealzheimer.org/les-symptômes/les-symptômes-comportementaux/181)

Publications

Articles de revues internationales

1. *An Efficient And Sparse Approach For Large Scale Human Action Recognition In Videos*,
Beaudry Cyrille, Renaud Péteri, Mascarilla Laurent, Machine Vision and Applications. (En cours de révisions)

Articles de revues nationales

1. *Analyse multi-échelle de trajectoires de points critiques pour la reconnaissance d'actions humaines*,
Beaudry Cyrille, Renaud Péteri, Mascarilla Laurent, Traitement du Signal, Octobre 2015.

Conférences internationales avec comité de lecture

1. *Human Activity recognition in the semantic simplex of elementary actions*,
Beaudry Cyrille, Renaud Péteri, Mascarilla Laurent, Proc. British Machine Vision Conference 2015 (BMVC 2015), 7-10 September 2015, Swansea, UK.

2. *Action recognition in videos using frequency analysis of critical point trajectories*,
Beaudry Cyrille, Renaud Péteri, Mascarilla Laurent, IEEE International Conference on Image Processing 2014 (ICIP 2014), 27-30 October 2014, Paris, France.

Conférences nationales avec comité de lecture

1. *Caractérisation de trajectoires d'activités humaines dans le simplexe sémantique*,
Beaudry Cyrille, Renaud Péteri, Mascarilla Laurent, XXVe colloque GRETSI (traitement du signal et des images), 8-11 September 2015, Lyon, France.

2. *Reconnaissance d'actions dans des vidéos par caractérisation fréquentielle des trajectoires de points critiques*,
Beaudry Cyrille, Renaud Péteri, Mascarilla Laurent, 19th RFIA conference, 2-4 July 2014, Rouen, France.

Liste des algorithmes

Liste des tableaux

Table des figures

Bibliographie

[mex, 2015] (2015). 28

[Aggarwal et al., 2004] Aggarwal, G., Chowdhury, A., and Chellappa, R. (2004). A system identification approach for video-based face recognition. In *Proc. Int. Conf. Pattern Recognition*, volume 4, pages 175–178. 137

[Aggarwal and Ryoo, 2011] Aggarwal, J. and Ryoo, M. (2011). Human activity analysis : A review. *ACM Comput. Surv.*, 43(3) :16–43. 34, 36, 131

[Allen and Ferguson, 1994] Allen, J. F. and Ferguson, G. (1994). Actions and events in interval temporal logic. *Journal of Logic and Computation*, 4 :531–579. 133

[Avila et al., 2013] Avila, S., Thome, N., Cord, M., Valle, E., and De A. AraúJo, A. (2013). Pooling in image representation : The visual codeword point of view. *Computer Vision and Image Understanding*, 117(5) :453–465. 39

[Baker et al., 2007] Baker, S., Scharstein, D., Lewis, J. P., Roth, S., Black, M. J., and Szeliski, R. (2007). A database and evaluation methodology for optical flow. In *Proc. Int. Conf. Computer Vision*. 100

[Bay et al., 2008] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3) :346–359. 41

[Bellet et al., 2013] Bellet, A., Habrard, A., and Sebban, M. (2013). A survey on metric learning for feature vectors and structured data. *Comput. Research Repository*. 63

[Berndt and Clifford, 1994] Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *Proc. Int. Conf. Knowledge Discovery and Data Mining*, pages 359–370. 34

[Bhatia et al., 2013] Bhatia, H., Norgard, G., Pascucci, V., and Bremer, P.-T. (2013). The helmholtz-hodge decomposition. *IEEE Transactions on Visualization and Computer Graphics*, 19(8) :1386–1404. 71

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3 :993–1022. 140, 141, 143, 144, 146

[Bobick and Davis, 2001] Bobick, A. F. and Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(3) :257–267. 31, 32

[Boujut et al., 2013] Boujut, H., Buso, V., Benois-Pineau, J., Gaëstel, Y., and Dartigues, J.-F. (2013). Visual saliency maps for studies of behavior of patients with neurodegenerative diseases : Observer's versus Actor's points of view. In *Innovation in Medicine & Healthcare*, page 5, Piraeus, Greece. 200

[Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). LIBSVM : A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2 :1–27. 91, 95

[Corpetti et al., 2002] Corpetti, T., Memin, E., and Perez, P. (2002). Dense estimation of fluid flows. *IEEE Trans. Pattern Anal. Machine Intell.*, 24(3) :365–380. 197

[Cuturi, 2011] Cuturi, M. (2011). Fast global alignment kernels. In Getoor, L. and Scheffer, T., editors, *ICML*, pages 929–936. Omnipress. 39

[Delhumeau et al., 2013] Delhumeau, J., Gosselin, P.-H., Jégou, H., and Pérez, P. (2013). Revisiting the VLAD image representation. In *ACM Multimedia*, Barcelona, Spain. 39

[Dollar et al., 2005] Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65 – 72. 39, 41, 54, 144

[Doretto et al., 2003] Doretto, G., Chiuso, A., Wu, Y., and Soatto, S. (2003). Dynamic textures. *Int. J. Computer Vision*, 51(2) :91–109. 137

[Efros et al., 2003] Efros, A. A., Berg, A. C., Mori, G., and Malik, J. (2003). Recognizing action at a distance. In *Proc. Int. Conf. Computer Vision*, ICCV '03, page 726, Washington, DC, USA. 35

[Emonet et al., 2014] Emonet, R., Varadarajan, J., and Odobez, J.-M. (2014). Temporal Analysis of Motif Mixtures using Dirichlet Processes. *IEEE Trans. Pattern Anal. Machine Intell.* 141

[Fan et al., 2014] Fan, H., Cao, Z., Jiang, Y., Yin, Q., and Doudou, C. (2014). Learning deep face representation. *Comput. Research Repository*. 199

[Fan et al., 2008] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR : A library for large linear classification. *Journal of Machine Learning Research*, 9 :1871–1874. 93

[Fei-Fei et al., 2007] Fei-Fei, L., Fergus, R., and Perona, P. (2007). Learning generative visual models from few training examples : An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1) :59–70. 118

[Fischler and Bolles, 1981] Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus : A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6) :381–395. 81

[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 *IEEE Int. Workshop on Automatic Face and Gesture Rec.*, pages 272–277. 35

[Gorelick et al., 2007] Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. *IEEE Trans. Pattern Anal. Machine Intell.*, 29(12) :2247–2253. 23, 31

[Harris and Stephens, 1988] Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *In Proc. of Fourth Alvey Vision Conference*, pages 147–151. 40, 44

[Hasan and Roy-Chowdhury, 2015] Hasan, M. and Roy-Chowdhury, A. (2015). A continuous learning framework for activity recognition using deep hybrid feature models. *IEEE Transactions on Multimedia*, (99) :1–1. 199

[Hastie et al., 2009] Hastie, T., Rosset, S., Zhu, J., and Zou, H. (2009). Multi-class AdaBoost. *Statistics and Its Interface*, 2(3) :349–360. 93, 151

[Hofmann, 1999] Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proc. Int. Conf. Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA. ACM. 140, 142

[Hongeng et al., 2004] Hongeng, S., Nevatia, R., and Bremond, F. (2004). Video-based event recognition : Activity representation and probabilistic recognition methods. *Computer Vision and Image Understanding*, 96(2) :129–162. 134

[Horn and Schunck, 1981] Horn, B. K. P. and Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, 17 :185–203. 66, 100, 101

[Hu, 1962] Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *Information Theory, IRE Transactions on*, 8(2) :179–187. 31

[Ivanov and Bobick, 2000] Ivanov, Y. A. and Bobick, A. F. (2000). Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. Pattern Anal. Machine Intell.*, 22(8) :852–872. 132

[Jain et al., 2013] Jain, M., Jégou, H., and Bouthemy, P. (2013). Better exploiting motion for better action recognition. In *Proc. Conf. Comp. Vision Pattern Rec.*, Portland, États-Unis. 81, 82

[Ji et al., 2014] Ji, X., Zhou, L., and Li, Y. (2014). Human action recognition based on adaboost algorithm for feature extraction. In *Proc. IEEE Int. Conf. Computer and information technology*, pages 801–805. 199

[Karaman et al., 2014] Karaman, S., Benois-Pineau, J., Dovgalecs, V., Mégret, R., Pinquier, J., André-Obrecht, R., Gaëstel, Y., and Dartigues, J.-F. (2014). Hierarchical Hidden Markov Model in Detecting Activities of Daily Living in Wearable Videos for Studies of Dementia. *Multimedia Tools and Applications*, 69(3) :743–771. 131

[Ke et al., 2005] Ke, Y., Sukthankar, R., and Hebert, M. (2005). Efficient visual event detection using volumetric features. In *Proc. Int. Conf. Computer Vision*, pages 166–173, Washington, DC, USA. 41

[Khosla et al., 2012] Khosla, A., Zhou, T., Malisiewicz, T., Efros, A. A., and Torralba, A. (2012). Undoing the damage of dataset bias. In *Proc. Europ. Conf. Computer Vision*, pages 158–171. 124

[Koenderink and van Doorn, 1987] Koenderink, J. and van Doorn, A. (1987). Representation of local geometry in the visual system. *Biological Cybernetics*, 55(6) :367–375. 40

[Krakowski and Manton, 2007] Krakowski, K. A. and Manton, J. H. (2007). On the computation of the karcher mean on spheres and special orthogonal groups. In *in Proc. Workshop Robot. Math.* 138

[Krapac et al., 2011] Krapac, J., Verbeek, J., and Jurie, F. (2011). Modeling Spatial Layout with Fisher Vectors for Image Categorization. In *Proc. Int. Conf. Computer Vision*, pages 1487–1494, Barcelona, Spain. IEEE. 197

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 199

[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 *Proc. Int. Conf. Computer Vision*. 27, 122

[Lafferty and Lebanon, 2005] Lafferty, J. and Lebanon, G. (2005). Diffusion kernels on statistical manifolds. *J. Mach. Learn. Res.*, 6 :129–163. 141, 157

[Laptev, 2005] Laptev, I. (2005). On space-time interest points. *Int. J. Computer Vision*, 64(2-3) :107–123. 39, 40, 54, 56

[Laptev, 2013] Laptev, I. (2013). *Modeling and visual recognition of human actions and interactions*. Habilitation à diriger des recherches, Ecole Normale Supérieure de Paris - ENS Paris. 12

[Laptev et al., 2008] Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Proc. Conf. Comp. Vision Pattern Rec.*, pages 1 –8. 24, 39, 43, 63, 68, 88, 90, 172, 183, 184

[Lari and Young, 1990] Lari, K. and Young, S. J. (1990). The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer speech and language*, 4(1) :35–56. 132

[Lazebnik et al., 2006] Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features : Spatial pyramid matching for recognizing natural scene categories. In *Proc. Conf. Comp. Vision Pattern Rec.*, volume 2, pages 2169–2178. 39, 88, 89

[Li et al., 2013] Li, W., Yu, Q., Sawhney, H., and Vasconcelos, N. (2013). Recognizing activities via bag of words for attribute dynamics. In *Proc. Conf. Comp. Vision Pattern Rec.*, pages 2587–2594. 139, 182, 183, 184

[Liu et al., 2009] Liu, J., Luo, J., and Shah, M. (2009). Recognizing realistic actions from videos "in the wild". In *Proc. Conf. Comp. Vision Pattern Rec.*, pages 1996–2003. 25

[Lucas and Kanade, 1981] Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proc. Int. Conf. Artificial Intelligence*, volume 2, pages 674–679, San Francisco, CA, USA. 67, 76, 100, 101

[Manning et al., 2008] Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA. 142

[Marszalek et al., 2009] Marszalek, M., Laptev, I., and Schmid, C. (2009). Actions in context. In *Proc. Conf. Comp. Vision Pattern Rec.*, pages 2929–2936. 24, 44

[Mottaghi et al., 2014] Mottaghi, R., Chen, X., Liu, X., Cho, N.-G., Lee, S.-W., Fidler, S., Urtasun, R., and Yuille, A. (2014). The role of context for object detection and semantic segmentation in the wild. In *Proc. Conf. Comp. Vision Pattern Rec.* 81

[Murphy, 2002] Murphy, K. P. (2002). Dynamic bayesian networks : Representation, inference and learning. 131

[Murthy and Goecke, 2013] Murthy, O. and Goecke, R. (2013). Ordered trajectories for large scale human action recognition. In *Proc. Int. Conf. Computer Vision*, pages 412–419. 48, 49, 106, 109

[Nasrollahi et al., 2015] Nasrollahi, K., Guerrero, S., Rasti, P., Anbarjafari, G., Baro, X., J. Escalante, H., and Moeslund, T. (2015). *Deep Learning based Super-Resolution for Improved Action Recognition*. 199

[Niebles et al., 2010] Niebles, J., Chen, C.-W., and Fei-Fei, L. (2010). Modeling temporal structure of decomposable motion segments for activity classification. In *Proc. Europ. Conf. Computer Vision*, volume 6312, pages 392–405. 182

[Niebles et al., 2008] Niebles, J. C., Wang, H., and Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *Int. J. Comput. Vision*, 79(3) :299–318. 144, 146

[Nigam et al., 2000] Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Journal of Machine Learning Research*, 39(2-3) :103–134. 142

[Nijholt, 1980] Nijholt, A. (1980). A survey of normal form covers for context free grammars. *Acta Informatica*, 14(3) :271–294. 132

[Oliver et al., 2002] Oliver, N., Horvitz, E., and Garg, A. (2002). Layered representations for human activity recognition. In *Pro. Int. Conf. Multimodal Interfaces*, pages 3–8. 130, 131

[Oneata et al., 2013] Oneata, D., Verbeek, J., and Schmid, C. (2013). Action and Event Recognition with Fisher Vectors on a Compact Feature Set. In *Proc. Int. Conf. Computer Vision*, pages 1817–1824, Sydney, Australia. IEEE. 39

[Peng et al., 2014] Peng, X., Wang, L., Wang, X., and Qiao, Y. (2014). Bag of visual words and fusion methods for action recognition : Comprehensive study and good practice. *Comput. Research Repository*. 39, 95, 106, 199

[Péteri et al., 2010] Péteri, R., Fazekas, S., and Huiskes, M. J. (2010). DynTex : a Comprehensive Database of Dynamic Textures. *Pattern Recognition Letters*. 197

[Poppe, 2010] Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6) :976 – 990. 30

[Rabiner, 1989] Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *Proc. of the IEEE*, volume 77, pages 257–286. 130

[Raptis and Soatto, 2010] Raptis, M. and Soatto, S. (2010). Tracklet descriptors for action modeling and video analysis. In *Proc. Europ. Conf. Computer Vision*, pages 577–590, Berlin, Heidelberg. 45, 46, 49

[Reddy and Shah, 2013] Reddy, K. K. and Shah, M. (2013). Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5) :971–981. 26, 43, 63, 106

[Rousseau et al., 2010] Rousseau, S., Helbert, D., and Carré, P. (2010). Metric Tensor for Multicomponent Edge Detection. In *Proc. IEEE Int. Conf. Image Processing*, page 2, Hong-Kong, China. 57

[Ryoo and Aggarwal, 2009] Ryoo, M. S. and Aggarwal, J. K. (2009). Semantic representation and recognition of continued and recursive human activities. *Int. J. Comput. Vision*, 82(1) :1–24. 133

[Sapienza et al., 2014] Sapienza, M., Cuzzolin, F., and Torr, P. H. S. (2014). Feature sampling and partitioning for visual vocabulary generation on large action classification datasets. *Comput. Research Repository*. 199

[Schindler and Van Gool, 2008] Schindler, K. and Van Gool, L. (2008). Action snippets : How many frames does human action recognition require ? In *Proc. Conf. Comp. Vision Pattern Rec.*, pages 1–8. 152, 178

[Schmid et al., 2000] Schmid, C., Mohr, R., and Bauckhage, C. (2000). Evaluation of interest point detectors. *Int. J. Comput. Vision*, 37(2) :151–172. 55, 63

[Schuldt et al., 2004] Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions : a local svm approach. In *Proc. Int. Conf. Pattern Recognition*, volume 3, pages 32–36 Vol.3. 22

[Scovanner et al., 2007] Scovanner, P., Ali, S., and Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. In *Proc. Int. Conf. on Multimedia*, pages 357–360, New York, NY, USA. 39, 43

[Shi et al., 2013] Shi, F., Petriu, E., and Laganiere, R. (2013). Sampling strategies for real-time action recognition. In *Proc. Conf. Comp. Vision Pattern Rec.* 48, 49, 109

[Shirazi et al., 2012] Shirazi, S., Harandi, M., Sanderson, C., and Alavi, A. (2012). Clustering on grassmann manifolds via kernel embedding with application to action analysis. In *Proc. IEEE Int. Conf. Image Processing*, pages 781– 784, Florida, USA. 139

[Simonyan et al., 2013] Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep fisher networks for large-scale image classification. In *Advances in Neural Information Processing Systems 26*, pages 163–171. 199

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems 27*, pages 568–576. 199

[Siskind, 2011] Siskind, J. M. (2011). Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Comput. Research Repository*. 134

[Sultani and Saleemi, 2014] Sultani, W. and Saleemi, I. (2014). Human action recognition across datasets by foreground-weighted histogram decomposition. In *Proc. Conf. Comp. Vision Pattern Rec.*, pages 764–771. 124

[Sun et al., 2010a] Sun, D., Roth, S., and Black, M. (2010a). Secrets of optical flow estimation and their principles. In *Proc. Conf. Comp. Vision Pattern Rec.*, pages 2432–2439. 67, 100, 101

[Sun et al., 2010b] Sun, J., Mu, Y., Yan, S., and Cheong, L.-F. (2010b). Activity recognition using dense long-duration trajectories. In *Proc. IEEE Int. Conf Multimedia and Expo*, pages 322–327. 76

[Tabbone et al., 2006] Tabbone, S., Wendling, L., and Salmon, J. P. (2006). A new shape descriptor defined on the radon transform. *Computer Vision and Image Understanding*, 102(1) :42–51. 32, 33

[Tang et al., 2012] Tang, K., Fei-Fei, L., and Koller, D. (2012). Learning latent temporal structure for complex event detection. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1250–1257. 183, 184

[Tavenard et al., 2013] Tavenard, R., Emonet, R., and Odobez, J.-M. (2013). Time-sensitive topic models for action recognition in videos. In *Proc. IEEE Int. Conf. Image Processing*, pages 2988–2992. 141, 146

[Torralba and Efros, 2011] Torralba, A. and Efros, A. A. (2011). Unbiased look at dataset bias. In *Proc. Conf. Comp. Vision Pattern Rec.* 118, 121, 122

[Turaga and Chellappa, 2009] Turaga, P. and Chellappa, R. (2009). Locally time-invariant models of human activities using trajectories on the grassmannian. In *Proc. Conf. Comp. Vision Pattern Rec.*, pages 2435–2441. 137, 138, 139

[Turaga et al., 2011] Turaga, P., Veeraraghavan, A., Srivastava, A., and Chellappa, R. (2011). Statistical computations on grassmann and stiefel manifolds for image and video-based recognition. *IEEE Trans. Pattern Anal. Machine Intell.*, 33(11) :2273–2286. 137, 138

[Uesaka, 1984] Uesaka, Y. (1984). A new fourier descriptor applicable to open curves. *Electronics and Communications in Japan*, 67(8) :1–10. 162

[Ullah and Laptev, 2012] Ullah, M. M. and Laptev, I. (2012). Actlets : A novel local representation for human action recognition in video. In *Proc. IEEE Int. Conf. Image Processing*, pages 777–780. 45

[Van de Sande et al., 2010] Van de Sande, K., Gevers, T., and Snoek, C. (2010). Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Machine Intell.*, 32(9) :1582–1596. 43

[Van Overschee and De Moor, 1991] Van Overschee, P. and De Moor, B. (1991). Subspace algorithms for the stochastic identification problem. In *Proc. IEEE Int. Conf. Decision and Control*, volume 2, pages 1321–1326. 137

[Veeraraghavan et al., 2005] Veeraraghavan, A., Roy-Chowdhury, A., and Chellappa, R. (2005). Matching shape sequences in video with applications in human movement analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(12) :1896–1909. 137

[Vrigkas et al., 2014] Vrigkas, M., Karavasilis, V., Nikou, C., and Kakadiaris, A. (2014). Matching mixtures of curves for human action recognition. *Computer Vision and Image Understanding*, 119 :27–40. 47

[Wang et al., 2011] Wang, H., Klaser, A., Schmid, C., and Liu, C.-L. (2011). Action recognition by dense trajectories. In *Proc. Conf. Comp. Vision Pattern Rec.*, pages 3169 –3176. 46, 48, 49, 68, 76, 90, 111

[Wang et al., 2013] Wang, H., Kläser, A., Schmid, C., and Liu, C.-L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *Int. J. Computer Vision*, 103(1) :60–79. 49, 109

[Wang et al., 2009] Wang, H., Muneeb Ullah, M., Kläser, A., Laptev, I., and Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. In *University of Central Florida, U.S.A.* 42, 54, 55, 63, 109, 198

[Wang and Schmid, 2013] Wang, H. and Schmid, C. (2013). Action Recognition with Improved Trajectories. In *Proc. Int. Conf. Computer Vision*, pages 3551–3558, Sydney, Australie. 46, 80, 81, 106, 109, 197

[Wang et al., 2015] Wang, L., Qiao, Y., and Tang, X. (2015). Action recognition with trajectory-pooled deep-convolutional descriptors. *Comput. Research Repository.* 199

[Wang et al., 2007] Wang, Y., Sabzmeydani, P., and Mori, G. (2007). Semi-latent dirichlet allocation : A hierarchical model for human action recognition. In *Proc. Conf. Human Motion*, pages 240–254, Berlin, Heidelberg. 145, 146

[Weinland et al., 2006] Weinland, D., Ronfard, R., and Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2) :249–257. 32

[Weinzaepfel et al., 2013] Weinzaepfel, P., Revaud, J., Harchaoui, Z., and Schmid, C. (2013). Deepflow : Large displacement optical flow with deep matching. In *Proc. Int. Conf. Computer Vision*, pages 1385–1392. 101

[Willems et al., 2008] Willems, G., Tuytelaars, T., and Gool, L. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proc. Europ. Conf. Computer Vision*, pages 650–663, Berlin, Heidelberg. 39, 41, 54

[Wilson and Bobick, 2002] Wilson, A. D. and Bobick, A. F. (2002). Hidden markov models. chapter Hidden Markov Models for Modeling and Recognizing Gesture Under Variation, pages 123–160. World Scientific Publishing Co., Inc., River Edge, NJ, USA. 37

[Worgan et al., 2011] Worgan, S. F., Behera, A., Cohn, A. G., and Hogg, D. C. (2011). Exploiting petri-net structure for activity classification and user instruction within an industrial setting. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, pages 113–120, New York, NY, USA. ACM. 131

[Wu et al., 2004] Wu, T.-F., Lin, C.-J., and Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5 :975–1005. 152

[Wulff et al., 2012] Wulff, J., Butler, D. J., Stanley, G. B., and Black, M. J. (2012). Lessons and insights from creating a synthetic optical flow benchmark. In *Proc. Europ. Conf. Computer Vision*, pages 168–177. 101

[Xiao et al., 2014] Xiao, J., Ehinger, K., Hays, J., Torralba, A., and Oliva, A. (2014). Sun database : Exploring a large collection of scene categories. *Int. J. Computer Vision*, pages 1–20. 118

[Yamato et al., 1992] Yamato, J., Ohya, J., and Ishii, K. (1992). Recognizing human action in time-sequential images using hidden Markov model. In *Proc. Conf. Comp. Vision Pattern Rec.*, pages 379–385. 37

[Yuan et al., 2011] Yuan, J., Liu, Z., and Wu, Y. (2011). Discriminative video pattern search for efficient action detection. *IEEE Trans. Pattern Anal. Machine Intell.*, 33(9) :1728–1743. 28

[Zhang and Lu, 2002] Zhang, D. and Lu, G. (2002). A comparative study of fourier descriptors for shape representation and retrieval. In *Proc. Asian Conf. Computer Vision*, pages 646–651. 85

[Zhang et al., 2006] Zhang, J., Marszalek, M., Lazebnik, S., and Schmid, C. (2006). Local features and kernels for classification of texture and object categories : A comprehensive study. In *Proc. Conf. Comp. Vision Pattern Rec.*, pages 13–13. 93

[Zhou et al., 2015] Zhou, Y., Yu, H., and Wang, S. (2015). Feature sampling strategies for action recognition. *Comput. Research Repository*. 199

Analyse et reconnaissance de séquences vidéos d'activités humaines dans l'espace sémantique

Résumé Dans cette thèse, nous nous intéressons à la caractérisation et la reconnaissance d'activités humaines dans des vidéos. L'intérêt grandissant en vision par ordinateur pour cette thématique est motivé par une grande variété d'applications telles que l'indexation automatique de vidéos, la vidéo-surveillance, ou encore l'assistance aux personnes âgées.

Dans la première partie de nos travaux, nous développons une méthode de reconnaissance d'actions élémentaires basée sur l'estimation du mouvement dans des vidéos. Les points critiques du champ vectoriel obtenu, ainsi que leurs trajectoires, sont estimés à différentes échelles spatio-temporelles. La fusion tardive de caractéristiques d'orientation de mouvement et de variation de gradient, dans le voisinage des points critiques, ainsi que la description fréquentielle des trajectoires, nous permet d'obtenir des taux de reconnaissance parmi les meilleurs de la littérature.

Dans la seconde partie, nous construisons une méthode de reconnaissance d'activités en considérant ces dernières comme un enchainement temporel d'actions élémentaires. Notre méthode de reconnaissance d'actions est utilisée pour calculer la probabilité d'actions élémentaires effectuées au cours du temps. Ces séquences de probabilité évoluent sur une variété statistique appelée *simplexe sémantique*. Une activité est finalement représentée comme une trajectoire dans cet espace. Nous introduisons un descripteur fréquentiel de trajectoire pour classifier les différentes activités humaines en fonction de la forme des trajectoires associées. Ce descripteur prend en compte la géométrie induite par le simplexe sémantique.

Mots clés : Reconnaissance d'activités humaines, Analyse multi-échelle, Caractérisation fréquentielle de trajectoires, Simplexe sémantique.

Analysis and recognition of human activities in video sequences in the semantic space

Summary This thesis focuses on the characterization and recognition of human activities in videos. This research domain is motivated by a large set of applications such as automatic video indexing, video monitoring or elderly assistance.

In the first part of our work, we develop an approach based on the optical flow estimation in video to recognize human elementary actions. From the obtained vector field, we extract critical points and trajectories estimated at different spatio-temporal scales. The late fusion of local characteristics such as motion orientation and shape around critical points, combined with the frequency description of trajectories allow us to obtain one of the best recognition rate among state of art methods.

In a second part, we develop a method for recognizing complex human activities by considering them as temporal sequences of elementary actions. In a first step, elementary action probabilities over time is calculated in a video sequence with our first approach. Vectors of action probabilities lie in a statistical manifold called *semantic simplex*. Activities are then represented as trajectories on this manifold. Finally, a new descriptor is introduced to discriminate between activities from the shape of their associated trajectories. This descriptor takes into account the induced geometry of the simplex manifold.

Keywords : Human activity recognition, Multi-scale analysis, Frequency analysis of motion trajectories, Semantic simplex.

Laboratoire Mathématiques, Image et Applications
Avenue Michel Crépeau
17042 La Rochelle Cedex 01