N° d'ORDRE :05/2013- M/INF

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET

DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DES SCIENCES ET DE LA TECHNOLOGIE

« HOUARI BOUMEDIENNE »

FACULTE D'ELECTRONIQUE ET INFORMATIQUE

MEMOIRE

Présenté pour l'obtention du diplôme de MAGISTER

En : Informatique

Spécialité : Intelligence Artificielle et Base de Données Avancées

Par : Melle HIMRI KHADIDJA

Sujet

RECONNAISSANCE DES ACTIONS HUMAINES PAR CLASSIFICATION

Soutenu publiquement le 29/10/2013, devant le jury composé de :

Mme A. SERIR BENMERABET	*Prof*	*USTHB/FEI*	*Présidente*
Mr. S.LARABI	*Prof*	*USTHB/FEI*	*Directeur de Thèse*
Mr. H. AZZOUNE	*MCA*	*USTHB/FEI*	*Examinateur*
Mme N.TOUZENE BAHA	*MCA*	*USTHB/FEI*	*Examinatrice*
Mr. MEZIANE Abdelkrim	*D.R*	*CERIST*	*Examinateur*

Résumé

Depuis très longtemps, les chercheurs ont été fascinés par la capacité du système de vision humaine à percevoir l'espace qui l'entoure sans aucune difficulté. Avec la naissance de machines de plus en plus puissantes et les progrès techniques au niveau du traitement automatique des vidéos, une nouvelle discipline est apparue sous le nom : « vision par ordinateur ». Et lors de ces dernières années, le domaine de la reconnaissance des actions humaine a connu une très forte expansion.

Aujourd'hui, il est très difficile d'aborder le problème de la reconnaissance des actions dans toute sa diversité. Ainsi, un grand nombre d'actions nécessitent la reconnaissance des objets manipulés par ces dernières, leurs attributs, leurs états, et les changements qui les affectent.

Dans notre manuscrit, nous avons présenté les différentes représentations des caractéristiques : les images de l'historique du mouvement (MHI), Flot optique et les caractéristiques robustes accélérées (SURF) sur la base de données Weizman, qui se compose de dix différentes classes d'action: se pencher, courir, marcher, sauter, sauter en écartant les jambes et les bras , sauter sur une seule jambe, sauter en place, galoper sur le côté, sauter en agitant les deux mains, et sauter en agitant une seule main, afin d'extraire les caractéristiques les plus importantes d'une séquence vidéo et décrire les principales méthodes permettant de fournir ces caractéristiques.

Après l'étape d'extraction des caractéristiques, nous utilisons des techniques de reconnaissance des formes pour réaliser la classification, à savoir les Modèles de Markov Cachés(HMMs).

L'objectif de l'extraction des caractéristiques est de permettre l'exploitation du contenu de la vidéo à des fins d'interprétation (aide à fournir un diagnostic dans le domaine aérien, satellitaire, médical par exemple), de localisation et de reconnaissance (vidéosurveillance, contrôle robotique).

Notre meilleur taux de reconnaissance moyen est de 81.11 %. Les résultats obtenus sont assez bons par rapport à ce que nous trouvons dans la littérature, avec une représentation de caractéristiques plus simple et un classifieur HMM plus au moins rapide. Cependant, plusieurs perspectives sont envisagées pour améliorer ce taux, d'où une étude plus approfondie du choix des caractéristiques permettrait d'améliorer et renforcer notre choix de classifieur HMMS.

Mots-clefs: Extraction des caractéristiques, les images de l'historique du mouvement (MHI), flot optique et les caractéristiques robustes accélérées (SURF), classification des actions humaines, les Modèles de Markov Cachés(HMMs).

Abstract

For many years, researchers have been motived to understand the performance capability of the human visual system for perceiving and tracking the objects in the space around us without difficulty. With the introduction of more powerful machines and evolution of automatic video processing that known many scientific and technical advances these last years, a novel discipline in computer science has been appeared and have become known as «computer vision». The human action recognition domain is a challenging problem that has received considerable attention and shows very strong growth in recent years.

Most of applications require recognition of high-level activities, composed of multiple simple actions of persons.In this thesis we focus on the human action recognition problem and we introduce three different representations: Motion History Image(MHI), optical flow and Speeded Up Robust Features (SURF). For training and testing actions, we use Weizmann database that contains 9 actors, each actor performing 10 actions such as “bending”, “running”, “walking”, “jumping”, “jacks”, “pjumping”, “gallop-sideways”, “two-hand waving”, “one-hand waving”, and “skipping”. Our system based mainly on these features representations which allow extracting information from video sequences in order to recognize a human action.

After step of features extraction, in order to reach the classification we proposed a recognition algorithm, named Hidden Markov Model (HMM).
Understanding human activities in video has many potential applications including the video surveillance, robotic control and diagnosis in each of medical, and satellite area.

The experimental results show the strength of HMM algorithm to recognize human action for rate of 81.11 %.
We have obtained good results compared with what we found in the literature. Our classifier seems to solve speedily a recognition problem. However a many ideas of perspective can be proposed to increase this classification rate.
Thus, further study of feature selection plays an important role in human action recognition to confirm and strengthen our choice of HMMS classifier.

Keywords: Features extraction, Motion History Image, optical flow, Speeded Up Robust Features (SURF), human action recognition, Hidden Markov Models (HMMs).

Remerciements

Tout d'abord merci LE TOUT PUISSANT de m'avoir accordé la volonté et l'amour du savoir ……

Je ne peux commencer mes remerciements sans évoquer la personne qui sans elle ce travail n'aurait pu aboutir mon Directeur de thèse le Professeur Mr S. LARABI qui m'a tout d'abord supporté et soutenu tout le long de mon parcours de magister.

Mes remerciements s'adressent également aux membres du jury, le professeur A. SERIR qui m'a fait l'honneur de présider le jury & aux docteurs N. BAHA et H.AZZOUNE pour avoir accepté d'être membre de jury et sans oublier bien sûr notre invité d'honneur Mr Abdelkrim MEZIANE.

Enfin, je voudrais exprimer mes plus profonds remerciements et ma plus grande gratitude à ma famille et surtout à SALOHTI, ma maman, mon papa à et à notre BEAU GOSSE Habibo……OUPS j'ai failli oublier les petits bouts de choux de notre famille.

Même si j'ai le soleil je ne renoncerai jamais à toi ma chandelle,

tu as illuminé ma vie pendant 16 ans... Je t'adore ma reine rouge NADJIBA <3

KHADIDJA

TABLE DES MATIERES

CHAPITRE 3: Méthode proposée : Approche et résultats de tests

LISTE DES FIGURES

Introduction Générale

Au cours des dix dernières années, les ordinateurs ont influencé nos vies d'une manière fondamentale. Ils effectuent des calculs intenses et répétitifs sur des bases de données très larges. De cette manière ils ont étendu nos possibilités de travailler et communiquer. Avec les progrès technologiques récents, les données vidéo sont devenues de plus en plus accessibles et jouent un rôle de plus en plus important dans notre vie quotidienne. Cependant, malgré leur importance croissante, les possibilités de les analyser d'une façon automatisée sont plutôt limitées. Les systèmes de vision par ordinateur sont loin d'être comparables à la vision humaine. Celle-ci a suscité l'intérêt de nombreux scientifiques depuis déjà très longtemps, des recherches théoriques et expérimentales ont été conduites pour comprendre l'anatomie et le fonctionnement du cerveau dans son ensemble et de la partie visuelle particulièrement. Une structure très complexe a été découverte et qui est loin de leur avoir révélé tous ses secrets.

Dès les années 60, un certain nombre de scientifiques se sont attaqués au problème de la vision d'un point de vue quantitatif : est-il possible de construire un modèle computationnel pour la perception visuelle ? En d'autres termes, c'est créer un modèle qui vu de l'extérieur possède des propriétés semblables au système visuel humain.

La vision artificielle atteint aujourd'hui une certaine fiabilité, et est utilisée dans différents secteurs tels que :

- le secteur grand public (caméscopes, photographies numériques, webcam, cellules photographiques de téléphones portables, etc.),
- le secteur automobile (anticollision, détecteurs d'obstacles, etc.)
- les applications demandant des vitesses de lecture particulièrement élevées (test de crashs, analyse d'explosion, etc.)
- la vidéo surveillance qui permet de minimiser les risques d'intrusion et les risques d'incendies dans les établissements industriels.
- la défense et l'imagerie médicale.

L'analyse du mouvement humain reste un sujet complexe de par la variété des situations et l'anatomie de l'individu. En effet, le corps humain est un objet 3D, doté d'articulations comportant plusieurs degrés de liberté et capable d'effectuer une infinité de mouvements à vitesse variable. Cette complexité soulève un problème intéressant dans la vision par ordinateur celui de trouver des représentations simplifiées qui permettent d'analyser et d'interpréter toutes les informations qu'elles contiennent.

Les systèmes de vision en question permettent leur extraction d'une scène à partir de caméras. D'un point de vue technique, le fait de travailler avec des images, ou des séquences d'images dans le cas de la vidéo, demande la manipulation de grandes quantités d'informations, qui prises à un instant, doivent subir successivement de nombreux traitements pour obtenir le résultat final. Le système doit aussi gérer le fait que les ressources telles que les caméras soient distribuées et non centralisées afin de recueillir le maximum de vues et permettre une vision plus large sur l'environnement en question.

Le travail traité dans ce mémoire concerne la reconnaissance des actions moyennant la classification. Concrètement, il est présenté en trois chapitres :

Dans le premier chapitre, nous présentons un état de l'art des principales méthodes proposées dans ce domaine dont les représentations sont basées soit sur l'apparence, le volume, le flux optique ou sur les points d'intérêt. Nous verrons que le problème posé est loin d'être résolu et reste encore parmi les thèmes abordés dans les conférences scientifiques de la communauté de vision artificielle.

Le second chapitre est consacré aux outils de traitement d'images permettant l'extraction de primitives pertinentes pour les images ou la vidéo. Nous donnerons avec détails le flot optique, et le descripteur SURF qui sont utilisées abondamment en imagerie.

Le troisième chapitre est dédié à notre proposition de solution. Nous exposerons dans l'ordre chronologique les travaux menées pour d'abord définir les caractéristiques à extraire de la vidéo et ensuite leur classification par les modèles de Markov cachés(HMMs). Nous présentons et discutons aussi les tests effectués sur la base de données Weizmann. Nous finissons par une conclusion dans laquelle un ensemble de perspectives sont suggérés.

Chapitre 1 : Reconnaissance d'Actions Humaines : Etat de l'Art

1.1 Introduction

Depuis l'antiquité, le mouvement humain a fait l'objet d'études approfondies. En effet, en 1887, le photographe américain d'origine anglaise Eadweard Muybridge en donne la preuve en réalisant une série d'instantanés photographiques décomposant les mouvements d'un cheval au galop (voir figure 1.1 (a)). C'est le premier à s'être penché sur l'analyse du mouvement de l'animal et de l'homme [MUY 79] avant de focaliser sur des mouvements tels la marche humaine (voir figure 1.1 (b)).

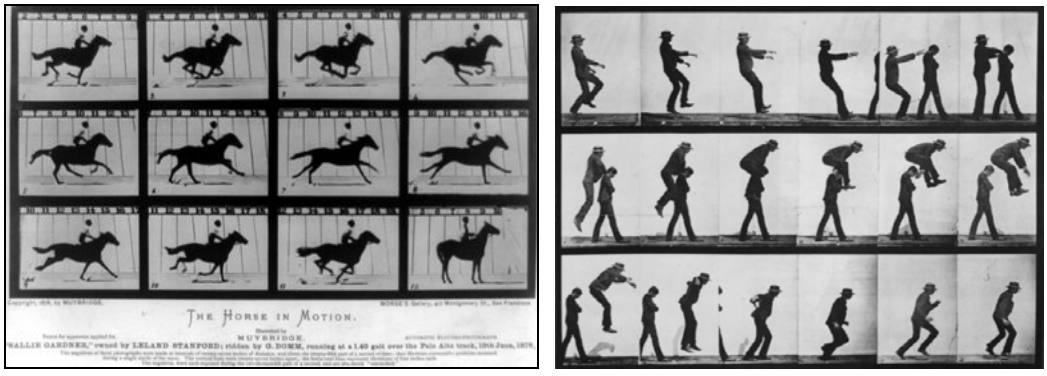


Figure 1.1 : Mouvements décomposés d'un cheval au galop (a) et saut d'un Homme (b) [MUY 79].

La recherche menée sur l'interactivité homme-machine a eu des développements considérables depuis que E. E. Sutherland a introduit la manipulation d'objets virtuels [SUT 63].

L'explosion technologique et en particulier les interfaces dites immersives (gants de données, caméras numériques, visiocasques, capteurs, etc.) ont largement contribué à enrichir les études dans ce domaine.

Avant d'entamer le sujet de « La Reconnaissance de l'Action Humaine », nous définissons tout d'abord l'action humaine et son identification par la vision artificielle.

1.2 Analyse du Mouvement Humain

1.2.1 Introduction

Parmi les mécanismes reliés à la vision, l'analyse du mouvement du corps humain (marche, saut, gestes, etc.) est l'une des plus importantes. C'est un sujet complexe de par la variété des situations et l'anatomie du corps humain. De nombreuses méthodes ont été réalisées par la communauté vision par ordinateur afin de reconnaître les mouvements du corps humain. La littérature sur ce thème est très abondante, plus de 350 publications durant cette dernière décennie [MOE 06], nous pourrons se reporter par exemple sur [AHM 06] [GAV 99][HU 04][MOE 01].

1.2.2 Action humaine

Dans le domaine de la vision par ordinateur, il parait naturel de tirer parti d'une analyse du mouvement afin d'interpréter le contenu des scènes observées. Le comportement humain se définit par ses mécanismes d'action. Une action est une séquence de mouvements qui sont causés par une raison, elle peut être longue ou courte, mais essentiellement elle n'est pas terminée tant que le but n'est pas atteint. Les êtres humains sont capables de reconnaître facilement les actions humaines. Par conséquent, le développement des techniques qui donnent des capacités similaires sur des ordinateurs est une tâche difficile et délicate.

1.2.3 Domaine d'application

La reconnaissance d'action humaine est abordée par le domaine de la vision par ordinateur avec beaucoup d'idées, elle est au cœur de nombreuses applications; telles que les interfaces homme-machine, l'interprétation de scènes pour la vidéo surveillance, l'indexation dans de grandes bases de vidéos, la recherche vidéo sur différents types d'actions, etc. C'est un domaine de recherche fascinant et insuffisamment approfondi, car la dynamique du corps humain en mouvement est illimitée. Une telle reconnaissance a pour objectif de reconnaître les actes de l'être humain à partir de vidéos (séquences d'images).

Figure 1.2 : Exemples de domaines d'application de la reconnaissance des actions humaines

1.3 Etat de l'art de la reconnaissance des actions

La reconnaissance d'actions humaines, qui donne lieu à une grande diversité d'approches, vise à comprendre la structure des mouvements de l'homme à partir de séquences vidéo, en classant les actions dans des catégories connues, comme la marche, le saut, etc. Généralement, la reconnaissance d'un objet se base sur l'extraction d'informations intéressantes et pertinentes de l'image. Par conséquent, l'extraction des caractéristiques nécessitera la mise en œuvre d'algorithmes robustes et efficaces. Et les méthodes proposées dans la reconnaissance d'actions humaines dans les vidéos sont très variées. Elles peuvent être classées en quatre grandes catégories:

- La méthode des représentations basées sur l'apparence,
- La méthode des représentations basées sur le flux optique,
- La méthode des représentations basées sur les points d'intérêt,
- La méthode des représentations basées sur le volume.

Nous détaillons les différentes méthodes citées précédemment comme suit :

1.3.1 Méthode des représentations basées sur l'apparence

La représentation basée sur l'apparence est l'une des techniques utilisées dans la littérature pour la reconnaissance d'objet, plus précisément dans la reconnaissance d'actions humaines. C'est une étape clé pour un grand nombre d'applications de Vision par Ordinateur.

Cette représentation est un indice pour mettre en correspondance les objets au cours du temps dans une séquence d'images. Elle peut être une caractéristique de couleur, de forme ou de texture. Son principe consiste à représenter chaque objet de la scène par un ensemble d'images d'intensité prises à partir de plusieurs points de vue et avec toutes les directions d'illumination possibles. Sa méthodologie générale est de faire un apprentissage aux modèles d'apparence du corps humain ou de la main et les faire correspondre correctement aux images dans les séquences vidéo pour la reconnaissance d'actions et gestes [BOB 01] [YAM 92][YU 05]. Parmi les premières applications avec modèle d'apparence [AHM 06][HOG 83][PIE 05] [TAR 95], nous citons le système de l'université de LEEDS d'Adam Baumberg [BAU 95] [HOG 83].

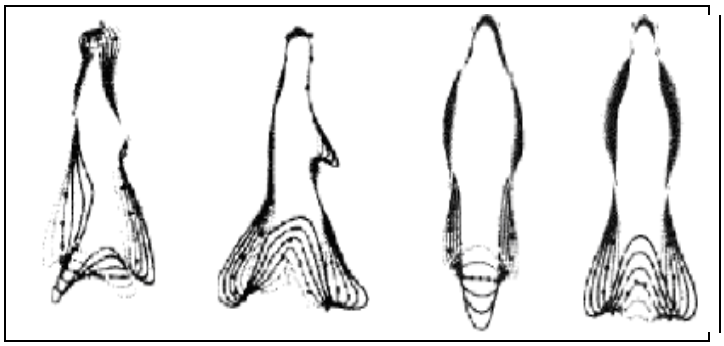


Figure 1.3 : Modèle d'apparence 2D du corps humain [HOG 83].

De nombreuses recherches ont été menées pour développer l'aspect théorique et pratique de cette technique. Ce problème a été résolu par l'analyse en composante principale (ACP), programmation linéaire (PL), modèle de Markov caché [DAV 03][JIA 06][RAH 06][YAM 92][YU 05].

Dans [YAM 92], les auteurs analysent les mouvements de tennis grâce aux caractéristiques des blobs de mouvements, de couleur et de texture. Ils utilisent ensuite ces caractéristiques pour classifier les différents types de mouvements en utilisant le Modèle de Markov Caché (HMM) et en s'appuyant sur des techniques de quantification vectorielle sur des séquences d'images binaires. Malgré que toutes les informations relatives à la dynamique de l'action sont contenues dans le volume 3D, nul besoin d'extraire le mouvement dans les différentes images de la séquence, étape délicate à réaliser vu la non-rigidité du corps humain. Ces méthodes présentent aussi d'autres problèmes, comme la disparition de mouvement qui se produit dans les régions de la silhouette et aussi la variante information contour/bord l'information qui surgit quand l'arrière-plan ou les vêtements sont d'une haute fréquence au lieu d'une basse fréquence (comme dans la plupart des scènes naturelles) [BOB 01]. Sachant que l'inconvénient majeur des approches basées sur des automates est que le modèle de geste doit être modifié dès qu'un nouveau geste doit être reconnu. De plus, la complexité informatique de telles approches est généralement énorme puisque c'est proportionnel aux gestes à reconnaitre qui n'est pas le cas pour des méthodes basées sur d'autres outils [BEC 09].

Mokhber et al. [MOK 08] ont développé une technique de reconnaissance d'actions où les points binaires forment dans l'espace 3D un volume qui est caractérisé par ses moments géométriques 3D. Les auteurs ont utilisé une gaussienne pour modéliser le vecteur de caractéristique. Les taux de reconnaissance sont acceptables. Il y a en effet certaines limites telle que la confusion entre classes aussi le nombre de séquence d'apprentissage semble insuffisant.

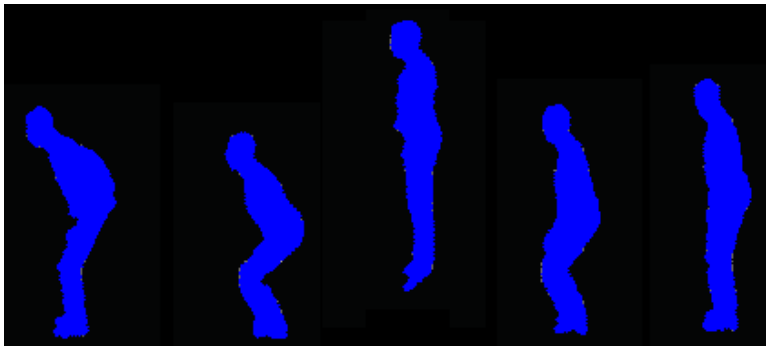


Figure 1.4 : Séquence d'image d'une personne en mouvement [MOK 08]

Parmi les pionniers de ces approches, nous pouvons citer la méthode proposée par Bobick et Davis [BOB 01] approuvée par de nombreux auteurs. Ils proposent une représentation spatio-temporelle d'activités humaines ou ils définissent les images d'énergie de mouvement «MEI» et les images de l'historique du mouvement «MHI» (chaque valeur du pixel est une fonction de l'historique temporel du mouvement) pour la représentation et la reconnaissance du mouvement humain (voir figure 1.5). Les images obtenues sont décrites avec les invariants de Hu [HU 62], et les gestes sont classifiés en utilisant la distance de Mahalanobis.

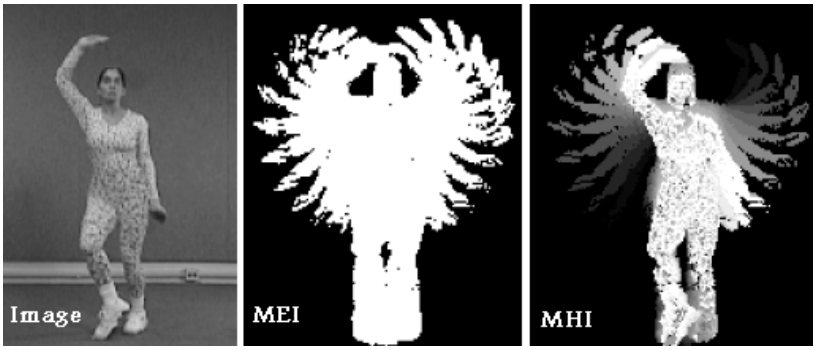


Figure 1.5 : Une image, son MEI et son MHI [BOB 01]

Les auteurs utilisent la représentation MHI comme une base des histogrammes de mouvement. Ainsi, ils génèrent le mouvement entre fenêtres en faisant la différence successive des images binaires de la silhouette de la personne. La différentiation entre images cause deux problèmes :

- Premièrement est que nous ne pouvons pas indiquer la magnitude ou la direction du mouvement mais seulement sa présence.
- Deuxièmement, il est difficile d'enlever le mouvement non désiré (exemple quand la personne agite sa tète).

Davis et Tyagi [DAV 03] ont présenté une méthode de reconnaissance d'action basée sur le concept d'inférence fiable. Cette approche est formulée dans une structure probabiliste utilisant un classement a posteriori afin de vérifier une entrée avant de l'affecter à la classe d'actions appropriées. Ainsi ils montrent qu'un taux d'erreur inférieur de la loi de Bayes peut être atteint comparant a d'autres méthodes probabilistes à savoir l'approche du maximum de vraisemblance (ML) et MAP (maximum a postériori). Dans le domaine de la reconnaissance d'action, la majorité des travaux ont été focalisés sur l'utilisation des modèles d'espace-état, ces modèles (espace-état) sont présentés avec les algorithmes permettant d'en estimer les variables cachées et les paramètres, notamment le filtre de Kalman et l'algorithme Expectation Maximisation. Cependant, ces méthodes utilisent des modèles non-linéaires et n'ont pas de solution au problème de forme fermée [MEN 09].

Black et al. [BLA 98] ont construit un modèle des changements d'apparence dans une séquence d'image, à l'aide de mélanges de lois. La méthode développée permet de différencier quatre sources de changements d'intensité : l'évolution de la forme des objets, les variations d'illumination, la spécularite, les changements iconiques. Parmi les inconvénients d'une telle technique est que nous avons besoin d'un sous-espace d'apprentissage «base propre» d'un objet à chaque point de vue avant le suivi [ROS 04]. D'autre part l'utilisation d'une modélisation complexe de la variation de luminance nécessite d'estimer un grand nombre de paramètres et conduit à des solutions peu stables. Notons aussi que les caractéristiques ne peuvent pas être extraites pour des applications réelles, ce qui diminue la performance de processus de la reconnaissance.

Ramanan et Forsyth [RAH 03] ont abordé le problème de la classification automatique de la structure et l'apparence du modèle humanoïde par groupement,

l'idée est de créer des vecteurs de caractéristiques en initialisant chaque point de ces caractéristiques en appliquant l'algorithme Mean-shift. Dans cet article les membres du corps sont détectés et regroupés par une approche ascendante «bottom up » (voir figure 1.6). L'inconvénient principal de cette méthode est la difficulté de la localisation indépendante des membres (têtes, mains, bras, jambes). Dans cette classe de méthodes, les actions sont représentées par des grammaires dont les éléments terminaux sont les postures d'un modèle anatomique du corps humain.

Figure 1.6 : Localisation des membres par approche ascendante [RAH 03].

Dans [FEL 05] les auteurs ont proposé des modèles d'apparence locaux basés sur les membres pour reconnaître les personnes en mouvement. Le principe repose sur la soustraction d'arrière-plan. Parmi les inconvénients de cette méthode est qu'elle est relativement coûteuse en temps de calcul, et ne peut se faire qu'en limitant l'espace de recherche des hypothèses. Aussi c'est une méthode qui ne prend pas en compte des auto-occultations.

Jiang et al. [JIA 06] ont aussi utilisé la représentation à base d'apparence pour la reconnaissance d'actions en cherchant des positions statiques. Ils ont étudié les modèles géométriques des mouvements du corps humain dans des séquences d'images, et ont proposé un algorithme fondé sur la méthode de relaxation successive et ont résolu le problème de correspondance comme problème de minimisation de l'énergie, après transformation du problème non convexe en un problème convexe. Cependant, leur approche est formulée comme une optimisation sur les coefficients d'interpolation associés à ces points convexes.

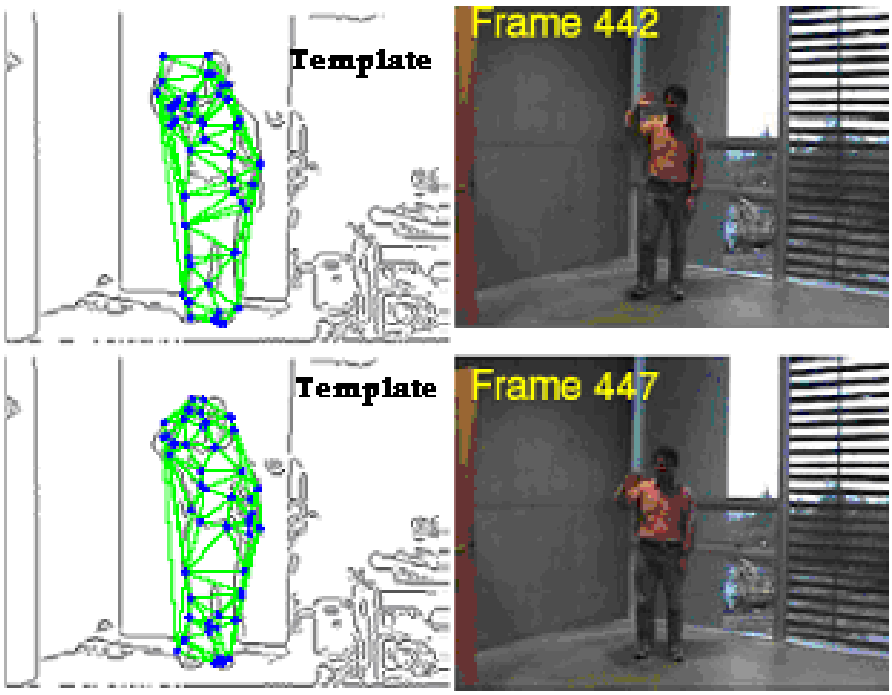


Figure 1.7 : Séquence d'images et de modèles [JIA 06]

Mikolajczyk et Uemura proposent une approche pour la reconnaissance d'action basée sur un vocabulaire de caractéristiques des mouvements apparents locaux et sur la rapidité dans la recherche approximative dans un grand nombre d'arbres [MIK 11]. Un grand nombre de caractéristiques avec des vecteurs de mouvement associés sont extraites des données vidéo et sont représentées par de nombreux arbres. Plusieurs détecteurs de points d'intérêt sont utilisés pour fournir des fonctionnalités pour chaque trame. Ces vecteurs sont estimés en utilisant le flux optique et une mise en correspondance du descripteur de base. Les caractéristiques sont combinées avec la segmentation d'image pour estimer les homographies dominantes, et ensuite séparées en des caractéristiques fixes et mobiles en dépit du mouvement de la caméra. Il a été démontré que cette méthode est robuste aux variations de l'apparence, le mouvement de caméra, changement d'échelle, des actions asymétriques, l'encombrement de l'arrière-plan et de l'occlusion.

1.3.2 Méthode des représentations basées sur le volume

L'œil humain est capable de reconnaître rapidement une multitude d'objets différents avec des variations de point de vue, d'illumination et de forme, même dans les contextes les plus inhabituels. C'est un domaine qui peut être défini par l'ensemble des techniques qui permettent d'explorer, d'extraire, de modéliser des données à dimensions élevées sous une forme graphique compréhensible.

Dans ces représentations, les actions humaines sont des représentations volumiques 3D de vidéos. En général, ces approches ont pour but la classification des séquences de vidéo courte d'actions humaine. Afin d'améliorer la performance de la classification l'extraction des caractéristiques à partir de la silhouette et du mouvement est appliquée pour résoudre le problème de la reconnaissance d'actions. La littérature sur ce thème est très abondante, parmi celles-ci nous pouvons citer :

Chomat et al. [CHO 98] ont proposé une représentation par l'arbre quaternaire basée sur l'analyse spatio-temporelle d'une séquence d'image. La séquence spatio-temporelle de pixels dans une région d'une séquence d'image est projetée sur une base de filtres spatio-temporels réalisée avec les filtres de Gabor. Chaque activité est caractérisée par un histogramme à douze dimensions qui donne une estimation de la densité de probabilité nécessaire à un processus de reconnaissance basé sur une règle de Bayes. L'avantage des modèles de mise en correspondant est la basse complexité informatique et la simplicité de la mise en œuvre. Cependant, elle est fréquemment plus sensible au bruit et aux variations de la durée de mouvements et elle dépend du point de vue [WAN 02].

Figure 1.8 : Les formes de l'espace-temps "Space-time" pour les actions Saut, Marche et Courir [GOR 07]

Yilmaz et Shah [YIL 05] ont construit un volume spatio-temporels « spatio-temporal volume » (STV), et ont présenté une méthode de soustraction de fond pour la description des actions. En effet, ils se sont basés sur l'extraction des silhouettes par soustraction de fond « l'arrière-plan ». Bien que leur méthode soit robuste aux changements de point de vue, elle ne peut pas distinguer des actions dont les mouvements sont parallèles à la caméra, comme le déplacement d'une main en arrière et en avant [ROH 09].

Figure 1.9 : Silhouette d'action de marche [YIL 05]

Pierobon et al. [PIE 05] ont proposé une méthode d'extraction d'information basée sur la représentation volumique 3D. Les auteurs ont utilisé l'algorithme DTW « Dynamic Time Warping » afin de calculer les distances entre les mouvements. Cependant l'inconvénient de cette approche réside dans le calcul complexe et la fiabilité des différents algorithmes utilisés, c'est une technique extrêmement coûteuse en temps de calcul.

Weinland et al. [WEI 06] ont proposé une méthode fondée sur l'analyse de Fourier dans une représentation volumique 3D, le principe était de calculer des descripteurs du mouvement permettant de reconnaître les actions de façon indépendante du point de vue et de la personne qui exécute l'action. Pour cela ils ont adapté la technique des images d'histoire du mouvement sous la forme de volumes d'histoire du mouvement. Afin d'avoir une représentation invariante au point de vue, la trace du volume occupé par l'acteur est calculée dans une représentation cylindrique autour d'un axe vertical partant du centre de gravité de l'acteur. Cependant les caractéristiques de mouvement employées sont relativement complexes, ce qui implique un coût important de résultat en reconstruction des caractéristiques.

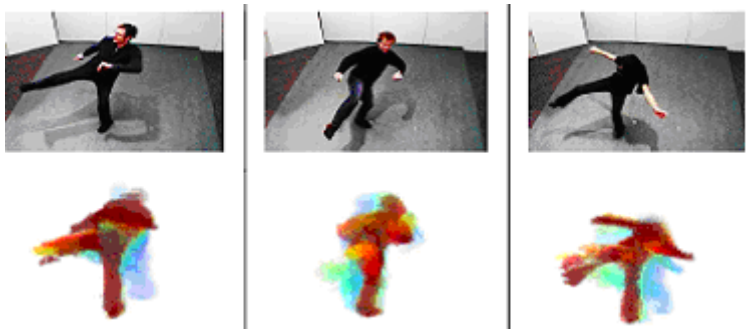


Figure 1.10 : Exécution de l'action d'un coup de pied par différents acteurs [WEI 06]

La représentation volumique constitue une alternative prometteuse pour résoudre les problèmes de robustesse que connaît la représentation 2D. Cependant, elle n'a pas encore atteint une certaine maturité, à cause notamment de la lourdeur du processus d'acquisition, et de la non disponibilité de grandes bases de données volumiques, à accès libre, afin de tester et d'évaluer les techniques élaborées.

1.3.3 Méthode des représentations basées sur le flux optique

La notion du flux optique a été abordée en 1955 par Gibson [GIB 55]. Dans [GIB 55][WAR 88], le flux optique est défini comme les modifications temporelles du profil d'intensités lumineuses dans des directions différentes au point d'observation en mouvement. L'intensité de l'image est une fonction continue de position et de temps I(x, t) en supposant que l'intensité des régions

d'images ne change pas en mouvement. Cette hypothèse peut rencontrer des difficultés avec les modèles lumineux, spéculaires, etc. Plusieurs auteurs ont entamé cette représentation dans le domaine de la détection de mouvement, avec la méthode proposée par Elzein et al. [ELZ 03] fondée sur le principe du flux optique [HOR 81]. Le but est de définir des zones d'intérêt en cherchant les régions de l'image contenant le mouvement. La finalité de la méthode revient à détecter les collisions potentielles.

Afin d'étudier le mouvement de l'être humain Yacoob et Davis [YAC 96] ont analysé le flux optique en utilisant des règles heuristiques au lieu d'un modèle physique. Ils ont tracé des traits caractéristiques sur la première image puis ils ont suivi ces traits caractéristiques tout au long de la séquence. C'est le flux optique qui détermine le suivi de l'évolution des points particuliers (points de fort gradient d'intensité) et des zones d'intérêt afin d'analyser des expressions faciales.

Cutler et Turk [CUT 98] proposent l'utilisation du flux optique pour la reconnaissance de sept gestes de bras. Ils considèrent le fond statique et les changements de luminosité faibles. Le flux optique calculé entre deux images successives est segmenté en zones par l'algorithme K-mean. Ils utilisent la taille et le déplacement de taches dans l'image pour reconnaître le mouvement. Une difficulté principale associée à l'extraction des caractéristiques à bas niveau est que la main doit être localisée avant l'extraction des caractéristiques. La localisation de mains dans des scènes arbitraires a prouvé la difficulté de la tâche [DER 04]. Cependant, ces approches restent très sensibles à des changements significatifs de l'apparence des visages, (par exemple yeux fermés, bouche ouverte, etc.).

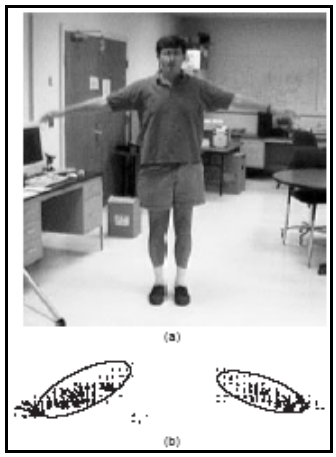


Figure 1.11 : Action de mouvements des mains (a) et direction du flux optique(b)

Li a proposé dans [LI 07] un descripteur de mouvement par flux optique basé sur les histogrammes orientés. Il a utilisé les modèles de Markov caché HMM pour la reconnaissance d'actions humaines. La dimension de l'histogramme orienté est réduite en utilisant les ACP. Les HMM sont utilisés avec la topologie de six états cachés et chaque observation est modelée en utilisant les mélanges de trois densités Gaussiennes. Quelques séquences ne sont pas classées à cause du haut degré relatif à la similitude entre ces séquences, comme 'running' et 'jumping forward'.

Ahmad et Lee ont présenté une méthode pour la reconnaissance d'actions humaines à partir des séquences d'image dans différents angles d'observation (capturés à trois angles d'observation) en utilisant les composants Cartésiens de vélocité du flux optique et le vecteur caractéristique de la silhouette [AHM 06]. La représentation de chaque action est effectuée en utilisant un ensemble de HMM multidimensionnels. Afin de réduire l'espace dimensionnel élevé des caractéristiques, ils ont utilisé la technique de l'ACP. Pour évaluer les actions, la base de données de geste d'Université de la Corée [HWA 06] était utilisée, où sa performance a été testée dans [HWA 05].

Les résultats expérimentaux montrent que l'algorithme peut reconnaître l'action humaine dans n'importe quelle direction d'observation. Cependant, leur méthode compte sur les silhouettes qui sont extraites par la soustraction de fond, en cas de collusion nous aurons une silhouette mal structurée.

Figure 1.12 : Différentes directions du flux optique pour une séquence d'images

Les représentations basées sur le flux optique sont des techniques robustes aux bruits et aux observations aberrantes, fournissent des champs de déplacement denses « robuste aux changements de luminosité » seulement l'inconvénient majeur de l'utilisation du flux optique est la somme importante de calculs à réaliser pour l'estimation du mouvement. Dans [HOL 10], Holte et al traitent dans leur approche les mouvements gestuels à partir d'un point de vue de 0° et testés sur des différents de points de vue 0° et ± 45°. Le mouvement est détecté par flot optique 2D estimé dans l'image, mais étendu à des données 3D capturées par caméra qui produit à la fois une carte de profondeur ainsi qu'une intensité à l'image d'une scène. Aussi, ils s'intéressent au contexte du mouvement qui se transforme en une représentation de vue invariante en utilisant des fonctions sphériques, donnant alors une représentation harmonieuse au contexte du mouvement. Quant au système de reconnaissance gestuelle, ils ont également abordé le problème en ignorant le temps du déroulement du geste, c'est à dire dans le cas ou nous ignorons son début et sa fin, qui se fait par la reconnaissance d'un geste non à travers une approche basée sur la trajectoire, mais sur un classificateur probabiliste de distance utilisée pour identifier au mieux un geste donné.

1.3.4 Méthode des représentations basées sur points d'intérêt

Les points d'intérêt sont des points particuliers de l'image importants soit pour un traitement particulier, soit pour des raisons d'optimisation dont l'objectif est de traiter quelques points au lieu de traiter la totalité de l'image. La littérature est abondante en détection de points d'intérêt à partir d'images. Le développement de

mise en correspondance d'image en utilisant un ensemble de points caractéristiques était marqué par le travail de Moravec [MOR 81] sur la mise en correspondance stéréo en utilisant un détecteur de coin pour choisir des points d'intérêt. Le détecteur Moravec a été amélioré par Harris et Stephens [HAR 88] dont l'objectif est de le rendre plus robuste.

Zhang et al. [ZHA 95] ont montré qu'il était possible d'utiliser Harris corners sur une longue variation d'image en utilisant une fenêtre de corrélation autour de chaque coin pour choisir les parties les plus semblables. En même temps, une autre approche similaire a été développée par Torr [TOR 95] pour la mise en correspondance de mouvement à longue variation. Schmidet et Mohr [SCH 97] ont montré que l'invariante de points de vue des caractéristiques locales pourrait être étendue aux problèmes de reconnaissance d'image générale. Les auteurs ont aussi utilisé les détecteurs de Harris au niveau des angles pour détecter les points d'intérêt, mais au lieu d'utiliser une fenêtre de corrélation, ils ont utilisé un descripteur invariant aux rotations des régions d'image. Le détecteur de point d'intérêt de Harris était largement utilisé dans les applications de reconnaissance d'objet et appliqué à la modélisation et la reconnaissance d'actions dans le spatio-temporel.

Taylor [TAY 00] a proposé une approche intéressante afin d'estimer la position du corps humain à partir de simples images orthographiques « cinématiques orthographiques » et de points articulaires étiquetés. Parmi les contraintes de cette approche, les points articulaires sont donnés manuellement par l'utilisateur et non extraits automatiquement. Aussi, cette technique suppose que toutes les articulations se trouvent dans l'image, ce qui n'est pas toujours réalisé à cause des occultations.

Figure 1.13 : Reconstruction obtenue selon les points de correspondance de l'image [TAY 00]

Song et al. [SON 03] ont utilisé une représentation basée sur les points d'intérêt pour détecter l'action humaine (voir figure 1.14). Ils ont également utilisé des modèles d'apprentissage pour déduire la forme des distributions représentées par les bords du graphe [SON 00][SON 03]. Aussi d'après [FAN 04] c'est une technique qui représente la position et la vélocité de chaque partie du corps par rapport à ses «voisins» dans le graphe triangulé. Tant que toutes les parties sont observées tout fonctionne bien, cependant, quand une partie manque sa position relative de ses voisins graphique elle ne peut pas être calculée et il faut revenir à des positions absolues [FAN 04].

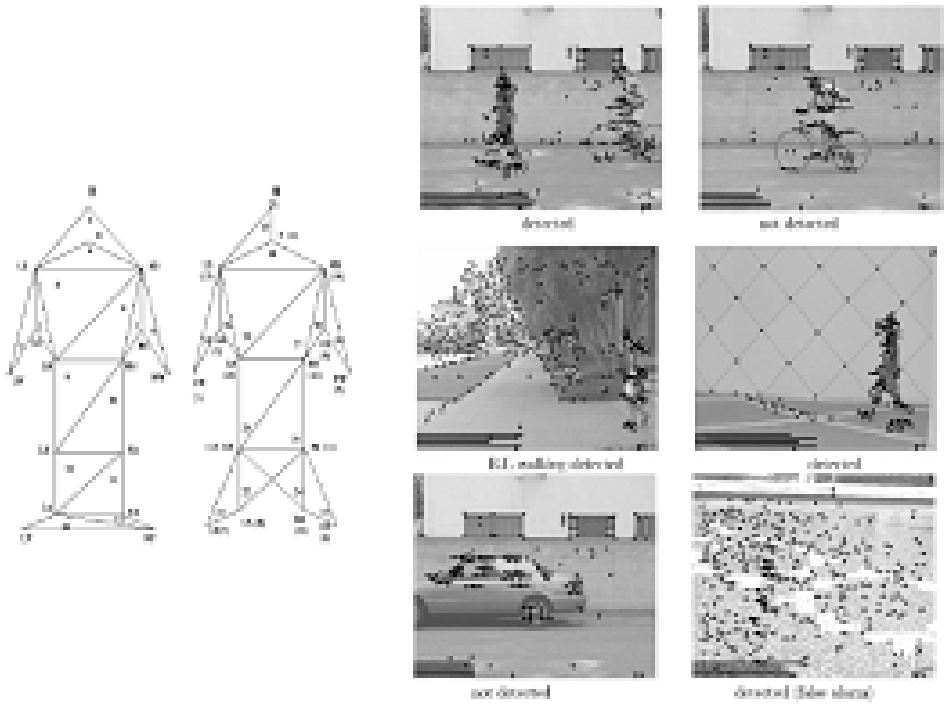


Figure 1.14 : Apprentissage non supervisé du mouvement humain [SON 03]

Lowe [LOW 04] utilise la description à l'aide d'orientation du gradient autour de points d'intérêt. La recherche de ces derniers est effectuée en multi-résolution, ce qui permet de définir une taille variable pour leur voisinage. La taille dépend de l'échelle sur laquelle le point d'intérêt est détecté. L'image est ainsi caractérisée par un ensemble d'histogrammes locaux.

L'inconvénient d'utiliser les filtres de difference-of-Gaussian (DOG) où l'approximation de Laplacians of Gaussian LOG comme des détecteurs de

caractéristiques est que la reproductibilité n'est pas optimale car non seulement ils fonctionnent en déterminant les descripteurs du blob qui sont estimés grâce à ses dimensions, le nombre de pixels, sa vitesse et son orientation moyenne, mais aussi en produisant des gradients élevés dans une seule direction. Pour cette raison, la localisation des caractéristiques peut ne pas être très précise [LEW 06].

Les points d'intérêt ont été étendus au niveau spatio-temporel par Laptev et al. [LAP 05]. Ils ont développé une technique "Space Time Interest Points" où les points d'intérêt sont décrits par un jet local au troisième ordre spatio-temporel. Les points d'intérêt correspondent aux maxima locaux de Harris (voir figure 1.15). L'inconvénient d'une telle technique est que le détecteur de Harris est bien souvent incapable de détecter des angles obtus. D'autre part, un coût élevé de calcul pour détecter un petit nombre de points d'intérêt.

Figure 1.15 : Détection des points d'intérêts dans une séquence d'image [LAP 05].

Pour résumer, les points d'intérêt offrent de nombreux avantages. De tels points sont supposés hautement distincts de leur voisinage et riches en terme d'information. Mais ils font plus que réduire l'information à l'essentiel, par leur faculté à être facilement détectés dans diverses situations. Dans [BRE 11], Bregonzio et al. exposent une méthode de représentation d'actions qui diffère sensiblement de la représentation des points d'intérêt existant, basée sur le fait que seule l'information de la distribution de ces derniers est exploitée. En particulier, les caractéristiques de ces points d'intérêt accumulés sur plusieurs échelles temporelles sont extraites à partir d'une vidéo et décrites à l'aide de descripteurs basés sur l'apparence. Puisque la représentation de la distribution spatio-temporelle proposée contient des informations différentes mais complémentaires ils ont formulé une méthode de fusion-métrage basée sur le noyau d'apprentissage

multiple. Les expériences utilisant les bases de données KTH et des ensembles de données Weizmann ont démontré que cette approche surpasse les méthodes les plus actuelles, en particulier sous l'occlusion et les variations de l'angle de vue, les vêtements et conditions de transport.

1.4 Conclusion

Malgré que ces approches soient à l'origine des meilleurs systèmes de reconnaissance d'action humaine actuels, nous pouvons constater que beaucoup de ces techniques souffrent du problème de la représentation des caractéristiques qui est une tâche difficile. Nous avons abordé dans ce chapitre les différents types de représentations qui fournissent un outil fiable pour de nombreuses tâches de vision.

La représentation MHI (Motion History Image) abordée par [BOB 01] qui rapporte toujours l'information de mouvement nous semble tout à fait utile pour divers types de mouvements malgré le problème de ne pas pouvoir indiquer la magnitude ou la direction du mouvement. Pour pallier ce problème, nous proposons d'utiliser en plus le flux optique pour bénéficier de ses avantages.

1.5 Références

[AHM 06]: M. Ahmad, S. Lee, Human Action Recognition using Multi-view Image Sequences Features, in: International Conference on Automatic Face and Gesture Recognition, Southampton, UK, April 10–12, 2006.

[BAU 95]: A. Baumberg, "Learning Deformable Models for Tracking Human Motion" PhD thesis, Sch. of Comp. Stu., University of Leeds, UK, 1995.

[HOR 81]: B. K.P. Horn et B. G. Schunck, " Determining optical flow" Artificial Intelligence, 17 :185–203, 1981.

[BEC 09]: M. Bécha Kaäniche, "Human gesture recognition", Thesis l'INRIA 2009.

[BHU 08]: A. Bhusnurmath, "Applying Convex Optimization Techniques to Energy Minimization Problems in Computer Vision", thesis 2008.

[BLA 98]: Michael J. Black, David J. Fleet, Yaser Yacoob, "A framework for modeling appearance change in image sequences", IEEE International Conference on Computer Vision, pp. 660-667, janvier 1998.

[BOB 01]: A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," IEEE Trans. vol. 23, no. 3, pp. 257–267, 2001.

[BOR 88]: G. Borgefors. Hierarchical chamfer matching: A parametric edge matching algorithm. PAMI, 10(6):849–865, November 1988.

[BRE 11]: M. Bregonzio, T. Xiang, S. Gong, "Fusing Appearance and Distribution Information of Interest Points for Action Recognition", 2011 Elsevier

[CHO 98]: O. Chomat, J.L. Crowley, Recognizing motion using local appearance, Inter Symposium on Intelligent Robotic, University of Edinburgh, 1998.

[CUT 98]: R. Cutler et M. Turk, "View-based Interpretation of Real-time Optical Flow for Gesture Recognition," Proc. 1998 IEEE Conf. on Automatic Face and Gesture Recognition, April 14-16, 1998, Nara, Japan.

[DAV 03]: J. Davis and A. Tyagi. A reliable-inference framework for recognition of human actions. In IEEE International. Conference on Advanced Video and Signal Based Surveillance, pages 169–176 ,2003.

[DER 04]: Konstantinos G. Derpanis,"A Review of Vision-Based Hand Gestures", Review 2004.

[ELZ 03]: H. Elzein, Sridhar Lakshmanan, and Paul Watta. A motion and shape-based pedestrian detection algorithm. In Intelligent Vehicles Symposium, 2003, pages 500–504, June 2003.

[FAE 02]: B. Fasela;*, Juergen Luettin, "Automatic facial expression analysis: a survey", 2002 Pattern Recog. Society. Published by Elsevier Science.

[FAN 04]: Claudio Fanti Lihi Zelnik-Manor Pietro Perona, "Hybrid Models for Human Motion Recognition", IEEE 2004.

[FEL 05]: P.F. Felzenszwalb, D.P. Huttenlocher, Pictorial Structures for Object Recognition, Int. J. Comput. Vis., Vol. 61, Issue 1, pp. 55-79, 2005.

[GAV 99]: D.M. Gavrila, The visual analysis of human movement: a survey, Computer Vision and Image Understanding 73 (1) (1999) 82–98.

[GOR 07]: L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, Actions as Space-Time Shapes, IEEE Trans Pat. Anal. Int., vol. 29, n° 12, 2007.

[GIB 55]: J. J. Gibson, The perception of the visual world. Houghton Mifflin, 1955.

[HAR 88]: Harris, C. and Stephens, M. 1988. A combined corner and edge detector. In Fourth Alvey Vision Conf., Manc., UK, pp. 148-151.

[HU 62]: M.-K. Hu, "Visual pattern recognition by moment invariants," IRE Trans. Inf. Theory, vol. 8, no. 2, pp. 179–187, Feb. 1962.

[HOG 83]: D . Hogg, Model-based vision: A program to see a walking person ». In Image and Vision Computing, Vol. 1, pages 5-20, 1983.

[HOL 10]: M.B. Holte , T.B. Moeslund, P. Fihl, "View-invariant Gesture Recognition Using 3D Optical Flow and Harmonic Motion Context", 2010 Elsevier.

[HU 04]: W. Hu, T. Tan, L. Wang, S. Maybank, A survey on visual surveillance of object motion and behaviors, Trans. on Sys., Man, and Cybernetics_Part C: Applications and Reviews 34 (3) (2004) 334–352.

[HWA 05]: B. -W. Hwang, S. Kim, and S.-W. Lee, "2D and 3D full-body gesture database for analyzing human gestures,"Advances in Intelligent Computing, Lecture Notes in Computer Science, Vol. 3644, pp. 611-620, August 2005.

[HWA 06]: B. -W. Hwang, S. Kim, and S.-W. Lee, "A Fullbody gesture database for automatic gesture recognition", Proc. 7th IEEE International Conference on Face and Gesture Recognition, Southampton,UK, April 2006. The KU Gesture Database, http://gesturedb.korea.ac.kr/.

[JIA 06]: H. Jiang, M. Crew, and Z. Li, "Successive convex matching for action detection," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006.

[LAP 05]: I. Laptev et T. Lindeberg, On Space-Time Interest Points, International Journal of Computer Vision 64 (2/3), Springer Science , Business Media, pp 107–123, 2005.

[LI 07]: X. Li, "HMM based action recognition using oriented histograms of optical flow field", Elec. Letters, 10th May 2007 Vol. 43 No. 10.

[LEW 06]: M. S. Lew, N. Sebe, C. Djeraba, "Content-Based Multimedia Information Retrieval: State of the Art and Challenges",2006

[LOW 04]: David Lowe. Distinctive image features from scale-invariant keypoints. Intern. Journal of Comp. Vision, 60(2) :91–110, 2004.

[MEN 09]: Hongying Meng, Nick Pears, " Descriptive temporal template features for visual motion recognition", 2009 Elsevier .

[MIK 11]: K. Mikolajczyk , H. Uemura, "Action Recognition with Appearance-Motion Features and Fast Search Trees", 2011 Elsevier

[MOK 08]: A. Mokhber, C. Achard, M. Milgram "Recognition of human behavior by space-time silhouette characterization", elsevier, Pattern Recognition N°29 pp 81–89, 2008.

[MOR 81]: Hans P. Moravec. " 3D graphics and the wave theory ", Proceedings of the 8th annual conference on Computer graphics and interactive techniques, pages 289–296, New York, NY, USA, 1981. ACM.

[MOR 06]: Greg Mori, Jitendra Malik , "Recovering 3D Human Body Conf. Using Shape Contexts", IEEE Trans Pat Anl, vol. 28, n°. 7, 2006.

[MUY 79]: E Muybridge. Muybridge's complete human and animal locomotion: all 781 plates from the 1887 animal locomotion, vol 1. Dover Pub, 1979.

[MOE 01]: T. Moeslund, E. Granum, "A survey of computer vision based human motion capture", Computer Vision and Image Understanding vol 3 n°81 pp 231–268, 2001.

[PIE 05]: Pierobon, M. Marcon, M., Sarti, A., Tubaro, S.: Clustering of human actions using invariant body shape descriptor and dynamic time warping IEEE International Conference on Advance Video and Signal Based Video and Signal-Based Surveillance pp 22-27, 2005

[RAH 06]: M. Rahman and A. Robles-Kelly. A Tuned Eigenspace Technique for Articulated Motion Recognition.In European Conference on Computer Vision, Graz, Austria, May 7-13, 2006.

[RAH 03]: D. Ramanan and D. A. Forsyth, Finding and tracking people from the bottom up, In IEEE Conf. on Comp. Vision and Pattern Recognition (CVPR'03), Vol. 2 pages 467–474, Madison, USA, 2003.

[ROH 09]: Myung-Cheol Roh, Ho-Keun Shin, Seong-Whan Lee , "View-independent human action recognition with Volume Motion Template on single stereo camera", 2009 Elsevier B.V.

[ROS 04]: D. A. Ross, J. Lim, M. Yang, Adaptive Probabilistic Visual Tracking with Incremental Subspace Update. European Conference on Computer Vision N°2 p 470-482,2004.

[SCH 97]: Schmid, C., and Mohr, R. 1997. Local grayvalue invariants for image retrieval. IEEE Trans. On Pattern Anal. & Mac. Int., 19(5):530-534.

[SON 00]: Y. Song, X. Feng, and P. Perona, "Towards detection of human motion," in IEEE Conf. on Comp. Vision & Pattern Rec, pp. 810–817, 2000.

[SON 03]: Y. Song, L. Goncalves, and P. Perona, "Unsupervised learning of human motion," IEEE T. Pattern Analysis and Machine Intelligence, vol. 25, no. 7, pp. 814–827, July 2003.

[SUT 63]: E.E. Sutherland, Sketchpad: The First Interactive Computer Graphics. Ph.D. Thesis, 1963. Mass. Institute of Technology, 1963.

[TAR 95]: Tarr M.J., Bülthoff H.H., "Is human object recognition better described by geon structural decriptionsor by multiple views?" Journal of experimental psychology: human perception and performance, vo 21, pp 1494-1505, 1995.

[TAY 00]: C.J Taylor, Reconstruction of Articulated Objects from Point Correspondences in a single Image, Computer. Vision and Image Understanding Volume 80, No. 3 pp 349-363 2000.

[TOR 95]: P. Torr, Motion Segmentation and Outlier Detection, Ph.D. Thesis, Dept. of Engineering Science, University of Oxford, UK, 1995.

[WAN 02]: Liang Wang, Weiming Hu, Tieniu Tan," Recent developments in human motion analysis", 2002 Pattern Recognition. Pub by Elsevier Science Ltd.

[WAR 88]: W. H. Warren, M. W. Morris et M. Kalish, "Perception of translational heading from optical flow", J Exp Psychol Hum Percept Perform, 14(4):646–660, Nov 1988.

[WEI 06]: D. Weinland, R. Ronfard, E. Boyer, Free Viewpoint Action Recognition using Motion History Volumes, Computer Vision and Image Understanding, Volume 104, Issues 2-3, November-December 2006, Pages 249-257

[YAC 96]: Yaser Yacoob, Larry S. Davis. " Recognizing human facial expressions from long image sequences using optical flow. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 6 n°18, pp 636-642, 1996.

[YAM 92]: J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time sequential image using hidden markov model," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1992.

[YIL 05]: Yilmaz, A., Shah, M.: Actions as objects: a novel action representation. In: Proc. IEEE Conf. Computer Vision and Pattern Rec, pp. 984–989 (2005).

[YU 05]: H. Yu, G.-M. Sun, W.-X. Song, and X. Li. Human Motion Recognition Based on Neural Networks.In International Conference on Communications, Circuits and Systems, volume 2, pages 982–989, Hong Kong, China, 2005.

[ZHA 95]: Zhang, Z., Deriche, R., Faugeras, O., and Luong, Q.T. 1995. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. Artificial Intelligence, 78:87-119.

[ZHA 05]: Y. Zhang et Q. Ji, "Active and Dynamic Information Fusion for Facial Expression Understanding from Image Sequences ", IEEE Trans Pat Anl , vol 27, n° 5, 2005.

Chapitre 2 : Représentation d'Actions Humaines : Images de l'historique du mouvement (MHI) et Flux Optique

2.1 Introduction

L'estimation de mouvement a été un problème classique dans la vision par ordinateur. De nombreuses techniques ont été proposées pour le décrire dans une scène comme nous l'avons montré dans le chapitre précédent. De l'état de l'art présenté précédemment, nous retenons l'utilisation des MHI et les méthodes associées aux textures temporelles « Flux optique » qui donnent une information du mouvement global sur les séquences. Nous allons décrire ces deux approches retenues pour la caractérisation du contenu d'une séquence. Le problème consiste en la définition de paramètres idéaux censés être à la fois intuitifs et représentatifs qui permettent une meilleure classification et reconnaissance d'actions humaines.

Dans ce chapitre, nous détaillerons ces deux outils utilisés dans la littérature pour la représentation d'action humaine. Nous commencerons par les modèles temporels présentés par A.F. Bobick et al. et pour finir, nous aborderons ensuite la technique du flux optique.

2.2 Définition des images de l'historique du mouvement

Un modèle (template) a pour objectif la formation d'une image statique du mouvement étudié, afin d'en extraire des caractéristiques pour la classification et

la reconnaissance. Parmi les méthodes basées sur les images modèles, citons les travaux de Bobick et Davis [BOB 01] qui reconnaissent différents mouvements à l'aide de modèle spatio-temporels. Ils construisent pour cela l'image binaire de l'énergie de mouvement appelée MEI (Motion Energy Image) d'une part et l'image scalaire de l'historique du mouvement MHI ("Motion-History Image") d'autre part.

La première représente la localisation du mouvement dans une séquence d'images tandis que dans la seconde, l'intensité d'un pixel qui indique si le mouvement en ce point est récent ou non [LAT 99], autrement dit, les pixels sont des valeurs correspondant à l'âge du mouvement [RIC 08].

Figure 2.1: Des images représentatives de séquences originales sont regroupées sur la rangée du haut et les MHI correspondantes sur la rangée du bas [VEN 02].

La MHI a été définie dans [BOB 96][DAV 97], et formalisée de manière récursive comme suit : Soit I(x, y, t) une séquence d'images et D(x, y, t) la séquence d'images binaires indiquant les régions en mouvement. D est en général obtenu par différentiation des images successives.

2.2.1 Motion History Image (MHI)

La représentation MHI est formée simplement de la réunion temporelle des cartes binaires des zones mobiles détectées dans les images d'une séquence [VEN 02]. La MHI donne la priorité au dernier mouvement observé en un point.

Ainsi, lorsqu'un point est détecté en mouvement à l'instant t, toute information en ce point, aux temps antérieurs est effacée. La MHI donne en chaque point l'indication du dernier instant où il y a eu un mouvement. L'intensité de pixel est en fonction de l'historique de mouvement à cet endroit, où la valeur lumineuse correspond au plus récent [DAV 99].

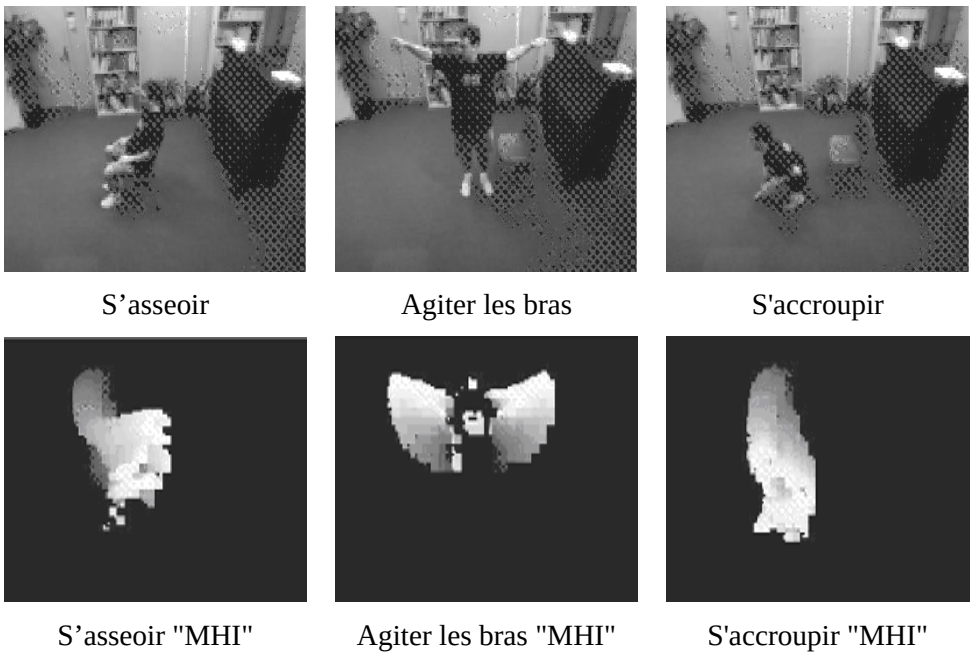


Figure 2.2 : MHI correspondant à une séquence "asseoir" [BOB 97].

La MHI contient en plus une information temporelle: la luminosité des pixels en mouvement détectés décroit avec le temps. Nous observons la même silhouette que dans les MEI mais avec une sorte de trainée lumineuse [BAI 01]. La définition des MHIs est donnée par :

$$\mathrm{MHI}(x,y,t)=\begin{cases}\tau & \text{si } D'(x,y,t)=1\\ \max(0,H(x,y,t-1)-1) & \text{sinon}\end{cases} \qquad (2.1)$$

Dans cette équation, τ est l'âge considéré dans l'historique. L'historique au point (x, y) est affecté à la valeur t si un mouvement est détecté en (x, y), c'est-à-dire si D(x,y,t)=1. Sinon, l'age du mouvement augmente et la valeur de H(x,y,t) est décrémentée. L'age du mouvement en (x, y) est en fait déterminé par :

$$\left|H_\tau(x,y,t)-\tau\right| \qquad (2.2)$$

2.2.2 Motion Energy Image (MEI)

La représentation MEI résulte d'une sommation temporelle des cartes binaires des zones en mouvement [VEN 02]: chaque pixel correspondant à un mouvement dans l'image est considéré comme un pixel d'intérêt.

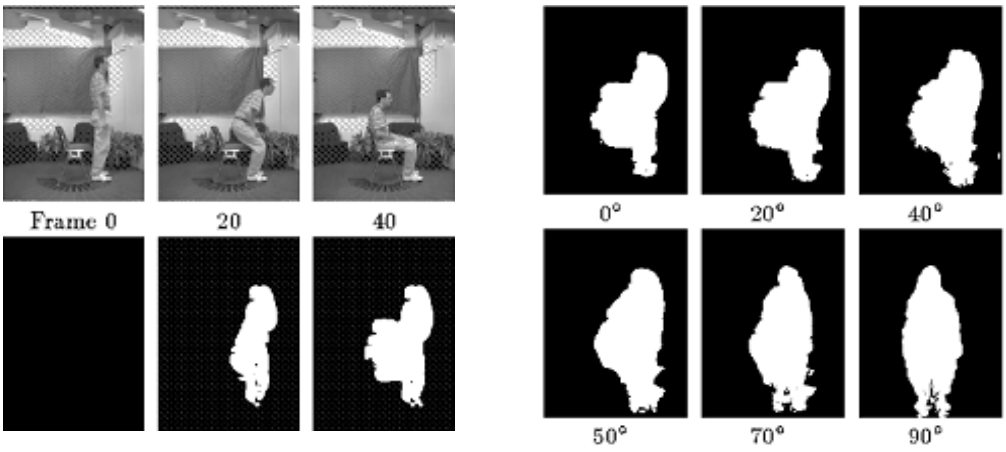
Les MEI visualisent les pixels qui ont bougé dans les τ dernières images (Figure 2.3). Pour cela, nous utilisons la quantité D(x,y,t) qui est définie par seuillage du flux optique à l'instant t en x,y. D(x,y,t)=1 si le point (x,y) est en mouvement à l'instant t et D(x,y,t)= 0 .

L'image E d'énergie de mouvement est alors définie par le seuillage de l'image H à zéro, c'est à dire :

$$E_\tau(x,y,t)=\bigcup_{i=0}^{\tau} D(x,y,t-i)=\begin{cases}1 & \text{si } H_\tau(x,y,t)>0\\ 0 & \text{sinon}\end{cases} \qquad (2.3)$$

$$MEI(x,y,t)=\bigcup_{i=0}^{\tau-1} D'(x,y,t-i) \qquad (2.4)$$

Prenons l'exemple d'une personne assise, comme illustré par la figure 2.3. La rangée du haut contient des images clés d'une séquence vidéo. La rangée du bas montre une cumulation des images binaires (MEI).



(a) (b)

Figure 2.3 : Série de MEI associés à une action asseoir vue de plusieurs angles [BOB 97]

Notons que dans le cas de la MEI, les points de valeurs maximales sont ceux qui ont le plus souvent bougé dans la séquence considérée. Dans le cas d'une camera fixe, les valeurs supérieures à zéro de ces deux représentations (MHI, MEI) forment l'aire balayée par l'objet mobile lors de son déplacement dans la séquence en question.

Dans [DAV 99], une analyse de gradient de la MHI calculé à l'aide de masques de Sobel a été menée. En chaque point de l'image, une direction locale du mouvement est déterminée et les histogrammes localisés sont construits afin de décrire l'information directionnelle du mouvement dans la séquence d'images. Dans [BRA 00], l'information d'orientation locale est intégrée par sommation pondérée et permet le calcul de directions moyennes du mouvement dans certaines régions de l'image. Enfin une variation de la MHI est proposée dans [BRA 00] pour permettre une certaine normalisation temporelle de la représentation. La tMHI (timed Motion History Image) est une représentation relativement indépendante du nombre d'images par seconde dans le flux vidéo et de la durée de réalisation d'un mouvement.

La tMHI est une séquence d'images résumées en une seule image représentant le déroulement du mouvement où l'intensité de chaque pixel est en fonction de l'historique du mouvement. Cette représentation permet à la fois de visualiser le mouvement et de fournir une information globale et locale de ce dernier, qui peut

être utilisée dans la reconnaissance de configurations dynamiques. Une implémentation de ces techniques a été réalisée dans la librairie CVLib d'INTEL [DAV 99].

Figure 2.4: Time Motion History Image [KHA 09]

2.3 Flot optique

Ce que nous appelons usuellement flot optique est le mouvement apparent, c'est-à-dire la projection de vecteurs vitesses des objets de l'espace sur un plan d'image [RAN 05]. Il est calculé à partir des seules variations d'intensité des pixels dans la séquence et permet d'associer à chaque pixel un vecteur vitesse 2D représentant l'estimation du mouvement.

Dans chaque image d'une séquence vidéo, chaque point suivi jusqu'alors est recherché par corrélation autour de la position qu'il avait sur l'image précédente; s'il n'est pas trouvé dans plusieurs images consécutives, il sera rejeté et remplacé par un nouveau point d'intérêt extrait là où il y en a peu. Les vecteurs de déplacement entre les points appariés forment le champ du flot optique [ALM 07]. Efros et al. [EFR 03] utilisent des descripteurs de mouvement basés sur le flot optique ainsi qu'une mesure de similarité afin de reconnaître des actions humaines en basse résolution.

L'estimation de mouvement est un problème fondamental dans le traitement d'image appliqué sur des séquences vidéo. Plusieurs méthodes ont été proposées parmi lesquelles nous distinguons trois grandes familles principales, que nous présentons comme suit :

2.3.1 Méthodes différentielles

Les méthodes différentielles sont basées sur les gradients spatiaux et temporels de l'intensité lumineuse des pixels. Elles s'appuient sur l'équation de contrainte du mouvement apparent issu d'un développement de Taylor de l'équation :

$$\frac{d_i(x_1, x_2, t)}{dt} = 0 \qquad (2.5)$$

où x1 et x2 sont les variables spatiales, t est la variable temporelle et i(x1,x2,t) l'intensité du pixel de coordonnées (x1,x2) dans l'image acquise à l'instant t.

$$\frac{\partial_i(x_1, x_2, t)}{\partial x_1} d_1 + \frac{\partial_i(x_1, x_2, t)}{\partial x_2} d_2 + \frac{\partial_i(x_1, x_2, t)}{\partial t} d_t = 0 \qquad (2.6)$$

Parmi les différentes variantes proposées, certaines sont basées sur des dérivées du premier ordre avec ou sans contrainte de régularisation sur le champ de vecteurs vitesses [LUC 84][HOR 81].

Il est également possible d'utiliser des dérivées d'ordre supérieur, mais aussi réduire la sensibilité des calculs numériques en utilisant des contraintes de régularisation locales [URA 88] ou globales [NAG 87].

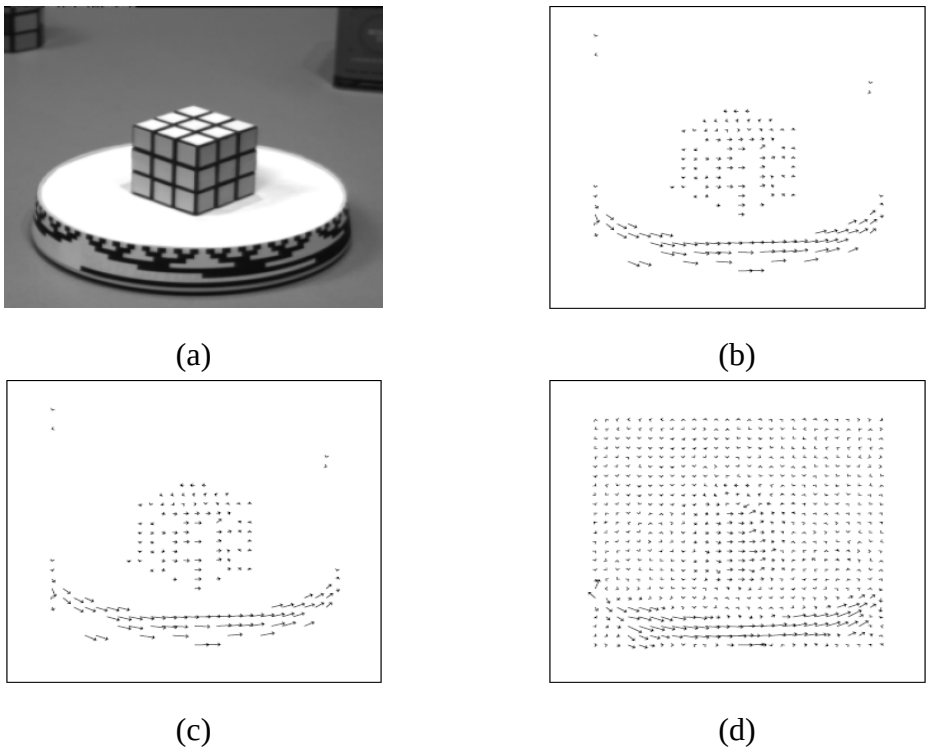


Figure 2.5: Images issues de [BAR 92]. (a) Une image de la séquence d'images expérimentales du cube de Rubik, champ 2-D de mouvement estimé avec (b) une méthode basée sur la technique de Horn et al. [HOR 81], (c) la méthode de Lucas et al. [LUC 81], (d) une technique basée sur la régularisation globale du champ proposée par Nagel [NAG 86].

2.3.2 Méthodes de mise en correspondance (block-matching techniques)

Ces méthodes tentent d'estimer le mouvement d'une région de l'image courante en minimisant la distance avec une région candidate de l'image suivante. En général, cette mesure de similarité est obtenue par une somme des différences inter-pixels au carré (Sum-Squared Difference - SSD).

Sachant qu'un test exhaustif de toutes les régions possibles est très coûteux en temps de calcul, de nombreux algorithmes «rapides» ont été proposés. Anandan préconise une approche de type multi-résolution en utilisant une décomposition pyramidale de l'image [ANA 89]. Le déplacement est estimé itérativement en commençant par le niveau de résolution le plus grossier. Dans [KOG 81], Koga propose une méthode de recherche rapide (logarithmic search) permettant d'estimer le déplacement d'un bloc en suivant la direction de la moindre déformation. Dans le même esprit, nous pouvons également citer la technique de recherche en trois étapes utilisée dans le codeur vidéo H.263 [ITU 95]. Le principe de base des méthodes de mise en correspondance de blocs est de découper l'image de référence en blocs de pixels, également appelés des régions d'intérêt (ROI). Pour chaque bloc, une zone de recherche est définie dans l'image cible [BAS 08].

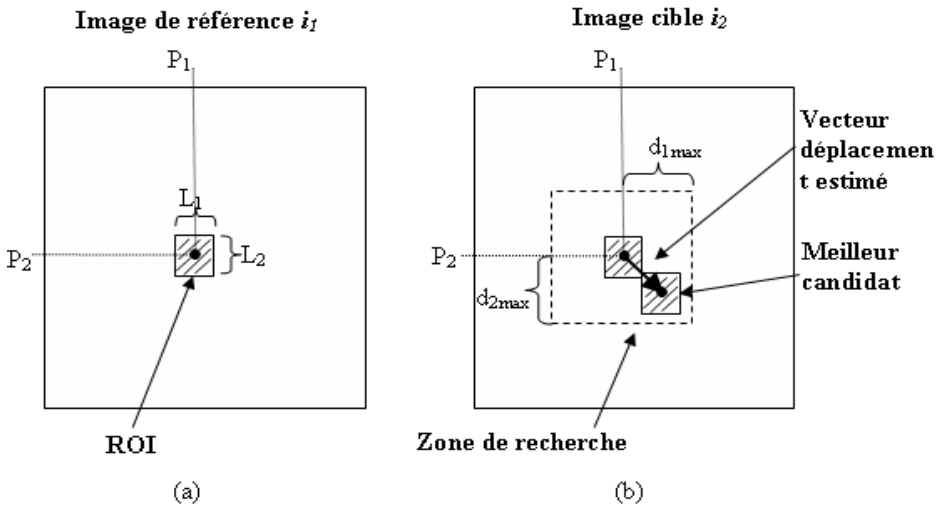


Figure 2.6 : Principe des méthodes de mise en correspondance de blocs. (a) Une région d'intérêt de taille L1×L2 est considérée dans l'image de référence. (b) Une zone de recherche est considérée dans l'image cible.

La ROI est recherchée dans cette zone et le meilleur candidat (le plus ressemblant à la ROI suivant un critère donné) est retenu. A noter que la taille de la zone de recherche impose le déplacement maximum autorisé pour la ROI courante [BAS 08].

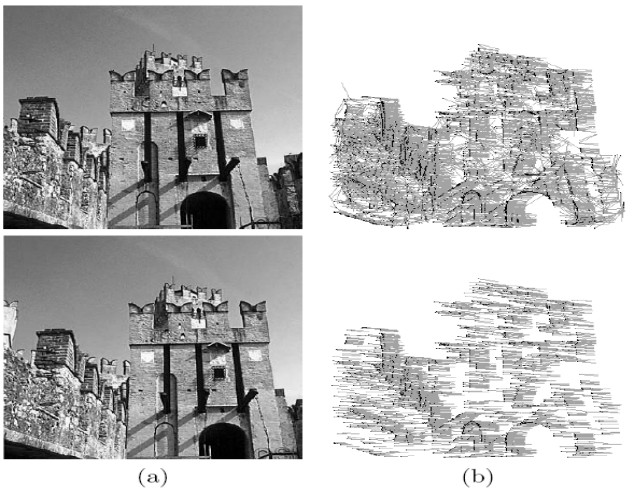


Figure 2.7 : Les images stéréos (a) et Les flots optiques selon la méthode de mise en correspondance (b) [SUV 06].

2.3.3 Méthodes fréquentielles

Les méthodes fréquentielles sont fondées sur une caractérisation du mouvement dans le domaine des fréquences [ADE 85]. Elles ont pour origine des recherches concernant la vision des mammifères, mettant en évidence la présence de cellules simples dans l'aire corticale V1 [PAL 85] et de cellules complexes dans l'aire MT [NEW 83], qui se comportent comme des filtres passe bande spatiotemporels. Elles présentent de nombreux atouts tels que la simplicité de l'interprétation physique ou la cohérence avec d'autres traitements (compression, restauration d'images) utilisant les mêmes modèles. Chaque filtre a la forme d'une paire d'ellipsoïdes (Gaussienne 3D) [PEL 96].

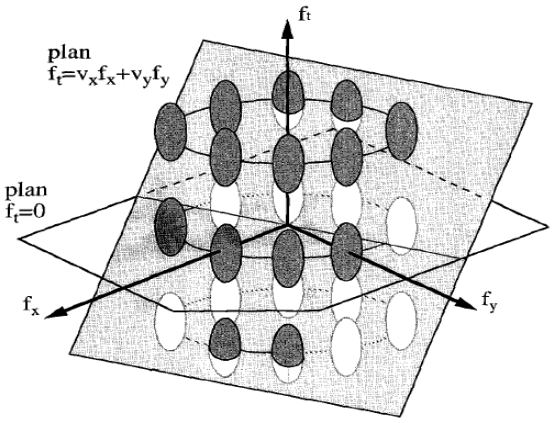


Figure 2.8 : Spectres de 12 filtres spatio-temporels orientés de Gabor pour une seule gamme de vitesse [PEL 96].

Les méthodes fréquentielles utilisent des bancs de filtres passe-bande permettant de décomposer le signal d'entrée selon l'échelle, la vitesse et l'orientation. Dans [HEE 88], Heeger analyse l'énergie à la sortie de filtres de Gabor pour estimer les vitesses. Dans [FLE 90], la méthode dite de filtrage spatio-temporel est proposée par Fleet et Jepson [JEP 90]. Cette méthode d'analyse fréquentielle locale suppose que le champ de flot optique est constant sur le support des filtres.

D'après l'étude détaillée menée dans [BAR 94], les techniques les plus fiables et les plus précises sont la méthode différentielle du premier ordre de Lucas et Kanade dont la méthode dite pyramidale est détaillée ci-dessous et la méthode de phase de Fleet et Jepson. La description algorithmique et mathématique sont illustrées par [BOU 00] et [SHI 94] dans ce qui suit :

2.3.3.1 Enoncé du problème

Soient I et J deux images 2D dans les niveaux de gris. Les deux quantités I (x) = I (x, y) et J (x) = J (x, y) sont alors la valeur de gris des deux images dont l'emplacement $x = [x\ y]^T$, où x et y sont les coordonnées d'un pixel de l'image générique x. L'image I est référencée comme la première image, et l'image J la seconde image.

Considérons maintenant un point $u = [u_x\ u_y]^T$ sur la première image I. L'objectif du suivi de caractéristiques est de trouver l'emplacement $v = u + d = [u_x + d_x\ \ u_y + d_y\ ]^T$ sur la seconde image J pour que I(u) et J (v) soient similaires. Le vecteur $d = [d_x\ d_y]^T$ est la vélocité de l'image en x, aussi connu sous le nom de flux optique à x.

Nous notons que w_x et w_y deux entiers. Nous définissons la vélocité d'une image comme étant le vecteur qui minimise la fonction résiduelle e défini comme suit:

$$\varepsilon(d) = \varepsilon(d_x, d_y) = \sum_{x=u_x-w_x}^{u_x+w_x} \sum_{y=u_y-w_y}^{u_y+w_y} (I(x,y) - J(x+d_x, y+d_y))^2 \qquad (2.7)$$

2.3.3.2 Représentation d'image sous forme pyramidale

La notion de pyramide a été introduite pour la première fois dans le domaine de l'analyse d'images par Tanimoto et Pavlidis [TAN 75]. Ces méthodes, aussi appelées approches multi-résolutions, permettent la description hiérarchique de l'information originale, c'est-à-dire la représentation d'un ensemble d'images aux différentes résolutions.

Un des intérêts majeurs de cette approche est de permettre un retour arrière à l'état précédent, car tous les niveaux de la pyramide peuvent être conservés, et de pouvoir faire une analyse à différentes résolutions. La structure pyramidale permet de faciliter l'extraction d'informations.

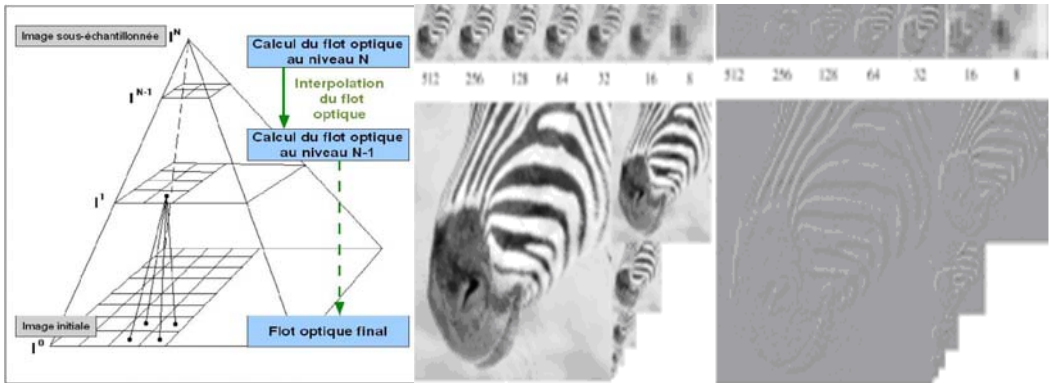


Figure 2.9 : Implémentation pyramidale d'une méthode de calcul du flot optique [MAR 07]

Nous allons définir la représentation pyramide d'une image général I de taille N_x x N_y. Prenons I^0 =I l'image au niveau zéro. Cette image est essentiellement l'image haute résolution. Notons N^{L-1} x et N^{L-1} y la largeur et la hauteur de l'image I^{L-1}.

L'image I^{L-1} est alors définie comme suit:

$$I^L(x,y) = \frac{1}{4} I^{L-1}(2x,2y) + \frac{1}{8}(I^{L-1}(2x-1,2y) + I^{L-1}(2x+1,2y) + I^{L-1}(2x,2y-1) + I^{L-1}(2x,2y+1)) + \frac{1}{16}(I^{L-1}(2x-1,2y-1) + I^{L-1}(2x+1,2y+1) + I^{L-1}(2x-1,2y+1) + I^{L-1}(2x+1,2y+1)) \quad (3.8)$$

Nous savons que la valeur L^M est la hauteur de la pyramide. Les valeurs pratiques de L^M sont 2, 3,4. Quant aux tailles d'image typique, il n'y a aucun sens d'aller au-dessus d'un niveau 4.

Alors, l'équation (2.8) n'est définie que pour des valeurs de x et y telles que :

$$0 \le 2x \le n_x^{L-1} - 1 \quad \text{et} \quad 0 \le 2y \le n_y^{L-1} - 1 \quad (2.9)$$

Par conséquent, la largeur n_x^L et la hauteur n_y^L de I^L sont les plus grands nombres entiers qui répondent aux deux conditions :

$$n_x^L \le \frac{n_x^{L-1}+1}{2} \tag{2.10}$$

$$n_y^L \le \frac{n_y^{L-1}+1}{2} \tag{2.11}$$

Pour L = 0,..., L_m, définir $u^L = \left[u_x^L \; u_y^L\right]$, les coordonnées correspondant au point u sur l'image pyramidale IL. Suite à notre définition de la représentation pyramidale les équations (2.8), (2.10) et (2.11), les vecteurs u^L sont calculés comme suit:

$$u^L = \frac{u}{2^L} \tag{2.12}$$

Supposons que $g^L = \left[g_x^L \; g_y^L\right]^T$ une proposition initiale de flux optique au niveau L, g^L est disponible suite aux calculs effectués à partir de niveau L_m au niveau L +1. Et afin de calculer le flux optique au niveau L, il est nécessaire de trouver le vecteur de déplacement $d^L = \left[d_x^L \; d_y^L\right]^T$ qui minimise la nouvelle image correspondant à la fonction d'erreur ε^L :

$$\varepsilon^L(d^L) = \varepsilon^L(d_x^L, d_y^L) \sum_{x=u_x^L-w_x}^{u_x^L+w_x} \sum_{y=u_y^L-w_y}^{u_y^L+w_y} (I^L(x,y) - J^L(x+g_x^L+d_x^L, y+g_y^L+d_y^L))^2 \tag{2.13}$$

De cette façon, le vecteur flux $d^L = \left[d_x^L \quad d_y^L\right]^T$ est petit et donc facile à calculer en utilisant la standard étape de Lucas et Kanade. Le résultat de ce calcul se propage à la vitesse supérieure L-1 en passant à la nouvelle proposition initiale g^{L-1} d'expression:

$$g^{L-1} = 2(g^L + d^L) \tag{2.14}$$

L'algorithme est initialisé en définissant la proposition initiale de niveau L_m g^{lm} à zéro :

$$g^{L_m} = [0\ 0]^T \tag{2.15}$$

- La solution finale du Flux optique d (voir équation (2.7)) est alors disponible après le calcul détaillé de flux optiques:

$$d = g^0 + d^0 \tag{2.16}$$

Notez que cette solution peut être exprimée sous la forme suivante :

$$d = \sum_{L=0}^{L_m} 2^L d^L \tag{2.17}$$

- Calcul itératif de flux optique (Lucas-Kanade itératif).

  Décrivons maintenant le calcul du flux optique.

  A chaque niveau L dans la pyramide, l'objectif est de trouver le vecteur d^L qui minimise la fonction définie dans l'équation (2.13). Nous y procédons de la même façon pour tous les niveaux L, les nouvelles images A et B sont définies comme suit:

$$A(x, y) \doteq I^L(x, y) \tag{2.18}$$

$$B(x, y) \doteq J^L(x + g_x^L, y + g_y^L) \tag{2.19}$$

- Changeons le nom du vecteur de déplacement $\bar{v} = [v_x \quad v_y]^T = d^L$, ainsi que le vecteur p = Position de l'image $[p_x \quad p_y]^T = u^L$. Suite à cette nouvelle notation, l'objectif est de trouver le vecteur de déplacement $\bar{v} = [v_x \quad v_y]^T$ qui minimise notre fonction:

$$\varepsilon(\bar{v}) = \varepsilon(v_x, v_y) = \sum_{x=p_x-w_x}^{p_x+w_x} \sum_{y=p_y-w_y}^{p_y+w_y} (A(x, y) - B(x + v_x, y + v_y))^2 \tag{2.20}$$

- Un itératif standard de l'algorithme de Lucas-Kanade peut être appliqué. À l'optimum, la première dérivée de ε par rapport à $\overline{v}$ est égale à zéro:

$$\left.\frac{\partial \varepsilon(\overline{v})}{\partial \overline{v}}\right|_{\overline{v}=\overline{v}_{opt}} = [0 \;\; 0] \tag{2.21}$$

- Après l'expansion de la dérivée, nous obtenons:

$$\frac{\partial \varepsilon(\overline{v})}{\partial \overline{v}} = -2 \sum_{x=p_x-w_x}^{p_x+w_x} \sum_{y=p_y-w_y}^{p_y+w_y} (A(x,y) - B(x+v_x, y+v_y)) \cdot \begin{bmatrix} \frac{\partial B}{\partial x} & \frac{\partial B}{\partial y} \end{bmatrix} \tag{2.22}$$

- Maintenant en remplaçant B(x + vx , y + vy) par son développement de Taylor de premier ordre sur le point $\overline{v} = [0 \;\; 0]^T$ nous aurons :

$$\frac{\partial \varepsilon(\overline{v})}{\partial \overline{v}} \approx -2 \sum_{x=p_x-w_x}^{p_x+w_x} \sum_{y=p_y-w_y}^{p_y+w_y} \left(A(x,y) - B(x,y) - \begin{bmatrix} \frac{\partial B}{\partial x} & \frac{\partial B}{\partial y} \end{bmatrix} \overline{v} \right) \cdot \begin{bmatrix} \frac{\partial B}{\partial x} & \frac{\partial B}{\partial y} \end{bmatrix} \tag{2.23}$$

- Notez que la quantité A (x, y)-B (x, y) peut être interprétée comme l'image temporelle dérivée au point [x y] T:

$$\delta I(x,y) \doteq A(x,y) - B(x,y) \tag{2.24}$$

- En remarquant que la matrice $\begin{bmatrix} \frac{\partial B}{\partial x} & \frac{\partial B}{\partial y} \end{bmatrix}$ est simplement le vecteur gradient de l'image. Faisons un léger changement de notation:

$$\nabla I = \begin{bmatrix} I_x \\ I_y \end{bmatrix} \doteq \begin{bmatrix} \frac{\partial B}{\partial x} & \frac{\partial B}{\partial y} \end{bmatrix}^T \tag{2.25}$$

- Si un opérateur de différence centrale est utilisé pour les dérivées, les deux images vont être dérivées comme suit:

$$I_x(x,y) = \frac{\partial A(x,y)}{\partial x} = \frac{A(x+1,y) - A(x-1,y)}{2} \tag{2.26}$$

$$I_y(x,y) = \frac{\partial A(x,y)}{\partial y} = \frac{A(x,y+1) - A(x,y-1)}{2} \tag{2.27}$$

- Suite à cette nouvelle notation, l'équation (2.23) peut être écrite:

$$\frac{1}{2}\frac{\partial \varepsilon(\overline{v})}{\partial \overline{v}} \approx \sum_{x=p_x-w_x}^{p_x+w_x} \sum_{y=p_y-w_y}^{p_y+w_y} \left(\nabla I^T \overline{v} - \delta I\right) \nabla I^T \tag{2.28}$$

$$\frac{1}{2}\left[\frac{\partial \varepsilon(\overline{v})}{\partial \overline{v}}\right]^T \approx \sum_{x=p_x-w_x}^{p_x+w_x} \sum_{y=p_y-w_y}^{p_y+w_y} \left(\begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \overline{v} - \begin{bmatrix} \delta I\, I_x \\ \delta I\, I_y \end{bmatrix}\right) \tag{2.29}$$

- Notons que

$$G = \sum_{x=p_x-w_x}^{p_x+w_x} \sum_{y=p_y-w_y}^{p_y+w_y} \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \tag{2.30}$$

et

$$\overline{b} = \sum_{x=p_x-w_x}^{p_x+w_x} \sum_{y=p_y-w_y}^{p_y+w_y} \begin{bmatrix} \delta I\, I_x \\ \delta I\, I_y \end{bmatrix} \tag{2.31}$$

- L'équation (2.29) peut être écrite comme suit

$$\frac{1}{2}\left[\frac{\partial \varepsilon(\overline{v})}{\partial \overline{v}}\right]^T \approx G\,\overline{v} - \overline{b} \tag{2.32}$$

- Par conséquent, l'équation suivante, où le vecteur flux optique est optimum est :

$$\bar{v}_{opt} = G^{-1}\bar{b} \qquad (2.33)$$

Cette expression n'est valable que si la matrice G est inversible. Cela équivaut à dire que l'image A (x, y) contient des informations gradient dans les deux directions X et Y dans le voisinage du point P.
Soit k l'indice itératif, initialisé à 1 à la première itération. Nous allons décrire l'algorithme récursif:

- L'itération générale k=>1, supposons que les calculs précédents des itérations 1,2, ...,k-1 fournissent une estimation initiale $\bar{v}^{k-1} = \begin{bmatrix} v_x^{k-1} & v_y^{k-1} \end{bmatrix}^T$ pour le pixel de déplacement $\bar{v}$.
- Bk est la nouvelle image traduite selon cette estimation initiale $\bar{v}^{k-1}$

$$B_k(x, y) = B(x + v_x^{k-1},\ v_y^{k-1}) \qquad (2.34)$$

- L'objectif est alors de calculer le résiduel pixel de vecteur de mouvement $\bar{\eta}^k = \begin{bmatrix} \eta_x^k & \eta_y^k \end{bmatrix}$ qui minimise la fonction d'erreur :

$$\varepsilon^k(\bar{\eta}^k) = \varepsilon(\eta_x^k, \eta_y^k) \sum_{x=p_x-w_x}^{p_x+w_x} \sum_{y=p_y-w_y}^{p_y+w_y} (A(x, y) - B_k(x + \eta_x^k, y + \eta_y^k))^2 \qquad (2.35)$$

- La solution de cette minimisation peut être calculée grâce à une étape de calcul de Lucas-Kanade du flux optiques (Équation (2.33)) :

$$\bar{\eta}^k = G^{-1}\bar{b}_k \qquad (2.36)$$

Où le vecteur $\bar{b}_k$ est défini comme suit (aussi appelé vecteur de décalage de l'image):

$$\bar{b}_k = \sum_{x=p_x-w_x}^{p_x+w_x} \sum_{y=p_y-w_y}^{p_y+w_y} \begin{bmatrix} \delta I_k(x,y) & I_x(x,y) \\ \delta I_k(x,y) I_y(x,y) & I_y(x,y) \end{bmatrix} \tag{2.37}$$

- Alors l'image de déférence devient :

$$\delta I_k(x,y) = A(x,y) - B_k(x,y) \tag{2.38}$$

Une fois que le flux optique résiduel $\bar{\eta}^k$ est calculé par l'équation 28, une nouvelle proposition de déplacement de pixel $\bar{v}^k$ est calculée pour la prochaine itération de l'étape k + 1:

$$\bar{v}^k = \bar{v}^{k-1} + \bar{\eta}^k \tag{2.39}$$

- La première itération (k = 1) de la proposition initiale est initialisée à zéro:

$$\bar{v}^0 = [0 \;\; 0]^T \tag{2.40}$$

- En supposant que K itérations ont été nécessaires pour parvenir à une convergence, la solution finale pour le vecteur de flux optique est :

$$\bar{v} = d^L = \bar{v}^k = \sum_{k=1}^{K} \bar{\eta}^k \tag{2.41}$$

Donc le résumé de l'algorithme pyramidal de Lucas-Kanade est le suivant

2.3.3.3 Algorithme pyramidal de Lucas-Kanade

Début

// Représentation de la construction de la pyramide

I et J : $\{I^L\}_{L=0,..,L_m}$ et $\{J^L\}_{L=0,..,L_m}$

Initialisation de la pyramide : $g^{L_m} = \left[g_x^{L_m} \; g_y^{L_m}\right]^T = [0\ 0]^T$;

Pour L :=L_m à 0 Faire

Début

Position du point u dans l'image : $u^L = \left[p_x \; p_y\right]^T = u / 2^L$;

Dérivée de I^L par rapport à x : $I_x(x,y) = \dfrac{I^L(x+1,y) - I^L(x-1,y)}{2}$;

Dérivée de I^L par rapport à y : $I_y(x,y) = \dfrac{I^L(x,y+1) - I^L(x,y-1)}{2}$;

Matrice Spatial de Gradient :

$$G = \sum_{x=p_x-w_x}^{p_x+w_x} \sum_{y=p_y-w_y}^{p_y+w_y} \begin{bmatrix} I_x^2(x,y) & I_x(x,y)I_y(x,y) \\ I_x(x,y)I_y(x,y) & I_y^2(x,y) \end{bmatrix} ;$$

Initialisation de L-K : $\bar{v}^0 = [0\ \ 0]^T$;

Pour k :=1 à K faire (seuil de précision $\left\|\bar{\eta}^k\right\|$)

Début

Calcul d'image différence:

$\delta I_k(x,y) = I^L(x,y) - J^L(x + g_x^L + v_x^{k-1} , y + g_y^L + v_y^{k-1})$;

Vecteur de décalage d'image :

$$\bar{b}_k = \sum_{x=p_x-w_x}^{p_x+w_x} \sum_{y=p_y-w_y}^{p_y+w_y} \begin{bmatrix} \delta I_k(x,y) & I_x(x,y) \\ \delta I_k(x,y)I_y(x,y) & I_y(x,y) \end{bmatrix} ;$$

Calcul du Flot optique (Lucas-Kanade): $\bar{\eta}^k = G^{-1}\bar{b}_k$;

Evaluer l'itération suivante: $\bar{v}^k = \bar{v}^{k-1} + \bar{\eta}^k$;

Fin // la boucle k

Flot optique final de niveau L : $d^L = \bar{v}^k$;

Evaluer l'itération suivante: $g^{L-1} = \left[g_x^{L-1} \quad g_y^{L-1}\right]^T = 2(g^L + d^L)$;

Fin // la boucle L

Flot optique final : $d = g^0 + d^0$;

Emplacement du point J : $v = u + d$;

FIN.

2.4 Conclusion

Dans ce chapitre, nous avons abordé les deux représentations MHI et flux optique pour faire la reconnaissance d'action humaine. Tout d'abord, nous avons procédé par l'extraction des caractéristiques issues de ces deux dernières représentions afin de pouvoir les combiner pour avoir une vue plus claire et plus précise du problème dit "reconnaissance", de manière plus formelle. Il s'agit du problème de la classification d'activités humaines dans le but d'identifier ces actions, définies par un ensemble de caractéristiques.

La méthode d'estimation que nous avons choisie repose à la fois sur la détermination du flux optique et sur l'image historique de mouvement. Le choix de caractéristiques est une étape importante qui conditionnera les performances du descripteur du mouvement

2.5 Références

[ADE 85]: E .H. Adelson and J.R. Bergen, "Spatiotemporal energy models for the perception of motion", Journal of the Optical Society of America A., vol , n°2, 1985, pp.284-299 .

[ANA 89]: P.Anadan, Computational framework and an algorithm for the measurement of visual motion, Int. J. Comput Vision, 1989, Vol. 2, n° 1, p. 283-310.

[BAR 92]: J.L.Barron, D.J. Fleet, Beauchemin, "Performance of optical flow techniques", International Journal of Computer Vision, 1992, Vol. 12, n° 1, p. 43-77.

[BOB 96]: A. Bobick and J. Davis, "An Appearance-based Representation of Action", IEEE International Conference on Pattern Recognition, August 1996, pp. 307-312.

[BOB 01]: A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates", IEEE Trans. vol. 23, no. 3, pp. 257–267, 2001.

[BRA 00]: Gary R. Bradski, James Davis, "Motion segmentation and pose recognition with motion history gradients", IEEE Workshop on Applications of Computer Vision, pp. 238-244, Palm Springs, 2000.

[BOU 00]: Jean-Yves Bouguet, "Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm",Intel Corporation, 2000

[DAV 97]: J. Davis and A. Bobick, "The Representation and Recognition of Action Using Temporal Templates", IEEE Conference on Computer Vision and Pattern Recognition, June 1997, pp. 928-934.

[DAV 99]: James W. Davis, Gary Bradski, "Real-time motion template gradients using Intel CVLib" ICCV Workshop on Frame-rate Vision, 1999.

[EFR 03]: Efros A.A., Berg A.C., Mori G., Malik J., Recognizing action at a distance, IEEE International Conference on Computer Vision, Nice, France.

[FLE 90]: D. J. Fleet,A. D. Jepson, Computation of component image velocity from local phase information, International Journal of Computer Vision, 1990, Vol. 5, n° 1, p. 77-104.

[HEE 87]: D. J. Heeger, Optical flow from spatiotemporal filters, Proceedings of Proc. 1st Int. Conf. Computer Vision, 1987, p. 181-190.

[HOR 81]: B. K. P. Horn,B. G. Shunck, Determining optical flow, Artificial intelligence, 1981, Vol. 17, n° p. 185-203.

[KHA 09]: Muhammad Jamil Khan and Hafiz Adnan Habib, "Video Analytic for Fall Detection from Shape Features and Motion Gradients", WCECS 2009, October 20-22, San Francisco, USA, 2009.

[KOG 81]: T. Koga, K. Iinuma, A. Hirano, Y. Iijima,T. Ishiguro, Motioncompensated interframe coding for video conferencing, Proceedings of National Telecommunications Conference: Innovative Telecommunications - Key to the Future, New Orleans, LA, USA, 1981, p. G5.3.1-5.3.5.

[LUC 81]: B. Lucas,T. Kanade, An iterative image registration technique with an application to stereo video, Proceedings of DARPA Image Understanding Workshop, Washington, DC., 1981, p. 121-130.

[MAR 07]: Julien MARZAT, "Estimation temps réel du Flot Optique", 2008.

[NAG 86]:H. Nagel,W. Enkelmann, An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, Vol. 8, n° 1, p. 565-593.

[NEW 83]: W.T. Newsome, M. S. Gizzi and J. A. Movshon, « Spatial and temporal properties of neurons in macaque MT », Invest. Ophthalmol. Vis. Sci. Suppl.24, 1983, 106 p.

[PAL 85]: J .McLean—Palmer, J. Jones and L. Palmer, « New degrees of freedom in the structure of simple receptive fields », Invest. Ophthalmol. Vis. Sci . Suppl . 26, 1985.

[PEL 96]: Denis PELLERIN, Adrian SPINEI, Anne GUERIN-DUGUEOptical Flow Based on Combined Gabor Filters, 1996.

[SHI 94]: Jianbo Shi and Carlo Tomasi, \Good features to track", Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn., pages 593/600, 1994.

[SUV 06]: Nikom Suvonvorn, "Mise en Correspondance d'Images pour l'Analyse du Mouvement et la Stéréovision", these, 2006.

[RIC 08]: V. Ricquebourg, L. Delahoche, B. Marhic, D. Menga, A.M. Jolly-Desodt, "Une approche non-probabiliste pour la fusion multi-capteurs dans l'habitat communicant ", 2008.

[TAN 75]: S. Tanimoto & T. Pavlidis, "A HierarchicalData Structure for Picture Processing", Comp Grap and Image Proc, vol 4, pp104-119, 1975.

[VEN 02]: Emmanuel Veneau, " Macro-segmentation multi-critère et classification de séquences par le contenu dynamique pour l'indexation vidéo", thèse, 2002.

Chapitre 3 : Méthode proposée : Approche et résultats des tests

3.1 Introduction

Il s'agit de proposer les caractéristiques dominantes qui permettent la reconnaissance de l'action en tant que composante de l'activité. Notre approche consiste à combiner différentes représentations et descripteurs.

Tout d'abord, nous commençons par étudier la possibilité d'utilisation des directions du mouvement de la silhouette pour la reconnaissance de l'action. En effet, toute action est caractérisée par un ensemble de mouvement. L'idée de départ est-il possible de regrouper ces directions pour constituer des directions dominantes, chacune ayant une trajectoire déterminée en fonction de l'action considérée. Le flot optique est calculé sur l'image et sur l'image historique. N'ayant pas pu obtenir le résultat escompté en raison de la sensibilité du flot optique, nous avons étudié et utilisé le descripteur SURF afin de caractériser les régions d'intérêt en mouvement. Pour chacune de ces régions, la trajectoire en position et en direction dominante est sauvegardée et utilisée ensuite pour la classification des actions.

Dans ce qui suit, nous allons donner analyse de toutes les méthodes retenues pour extraire les caractéristiques nécessaires à la l'étape de classification

3.2 Approche 1 : Approche basée sur le regroupement de vecteurs du flot optique

Une action de l'humain est caractérisée par les mouvements effectués par ses membres (mains, bras, jambes, …). Pour caractériser ces mouvements, les vecteurs de mouvement (flot optique) peuvent être extraits et suivis le long de la séquence d'images.

Ce que nous proposons est d'étudier la nature du regroupement de ces vecteurs dans une image et leur évolution dans le temps en fonction des actions. L'objectif est de pouvoir trouver des primitives pouvant discriminer les actions. La démarche proposée est la suivante :

- Extraire les vecteurs de mouvement (flot optique) pour chaque image de la séquence vidéo
- Représentation des vecteurs sur une Image d'Historique du Mouvement (MHI).
- Regroupement des vecteurs de même direction (Clustering) sur la MHI
- Inférer les primitives permettant la classification de l'action.

La mise en œuvre de cette approche a nécessité de traiter les problèmes suivants :

3.2.1 Calcul des vecteurs de mouvement

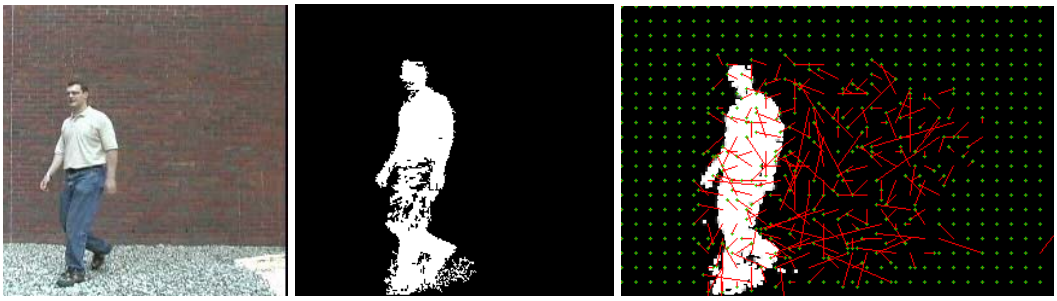
Cette étape nécessite la mise en correspondance de blocs obtenus suite à la subdivision de l'image en blocs non chevauchés de tailles identiques [PAC 97].

Le principe est de trouver le meilleur bloc dans l'image courante comparé à un bloc de l'image de référence. Le meilleur bloc est obtenu par un algorithme qui minimise un critère de similarité.

Le calcul des vecteurs de mouvement concernera uniquement les objets en mouvement pour éliminer le traitement des vecteurs dû aux bruits et optimiser le temps de traitement. En effet, se concentrer sur les blocs de l'avant plan permettra de réaliser l'appariement des blocs sur des régions restreintes.

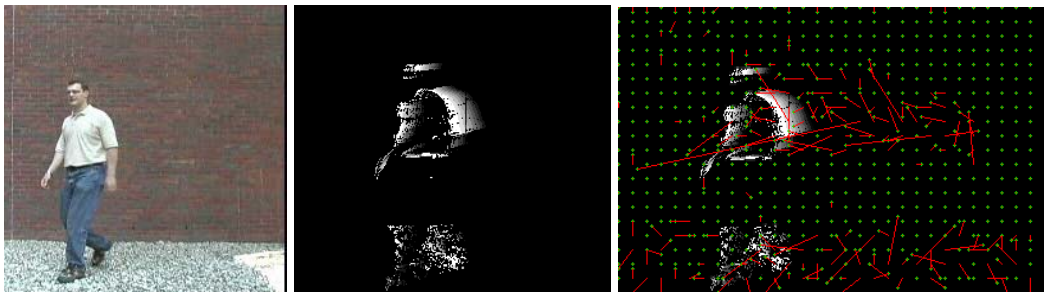
Il concernera deux types d'images:

- L'image de l'avant plan (Foreground) obtenue suite à la soustraction de l'image de référence (Background) et l'image courante (voir figure 3.1),
- L'image de l'historique de mouvement (MHI) obtenue comme cumul des images calculées comme différence de chaque paire d'images successives (voir figure 3.2).



(a) (b) (c)

Figure 3.1 : (a) Image courante, (b) Résultat de soustraction de la dernière image de la séquence avec l'arrière-plan, (c) Cumul des vecteurs de mouvement localisés sur la séquence d'images (en couleur rouge)



(a) (b) (c)

Figure 3.2 : Résultat de mise en correspondance de blocs appliqué à l'image MHI

Une image MHI est une image qui accumule les changements durant un intervalle de temps donné pour chaque pixel d'une image (voir la figure 3.2(b)). Les pixels de l'avant-plan les plus récents se voient assigner une couleur claire tandis que ceux appartenant au passé sont progressivement assombris. Comme cette méthode fait la mise en correspondance entre les blocs des deux images successives, nous générons des vecteurs de mouvement pour les deux images successives.

Pour le second type d'images (MHI), les vecteurs de mouvement sont calculés sur les différences entre chaque paire d'images successives. Le cumul des vecteurs de mouvement calculés sur la même image donne le résultat illustré par la figure 3.2(c).

La mise en correspondance de blocs (block matching) suppose la subdivision de l'image en blocs de tailles identiques appelés macro blocs. Chacun des blocs est comparé aux macros blocs situés dans la zone de recherche de l'image de référence. La sélection se fait en minimisant la mesure de similarité calculée comme étant la somme des valeurs absolues des différences entre les valeurs de pixels correspondants.

A noter que pour l'algorithme de mise en correspondance de blocs fournit des vecteurs de mouvement reliant les blocs et non des points d'intérêt. Ceci est dû au fait que la direction du vecteur de mouvement relie le pixel central du bloc de l'image au pixel central apparié du bloc sur l'image suivante. Les vecteurs de déplacement obtenus sont assez cohérents sur la totalité des tests opérés (voir exemples des figures 3.1(c), 3.2(c)).

Le regroupement de ces vecteurs ne peut donner une information caractérisant l'action pour les raisons suivantes (voir figure 3.3) :

(a)- Différences d'orientations des vecteurs de mouvement dans une même région

(b)- Taille du bloc à considérer pour le regroupement. En effet, plus cette taille est importante, plus il y a des vecteurs avec différentes directions. Si la taille est réduite, nous obtenons plus de directions dominantes de régions qui ne définissent pas en réalité le mouvement de l'action.

(c)- En cas de calcul de la résultante, plusieurs vecteurs dus au bruit sont présents dans les blocs et ne contribuent pas à calculer correctement le vecteur de direction dominant de la région.

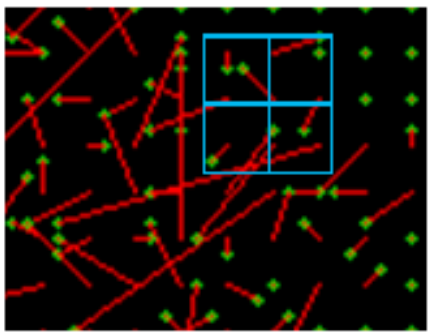


Figure 3.3 : Variation de la taille du bloc pour le calcul de la direction dominante des vecteurs du flot optique

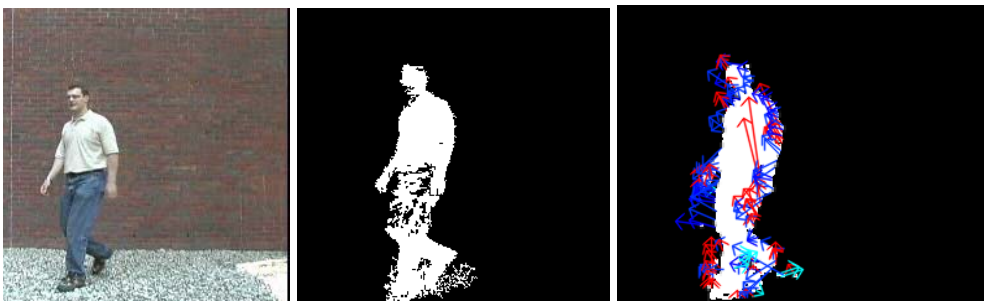
L'utilisation de MHI ne s'avère pas une solution pour faciliter l'étape de distinction des différents mouvements, en plus, la carte des vecteurs de déplacement est trop bruitée car un petit bloc sans texture peut corréler avec une multitude de voisins.

Il s'avère donc nécessaire d'améliorer le calcul du flot optique. Pour ce faire, nous utiliserons la technique de Lucas–Kanade (Pyramidal Lucas–Kanade) qui applique les traitements sur une multitude d'images obtenus suite au changement de résolution (pyramide d'images).

3.2.2 Calcul du flot optique moyennant la Pyramide de Lucas–Kanade

Une hauteur L_m de la pyramide est d'abord définie (Fixée pour nos tests à 4). A chaque niveau de la pyramide, nous sous-échantillonnons l'image d'un facteur 2 pour les deux images successives considérées. Le niveau zéro correspond à l'image initiale, le niveau L_4 correspond au niveau le plus grossier. A ce niveau (L_4), nous appliquons l'algorithme Lucas-Kanade [LUC81] sur l'image puis nous réitérons le calcul à une résolution moindre en partant du flot du résultat précédent. Réappliquons l'algorithme à nouveau et ainsi de suite jusqu'au niveau 0 qui correspondra à l'image initiale. Nous récupérons alors le flot optique final.

C'est un algorithme qui sert de référence dans le domaine du flot optique est disponible dans la librairie OpenCV d'Intel [BOU 00]. Nous appliquons cette méthode sur les deux résultats obtenus en soustrayant l'arrière-plan et en utilisant Motion History Image comme le montrent respectivement les figures 3.4 et 3.5.



(a) (b) (c)

Figure 3.4: Résultat du calcul du flot optique moyennant la pyramide Lucas-Kanade (utilisant soustraction d'arrière-plan). (c) concerne une paire d'images. Les vecteurs relient les points de Harris.

(a) (b) (c)

Figure 3.5: Résultat du calcul du flot optique moyennant la pyramide Lucas–Kanade (utilisant MHI)

Afin d'analyser les déplacements, nous nous intéresserons à leur orientations qui seront classés par différentes divisions d'angles. Pour représenter les différents sens du mouvement sur les figures 3.4 et 3.5, nous avons attribué à chacun d'eux une couleur (la couleur rouge pour le mouvement vertical, le bleu foncé indiquera le sens du mouvement horizontal et le bleu clair le sens contraire).

L'angle θ d'orientation de chaque vecteur de mouvement est calculé pour un point "p" dans l'image à l'instant "t", en utilisant le point q estimé par l'algorithme Pyramidal Lucas-Kanade sur l'image suivante (l'instant "t+1") (voir figure 3.6).

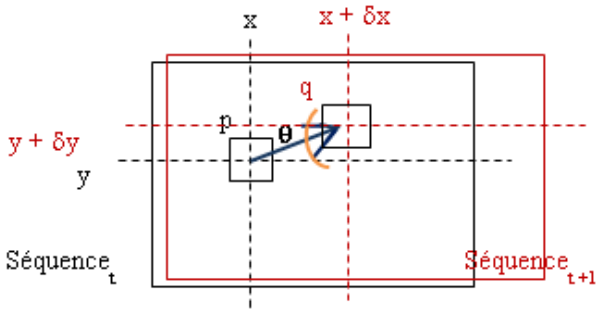


Figure 3.6: Orientation d'un vecteur de mouvement

L'attribution des différentes directions du mouvement représentées par l'angle θ est inspirée du principe de la rose du vent. Dans notre cas : l'arrière est représenté par l'Est, l'avant par l'Ouest, le haut par le Nord et enfin le bas par le Sud. A chaque essai, l'intervalle entre les points cardinaux a été ajusté jusqu'à déterminer l'angle le plus approprié à la direction.

Dès lors, chaque ensemble de θ approprié à une direction contribuera à la discrimination des classes des différentes directions du mouvement. Et la direction de l'activité globale est celle de la direction dominante de l'activité dans une vidéo.

Le principe repose sur l'utilisation de modèles de direction et de magnitude pour représenter des actions sans passer par la détection des membres du corps humain. L'approche consiste à extraire les magnitudes et les orientations de mouvement principales dans chaque bloc de la scène. Elle s'appuie sur les vecteurs de flot optique en tant que caractéristiques permettant de construire le modèle associé à une séquence.

3.2.3 Outils de classification

La reconnaissance des formes traite du problème de la prise de décision dans des problèmes de classements [AMB97]. Pour le problème traité, il s'agit de la classification d'actions d'individus, définies par un ensemble de caractéristiques, parmi un ensemble de classes préalablement connues, définies ou non.

Un problème de reconnaissance des formes nécessite alors de [SHU96] :

- définir les paramètres constituant le vecteur forme x, représentatif de l'état du système ; la dimension de x est celle de l'espace des caractéristiques ;
- définir l'ensemble des états ou classes connus pour lesquels nous disposons d'informations : modèle probabiliste de comportement, ensemble de vecteurs d'échantillons, etc. ;
- construire une règle de décision qui, à un vecteur forme x, associe soit la décision d'affecter une classe, soit la décision de rejeter toutes les classes connues, soit une non décision.

Parmi les approches de reconnaissance des formes, nous distinguons les méthodes supervisées pour lesquelles les caractéristiques des classes sont a priori connues et les méthodes non supervisées qui tendent à faire d'elles-mêmes l'apprentissage de ces caractéristiques.

Parmi ces méthodes, certaines offrent la prise en compte des informations contextuelles, ce qui les positionne donc dans un processus de segmentation et non dans le cadre de la discrimination ou de la classification.

Enfin, chaque méthode de reconnaissance des formes peut être décrite relativement à un fondement théorique, parmi lesquelles nous retrouvons la théorie des probabilités, la théorie des sous-ensembles flous et la théorie de l'évidence ou théorie des fonctions de croyance.
Parmi les techniques de classification les plus utilisées nous citons (voir l'annexe 1):

- Déformation temporel dynamique (DTW : Dynamic Time Warping) ;
- Les modèles de mixture de gaussiennes (GMM : Gaussian Mixture Models) ;
- Les modèles de Markov cachés (HHM : Hidden Markov Models).
- Les réseaux de neurones.

3.2.4 Tests de validation

Base de données WEIZMANN

La base de données Weizmann [GOR 07] sur laquelle nous avons testé nos algorithmes, se compose de dix différentes classes d'action: se pencher, courir, marcher, sauter, sauter en écartant les jambes et les bras, sauter sur une seule jambe, sauter en place, galoper sur le côté, sauter en agitant les deux mains, et sauter en agitant une seule main. Dans cette dernière, chaque classe d'action est effectuée une fois (parfois deux) par 9 sujets issus de 93 séquences vidéo au total.

L'arrière-plan dans les vidéos étant homogène et statique, Blank et al [BLA 05] préconisent un test à l'aide de l'algorithme leave-one-out de validation croisée pour approuver la performance des algorithmes et le choix des attributs utiles, ces méthodes procèdent par une subdivision de l'ensemble de données en différents sous-ensembles. Cette subdivision permet de fixer les paramètres des classes à

l'aide d'un plus grand nombre d'échantillons récupérés sur tous les sous-ensembles sauf quelques-uns et de les tester sur les sous-ensembles exclus afin d'obtenir un taux de classification élevée.

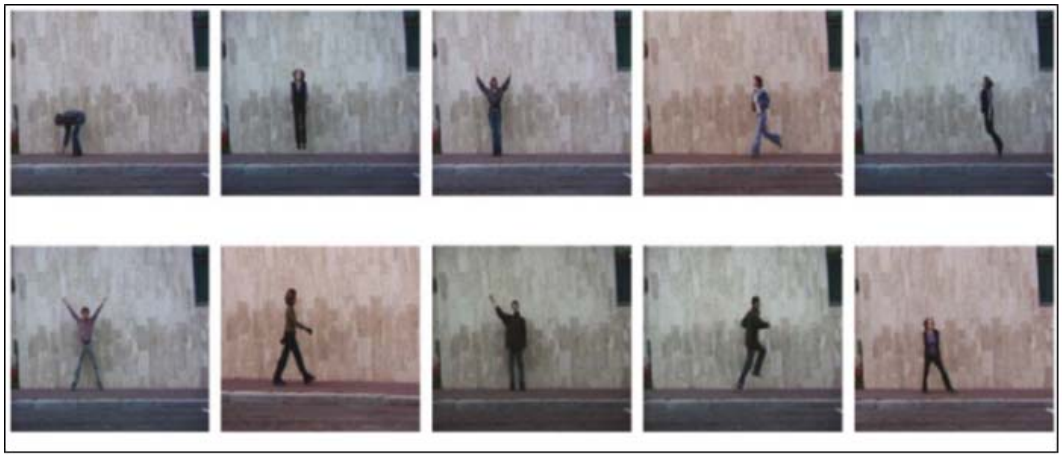


Figure 3.7 : Exemple de pose prise de différentes classes d'actions dans différents scenarios de la base WEIZMANN) : Se pencher, Sauter en place, Lever les deux mains, Courir, Sauter, Sauter en écartant les jambes et les bras, Marcher, Lever une seule main, Sauter sur une seule jambe, Galoper sur le côté.

Caractéristique de corpus

Les caractéristiques de notre base sont représentées en deux différentes parties : L'ensemble d'apprentissage et l'ensemble du test.

Les fichiers de la base sont des fichiers ou chacun est attribué à une caractéristique qui à son tour est caractérisé par une suite de vecteurs tirés de chaque action. Afin d'extraire les caractéristiques les plus importantes d'une vidéo, nous avons utilisé des méthodes d'analyse et d'estimation du mouvement. Dans ce qui suit, nous allons recenser les différentes caractéristiques qui peuvent être extraites des séquences vidéo et décrire les principales méthodes permettant de fournir ces dernières.

Le processus que nous avons suivi est le suivant :

- Tout d'abord, nous avons appliqué la soustraction entre l'arrière-plan de l'image courante pour obtenir l'avant plan représenté par une image binaire. Ce procédé permet la détection d'objet en mouvement en utilisant la théorie de « Background Gaussian Mixture substraction » [PIC 04]. Cela consiste à

construire un modèle d'arrière-plan de la vidéo puis de soustraire l'image courante de ce fond pour ne conserver que les objets qui ne font pas partie intégrante de l'arrière-plan, à savoir les éléments mobiles.

- En parallèle, nous avons utilisé la méthode de détection du mouvement « Motion History Image » pour regrouper les mouvements effectués au cours du temps en une seule image.
- Ensuite, nous avons calculé le flot optique en utilisant les deux méthodes de calcul (Block-Matching Et Pyramidal Lucas–Kanade) sur les résultats de la soustraction.

Les directions de mouvements pour chaque action de la base de tests Wiezmann [GOR 07] sont données par les histogrammes illustrés par les figures suivantes.

Action 1 : Boxer

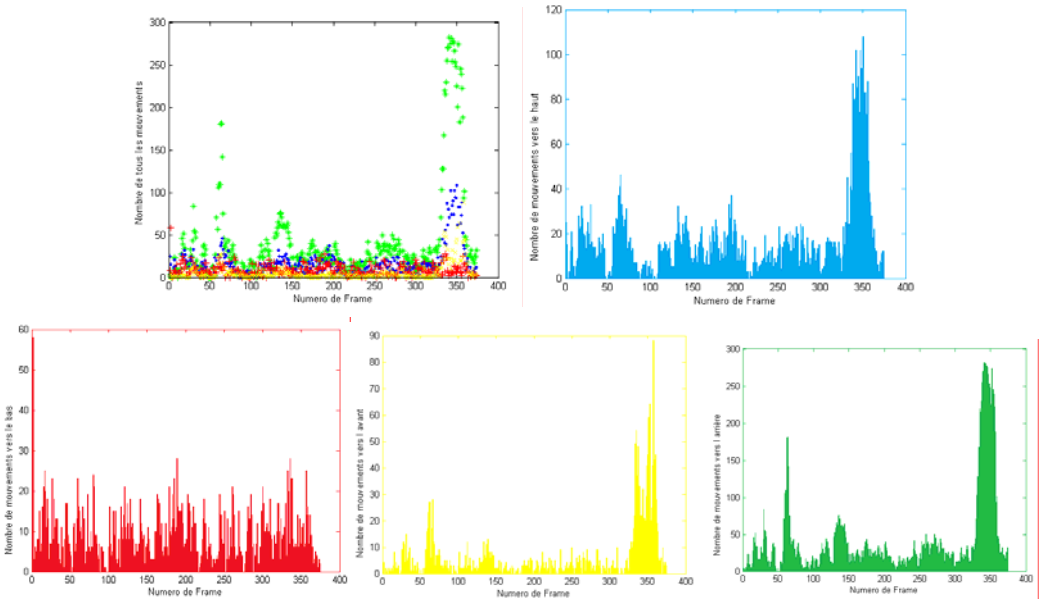


Figure 3.8 : Représentations graphiques des différentes orientations des mouvements de l'action « Boxer » où les couleurs Bleu, Rouge, Jaune, Vert sont associées respectivement aux mouvements vers le haut, le bas, l'avant et l'arrière

Si nous prenons ces histogrammes graphe par graphe, nous remarquons que pour l'action « boxer » les mouvements vers le haut, l'avant et l'arrière sont beaucoup plus présents et atteignent leur maximum au niveau de la 350ème image, alors que le mouvement vers le bas est au même niveau sur l'ensemble des séquences.

Action 2: Applaudir (Hand Clapping)

Pour le deuxième cas (figure 3.9) dont l'action de l'acteur consiste à taper des mains, nous avons suivi le même procédé. Tout d'abord, nous avons commencé par interpréter les différents graphes; les histogrammes du mouvement vers l'arrière (vert) et du mouvement vers l'avant (jaune) démontrent qu'ils se démarquent sur quelques séquences uniquement et non de façon régulière en comparaison avec les graphes du mouvement vers le bas et vers le haut, ou sur plusieurs séquences le nombre de mouvement croit.

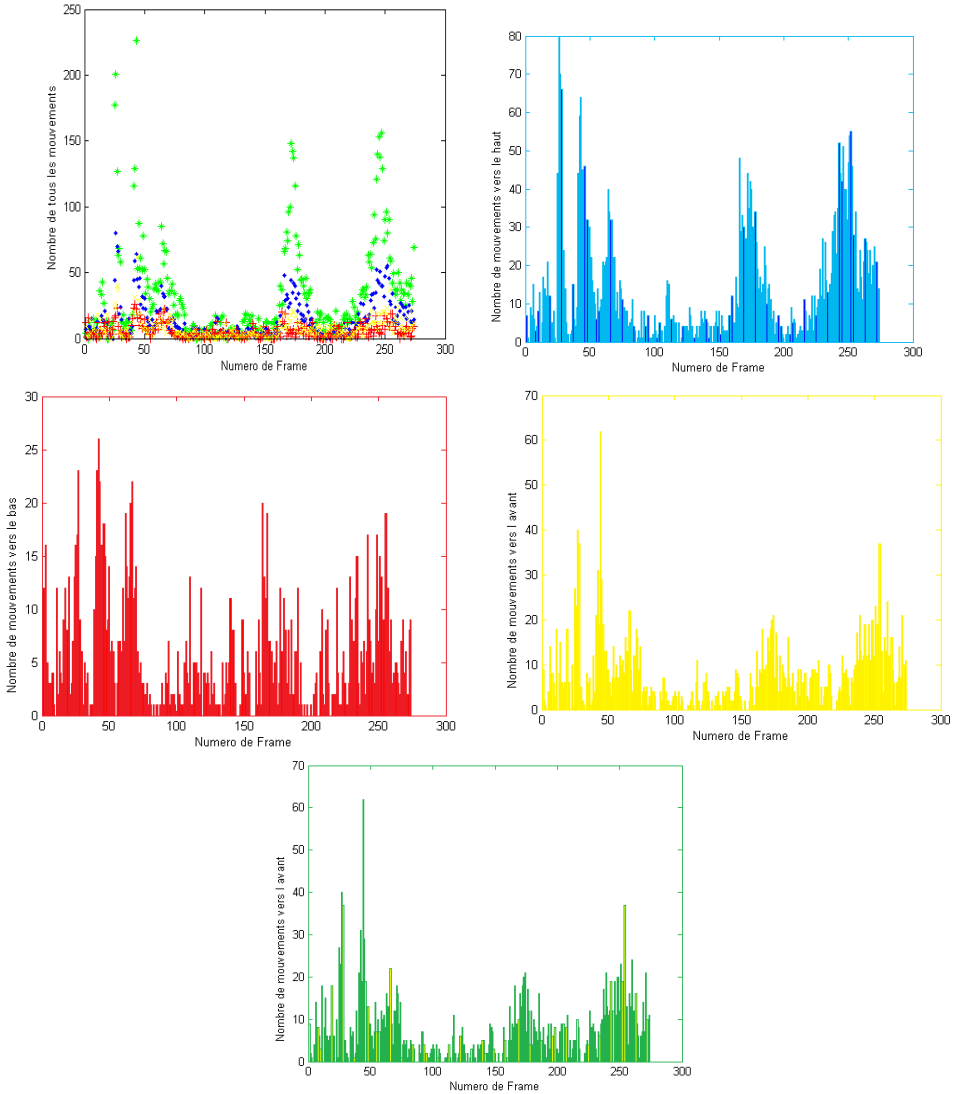


Figure 3.9 : Représentations graphiques des différentes orientations des mouvements de l'action « applaudir »

Action 3: Soulever les bras (Hand waving)

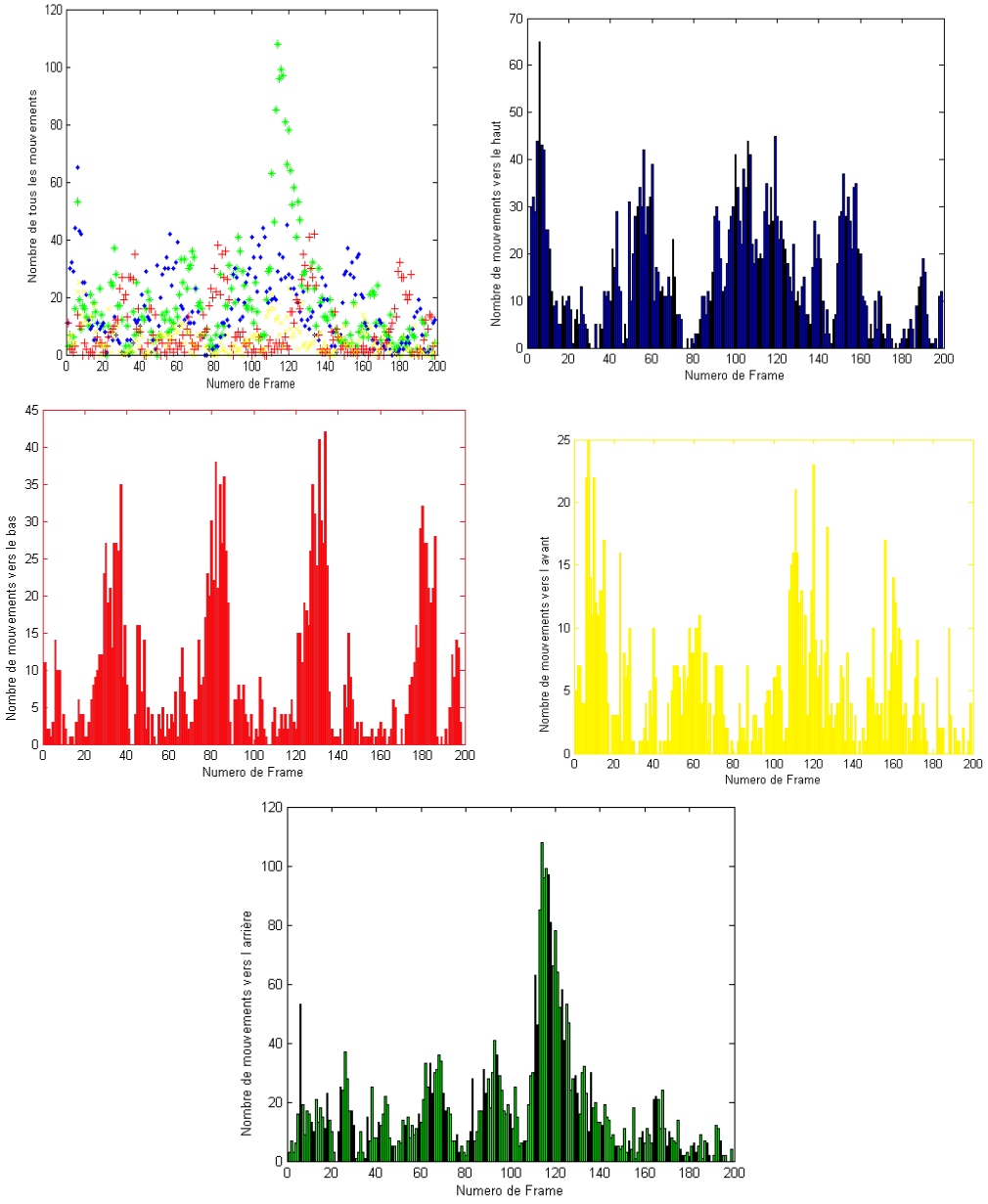


Figure 3.10 : Représentations graphiques des différentes orientations des mouvements de l'action « Soulèvement des bras »

Il en ressort que sur le graphe du mouvement vers l'arrière, le nombre de mouvements atteint un nombre assez important que dans la séquence 120, le graphe de mouvements vers le bas ne donne pas des informations quant à la présence de ces mouvements dans cette action, contrairement aux graphes des mouvements vers le haut et vers le bas, là ou le mouvement dans l'un des deux sens

vers le haut ou vers le bas atteint un nombre assez important, l'autre diminue et c'est un résultat assez logique du fait que les bras ne peuvent pas être dans ces deux directions contraires en même temps.

Synthèse :

Nous présentons sur la figure 3.11 les histogrammes par type de mouvement et par action.

Boxer Applaudir Soulever les bras

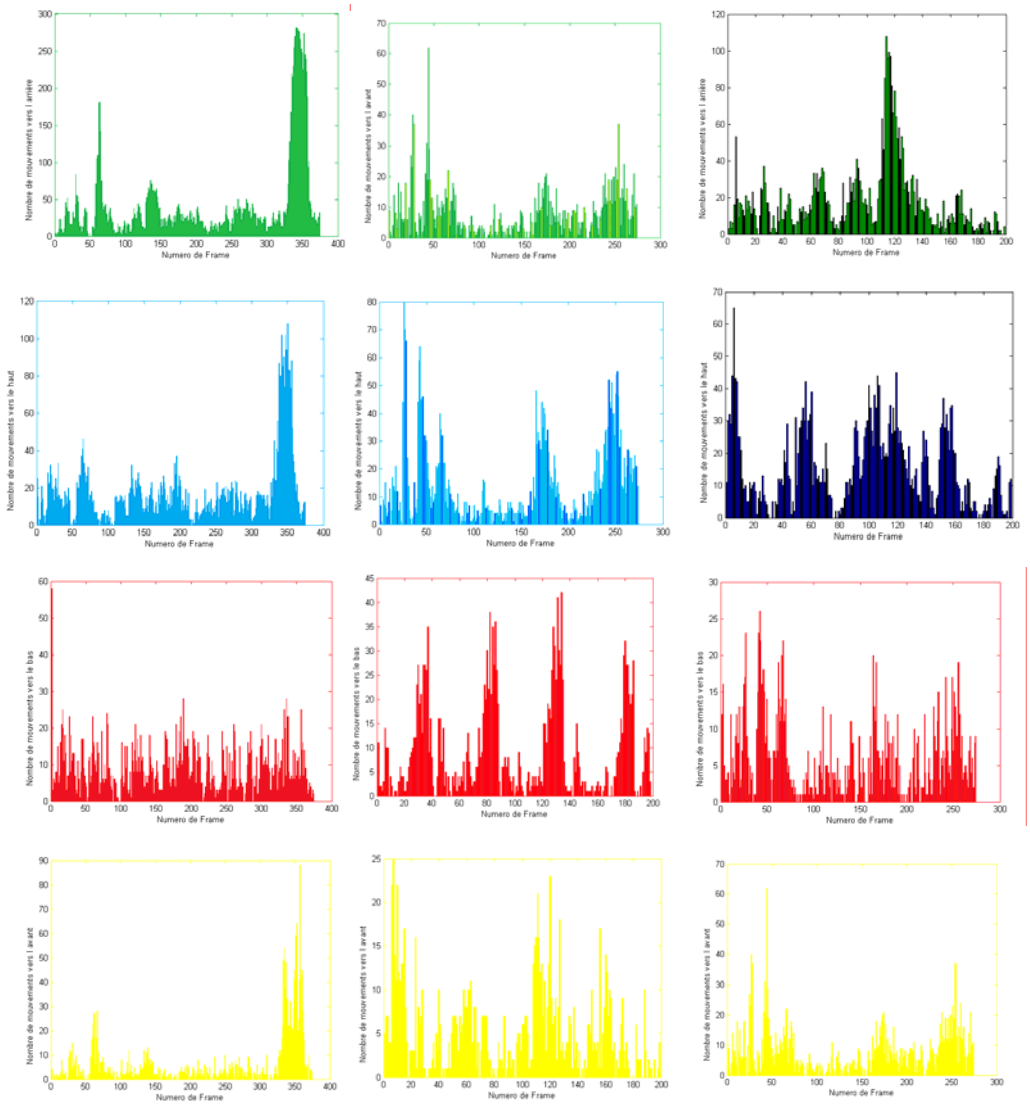


Figure 3.11 : Histogrammes par type de mouvement

L'étude des histogrammes des différents mouvements pour différentes actions a fait ressortir que le nombre de vecteurs par orientation n'est pas discriminant. En effet, sur la figure 3.11, nous constatons que pour les mouvements vers le bas (couleur rouge) la variation du nombre de vecteurs est quasiment identique pour les trois actions. Le même constat est valable pour les autres types de mouvement.

Avec un nombre d'actions élevé à reconnaître, la discrimination des histogrammes s'est avéré une tâche complexe et non soluble.

D'autre part, l'algorithme pyramidale Lucas et Kanade n'est pas adaptable aux grands mouvements (présence de confusion en cas de mouvements simultanés de bras et de jambes) et non adaptable aussi aux mouvements rapides (cas d'une action « courir » par exemple).

Aussi, le fait de lisser les détails (l'effet du filtre utilisé par la pyramide gaussienne) et permet de diminuer la précision du flot optique.

De ce qui précède, nous concluons que cette méthode reste inadaptée dans le cadre de notre étude.

La prise en considération de chaque partie a été envisagée en intégrant le descripteur SURF comme élément pour décrire les mouvements.

3.3 Approche 2 : Approche basée sur le descripteur SURF

Le descripteur SURF est basé en partie sur SIFT proposé en 2006 [BAY 06], permet de fournir des caractéristiques robustes et accélérées. Ce descripteur analyse les points d'intérêt détectés sur l'image pour en extraire des informations.

La première étape dans le calcul des caractéristiques locales consiste à détecter les endroits saillants tels que les coins et les jonctions. A partir des régions voisines de ces points d'intérêt, les caractéristiques de l'image sont ensuite calculées, ce qui donne un descripteur pour chacun d'eux. Les points correspondants entre deux images peuvent ensuite être trouvés en comparant ces descripteurs.

Notre choix s'est porté sur le descripteur SURF pour les raisons suivantes (Une présentation des principaux détecteurs de points d'intérêt est présentée en annexe 2):

- SURF permet la détection à la fois des points d'intérêt et de leurs descripteurs associés.
- Chaque point d'intérêt correspond à un endroit caractéristique codé par un descripteur local appelé SURF. Ces points d'intérêt et leurs descripteurs ont l'avantage d'être robustes au bruit, rapides en calcul, et précis. Leur utilisation a permis d'avoir de meilleurs résultats en vision par ordinateur [BEN 12] [LIV 12][NOG 10][NOG 12][WAN 12][ZHA 11] .
- Ces points d'intérêt sont des points qui contiennent beaucoup d'informations relatives à l'action. Ce sont des points aux voisinages auxquels l'image varie significativement dans la direction dominante.

Le principe de base de notre approche consiste à calculer pour chaque image d'une séquence vidéo les descripteurs SURF associés aux régions d'intérêt. Nous nous intéressons aux régions qui présentent un mouvement durant la production d'une action. Il s'agit ensuite de faire un suivi de ces régions le long de la séquence d'images. Si la région est en déplacement, nous nous intéressons à sa trajectoire : la direction dominante et les positions seront utilisées comme caractéristiques de l'action.

Ces caractéristiques sont alors utilisées pour la classification des actions moyennant l'outil de classification HMM.

3.3.1 Calcul du descripteur SURF

Le processus du calcul du descripteur SURF est le suivant :

- **Extraction des points SURF**
  - Cette étape consiste à rechercher les maxima de *det(H)*

$$H(x,\sigma) = \begin{bmatrix} L_{xx}(x,\sigma) & L_{xy}(x,\sigma) \\ L_{xy}(x,\sigma) & L_{yy}(x,\sigma) \end{bmatrix} \quad (3.1)$$

$$L_{xy} = \frac{\partial^2 G_\sigma}{\partial x \partial y} * I \qquad (3.2)$$

Bay et al. [BAY 06] approximent les dérivées secondes des gaussiennes par des filtres plus simples présentés par la figure 3.12. Le calcul du déterminant de la matrice H(x,σ) devient alors :

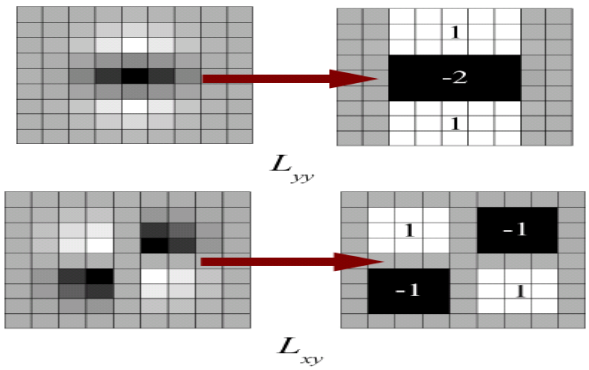


Figure 3.12 : Application du filtre gaussien

$$\det(H_{approx}) = D_{xx}D_{yy} - (0.9D_{xy})^2 \qquad (3.3)$$

- Accélération du calcul par Image intégrale

  Utilisation d'image intégrale est introduite pour accélérer le temps de calcul. L'intégrale d'image est définie comme suit :

$$I_\Sigma(x,y) = \sum_{i=0}^{i \le x} \sum_{j=0}^{j \le y} I(x,y) \qquad (3.4)$$

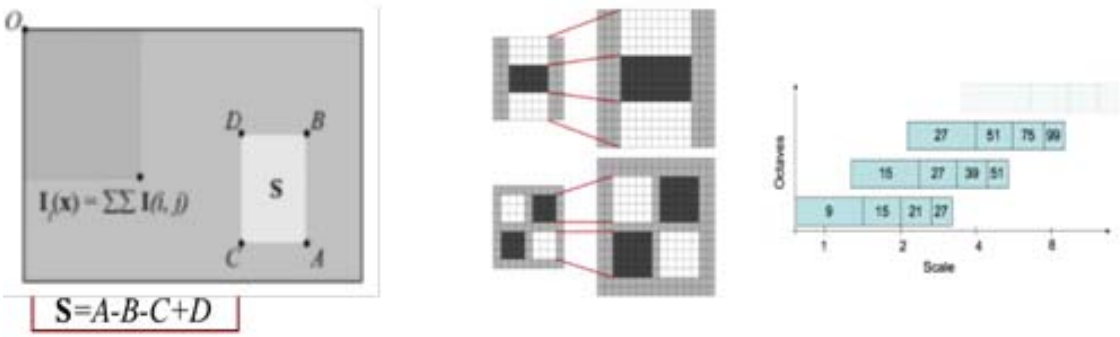


Figure 3.13 : Intégrale d'image

Le descripteur proposé peut se calculer de manière rapide à partir d'intégrale d'image. Un tel principe a déjà été utilisé dans [VIO 01] pour le calcul rapide d'ondelettes de Haar ou plus récemment dans [TUZ 08] pour le calcul rapide de matrices de covariances ou dans [POR 05] pour le calcul d'histogrammes.
Il nécessite l'interpolation de la position (x,y,σ) du point d'intérêt, dont le but est d'avoir la position à l'échelle 'S' en coordonnées réelles (non entières). Et en multipliant par le facteur d'échelle, le pixel exact à toutes les échelles est calculé.

- **Calcul de l'orientation des points détectés**.

  L'orientation est calculée comme suit :

- **Application des ondelettes de Harr pour** les directions x et y au niveau des pixels voisins. Les réponses d_x et d_y sont obtenues:

Figure 3.14 : Calcul des réponses des ondelettes de Haar

- Calcul d'un vecteur d'orientation à partir des sommes des réponses d_x et d_y au sein d'une fenêtre "angulaire".
  Chaque point d'intérêt est caractérisé par la distribution des orientations du gradient d'intensité dans 16 quadrants 4x4 autour du point considéré. Les orientations sont quantifiées comme étant correspondante au plus long de ces vecteurs suivant 8 valeurs [BOU 11].

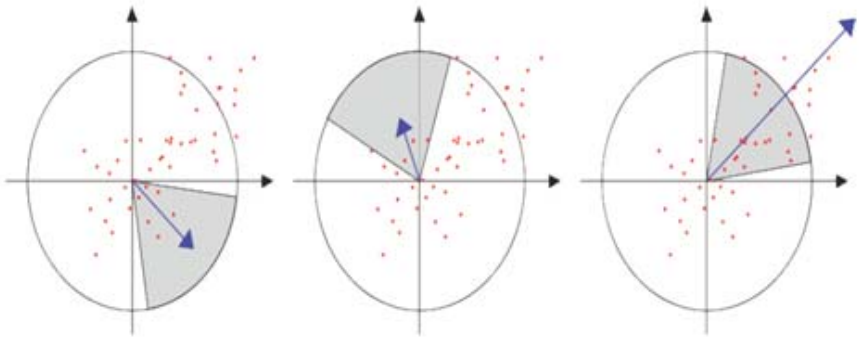


Figure 3.15 : Orientations du gradient

- "Rotation" pour se trouver dans la direction de l'orientation du point d'intérêt calculé précédemment

- Calcul des réponses dx et dy des ondelettes de Haar respectivement pour les directions x et y au niveau de pixels voisins
- Somme de ces réponses dx et dy mais aussi de |dx| et |dy| au sein de fenêtres situées de part et d'autres du point d'intérêt.

Un vecteur obtenu pour chaque fenêtre

$$v_{subregion} = [\sum dx, \quad \sum dy, \quad \sum |dx|, \quad \sum |dy|] \quad (3.5)$$

- Concaténation de ces vecteurs en un seul vecteur : le descripteur SURF

Le vecteur SURF est constitué de 64 éléments alors que le descripteur SIFT est constitué de 128 éléments.

3.3.2 Extraction des caractéristiques

Notre objectif est d'apparier les points caractéristiques entre 2 séquences successives de la même action. Les points d'intérêt représentant une région d'une séquence sont détectés et utilisés ensuite pour identifier les parties les plus correspondantes. Sur la figure 3.16, chaque point d'intérêt est associé à une région indiqué avec un cercle dont le rayon est estimé en fonction de l'échelle calculée.

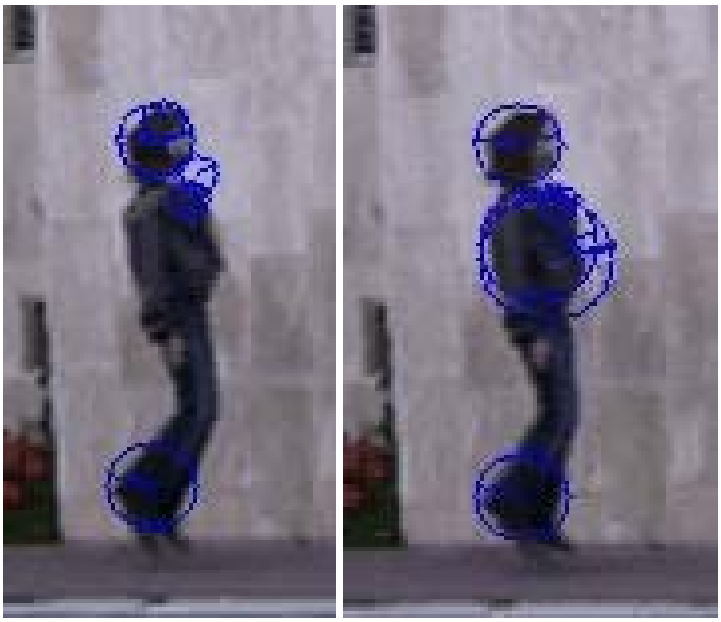


Figure 3.16: Résultat du calcul du descripteur SURF appliqué à une paire d'images d'une action.

Le calcul du descripteur SURF permet d'évaluer pour chaque zone de l'image les descripteurs des pixels et l'orientation.

Une façon simple de localiser le mouvement pour la classification des actions et de retrouver pour chaque région un point d'intérêt représentant le centre de gravité de la région considérée et de lui associer en plus de ses caractéristiques, la direction dominante de la région.

La taille du vecteur descripteur est un critère important qui intervient dans l'utilisation des algorithmes de classification. Ce vecteur caractéristique est de taille invariante selon le nombre de catégories des points d'intérêts dans l'image.

L'algorithme que nous proposons et illustré par l'organigramme de la figure 3.17 consiste à recalculer, pour deux zones de mouvement en position générale, deux nouveaux points représentatifs. Ce qui permet d'avoir comme propriété un résumé de la zone où se déroule le mouvement (en origine cette zone est caractérisée par plusieurs points d'intérêts détectés lors du déroulement de descripteur SURF) déclenché qui tourne autour de ce point pendant tout son trajet (mouvement).

En pratique, cette contrainte proposée n'est pas toujours vérifiée, le point associé doit être le plus robuste possible afin d'être reconnu même si la position et l'orientation du mouvement diffèrent entre les deux zones à comparer. De même, le changement d'échelle (zoom) doit être pris en considération pour qu'il n'y ait pas d'occultations lors de la recherche d'une zone similaire, sans parler des changements de points de vue (orientation des caméras) qui altèrent ce processus.

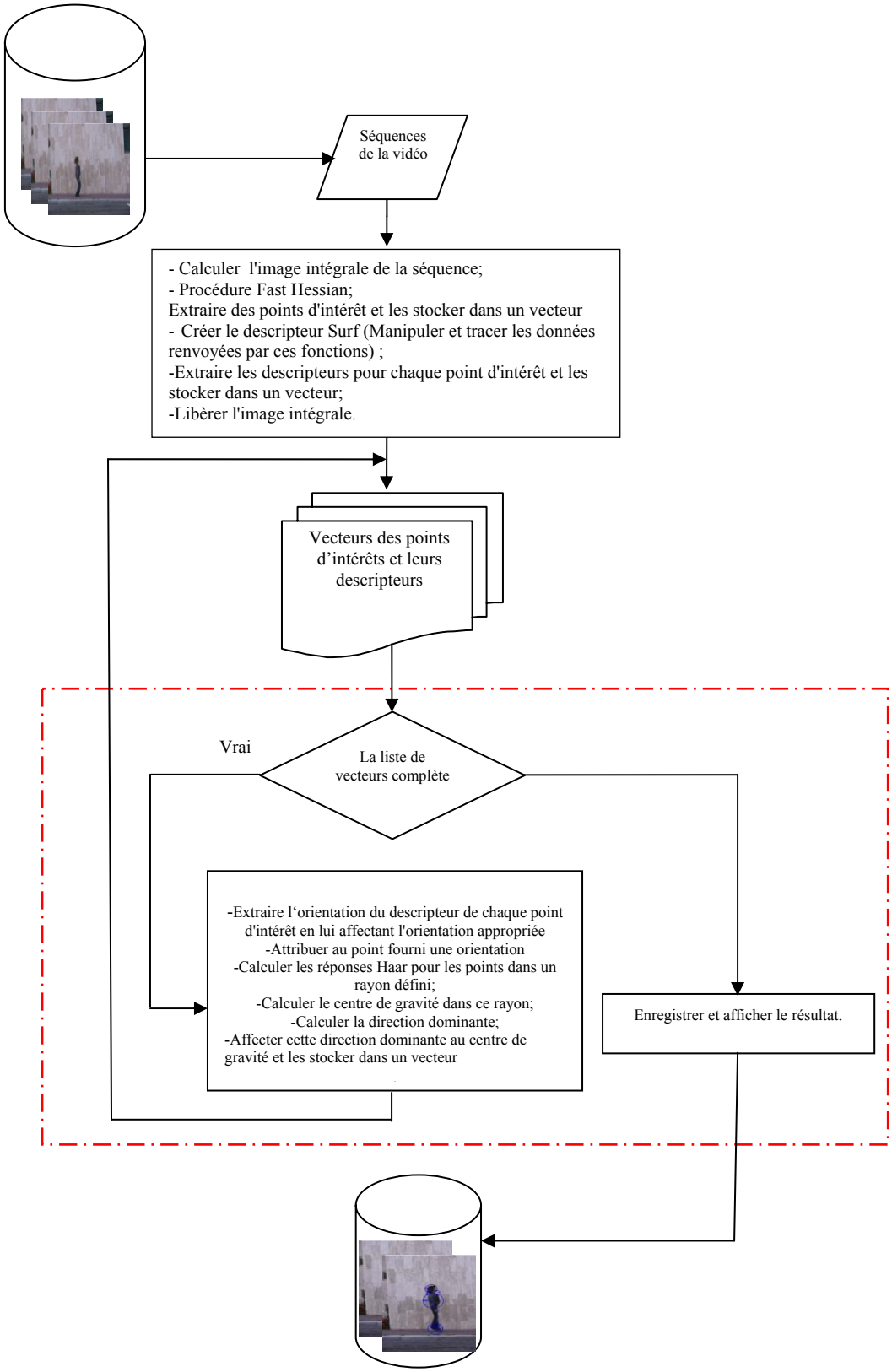


Figure 3.17 : Processus d'extraction des caractéristiques utilisant le SURF.

3.3.3 Classification des actions

La prochaine étape consiste à appliquer les modèles de Markov caché choisis pour la tâche de classification.

Le but de la phase de reconnaissance est de classer chaque action dans l'une des dix classes. Les traitements nécessaires sont relativement similaires à ceux utilisés dans l'étape d'apprentissage.

Le tableau 3.1 illustre l'application de la théorie des HMM à notre problème de reconnaissance qui s'est traduite par l'adaptation d'un outil existant basé sur la toolbox HMM développé par Kevin Murphy [MUR 05].

Une fois notre modèle construit, il est utilisable pour la reconnaissance de mouvements. La reconnaissance se fait en construisant un nouvel HMM au fur et à mesure de la lecture des données et en le comparant avec les HMM construits lors de l'apprentissage via l'algorithme de Viterbi. L'idée principale est de supposer qu'un mouvement et plus largement une série de caractéristiques est un HMM. Il s'agit ensuite d'utiliser les données d'apprentissage pour construire une chaine de Markov dans le but de reconnaitre le mouvement requête où la probabilité d'une séquence test est calculée par rapport à un ensemble d'apprentissage des modèles HMM où le HMM a un maximum de vraisemblance.

3.3.4 Tests et Discussion

Pour chaque opération, le tableau 3.1 illsutre le résultat de classification des actions sur la base de données de Weizmann. Pour chaque action, en nombre de 10, nous donnons le numéro de l'activité reconnu et la valeur du maximum de vraisemblance calculé.

Table 3.1: Résultat de classification de chaque activité pour les 9 opérateurs

Action	Sujet1	Sujet2	Sujet3	Sujet4	Sujet5	Sujet6	Sujet7	Sujet8	Sujet9
1	1, -0.011	1, -0.018	1, -0.001	1, -0.010	1, -0.023	1, -0.006	1, -0.002	1, -0.021	1, -0.013
2	2, -0.077	2, -0.034	2, -0.176	2, -0.137	2, -0.011	2, -0.111	2, -0.234	2, -0.004	2, -0.09
3	3, -0.036	7, -0.006	3, -0.023	3, -0.009	3, -0.007	3, -0.008	3, -0.025	3, -0.006	3, -0.030
4	2, -0.011	4, -0.026	4, -0.031	4, -0.027	4, -0.025	4, -0.023	4, -0.031	4, -0.025	2, -0.005
5	5, -0.001	5, -0.016	5, -0.021	5, -0.019	5, -0.015	5, -0.010	5, -0.027	5, -0.013	7, -0.006
6	6, -0.018	6, -0.010	6, -0.025	6, -0.015	6, -0.012	6, -0.013	6, -0.024	6, -0.013	6, -0.011
7	7, -0.021	7, -0.001	7, -0.024	7, -0.024	7, -0.0008	7, -0.025	5, -0.029	7, -0.002	7, -0.027
8	8, -0.039	8, -0.019	8, -0.010	8, -0.026	8, -0.016	8, -0.026	8, -0.013	8, -0.018	8, -0.040
9	10, -0.01	2, -0.01	4, -0.026	9, -0.151	2, -0.015	9, -0.129	4, -0.020	2, -0.020	10, -0.001
10	2, -0.037	2, -0.004	10, -0.01	10, -0.002	2, -0.005	10, -0.01	10, -0.006	2, -0.007	2, -0.026

Le pourcentage de classification des actions a été évalué à 70 %, 70%, 90%, 100%, 80%, 100%, 80%, 80%, 60% respectivement pour les actions : 1, 2, 3, 4, 5, 6, 7, 8, 9 et 10.
Nous remarquons aussi que les actions 9 et 10 sont souvent mal classifiées. L'action 9 est reconnue comme étant l'action 2, 4 ou 10.Alors que l'action 10 est reconnue comme étant l'action 2.

Il est intéressant de remarquer que plus nous effectuons des itérations, plus le classifieur est performant. Cela montre que même si les descripteurs SURF fournissent des caractéristiques moins précises, le classifieur HMM arrive à classer correctement la plupart des actions malgré le bruit des données dans le descripteur.

Il existe des cas de confusion entre l'ensemble de mouvements, que nous citons :

Pour l'opération 1 :

-Sauter en place (action 4) et sauter en écartant les jambes et les bras (action 2),
-Sauter en agitant une main (action 9) et sauter en agitant les deux mains (action 10),
-Sauter en agitant les deux mains (action 10) et Sauter en écartant les jambes et les bras (action 2).

Pour l'opération 2 :

-Sauter (action 3) et Sauter sur une seule jambe (action 7),
-Sauter en agitant une main (action 9) et Sauter en écartant les jambes et les bras (action 10),
-Sauter en agitant les deux mains (action 10) et Sauter en écartant les jambes et les bras (action 2).

P*our l'opération 3 :*

-Sauter en agitant une main (action 9) et Sauter en place (action 4).

Pour l'opération 5 :

-Sauter en agitant une main (action 9) et Sauter en écartant les jambes et les bras (action 2),
-Sauter en agitant les deux mains (action 10) et Sauter en écartant les jambes et les bras (action 2).

Pour l'opération 7 :

-Sauter sur une seule jambe (action 7) et courir (action 5),
-Sauter en agitant une main (action 9) et Sauter en place (action 4).

Pour l'opération 8 :

-Sauter en agitant une main (action 9) et Sauter en écartant les jambes et les bras (action 2),
-Sauter en agitant les deux mains (action 10) et Sauter en écartant les jambes et les bras (action 2).

Pour l'opération 9 :

-Sauter en agitant une main (action 9) et sauter en agitant les deux mains (action 10),
-Sauter en agitant les deux mains (action 10) et Sauter en écartant les jambes et les bras (action 2).

En fait, le mouvement "sauter en agitant une main (action 9) "est chevauché avec celui de "sauter en agitant les deux mains (action 10) ", et ces deux derniers peuvent être confondus avec le mouvement "Sauter en écartant les jambes et les bras (action 2) ", les autres actions sont classifiées avec succès. Concernant les traitements 4 et 6, le taux de classification était à 100%.

Comparaison à l'état de l'art

Nos résultats avec une moyenne de 81.11% comme taux de reconnaissance, comparés à certains travaux qui ont été appliqués à la base de Weizmann sont jugés acceptables même s'ils ont été surclassés. En effet, les taux 99.1%, 98.8%, 94.40%, 90%, sont atteints respectivement dans: Blank [BLA 05], Jhuang [JHU 07], Thurau [THU 08], Niebles [NIE 08].

Ceci est dû au fait que notre approche utilise un seul descripteur à savoir le descripteur SURF. L'intégration d'autres descripteurs pourra améliorer considérablement les résultats.

3.4 Conclusion

Dans ce chapitre nous avons exposé les résultats de notre investigation pour la recherche d'une solution au problème de reconnaissance des actions humaines. Nous avons d'abord exploré la technique du flot optique dont l'objectif de regroupements des vecteurs de mouvement pour chaque action. Les tests opérés ont montré la limite de cette méthode, ce qui nous a amené à utiliser le flot optique mais avec une pyramide gaussienne proposée par L. Kanade. La représentation des mouvements obtenus n'a pas été discriminante pour différencier les actions.

La troisième proposition concerne l'intégration du descripteur SURF pour la description des actions. La méthode proposée qui consiste à tenir compte des zones d'intérêt en mouvement a été validée sur la base de tests de Weizmann. Les résultats obtenus ont été comparés à l'état de l'art et jugés acceptables.

Le problème de reconnaissance d'actions reste ouvert et le problème de caractérisation des actions conditionne les résultats de classification.

3.5 Référence

[ACH 00] : C. Achard, E. Bigorgne, and J. Devars. "A sub-pixel and multispectral corner detector", International Conf. on Pattern Recognition, 2000.

[AMB 97] : C. Ambroise, "Introduction à la reconnaissance statistique des formes", Technical report,École des Mines de Paris, 1997.

[BAU 73] : L.E. Baum, "An Inequality and Associated Maximisation Technique in Statistical Estimation for Probabilistic Functions of Markov Processes", inequalities, vol. III Academic press New York 1973.

[BAY 06] : H. Bay, T. Tuylelaars, and L. Van Gool. Surf : "Speeded up robust features.European" Conference on Computer Vision, 2006.

[BEA 78] : P.R. Beaudet. "Rotational Invariant Image Operators". International Conference on Pattern Recognition, pp.579-583, 1978.

[BEN 12]: Rachid Benmokhtar, "Robust human action recognition scheme based on high-level feature fusion", springer 2012.

[BIM 88] : F. Bimbot, " Synthèse de la parole : du segments au règles, avec utilisation de la décomposition temporelle ", Thèse 1988.

[BLA 05] : M. Blank, L. Gorelick, E. Shechtman, M. Irani, et R. Basri, "Actionsasspace-time shape"s, ICCV,vol.2,2005.

[BOH 00] :L.N. Bohs, B.J. Geiman, M.E. Anderson, S.C. Gebhart and G.E. Trahey, "Speckle tracking for multidimensional flow estimation," Ultrason., vol. 38, pp. 369-375, 2000.

[BOU 00] : J.Y. Bouguet, "Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm",Intel Corporation, 2000.

[BOU 11] : J.M. BOUCHER,"Caractérisation et modélisation de la distribution spatiale de signatures locales dans les images : application à la classification d'image sonar de fonds marins", Thèse 2011.

[BOR 98] : J. Boreczky and L. Wilcox, " A hidden markov model framework for video segmentation using audio and image features", In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 6, pages 3741–3744, 1998.

[BRA 97] : M.Brand, N.Oliver, and A. Pentland,: "Coupled hidden Markov models for complex action recognition". Proc. Int. Conf. on Computer Vision and Pattern Recognition, 1997.

[BRA 08] : G.Bradski and A.Kaehler,"Learning OpenCV",2008.

[BRI 95] : J. S. Bridle, “ Optimization and search in speech and language processing ”, Survey of the state of the art in human language technology, 1995.

[CAM 07] : N. CAMELIN , "Stratégies robustes de compréhension de la parole basées sur des méthodes de classification automatique ", thèse, 2007.

[CES 98] : CESAC, UMR C5576, CNRS, Internationnal workshop on applications of artificial neural networks to ecological modelling, Toulouse, France, 14-17 décembre 1998.

[CHA 00] : A. Chardin, "Modèles énergétiques hièrarchiques pour la résolution des problèmes inverses en analyse", 2000.

[CHE 03] : F.-S .Chen., C.-M .Fu., C.-L.Huang, “Hand gesture recognition using a realtime tracking method and hidden Markov models”, Image and Vision Computing, 2003.

[CHO91] : H.S.Choi, D.R.Haynor and Kam, "Partial volume tissue classification of multichannel magnetic resonance images Amixel model ", IEEE Trans. Med. Image., Vol.10, pp395-407, 1991.

[DAV 93] : E.Davalop and P.Naim, "Les réseaux de neurones", Edition Masson 1993.

[DRE 82] : L.Drechler and H.Nagel, "On Selection of Critical Points and Local Curvature Extrema of Region Boundaries for Interframe Matching"., Inter-national Conference on Pattern Recognition, 542-544, 1982.

[GAR 98] : S. Garcia-Sallicetti, B. Dorizzi, and P. Gallinari, "A neural predictive system for online handwriting recognition", Pattern Recog., 1998.

[GOR 07]: L.Gorelick, M.Blank, E.Shechtman, M.Irani, R.Basri, "Actions as space-time shapes", IEEE Transactionon Pattern Analysis and Machine Intelligence, 29 (12),2007,pp2247–2253.

[HAR 88] : C. Harris and M. Stephens. A combined corner and edge detector. Alvey Vision Conference, pages 147-151, 1988.

[HIE 99] Hienz H., Bauer B., Kraiss K.F., “HMM-based continous sign language recognition using stochastic grammars”, 1999.

[HUA 01] : X. Huang, A. Acero, and H.-W. Hon. "Spoken Language Processing: A Guide to Theory, Algorithm and System Development", Prentice-Hall, Englewood Cliffs, N.J., 2001.

[INT 11] : INTEL CORPORATION, Open Source Computer Vision Library – OpenCV [en ligne]. Disponible sur : http://www.intel.com/research/mrl/research/opencv.

[JHU 07]: H. Jhuang, T. Serre, L.Wolf, and T. Poggio. A biologically inspired system for action recognition. ICCV, pp. 1-8, 2007.

[JOH 98]: JOACHIMS T., " Text categorization with support vector machines : learning with many relevant features ", Proceedings of ECML-98, 10th European Conference on Machine Learning, Chemnitz, DE, 1998, Springer Verlag, Heidelberg, DE, p. 137–142.

[JOD 94] : J-F.Jodouin, "Les Réseaux de neurones : Principes et définitions. Les Réseaux Neuromimétiques : Modèles et applications.Editions" Hermès, Paris.1993.

[KHA 02] : J. Kharroubi, " Etude de Techniques de Classement Machines à Vecteurs Supports pour la Vérification Automatique du Locuteur" , thèse , 2002.

[KIT 82] : Kitchen L. et Rosenfeld A., "Gray-Level Corner Detection", Pattern Recognition Letters, 95-102, 1982.

[KLA 10]:A. Klaser, "Learning human actions in videos", thèse ,2010.

[KOE 84] : J.Koenderink, "The Structure of Images", Biol. Cybern,363-370,1983.

[LEE 96]: C.Lee, Y.Xu, "Online, interactive learning of gestures for human/robotinterfaces", IEEE 1996.

[LIN 94]: Lindeberg, "Scale-Space Theory in Computer Vision", Kluwer Academic Publishers/Springer, Dordrecht, Netherlands, 1993.

[LIN 97] : Lindeberg and Garding "Shape-adapted smoothing in estimation of 3-D depth cues from affine distortions of local 2-D brightness structure", Proc. 3rd European Conf. on Computer Vision, pp. 389--400, Springer-Verlag 1997.

[LIV 12]: Micha Livnea, ,Leonid Sigala, Nikolaus F. Trojec, David J. Fleet, "Human attributes from 3D pose tracking ", Elsevier, 2012.

[LOE 04] : B.L. Loeding, S.Darkar, A.Parashar, A.Karshmer, "Progress in automated computer recognition of Sign Language", Springer-Verlag, 2003.

[LOW 04] : D. Lowe, "Distinctive image features from scale-invariant keypoints," Internatio-nal Journal of Computer Vision, vol. 60(2), pp. 91–110, 2003.

B. D. Lucas and T. Kanade (1981), *An iterative image registration technique with an application to stereo vision.* Proceedings of Imaging Understanding Workshop, pages 121--130

[MA 00] : J.Ma, W.Gao, J.Wu, C.Wang, "A continuous Chinese Sign Language recognition system", 2000.

[MAN 11] : G-B.Manuel, "Descripteurs 2D et 2D+t de points d'interet pour des appariements robustes", Thèse 2011.

[MAT 02] : J. Matas, O. Chum, M. Urban and T. Pajdla (2002). "Robust wide baseline stereo from maximally stable extremum regions" (PDF). British Machine Vision Conference: 384–393. (le détecteur de régions d'intérêt MSER)

[MER 04] : D.MERAD, "Reconnaissance 2D/2D et 2D/3D d'objets àpartir de leurs squelettes" thèse 2003.

[MIK 05]: K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors,"IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27(10), pp. 1615–1630, 2005.

[MOR 01] : P. Moravec. " 3D graphics and the wave theory ", Proceedings of the 8th annual conference on Computer graphics and interactive techniques, pages 289–296, New.

[MUR 05]: Hidden Markov Model (HMM) Toolbox for Matlab Written by Kevin Murphy, 1998,Last updated: 8 June 2005. http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html.

[NIE 08] J. C. Niebles, H.Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. Int'l: J. Computer Vision, 79(3):299–318, 2008.

[NOB 89] :J. A. Noble, Descriptions of Image Surfaces. PhD thesis, Department of Engi-neering Science, Oxford University, 1989.

[NOG 10]: Akitsugu Noguchi, Keiji Yanai, "Extracting Spatio-temporal Local Features Considering Consecutiveness of Motions", springer, 2010.

[NOG 12]: Akitsugu Noguchi, Keiji Yanai, "A SURF-Based Spatio-Temporal Feature for Feature-Fusion-Based Action Recognition", springer, 2012.

[NIL 02]: M.Nilsson, M.Ejnarsson, Speech Recognition using Hidden Markov Model, MEE-01-27

[PAC 97]: R. A. Packwood, M.K.Steliaros and G.R.Martin, Variable Size of Block Matching Motion Compensation for Object-Based Video Coding, IPA97, 15-17 July 1997 Conference Publication N° 443 P56.

[PAC 03] H. Packard, P. J. Moreno et S. Agarwal, "An Experimental Study of EM-Based Algorithms for Semi-Supervised Learning in Audio Classification" , 2003.

[PIC 90] J.Picone , IEEE Automatic Speech and Signal Processing Magazine,pp. 26-41,1990.

[POP 09]Ronald Poppe,"A survey on vision-based human action recognition",Image and Vision Computing ScienceDirect,2009 .

[PLA 88]: Plamondon R., Parizeau M., Signature verification from position, velocity and acceleration signals : A comparative Study, 9th International Conference on Pattern Recognition (ICPR'88), Rome (Italie), Vol. I, 1988, p260-265.

[POR 05] : F. Porikli, Integral Histogram : A Fast Way to Extract Histograms in Cartesian Spaces IEEE International Conference on Computer Vision and Pattern Recogni-tion (CVPR), Vol. 1, pp. 829-836, June 2005

[RAB 89] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol. 77, pp. 257–285, 1989.

[RAM 03] Ramamoorthy A. et al., "Recognition of dynamic hand gestures", Pattern Recognition, 2003.

[RAM 06] E. Ramasso, M. Rombaut, and D. Pellerin, "A Temporal Belief Filter improving human action recognition in videos," in IEEE Int. Conf. on Acoustics, Speech and Signal Processing, vol. 2, 2006, pp. 141–143.

[RED84]:. R. A. Redner, H. F. Walker, "Mixture densities, maximum vraisemblance and the EM algorithm ", SIAM Review, 26, pp195-239, 1983.

[ROS 10]: ROSTEN, E., PORTER, R., AND DRUMMOND, T. 2010. Faster and better: A machine learning approach to corner detection. IEEE Trans. Pattern Analysis and Machine Intelligence, 105–119.

[SAK 71] : H. Sakoe et S. Chiba, " A dynamic programming approach to continuous speech recognition" , Proceedings of the 7th International Conference on Acoustics, article 20C-13, 6 pp", 1971.

[SAN 95]: P. Santago and H. D. Gage, " Statistical models of partial volume effect ", IEEE Trans. Image Proc., Vol. 4, No. 11, pp1531-1540, 1995.

[SAN 98]: S. Sanjay-Gopal and T. J. Hebert, "Bayesian Pixel Classification Using Spatially Variant Finite Mixtures and the Generalized EM algorithm ", IEEE trans. Image Processing, Vol. 7, No.7, pp1014-1028, 1998.

[SHU 96] J. Schürmann. Pattern classification. A unified view of statistical and neural approaches.John Wiley et Sons, 1996.

[SCH 04] C.Schuldt, I.Laptev,B.Caputo,"Recognizing human actions:a local SVM approach", ICPR,vol.3,2004,pp.32–36.

[SHI 94] :J. Shi and C. Tomasi. Good features to track. IEEE Conference on Computer Vision and Pattern Recognition, pages 593-600, 1993.

[SHW 04] : S. Shalev-Shwartz, J. Keshet et Y. Singer, " Learning to Align Polyphonic Music", 2003.

[SIG 02] : Projet sigma2 Signaux, "Modèles et Algorithmes" , Rennes 2002.

[SMI 97] :S M Smith and M Brady, "SUSAN -- a new approach to low level image processing", International Journal of Computer Vision, vol. 23(1), 45-78, 1997.

[STA 95] Starner T., Pentland A., "Visual Recognition of American Sign Languageusing hidden Markov models", Dans: International Workshop on AutomaticFace and Gesture Recognition, Zurich, 1995, pp.189-193.

[THU 08]: C. Thurau and V. Hlavac. Pose primitive based human action recognition in videos or still images. CVPR, pp. 1-8, 2008.

[TOK 99] : K. Tokuda, T. Masuko and T. Kobayashi, " Hidden Markov Models based on multi-space probability distribution for pitch pattern modelling", In IEEE International Conference On acoustics, Speech and Signal Processing, Vol. 1, pp. 229–232, March, 1999.

[TUZ 08] Tuzel, F. Porikli, P. Meer, "Pedestrian Detection via Classification on Riemannian Manifolds", IEEE Trans-actions on Pattern Analysis and Machine Intelligence, Vol. 30, pp. 1713-1727, 2008

[VAL 91] : R. Vallet, « Application de l'identification des chaînes de Markov cachées aux communications numériques », Octobre 1991.

[VAP 64] :Vapnik V., Chevonenkis A., "A note on one class of perceptrons ", Automatic and Remote Control, vol. 25, 1964

[VAP 95] : V. Vapnik. The Nature of Statistical Learning Theory. Springer, N.Y., 1995.

[VAP 91] : V. N. Vapnik and A. Y. Chevonenkis. The necessary and sufficient conditions for consistency in the empirical risk minimisation method Pattern Recognition and Image Analysis, 1(3) : 283-305, (1991).

[VOG 98] : C. Vogler., D.Metaxas, "ASL recognition based on a coupling between HMMs and 3D motion analysis", 1996.

[WAN 95] : H Wang and M Brady, "Real-time corner detection algorithm for motion estimation", Image and Vision Computing, vol. 13:9, pp. 695-703, 1995.

[WAN 12]: Xiaogang Wang, Intelligent multi-camera video surveillance: A review , Elsevier, 2012.

[WHE 03] : Wheeler K.R. "Device control using gesture sensed form EMG",IEEE 2003.

[WID 90] : B. Widrow and M. A. Lehr; 30 years of adaptive neural networks: Perceptron, madaline and back-propagation, In Processin of the Institute of Electrical and Electronic Engineers, volume 78, pages 1415-1442, 1990.

[WIR 05] : Wirotius M., Ramel JY, Vincent N., Distance and Matching for Authentication by On-Line Signatures. IEEE Workshop on Automatic Identification Advanced Technologies. 17-18 October 2005, Buffalo, New York, USA. p230-235.

[ZHA 01] Zhang, S. Z. Li, M .Chengyuan, S. Heung-Yeung, and E. Chang, "Learning to
Boost GMM Based Speaker Verification", 2001.

[ZHA 11]: Hong Zhang, Lu Li, Wenyan Jia, John D. Fernstrom, Robert J. Sclabassi, Zhi-Hong Mao, Mingui Sun, "Physical activity recognition based on motion in images acquired by a wearable camera", Elsevier, 2011.

Conclusion générale

Le problème traité dans ce mémoire concerne la reconnaissance des actions de l'humain. Nous avons d'abord étudié les méthodes proposées dans la littérature pour la reconnaissance d'actions et principalement, nous avons accès sur les principales représentations utilisées par la communauté dans cet axe de recherche. Parmi ces représentations, nous avons initialement ciblé les techniques du le flot optique et la représentation MHI (Motion History Image) dont le but de localiser sur la séquence d'images les vecteurs de mouvement et de les exploiter pour la classification des actions.

En se basant sur ces deux caractéristiques, nous avons proposé une méthode de classification d'actions basée sur le regroupement des mouvements de même direction et leur évolution dans le temps.

Les tests opérés ont montré la limite de cette méthode en raison de la difficulté d'obtenir une discrimination des directions pour les différentes actions. Ceci nous a amené à utiliser le flot optique mais avec une pyramide gaussienne proposée par L. Kanade. La représentation des mouvements obtenus n'a pas été aussi discriminante pour différencier les actions.

La troisième proposition concerne l'intégration du descripteur SURF pour la description des actions. La méthode proposée qui consiste à localiser les zones d'intérêt en mouvement et d'utiliser leurs descripteurs SURF afin de caractériser les actions. La méthode a été implémentée et testée sur la base de tests de Weizmann. Les résultats obtenus ont été comparés à l'état de l'art et jugés acceptables.

Le problème de reconnaissance d'actions reste ouvert et le problème de caractérisation des actions conditionne les résultats de classification.

Le travail proposé reste à améliorer en intégrant d'autres descripteurs au descripteur SURF. Le regroupement des directions pourra être aussi amélioré en utilisant la connaissance a priori de la silhouette de l'humain.

Annexe 1 : Outils de classification

Alignement temporel dynamique

L'alignement temporel Dynamique, plus connu sous l'acronyme de DTW, *Dynamic Time Warping " est une application dans le domaine de la reconnaissance des formes [SAK 71]. Cet alignement est réalisé par une technique de programmation dynamique [BIM 88]"*, fondée sur un principe de comparaison de la forme à analyser avec un ensemble de formes stockées dans une base de référence (voir figure 1).

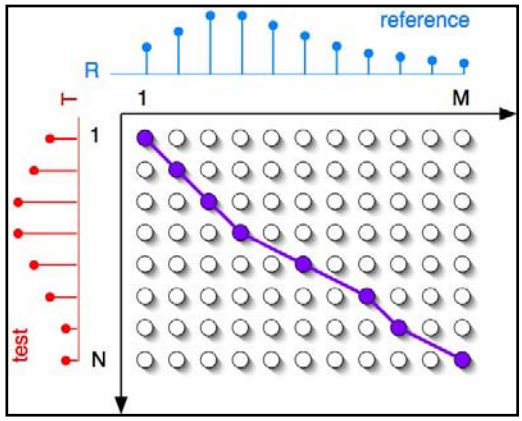


Figure 1 : Exemple d'un Alignement temporel dynamique (DTW) [HUA 01].

La figure 1 illustre le principe de l'alignement temporel dynamique qui consiste à calculer la distance (exemple : distance euclidienne, distance Manhattan, etc.) entre le vecteur de l'échantillon à tester et l'ensemble des vecteurs de références.

- N : nombre de vecteurs dans la séquence de test,
- M : nombre de vecteurs dans la séquence de référence.

Il suffit alors de déterminer le « meilleur » chemin par récurrence.

Le DTW peut ainsi être vue comme un problème de cheminement dans un graphe (voir figure 2).Il permet de comparer deux formes représentées par une suite ordonnée de points. Il attribue à chaque élément d'une courbe, le meilleur élément correspondant dans l'autre courbe relativement à une certaine métrique [PLA 88].

Ainsi, des points de chacune des deux courbes sont mis en correspondance avec une légère tolérance aux différences d'échantillonnage et de longueur. Elle détermine le meilleur chemin reliant le début et la fin des deux blocs de paramètres [BRI 95, KHA02].

La plupart du temps, la métrique utilisée est la distance spatiale euclidienne entre les points. Une fois la mise en correspondance effectuée, nous pouvons calculer la distance entre les deux courbes en ajoutant la somme des distances entre les couples de points correspondants [RAM 06].

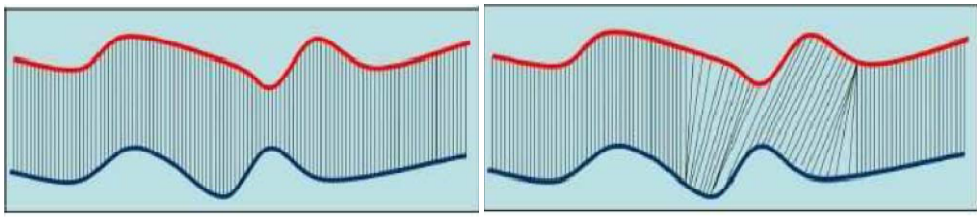


Figure 2: Illustration d'une mise en correspondance classique (a) et DTW (b)[WIR 05]

Les modèles de mixture de gaussiennes (GMM Gaussian Mixture Models)

Le modèle de mixture des gaussiennes constitue une des techniques de classification. L'approche de maximum vraisemblance (ML) est toujours utilisée [RED84] pour ajuster le modèle de mixture de gaussiennes 'GMM' le plus adapté aux données que nous désirons modéliser [CHO91] [SAN95] [SANJ98].

Pour traiter les formes, nous utilisons une hypothèse classique celle du modèle de mixture ou nous supposons que chaque classe, chaque région de la forme suit une distribution particulière. La distribution de probabilité associée à l'image est alors considérée comme étant un mélange de densités de probabilités.

Les modèles à mélanges gaussiens sont populaires comme modèle génératif dans beaucoup de domaines [PAC 03]. Ce modèle est caractérisé par un vecteur moyen, une matrice de covariance et le poids de chaque gaussienne [ZHA 01].
Ces paramètres sont déduits en utilisant l'algorithme de maximisation d'espérance (EM) pour former les GMMs. C'est-à-dire nous avons des exemples pour chacune des classes que nous voulons séparer.

Nous apprenons grâce à EM un modèle GMM pour chaque classe. Pour un exemple inconnu, nous calculons la vraisemblance pour chacun des modèles et nous décidons de celui qui ressemble le plus aux données de test en choisissant le modèle ayant la plus grande vraisemblance [SHW 04].

Modèles de Markov Cachés

Les Modèles de Markov Cachés (HMMS) sont un type de modèle statistique largement utilisés depuis quelques années dans la reconnaissance de la parole et récemment dans la reconnaissance de l'écrit, des gestes mais encore dans la reconnaissance des actions humaines.

Un modèle de Markov permet de représenter un système pouvant se trouver à un instant t dans un état pris parmi un ensemble de N états possibles. L'état dans lequel se trouvera le système à l'instant t+1 dépend uniquement de son état à l'instant t, le passage d'un état à un autre étant un processus stochastique [CAM 07]. Un tel modèle est défini par ses états, et ses probabilités de transitions, la figure ci-dessous donne l'exemple d'un modèle de Markov à trois états.

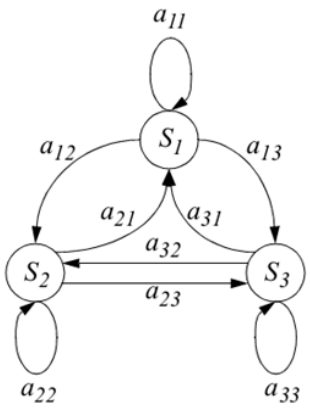


Figure 3: Modèle (ou chaîne) de Markov à 3 états

En outre, un HMMS est caractérisé par les éléments suivants:

• **N** : est le nombre d'états cachés du modèle, nous notons alors:

L'ensemble des états cachés **S= {s_1, s_2, …, s_N}**.

A l'instant t, un état est représenté par **q_t(q_t∈S)**.

• **M** : est le nombre de symboles distincts que nous pouvons observer dans chaque état.

Nous représentons par désagrément l'ensemble **V={v_1, v_2, …, v_M}**.

A l'instant t, un symbole observable est désigné par **O (O_t∈V)**.

•**Une matrice de probabilité de transition** : prenons **A=[a_{ij}]**, où **a_{ij}**est considérée comme étant la probabilité de transition de l'état i vers l'état j.

Dans le cadre d'un HMMS stationnaire du premier ordre, cette probabilité ne dépend pas de t, nous définissons **a_{ij}=P(q_{t+1}=s_j | q_t=s_i), 1 ≤ i, j ≤ N**.

• **Une matrice de distribution des probabilités** : prenons **B=[$b_j(k)$]**, associée à chaque état où **$b_j(k)$** est la probabilité d'observer le symbole vk en étant à l'état sj à l'instant t, nous définissons **$b_j(k)=P(o_t=v_k \mid q_t=s_j)$, $1\leq i \leq N$, $1 \leq k \leq M$**.

• **Un vecteur$\prod = [\pi i]$:** vecteur de distribution des probabilités de transitions initiales, où p i est la probabilité de commencer dans l'état i, nous définissons **$\pi i=P[\, q1=si \,]$** avec **$1\leq i \leq N$**

De ce qui précède, nous concluons que pour spécifier un HMMS, il faut spécifier les paramètres que nous avons présentés.

Dans un modèle de Markov classique, à chaque instant *t*, l'état du processus correspond à une observation unique.

Et lorsque le processus est dans un état donné S, y a possibilité de produire différentes observations selon différentes probabilités appelées probabilités d'émission.

Quand les états ne sont pas accessibles directement à l'observation, ces derniers sont appelés modèles cachés, d'où la nécessité de préciser les valeurs des probabilités d'émission des différents états.

– Les HMMs sont des modèles stochastiques qui prennent en considération la variabilité temporelle d'un phénomène. Leur utilisation était considérable dans la reconnaissance de la parole, permettant ainsi d'obtenir de très bonnes performances [RAB 89][PIC 90] qui ont conduit la communauté scientifique à les utiliser massivement pour la reconnaissance de signes dynamiques [LOE 04]. Chaque signe est modélisé par un HMM [STA 95][BRA 97][LEE 96][VOL 98][HIE 99][MA 00], cette méthode est utilisée pour reconnaitre les signes dans des conditions assez difficiles tels les systèmes de vision dans un milieu naturel [CHE 03][RAM 03] ou encore des électromyogrammes [WHE 03].

– Les trois problèmes fondamentaux des HMMs :Pour qu'un Modèle de Markov Caché fonctionne correctement, il faudra résoudre les trois problèmes de base, dès lors, ce dernier pourra être utilisé dans des conditions réelles.

Nous présentons ces problèmes comme suit :

1. Pour une séquence d'observations donnée $O=O_1, O_2,\ldots O_t$et un modèle λ il faut calculer la probabilité $P(O \mid \lambda)$ en définissant la variable $\alpha_1(i)$.

$\alpha_1(i) = P(O_1, O_2, \ldots O_t, Q_t = S_i \mid \lambda)$ $1 \leq i \leq N$ (1)

Pour trouver la probabilité d'une séquence d'observation $O = \{O_1, O_2, O_3. \ldots O_T\}$ et un modèle donné $\lambda = (\pi, A, B)$, nous utiliserons la probabilité d'évaluation [Rab89]. Dont le but est de trouver quels sont les modèles (en supposant qu'ils existent) qui ont probablement produit la séquence d'observation.
La méthode est d'évaluer chaque séquence possible des états de longueur T, puis ajoutez les ensembles.

$$P(O \mid \lambda) = \sum_{q_1, q_2, \ldots, q_T} \pi_{q_1} \prod_{t=2}^{T} a_{q_{t-1} q_t} b_{q_t}(o_t) \quad (2)$$

L'interprétation de cette équation est donnée dans [NIL 02][RAB 89], comme suit :

Initialement, à l'instant t=1, nous sommes à l'état q1 avec la probabilité π_{q_1}, nous générons le symbole o_1 avec une probabilité $b_{q_1}(o_1)$ en passant de l'instant t à t + 1, une transition de q1 à q2 se fera avec la probabilité $a_{q_1 q_2}$, ou le symbole o_2 doit être généré avec la probabilité $b_{q_2}(o_2)$, et le processus se poursuit jusqu'à ce que la transition à l'instant T soit exécutée.
Dans le cas ou le calcul devient très important, il devient impératif de le réduire en utilisant l'algorithme Forward, définit par :

$$\alpha_t(i) = P(o_1 o_2 \ldots o_t, q_t = i \mid \lambda) \quad (3)$$

La définition de $\alpha_t(i)$ est que la probabilité au temps t et dans l'état i, génère la séquence d'observation partielle de la première observation jusqu'à N observation à l'instant t, $O_1, O_2, \ldots O_t$.

$\alpha_{t+1}(i)$ peut être calculée en faisant la somme de la variable pour tous les états N à l'instant t multipliée par la probabilité de transition d'état correspondant et par la probabilité d'émission $b_{qt}(O_{t+1})$ et la procédure de calcul est la suivante :

- Initialisation :

  $\alpha_1(i) = \pi_i b_i(O_1)$ (4)

- Induction :

$\alpha_{t+1}(i) = b_i(O_{t+1}) \sum \alpha_t(j)\, \alpha_{i+j}$ $1 \leq t \leq T-1$ (5)

- Terminaison :

$P(O \mid \lambda) = \sum \alpha_T(i)$ (6)

2. Pour une suite d'observation O et un modèle λ il est important de trouver la suite d'états S1, S2,…ST qui génère O.

Il existe une probabilité discrète d'observation que nous pouvons utiliser. Cette alternative est moins compliquée dans les calculs, mais elle utilise une quantification vectorielle qui génère une erreur de quantification

La répartition couramment utilisée pour décrire les densités d'observation est celle de Gauss. Et pour représenter la fonction de densité de la probabilité d'observation continue, nous utilisons la matrice B, la moyenne μ et la variance Σ.

$$b_j(o_t) = \sum_{k=1}^{M} c_{jk} b_{jk}(o_t) \qquad , j = 1,2,..., N \tag{7}$$

$$\sum_{k=1}^{M} c_{jk} = 1 \qquad , j = 1,2,..., N \tag{8}$$

$$c_{jk} \geq 0 \qquad , j = 1,2,..., N, k = 1,2,..., M \tag{9}$$

Avec l'utilisation de matrices de covariance diagonales, et en raison de l'implémentation rapide et un calcul moins rapide, nous utilisons la formule suivante :

$$b_{jk}(o_t) = \frac{1}{(2\pi)^{D/2} (\prod_{l=1}^{D} \sigma_{jkl})^{1/2}} e^{-\sum_{l=1}^{D} \frac{(o_{tl} - \mu_{jkl})^2}{2\sigma^2_{jkl}}} \tag{10}$$

Sila récursiondécritepour calculerla variableforwardestfaite dans lesens inverse, nous obtiendrons lavariablebackward, qui est la suivante:

$$\beta_t(i) = P(o_{t+1} o_{t+2} ... o_T \mid q_t = i, \lambda) \tag{11}$$

3. Ajuster les paramètres d'un HMMS λ de telle façon à permettre de maximiser P(O | λ) pour une suite d'observation O, quoi qu'il n y ait pas une solution analytique qui pourrait prendre en charge cette maximisation, c'est pourquoi et pour y remédier, nous utilisons l'algorithme de Baum-Welch.

Le processus de ré-estimation, dans le cas de densités de probabilité continues, se déroule comme suit :

$b_j(O) = \sum c_{jm} G(O, \mu_{jm}, U_{jm})$ (12)

m : le nombre de gaussiennes

j : le numéro de l'état

c : le poids de la gaussienne m dans l'état j

G : une gaussienne de moyenne μet de matrice de covariance U

Par la variable de backwardβ:

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, \ldots\ldots O_T \mid Q_t = S_i, \lambda) \quad (13)$$

$$\beta_T(i) = 1 \quad (14)$$

$$\beta_t(i) = \sum a_{ij} b_{ij}(O_{t+1})\, \beta_{t+1}(j) \qquad 1 \le i \le N,\ 1 \le t \le T\text{-}1 \quad (15)$$

Ainsi que par ζ et γ tels que :

$$\zeta_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \hat{\beta}_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}) \hat{\beta}_{t+1}(j)} \quad (16)$$

$$\gamma_t(i) = \sum_{t=1}^{T-1} \zeta_t(i,j) \quad (17)$$

$\sum \zeta_t(i,j)$ peut être considéré comme étant le nombre attendu de transitions suivies de S_i vers S_j

De la même façon, $\sum \zeta_t(i,j)$ correspond au nombre de transitions suivies depuis S_i.

Dès lors, les formules de ré-estimation pour les transitions et les probabilités d'émission s'écrivent :

$$\pi_i = \gamma_t(i) \quad (18)$$

$$a_{ij}(i) = \frac{\sum_{t=1}^{T-1} \zeta_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (19)$$

$$\mu_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) o_t}{\sum_{t=1}^{T} \gamma_t(j,k)} \quad (20)$$

- Si nous répétons cette procédure, nous convergeons vers une probabilité maximum, en général au bout de 10 à 15 itérations.

Le but de la phase de reconnaissance est de classer chaque action dans l'une des dix classes. Les traitements nécessaires sont relativement similaires à ceux utilisés dans l'étape d'apprentissage.

Une fois notre modèle construit, il est utilisable pour la reconnaissance de mouvements. La reconnaissance se fait en construisant un nouvel HMM au fur et à mesure de la lecture des données et en le comparant avec les HMM construits

lors de l'apprentissage via l'algorithme de Viterbi. L'idée principale est de supposer qu'un mouvement et plus largement une série de caractéristiques est un HMM et d'utiliser les données d'apprentissage pour construire une chaine de Markov dans le but de reconnaitre le mouvement requête où la probabilité d'une séquence de test est calculée par rapport à un ensemble d'apprentissage des modèles HMM où le HMM à maximum de vraisemblance (pour chaque opération nous avons le numéro de l'activité déduit par le classifieur et le maximum de vraisemblance adéquate)représente la séquence de test. Plus cette valeur est élevée pour un élément des données d'entrainement, plus cet élément est proche du mouvement suivit actuellement.

Les machines à vecteur de support (SVM)

Pendant les années allant de 1986 à 1995, la théorie de l'apprentissage n'a pas fait de progrès notable, et ce malgré l'explosion du nombre des modèles connexionnistes, leurs variantes et les nombreuses applications du monde réel résolues. Et ce n'est que depuis 1995 que nous constatons un regain d'intérêt pour les approches théoriques et les ébauches de nouveaux résultats qui permettraient de mieux comprendre les processus d'apprentissage et appréhender les capacités de généralisation des modèles. Parmi les méthodes à noyaux, inspirées de la théorie statistique de l'apprentissage de Vladimir Vapnik, les Support Vector Machines (SVMs) constituent la forme la plus connue [VAP 91]. Ils sont des méthodes de classification binaire par apprentissage supervisé, qui furent introduits en 1995[VAP 95], bien que les premières publications sur le sujet datent des années 60 [VAP 64]. La plupart des systèmes de classification sont élaborés afin de résoudre des problèmes binaires. Parmi les modèles des SVMs, nous citons :

- ***Cas des données linéairement séparables***

Ce sont les plus simples des SVM car ils permettent de trouver facilement le classificateur linéaire. Dans le schéma qui suit, nous déterminons un hyperplan qui sépare les deux ensembles de points. Les points les plus proches sont utilisés pour la détermination de l'hyperplan (L'hyperplan optimal (1) avec la marge maximale et les échantillons entourés près des lignes pointillées des vecteurs supports (2).), sont appelés vecteurs de support (voir figure 4).

- ***Cas des données non linéairement séparables***

Dans la plupart des problèmes réels, il n'y a pas de séparation linéaire possible entre les données, le classificateur de marge maximale ne peut être utilisé car il fonctionne seulement si les classes de données d'apprentissage sont linéairement séparables.Si par exemple les données des deux classes se chevauchent sévèrement comme dans la figure5, aucun hyperplan séparateur ne sera satisfaisant (Classifieur non linéaire(3)). Et pour surmonter les inconvénients engendrés, dans ce cas, l'idée des SVM est de changer l'espace des données. La transformation non linéaire de ces données peut permettre une séparation linéaire des exemples dans un nouvel espace. Nous allons donc avoir un changement de dimension ; appelé « espace de re-description ».

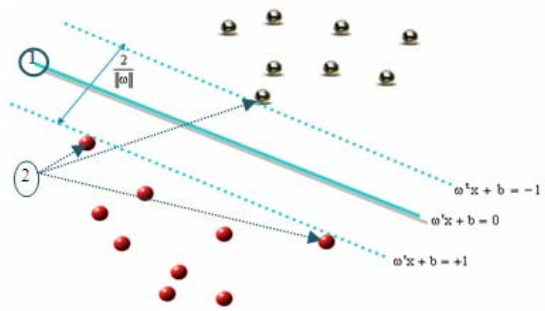


Figure 4 : Cas linéairement séparable

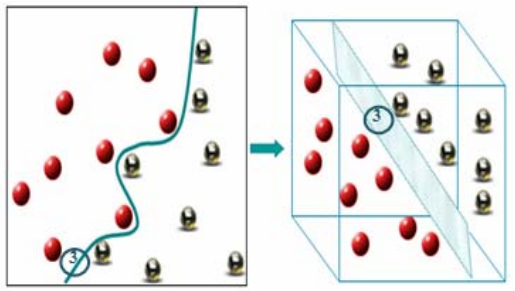


Figure 5 : Cas non linéairement séparable

Les Réseaux de neurones

Les réseaux de neurones [WID90] s'avèrent capables de traiter des problèmes complexes de reconnaissance de forme, ou de simulation de processus non linéaires et/ou dynamiques. Ils trouvent donc de nombreuses applications en ingénierie [GAR98], en physique appliquée et en environnement [CES98] où le nombre et la dimension des données ainsi que la non-linéarité des problèmes traités posent de nombreuses difficultés quant à l'utilisation des outils statistiques classiques.

Les réseaux de neurones ont la capacité de stocker des connaissances expérimentales acquises à partir d'un environnement formel. Les réseaux de neurones artificiels sont des réseaux fortement connectés de processeurs élémentaires fonctionnant en parallèle. Chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit. Toute structure hiérarchique de réseaux est évidemment un réseau.

Ils sont capables d'assurer de nombreuses tâches. Ils sont souvent utilisés pour leur capacité de classifieur et ils sont classiquement employés dans des problèmes d'approximation pour simuler des fonctions de transfert non-linéaires et multidimensionnelles et pour résoudre des problèmes d'inversion [CHA00].
Le premier modèle du neurone formel a été introduit dans les années quarante par Mac Culloch et Pitts, qui en s'inspirant de leurs travaux sur les modèles biologiques ; ils ont proposé le modèle suivant [JOD 94, DAV 93].

"*Un neurone formel fait une somme pondérée des potentiels d'actions qui lui proviennent des autres neurones, puis s'active suivant la valeur de cette sommation pondérée. Si cette sommation dépasse un certain seuil, le neurone est activé et transmet une réponse dont la valeur est celle de son activation. Si le neurone n'est pas activé, il ne transmet rien. "*

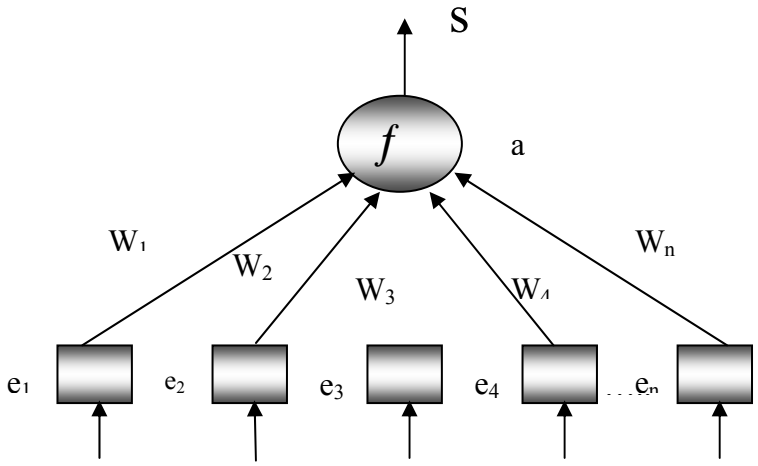


Figure 6 : Neurone formel

Le neurone formel est un processeur élémentaire (fonction algébrique), qui réalise une somme pondérée de ses entrées par des coefficients de connexion appelés poids synaptiques. A partir de cette valeur, une fonction d'activation (une fonction de transfert) détermine l'état de sortie du neurone qui va par la suite exciter les autres neurones qui lui sont connectés (voir figure 6)
Notons que :

$(ei)_{i=1..n}$: les entrées du neurone formel ;

Wi: les paramètres de pondération ;

a :la somme pondérée des entrées ; elle est calculée de la façon suivante: $a = \sum (w_i \circ e_i)$

f : la fonction d'activation.

S=f(a) : la sortie du neurone.

Référence

[BIM 88] : F. Bimbot, " Synthèse de la parole : du segments au règles, avec utilisation de la décomposition temporelle ", Thèse 1988.

[BRA 97] :M.Brand, N.Oliver, and A. Pentland,: "Coupled hidden Markov models for complex action recognition". Proc. Int. Conf. on Computer Vision and Pattern Recognition, 1997.

[BRI 95] : J. S. Bridle, " Optimization and search in speech and language processing ", Survey of the state of the art in human language technology, 1995.

[CAM 07] : N. Camelin, "Stratégies robustes de compréhension de la parole basées sur des méthodes de classification automatique ", thèse, 2007.

[CES 98] : CESAC, UMR C5576, CNRS, Internationnal workshop on applications of artificial neural networks to ecologicalmodelling, Toulouse, France, 14-17 décembre 1998.

[CHA00] : A. Chardin, "Modèles énergétiques hièrarchiques pour la résolution des problèmes inverses en analyse", 2000.

[CHE 03]: F.-S .Chen., C.-M .Fu., C.-L.Huang, "Hand gesture recognition using a realtime tracking method and hidden Markov models", Image and Vision Computing, 2003.

[CHO91]: H.S.Choi, D.R.Haynor and Kam, "Partial volume tissue classification of multichannel magnetic resonance images Amixel model ", IEEE Trans. Med. Image., Vol.10, pp395-407, 1991.

[DAV 93] : E.Davalop and P.Naim, "Les réseaux de neurones", Edition Masson 1993.

[GAR98] : S. Garcia-Sallicetti, B. Dorizzi, and P. Gallinari, "A neural predictive system for online handwriting recognition", Pattern Recog., 1998.

[HIE 99] :Hienz H., Bauer B., Kraiss K.F., "HMM-based continous signlanguage recognition using stochastic grammars", 1999.

[HUA 01] : X. Huang, A. Acero, and H.-W. Hon. "Spoken Language Processing: A Guide to Theory, Algorithm and System Development", Prentice-Hall, Englewood Cliffs, N.J., 2001.

[JOD 94] : J-F.Jodouin, "Les Réseaux de neurones : Principes et définitions. Les Réseaux Neuromimétiques : Modèles et applications.Editions" Hermès, Paris.1994.

[KHA 02] : J. Kharroubi, " Etude de Techniques de Classement Machines à Vecteurs Supports pour la Vérification Automatique du Locuteur" , thèse , 2002.

[LEE 96] :C.Lee, Y.Xu, "Online, interactive learning of gestures for human/robotinterfaces", IEEE 1996.

[LOE 04] : B.L. Loeding, S.Darkar, A.Parashar, A.Karshmer, "Progress in automated comp.recog.of Sign Language", Springer-Verlag, 2004.

[MA 00] :J.Ma, W.Gao, J.Wu, C.Wang, "A continuous Chinese Sign Language recognition system", 2000.

[MUR 05] : Hidden Markov Model (HMM) Toolbox for Matlab Written by Kevin Murphy, 1998,Last updated: 8 June 2005. http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html

[NIL 02] :M.Nilsson, M.Ejnarsson, Speech Recognition using Hidden Markov Model, MEE-01-27

[PAC 03] H. Packard, P. J. Moreno & S. Agarwal, "An Experimental Study of EM-Based Algorithms for Semi-Supervised Learning in Audio Classification" , 2003.

[PIC 90] J.Picone , IEEE Automatic Speech and Signal Processing Magazine,pp. 26-41,1990.

[PLA 88]: Plamondon R., Parizeau M., Signature verification from position, velocity and acceleration signals : A comparative Study, 9th Intern.Conf. on Pat.Rec. (ICPR'88), Rome (Italie), vol.I, p260-265, 1988.

[RAB 89] : L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol. 77, pp. 257–285, 1989.

[RAM 03] :A.Ramamoorthy et al., "Recognition of dynamic hand gestures", Pattern Recognition, 2003.

[RAM 06] E. Ramasso, M. Rombaut, & D. Pellerin, "A Temporal Belief Filter

improving human action recognition in videos," in IEEE Int. Conf. on Acoustics, Speech and Signal Processing, vol. 2, 2006, pp. 141–144.

[RED84] : R. A. Redner, H. F. Walker, "Mixture densities, maximum vraisemblance and the EM algorithm ", SIAM Review, 26, p.195-239, 1984.

[SAK 71] : H. Sakoe et S. Chiba, " A dynamic programming approach to continuous speech recognition" , Proceedings of the 7th International Conference on Acoustics, article 20C-13, 6 pp", 1971.

[SAN95] : P. Santago and H. D. Gage, " Statistical models of partial volume effect ", IEEE Trans. Image Proc., Vol. 4, No. 11, pp1531-1540, 1995.

[SAN98] : S. Sanjay-Gopal and T. J. Hebert, "Bayesian Pixel Classification Using Spatially Variant Finite Mixtures and the Generalized EM algorithm ", IEEE trans. Image Processing, vol. 7, No.7, pp1014-1028, 1998.

[SHW 04]: S. Shalev-Shwartz, J. Keshet& Y. Singer, " Learning to Align Polyphonic Music", 2004.

[STA 95]: Starner T., Pentland A., "Visual Recognition of American Sign Languageusing hidden Markov models", International Workshop on AutomaticFace and Gesture Recognition, Zurich, 1995, pp.189-194.

[VAP 64] :Vapnik V., Chevonenkis A., "A note on one class of perceptrons ", Automatic and Remote Control, vol. 25, 1964

[VAP 95] : V. Vapnik. The Nature of Statistical Learning Theory.Springer, N.Y., 1995.

[VAP 91] : V. N. Vapnik and A. Y. Chevonenkis. The necessary and sufficient conditions for consistency in the empirical risk minimisation method Pattern Recognition and Image Analysis, 1(3) : 283-305, (1991).

[ZHA 01] Zhang, S. Z. Li, M .Chengyuan, S. Heung-Yeung, and E. Chang, "Learning to Boost GMM Based Speaker Verification", 2001.

[WHE 03] : Wheeler K.R. "Device control using gesture sensed form EMG",IEEE 2003.

[WID90] : B. Widrow and M. A. Lehr; 30 years of adaptive neural networks: Perceptron, madaline and back-propagation, In Proc. of the Institute of Electrical and Electronic Engineers, vol. 78, p. 1415-1442, 1990.

[WIR 05] :Wirotius M., Ramel JY, Vincent N., Distance and Matching for Authentication by On-Line Signatures. IEEE Workshop on Automatic Identification Advanced Technologies. 17-18 October 2005, Buffalo, New York, USA.p230-235.

Annexe : Détecteurs de Points d'intérêt dans l'image

La figure 1 ci-dessous propose une classification chronologique des différents détecteurs [MAN 11].

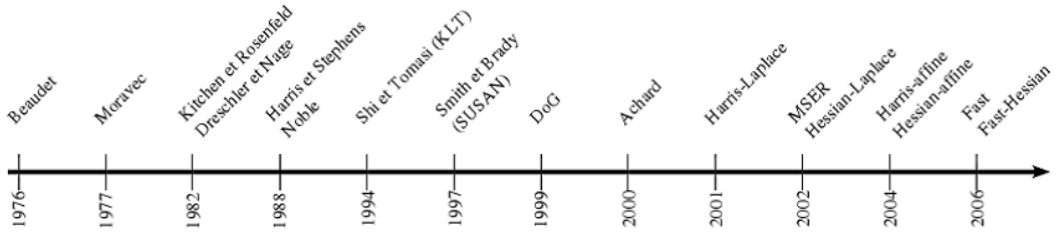


Figure 1 : Classification chronologique des détecteurs étudiés [MAN 11]

Ils sont regroupés en trois (03) groupes, que nous citons :

1. **Invariances aux transformations euclidiennes**

Dès 1976, de nombreuses méthodes de détection de points d'intérêt ont vu le jour. Certaines d'entre elles se basent sur l'utilisation d'une matrice Hessienne, d'autres s'appuient sur l'analyse du changement d'intensité local. Une description de ces différentes méthodes est donc nécessaire afin de détailler la construction et définir les relations qui existent entre elles.

- **Détecteurs : Beaudet, Dreschler-Nagel, Kitchen-Rosenfeld**

Ce sont des détecteurs qui se basent sur l'utilisation des dérivées secondes partielles, nous citons les détecteurs de Beaudet, de Kitchen-Rosenfeld ou encore de Dreschler-Nagel [BEA78][KIT 82][DRE 82]. Ces derniers s'appuient sur la matrice Hessienne.

- Beaudet dans [BEA 78] a proposé un opérateur invariant par rotation. Il provient du développement en série de Taylor de l'image, qui est considérée comme une surface d'élévation des intensités du niveau de gris. Cet opérateur correspond au déterminant de la matrice Hessienne. La détection est basée sur le seuillage des valeurs absolues des extrema de l'opérateur. L'angle n'est pas localisé précisément car il y a un déplacement engendré par la déviation standard du processus de filtrage. Par contre, cette approche est plus stable que celle de Rosenfeld que nous présenterons plus loin.

- Kitchen-Rosenfeld proposent dans [KIT 82] d'utiliser un opérateur appelé "cornerness", il n'est autre que le produit de la courbure par la norme du gradient, qui est une représentation explicite de la dérivée seconde directionnelle dans la direction orthogonale au gradient. Les maxima locaux de cet opérateur représentent les angles dans une image, qui détectés par cet opérateur sont très mal localisés notamment pour les coins dont l'angle est inférieur à 45°.
- La méthode, exposée dans [DRE 82], s'inspire fortement de celle de Beaudet et elle s'applique en calculant la courbure Gaussienne, et en sélectionnant les extrema de cette dernière (courbures positives et négatives (elliptique E et hyperbolique H)), c'est-à-dire, procéder par l'appariement de chaque max positif E (point elliptique) avec le point H (maximum hyperbolique), et enfin sélectionner le point T pour lequel la courbure principale passe par 0 entre E et H(Le point T ainsi trouvé ne correspond pas à la position exacte de l'angle, car cette position évolue sur la bissectrice de l'angle dans l'échelle spatiale). La précision de localisation des angles obtenue par cette méthode subit la même délocalisation que celle obtenue par l'approche de Beaudet.

- **Détecteurs : Harris, Noble, KLT, Achard**

Approche basée sur l'utilisation des dérivées premières partielles, elle permet d'observer les changements locaux de l'intensité utilisée (détecteurs de Harris [HAR 88], Noble [NOB 89], Shi et Tomasi [SHI 94], et Achard [ACH 00]). Le détecteur le plus courant, utilisé est sans doute le détecteur de Harris [HAR 88], proposé en 1988. Il est basé sur les valeurs propres de la matrice du second moment.

- **Détecteurs : Moravec**

L'un des premiers détecteurs de points d'intérêt à avoir été développé est le détecteur de Moravec [MOR 01]. Son principe est de rechercher la variation d'intensité lumineuse dans un grand nombre de directions. Pour vérifier qu'un point est un point caractéristique, nous utilisons un voisinage rectangulaire autour de ce point (une fenêtre 3x3, 5x5 ou 7x7, par exemple).

Cette fenêtre est décalée dans l'une des 8 directions parmi les directions horizontales, verticales et diagonales. Il existe de nombreux algorithmes différents de détection de coin qui permettent l'analyse du voisinage et/ou d'optimiser les temps de calculs : Wang et Brady [WAN 95], SUSAN (SmallestUnivalue Segment AssimilatingNucleus) [SMI 97], Trajkovic et Hedley, FAST (FeaturesfromAccelerated Segment Test)[ROS 10].

2. Invariances aux similitudes

Proposés en 1984 par Koenderink [KOE 84], ils permettent notamment, de gérer les différents changements d'échelles subis par l'image. L'auteur a mis en évidence le lien étroit entre un processus de restauration d'images et un processus de diffusion modélisé par des équations aux dérivées partielles. Cette EDP permet une diffusion isotrope, ce qui présente des inconvénients notamment au niveau de la qualité visuelle de l'image, en effet, il ne lisse pas uniquement le bruit mais il gomme aussi les contours, les rendant difficilement identifiables. L'analyse linéaire élimine le bruit mais introduit le flou [MER 04].

- **Détecteur Harris-Laplace**

Mikolajczyk et Schmid proposent [MIK 05] la création de détecteurs de caractéristiques robustes, avec des échelles invariantes et une répétabilité élevée, inventés par Harris-Laplace et Hessian-Laplace. Ils ont utilisé la mesure de Harris ou le déterminant de la matrice hessienne pour sélectionner l'emplacement et le Laplacien pour sélectionner l'échelle.

- **Détecteur basé sur des différences de gaussiennes (DoG)**

Pour mettre l'accent sur la vitesse, Lowe a proposé en [LOW 04] de rapprocher le Laplacien de Gaussiennes (LOG) par une différence de gaussiennes (DoG) filtre.Lowe propose en 1999, d'intégrer la notion d'espace d'échelles dans le calcul de différences de gaussiennes, afin de rendre plus stable la détection de points d'intérêt.

- **Détecteur fast-hessien**

Proposé en 2006 par Bay et al [BAY 06], c'est un détecteur basé sur une approximation du filtrage gaussien, il permet de diminuer les temps de calculs.Le détecteur fast-hessien nous fournit une liste de points d'intérêt, caractérisés par

leurs coordonnées et leur échelle locale. Il se base sur l'exploitation de la matrice Hessienne (équation 1) :

$$H(\boldsymbol{x},\sigma)=\begin{bmatrix}L_{xx}(\boldsymbol{x},\sigma) & L_{xy}(\boldsymbol{x},\sigma)\\ L_{xy}(\boldsymbol{x},\sigma) & L_{yy}(\boldsymbol{x},\sigma)\end{bmatrix} \quad (1)$$

Le fast -hessien a le meilleur taux de répétabilité par rapport aux autres descripteurs.

- Détection

- Le fast-hessien se base sur l'exploitation de la matrice Hessienne (équation 1), dont le déterminant se calcule de la façon suivante :

$$\det\big(H(x,y,\sigma)\big)=\sigma^2(\mathrm{L_{xx}(x,y,\sigma)L_{yy}(x,y,\sigma)}-\mathrm{L_{xy}^2(x,y,\sigma)}) \quad (2)$$

- En recherchant les maxima locaux de ce déterminant, nous établissons une liste de K points associés à une échelle, notée :

$$\{(\mathrm{x_k,y_k,\sigma_k});\mathrm{k}\in[\![0;\mathrm{K}-1]\!]\},\text{où: }(\mathrm{x_k,y_k,\sigma_k})=\operatorname{argmax}\det\big(H(x,y,\sigma)\big)$$

Les convolutions obtenues par des dérivées régularisées (par un filtrage gaussien) sont notées D_{xx}, D_{xy} et D_{yy}. Afin de garder la cohérence dans le filtrage, Bay et al. [BAY 06] utilisent un filtre approximé initial de taille 9x9 correspondant à un filtre gaussien d'écart type $\sigma = 1{,}2$. Cette cohérence est assurée par l'équation suivante :

$$\frac{\big|\mathrm{L_{xy}}(1,2)\big|\mathrm{F}|\mathrm{D_{xx}}|\mathrm{F}}{|\mathrm{L_{xx}}(1,2)|\mathrm{F}\big|\mathrm{D_{xy}}\big|\mathrm{F}}=0{,}912\cdots\approx 0{,}9 \quad (3)$$

- ou $|.|\mathrm{F}$ est la norme de Frobenius. Le hessien sera donc déterminé par l'équation suivante :

$$\det\big(H_{approx}\big)=D_{xx}D_{yy}-(0.9D_{xy})^2 \quad (4)$$

- Enfin, afin de gérer la multi-échelle, Bay et al. s'appuient sur un ensemble de masques de tailles croissantes (9 x 9, 15 x 15, 21 x 21, 27 x 27,...), dépendant de l'écart type de la gaussienne à approximer. Par exemple dans le cas d'un

filtre de taille 27 x 27, l'approximation correspond à une gaussienne possédant un $\sigma = \frac{27}{9} \times 1{,}2 = 3{,}6$.

- Ce détecteur permet notamment la détection de zones homogènes.

3. Invariances aux transformations affines et projectives

Introduit en 1994 par Lindeberg [LIN 94], puis détaillé en 1997 par Lindeberg et Garding [LIN 97]. Le détecteur utilisé est le Laplacien avec adaptation affine. Brièvement, les points d'intérêt détectés correspondent à des maximums locaux dans l'espace échelle du Laplacien normalisé en intensité. A ces points sont associés des cercles de rayon d'échelle correspondante, cercles transformés en ellipse par le processus d'adaptation.

- **Détecteurs Harris-affine et Hessian-affine**

Mikolajczyk et Schmid proposent [MIK 05] le Harris-affine et le Hessian-affine, basés sur l'extraction multi-échelles des points d'intérêt, et la détermination itérative d'une région locale circulaire.

Leurs travaux ont montré la pertinence des approches qui reposent sur l'extraction et la caractérisation de points d'intérêt. L'intérêt de ces approches réside notamment dans leurs propriétés d'invariance aux changements de contraste et aux transformations affines des images.

- **Détecteur MSER (Maximally Stable ExtremalRegions)**

Matas et al. proposent en 2002[MAT 02] une approche leur permettant d'extraire des régions d'intérêt en étant invariantes aux transformations affines qu'elles soient photométriques ou géométriques. Elle repose sur l'extraction de composantes connexes en utilisant un seuillage de l'image. Les régions extrémales de MSER, sont constituées de tous les pixels dont l'intensité est plus (ou moins) élevée que celle des pixels de son extérieur.

De l'étude des détecteurs existants et à partir des comparaisons publiées, nous pouvons conclure que les détecteurs de Hessian sont plus stables et reproductibles que leurs homologues de Harris.

Référence

[ACH 00] : C. Achard, E. Bigorgne, and J. Devars. "A sub-pixel and multispectral corner detector", International Conf. on Pattern Recognition, 2000.

[BAY 06]: H. Bay, T. Tuylelaars, et L. Van Gool. Surf : "Speeded up robust features.European" Conference on Computer Vision, 2006.

[BEA78] : P.R. Beaudet. "Rotational Invariant Image Operators". International Conference on Pattern Recognition, pp.579-583, 1978.

[DRE 82] :L.Drechler and H.Nagel, "On Selection of Critical Points and Local Curvature Extrema of Region Boundaries for Interframe Matching"., Inter-national Conference on Pattern Recognition, 542-544, 1982.

[HAR 88] : C. Harris and M. Stephens. A combined corner and edge detector.Alvey Vision Conference, pages 147-151, 1988.

[KIT 82] : Kitchen L. et Rosenfeld A., "Gray-Level Corner Detection", Pattern Recognition Letters, 95-102, 1982.

[KOE 84] :J.Koenderink, "The Structure of Images", Biol. Cybern,363-370,1984.

[LIN 94]: Lindeberg, "Scale-Space Theory in Computer Vision", Kluwer Academic Publishers/Springer, Dordrecht, Netherlands, 1994.

[LIN 97] :Lindeberg and Garding "Shape-adapted smoothing in estimation of 3-D depth cues from affine distortions of local 2-D brightness structure", Proc. 3rd European Conf. on Comp. Vision, pp. 389-400, Springer-Verlag 1997.

[LOW 04] : D. Lowe, "Distinctive image features from scale-invariant keypoints," Intern. Journal of Comp. Vision, vol. 60(2), pp. 91-110, 2004.

[MAN 11]: G-B.Manuel, "Descripteurs 2D et 2D+t de points d'interet pour des appariements robustes", Thèse 2011

[MAT 02] : J. Matas, O. Chum, M. Urban and T. Pajdla (2002). "Robust wide baseline stereo from maximally stable extremum regions" .British Machine Vision Conference: 384–393.

[MER 04] : D.Merad, "Reconnaissance 2D/2D et 2D/3D d'objets àpartir de leurs squelettes" thèse 2004.

[MIK 05] : K. Mikolajczyk et C. Schmid, "A performance evaluation of local descriptors,"IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27(10), pp. 1615–1630, 2005.

[MOR 01]: P. Moravec. " 3D graphics and the wave theory ", Proc. of the 8th annual conf. on Comp.grap.& interactive tech., p.s 289–296, 2001.

[NOB 89]:J. A. Noble, Descriptions of Image Surfaces. PhD thesis, Department of Engi-neering Science, Oxford University, 1989.

[ROS 10]:E. Rosten, R. Porter, &T. Drummond, "Faster and better: A machine learning approach to corner detection", IEEE Trans. Pattern Analysis and Machine Intelligence, 105–119, 2010.

[SHI 94]:J. Shi and C. Tomasi. Good features to track. IEEE Conference on Computer Vision and Pattern Recognition, pages 593-600, 1994.

[SMI 97]:S M Smith and M Brady, "SUSAN a new approach to low level image processing", Intern.Jour. of Comp. Vision, vol.23(1), 45-78, 1997.

[WAN 95]: H Wang and M Brady, "Real-time corner detection algorithm for motion estimation", Image and Vision Computing, vol. 13:9, pp. 695-703, 1995.