République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université des Sciences et de la technologie d'Oran.
« Mohamed BOUDIAF »

Faculté de Génie Électrique
Département d'Électronique

Mémoire en vue de l'obtention du diplôme de magistère

Spécialité : Ecole Doctorale NTIC, Signaux, Systèmes Intelligents et Robotique
Option: Systèmes Intelligents et Robotique

Présenté par :
Mme. BENHALLOU Khadidja Ep. BENACHENHOU

Interface Design for Human Pose Estimation

Soutenu le : 29 Avril 2015

Devant le jury composé de :

	Nom & Prénom	Grade	Etablissement
Président	Mr.OUAMRI Abdelaziz	Professeur	**USTO-MB**
Encadreur	Mr. BERRACHED Nasr Eddine	Professeur	**USTO-MB**
Examinateur	Mr. OUSSLIM Mohamed	Professeur	**USTO-MB**
Examinateur	Mr. LOUKIL Abdelhamid	MCA	**USTO-MB**

Résumé

L'objectif de cette thèse est de mettre en place un procédé d'estimation de la pose humaine, cette dernière étant par définition le processus de localisation des parties du corps, ainsi nous considérons dans notre travail les parties du corps suivantes : Tête, torse, bras gauches et droits ainsi que les jambes droites et gauches.

Ce domaine de recherche est considéré comme étant un grand challenge dans le domaine de la vison par ordinateur; Ayant comme objet d'intérêt le corps humain, nous sommes soumis à plusieurs contraintes liées à la nature articulée du corps humain, le nombre élevé du degré de liberté, l'apparence variée de l'humain, variation intra-classe, etc.

Dans l'optique de fournir un système d'estimation de pose humaine, nous tenons à relever le défi par l'adoption d'une approche discriminatoire basé modèle, combinant deux grands concepts de la vision par ordinateur : Détection de personne par le descripteur de l'**histogramme de gradient orienté** et l'approche **Structure Picturale**, s'inspirant ainsi par le concept « Objet Basé Modèle »: Chercher un objet articulé par la recherche de ses composantes, dans notre cas les parties du corps humain.

Nous nous inspirons de l'approche de Dalal &Triggs par l'utilisation de l'Histogramme de gradient orienté comme descripteur d'apparence. Permettant ainsi une classification plus fiable qui doit capturer les similitudes essentielles entre les objets de la même classe et les différences avec des objets de classes concurrentes en se basant sur l'apparence de l'objet. Ces descripteurs ont l'exclusivité de mieux représenter la structure interne d'un objet via l'information du gradient, permettant ainsi de surmonter les problèmes liés à l'apparence de l'objet : pose, l'éclairage, l'occlusion, texture de fond, etc. Les approches de la fenêtre coulissante, suppression des non-maxima et la pyramide descripteurs sont exploitées pour rechercher l'objet « personne » sur une image, via un **modèle** préalablement appris par une machine à vecteurs de support Linéaire. Cette détection est essentielle pour parer aux problèmes

d'apparence variée liée aux humains vu que le modèle exploité décrit l'allure générale de l'objet personne.

Les parties du corps humain sont modélisées via une structure d'arbre, afin de faciliter le processus d'inférence, où les nœuds de l'arbre codifient les emplacements des parties du corps humain via des variables latentes. Les modèles des différentes parties du corps sont apprises via une **Machine à vecteurs de supports Latents**.

Le même processus de détection appliqué pour le corps humain est appliqué aux différentes parties du corps humain sur l'image. Ainsi le score est compensé avec le score de détection du corps humain, considéré comme racine de l'arbre. Le score maximum de détection représente la meilleure configuration possible des parties du corps humain.

Enfin, nous validons nos propos par l'implémentation de l'approche proposée via l'application IDHPE : Interface Design for Human Pose Estimation, cette application comporte deux modules : détection de personne et estimation de la pose humaine, offrant un moyen d'évaluation et d'évolution dans les travaux futurs.

Mots-clés : discriminatoire, basé modèle, histogramme de gradient orienté , Structure Picturale , Fenêtre coulissante, pyramide de descripteurs, machine à vecteurs de supports linéaires , machine à vecteurs de supports latents, suppression des non-maxima.

ملخص

ان الهدف من هذه الرسالة هو تطوير طريقة لتقدير هيكل الإنسان، وهذا الأخير من خلال عملية تحديد موقع أجزاء الجسم، . الجذع ،الساقين واليدين,واثناء عملنا نعتبر أجزاء الجسم التالية: الرأس

ويعتبر هذا المجال من الأبحاث تحديا كبيرا في مجال الرؤية الحاسوبية. باعتبار جسم الانسان محورا بحاثنا نخضع لعدة معوقات لطبيعة المفصلية للجسم البشري، وهذا للعدد الكبير من درجات الحرية، ومظاهر متنوعة من البشر، والتباين داخل المجموعات، الخ

من أجل توفير نظام تقدير الإنسان ، نريد أن مواجهة هذا التحدي من خلال تبني نهج تمييزي قائم على نموذج ، والجمع بين اثنين من المفاهيم الرئيسية في عالم الرؤية بالكمبيوتر : الكشف عن شخص عن طريق الرسم البياني للميل الموجه و نهج البنية التصويرية استناد اعلى مفهوم بحث عن مفصلية كائن من خلال البحث عن مكوناته ، في حالتنا أجزاء من الجسم البشري

نستوحي عملنا من اعمال دلال و تريجس باستخدام الرسم البياني للميل الموجه لوصف المظهر. للسماح لتصنيف أكثر موثوقية يجب التقاط أوجه التشابه الأساسية بين كائنات من نفس الفئة والاختلاف مع الأشياء التي تنتمي الي المجموعة المنافسة وهذا بناء على المظهر. هذه الأوصاف تمثل البنية الداخلية للشئ باستخدام معلومات من التدرج ، وبالتالي التغلب على المشاكل مثل :الهيكل ،الإضاءة ، والملمس الخلفية، إلخ.كما نستخدم طريقة النافذة المنزلقة ، وقمع غيرالاقصى و الهرم للبحث عن "شخص" على الصورة عبر نموذج قد علم سابقا بالشعاع الدعم الآلي الخطي . هذا الكشف ضروري للتعامل مع مشاكل ظهور متنوعة .

أجزاء الجسم البشري ممثل كهيكل شجرة لتسهيل عملية الاستدلال ، حيث العقد من شجرة تقنن مواقع أجزاء من جسم الإنسان عن طريق المتغيرات الكامنة . و علمت نماذج من أجزاء الجسم المختلفة عن طريق شعاع الدعم الآلي الكامن .

اخيرا وللتحقق من صحة اعمالنا قمنا بتطبيقIDPHE هذا التطبيق يحتوي على وحدتين: كشف عن شخص والكشف عن اعظاء الجسم وكذلك توفير وسائل التقييم والتطور في العمل المستقبلي

الكلمات الرئيسية : التمييز القائم على نموذج ، الرسم البياني للميل الموجه ، البنية التصويرية ،انزلاق نافذة ،الهرم ، شعاع الدعم الآلي الخطي ، شعاع الدعم الآلي الكامن ، قمع غيرالاقصى.

Abstract

Our goal in this thesis is to present an effective approach to estimate a 2d human pose in image. Modeling human bodies (or articulated objects in general) in images is a long-lasting problem in computer vision, comparatively with rigid objects which can be reasonably modeled using several templates. Also, human bodies can vary greatly in appearance. Variations arise not only from changes in illumination and viewpoint, but also due to intra-class variability in shape and other visual properties.

We propose a discriminative approach based model, motivated by the concept of "search articulated object by his components", and the problem of "pedestrian detection with Histogram of Oriented Gradient proposed by Dalal & Triggs" ;

The Histogram of Oriented Gradient represent the internal structure of an object using the information of the gradient, there by overcoming the problems associated with the appearance of the object: pose, lighting, occlusion, texture, etc; Using the sliding window approach, the filter is applied at all positions and scales of an image followed by the non-maxima suppression.

Parts of the human body are modeled as a tree structure to facilitate the process of inference, where the nodes codify the locations of parts body as latent variables. Models of different body parts are learned via a Latent support vector machines.

The score is offset with the detection score of the filter which represent the general aspect of human body, considered as the root of the tree and learned using linear support vector machines and the detection score of the parts filter. The maximum score represent the best possible configuration of parts of the human body.

Finally, we validate our words by IDHPE Application: Interface Design for Human Pose Estimation, which contains two main modules: human detection and human pose estimation providing a tool of evaluation.

Keywords: discriminative, based model, histogram of oriented gradient , pictorial structure, sliding window, pyramid of descriptors, linear support vector machines, latent support vector machines, non-maxima suppression.

Table des matières

Liste des tableaux

Liste des diagrammes

Liste des figures

Remerciements

En préambule à ce mémoire, je souhaite adresser mes remerciements les plus sincères aux personnes qui m'ont apportée leur aide et qui ont contribué à l'élaboration de ce mémoire.

Je tiens à remercier sincèrement Monsieur N.BERRACHED, qui en tant que Directeur du laboratoire LARESI et encadreur de ma thèse, s'est toujours montré à l'écoute et très disponible tout au long de la réalisation de ce mémoire, ainsi pour l'inspiration, l'aide et le temps qu'il a bien voulu me consacrer et sans qui ce mémoire n'aurait jamais vu le jour.

Mes remerciements s'adressent également aux membres de mon jury, qui m'ont fait l'honneur d'évaluer mon travail.

Un remerciement particulier à mes chers parents Mohammed et Houria pour leur contribution, leur soutien et leur patience, à mon cher époux Salim pour son réconfort, à mes adorables sœurs Kawther, Noussiba et Ibtihel , à mes chères beaux-parents Fouzi et khadidja.

Enfin, j'adresse mes plus sincères remerciements à tous mes proches et amis, qui m'ont soutenue et encouragée au cours de la réalisation de ce mémoire.

Merci à tous et à toutes!

.

Introduction générale

La perception du mouvement humain, non verbale, est un aspect important de l'interaction homme-Robot : Pour permettre aux robots de devenir des collaborateurs fonctionnels dans la société, ils doivent être en mesure de prendre des décisions en fonction de leur perception de l'état de l'humain. De plus, les connaissances sur l'état de l'humain est crucial pour permettre aux robots d'apprendre des stratégies de contrôle et d'observation directe de l'homme.

L'activité de l'humain est caractérisée par la pose adoptée pour accomplir une tâche, la perception d'une activité humaine est étroitement liée par le processus d'estimation de la pose humaine étant définie par le processus d'estimation de la structure sous-jacente cinématique d'une personne à partir de capteurs [1]. Dans le domaine de la vision par ordinateurs, les caméras sont exploitées en tant que capteur.

L'estimation de la pose du corps humain est un grand challenge, dû au nombre important du degré de liberté à estimer. En outre, apparence variée dû aux vêtements, la forme du corps humain (auto occlusion), ainsi que le problème de projection du 3D sur une image plane 2D. Ces difficultés ont été abordées de plusieurs manières en fonction des données d'entrée fournies.

Les approches de l'estimation de la pose humaine peuvent être classées, dans une première étape, entre les méthodes basées ou non basées modèle : D'une part, les méthodes non basés modèle utilisent de l'apprentissage pour la mise en correspondance entre l'apparence et la pose du corps, conduisant à une performance rapide et des résultats précis pour certaines actions (ex. poses de la marche). Cependant, ces méthodes sont limitées par une première étape de soustraction de fond ou par la difficulté d'étendre les actions possibles. Par ailleurs, les méthodes basées modèle emploient la connaissance préalable sur la morphologie humaine.

Dans l'optique de fournir un système d'estimation de pose humaine, nous tenons à relever le défi per l'adoption d'une approche discriminatoire basé modèle, combinant deux grands concept de la vision par ordinateur : Détection de personne par le descripteur de l'Histogramme de Gradient Orienté et l'approche **Structure Picturale** [45], s'inspirant ainsi par le concept : Chercher un objet articulé par la recherche de ses composantes, dans notre cas les parties du corps humain.

Nous nous inspirons de l'approche de Dalal &Triggs pour l'utilisation de l'Histogramme de gradient orienté comme descripteur. Permettant ainsi une classification plus fiable qui doit capturer les similitudes essentielles entre les objets de la même classe et les différences avec des objets de classes concurrentes en se basant sur l'apparence de l'objet; Ces descripteurs ont l'exclusivité de mieux représenter la structure interne d'un objet via l'information du gradient, permettant ainsi de surmonter les problèmes liés à l'apparence de l'objet : pose, l'éclairage, l'occlusion, texture de fond, etc. . Les approches de la fenêtre coulissante, suppression du non-maxima et la pyramide de descripteurs sont exploitées pour rechercher l'objet « personne » sur une image, via un **modèle** préalablement appris par une machine de vecteurs à support Linéaire. Cette détection est essentielle pour parer aux problèmes d'apparence variée liée aux humains vu que le modèle exploité décrit l'allure générale de l'objet personne.

La représentation la plus naturelle pour un corps humain est la structure d'arbre où les emplacements des parties du corps humains sont représentés via des nœuds (variables latentes), permettant ainsi une inférence efficace ; Les modèles des différentes parties du corps sont apprises via une machine de vecteurs à support Latent.

Le modèle définit une distribution de probabilité a posteriori sur les poses humaines dans une image d'entrée. Pour une image donnée, nous échantillons la configuration du corps humain en entier. Les configurations échantillonnées sont

ensuite classées en fonction de leurs probabilités a posteriori et combinées par la suite par les configurations de la racine, dans notre cas le filtre représentant le corps en entier.

La notation finale de la pose du corps et la meilleure sélection de candidat est un problème difficile en lui-même.

Dans cette thèse, nous allons justifier notre démarche par un tour d'horizon dans la littérature de la détection de personnes et l'estimation de la pose humaine dans le **chapitre 1 État de l'art ,** nous exhibons nos motivations ainsi que nos contraintes par rapport à notre défi de l'estimation de la pose humaine dans le **chapitre 2 : Motivation et Challenge ,** étant donné que le choix d'un descripteur pour l'apparence représente est cruciale pour la réussite d'un système de détection de la pose humaine , nous présentons dans le **chapitre 3** une présentation de notre **descripteurs de l'histogramme de gradient orienté** , le **chapitre 4 : Champs aléatoires conditionnel (CRF) :** présentant une approche probabiliste de notre approche ainsi qu'une représentation graphique de notre modèle**,** le **chapitre 5** : **Détection des parties du corps humain** est une description détaillée du procédé de recherche et de localisation des différentes parties du corps humain

Et pour finir, nous validons nos propos par une phase d'expérimentation et évaluation illustrée dans le **chapitre6 : Implémentation et évaluation.**

Chapitre 1 : État de l’art

1.1. Introduction

L’estimation de la pose humaine, fait référence au processus d'estimation de la structure sous-jacente cinématique d'une personne à partir de capteurs [1]. Les approches basées sur la vision sont souvent utilisés pour fournir une telle solution, en utilisant des caméras en tant que capteurs [2]. L’estimation de la pose est une question importante pour beaucoup d’applications de vision par ordinateur , telles que l'indexation de la vidéo [3] , Le domaine de la télésurveillance [4] , la sécurité automobile [5] et l'analyse du comportement [6] , ainsi que d’autres telles que : L’interaction Homme-Machine [7] , [8] .

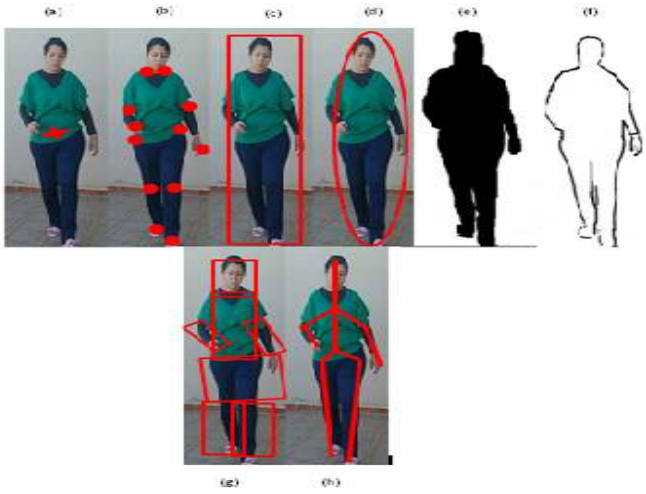
L’estimation de la pose du corps humain est un problème difficile, dû au nombre important du degré de liberté à estimer. En outre, apparence variée dû aux vêtements, la forme du corps humain (auto occlusion) , ainsi que le problème de projection du 3D sur une image plane 2D. Ces difficultés ont été abordées de plusieurs manières en fonction des données d'entrée fournies. Dans certain cas, l'information 3D peut être disponible en présence de plusieurs caméras sur la scène. A l’heure actuel, un certain nombre d’applications dans le domaine de l'estimation de la pose exploitant la profondeur dû à l’apparition des caméras de profondeur à faible coût [9].

Les approches de l’estimation de la pose humaine peuvent être classées, dans une première étape , entre les méthodes **basés ou non basés modèle** [10].D'une part, le modèle non basé modèle [11] , [12] est celui qui utilisent de l’apprentissage pour la mise en correspondance entre l'apparence et la pose du corps, conduisant à une performance rapide et des résultats précis pour certaines actions (ex. poses de la marche). Cependant, ces méthodes sont limitées par une première étape de soustraction de fond ou par la difficulté d'étendre les actions possibles. Par ailleurs,

les méthodes basées modèle emploient la connaissance préalable sur la morphologie humaine .

1.2. Représentation du corps humain

La représentation de d'objet se réfère à la façon dont l'objet est modélisé et localisé. Le domaine de la représentation du corps humain enregistre une grande variété d'approches et ceci selon le degré d'information à mettre en évidence : Nous distinguons , ceux qui utilisent un minimum d'informations extraites de l'objet , comme la couleur [13] , intensité [14] , des points caractéristiques [15], les histogrammes [16],modèles d'objet [17]. Les représentations de formes d'objets généralement utilisées pour suivi sont : des points , des formes géométriques primitives (par exemple, rectangle , ellipse) , silhouette , contour , forme articulée et modèles squelettiques [18] comme indiqué sur la figure 1-1 .



(a) Point, (b) points multiples, (c) forme géométrique (rectangle) , (d) forme géométrique (ellipse) (e) silhouette (f) contour, (g) forme articulée (h) forme squelettique

Figure 2-1Représentation de la forme d'objet

La représentation d'objet est liée aussi au domaine d'exploitation, par exemple si nous ciblons le domaine de la vidéosurveillance, où juste la détection de la cible est nécessaire, nous nous contenterons d'une détection du corps en forme de bloc.
Nous rappelons notre objectif de départ, qui est l'interaction homme-machine, sachant que si nous nous inspirons de nos différentes interactions avec nos homologues, nous remarquons qu'elles sont assurées à travers des gestes, ces derniers générés par nos membres du corps.
D'où notre recours à l'estimation de la pose humaine.
En se basant sur notre approche proposée pour l'estimation de la pose humaine où nous proposons de la dérouler en deux étapes : détection de la cible (personne) afin de réduire l'espace de recherche, estimation de la pose (localisation des parties du corps humain). Notre chapitre état de l'art est organisé comme suit :

I. Détection de personne
II. Estimation de la pose humaine

1.3 Détection des personnes

Notre but est de localiser une personne dans une vidéo (ou image) , la littérature permet de subdiviser ce domaine de détection en deux principales classes , celle qui exploitent l'information du fond et celle qui détectent directement la personne.

Les techniques de soustraction de fond cherchent le premier plan de l'objet de l'image et le classifient dans des catégories comme les humains, véhicules, etc., sur la base de la forme, la couleur, ou le mouvement ou autres caractéristiques.

Les techniques directes fonctionnent sur (à partir de caractéristiques extraites) une image ou vidéo et les classent comme humain ou non-humain. Vous pouvez également classer les techniques basées sur les caractéristiques qui sont utilisées pour classer une entrée donnée comme humain ou pas. Ces caractéristiques

comprennent la forme (forme de contours ou d'autres descripteurs), la couleur (de détection de la couleur de peau), mouvement, ou des combinaisons de ceux-ci.

Ci-dessous une liste des techniques basées détection directe.

Tableau 2-1Méthode basées détection directe

Article	Modèle Humain	Classificateur
Cutler and Davis [32]	Mouvement périodique	Similarité de mouvement
Utsumi and Tetsutani [33]	Forme	Distance
Gavrila and Giebel [34]	Modèle	Distance de Chamfer
Viola et al. [35]	Forme+mouvement	Cascade Adaboost
Sidenbladh [36]	Flux optique	SVM (RBF)
Dalal and Triggs [37]	Hist. de gradient	SVM (Linéaire)

2.3.1 Les techniques de soustraction de l'image de fond

2.3.1.1 Wren et al. [19]

Ils décrivent le système **Pfinder** en temps réel pour la détection et le suivi des humains. Le modèle de base utilise une distribution gaussienne dans l'espace YUV , à chaque pixel , et le modèle d'arrière-plan est continuellement mise à jour . La personne est modélisée en utilisant plusieurs blobs avec des composantes spatiales et de couleurs et du gaussien correspondant aux distributions. Tant que la personne est dynamiquement en mouvement les paramètres spatiaux sont constamment estimés avec un filtre de Kalman . Ensuite, pour chaque pixel d'image le procédé évalue, la probabilité qu'il fait partie de la scène de fond ou du blob. Chaque pixel est ensuite affecté à soit au blob ou à l'arrière-plan dans le maximum a posteriori

(MAP), suivi par l'application de simple opérations morphologiques. Après cette étape, les modèles statistiques pour le blob et la texture de fond sont mis à jour.

Les modèles de blobs de personnes sont initialisés à l'aide d'une détection de contour, étape qui tente de localiser la tête, les mains et les pieds. Le blob relatif au visage est initialisé avec la recherche de la couleur de peau. Ce système est conçu en vue de trouver un seul homme, et fait plusieurs hypothèses spécifiques au domaine .Il a été testé dans plusieurs scénarios HCI et est en temps réel.

2.3.1.2 Beleznai et al. [20]

Il traite de la différence d'intensité entre une trame d'entrée et une image de référence comme une distribution de probabilité multi-modale, et la détection de mode est effectuée en utilisant le calcul de changement de vitesse moyenne.

Cette procédure de détection est capable de localiser des personnes isolées sur l'image, mais pour séparer les humains partiellement occlus et groupés, un processus de validation basé sur un modèle est utilisé. Le modèle d'humain est très simple et se compose de trois zones rectangulaires. Au sein de chaque groupe d'humains, une configuration de maximum de vraisemblance de l'homme est identifiée.

2.3.1.3 Haga et al. [21]

Dans cet article, un objet en mouvement est classé comme personne sur la base de l'unicité spatiale du mouvement d'image (appelé critère F1 par les auteurs), l'unicité temporelle du mouvement humain (F2), et la continuité du mouvement temporel (F3). En premier lieu, l'objet en mouvement est détecté par soustraction du fond, puis, F1, F2, et F3 sont évaluées. L'unicité spatiale de mouvement de l'image est une mesure de l'uniformité de mouvement local dans une région. L'unicité temporelle est en conséquence définie dans le sens horaire. Un classificateur linéaire

sépare personne et non personne dans l'espace F1-F2-F3, utilisé aussi pour classer de nouvelles données d'entrée.

2.3.1.4 Eng et al [22]

Cet article propose une combinaison d'une approche ascendante axée sur la soustraction de fond et une approche descendante incorporant un modèle de forme humaine comme une solution aux problèmes de la détection de plusieurs personnes qui se chevauchent partiellement. Tout d'abord, un modèle de fond à base de région est construit sous l'hypothèse que chaque région a une distribution de probabilité gaussienne multivariée sur les couleurs.

Les modèles de base sont construits de manière simple au moyen d'un ensemble de trames d'arrière-plan qui sont séparés en blocs carrés en utilisant un algorithme k-means .Les pixels dans une nouvelle image sont comparés avec ce modèle d'arrière-plan et classés en premier plan ou arrière-plan. Les pixels non classés sont ajoutés au premier plan en utilisant un modèle basé couleur de la tête. Ensuite, une formulation bayésienne est appliquée sur la base d'un modèle simple de la tête et du corps sous forme de deux ellipses, et toutes les paires de têtes et de corps sont déterminées à partir du maximum a posteriori. Les expériences présentées dans ce document traitent uniquement un domaine spécifique pour la surveillance des piscines.

2.3.1.5 Elzein et al [23].

La méthode présentée dans ce papier détecte d'abord les objets en mouvement par le calcul du flux optique dans certaines régions sélectionnées par image différenciation. La vitesse du flux optique est alors utilisée pour calculer le temps de collision à un point de référence fixe dans l'image. Cela se fait parce que le but est de détecter les régions qui pourraient entrer en collision avec le véhicule sur lequel l'appareil est embarqué, qui est considérée comme un point référence. Les

pixels avec un petit temps de collision sont sélectionnés à l'aide d'un seuil, et les opérations morphologiques sont utilisées pour construire des groupes ou des blobs de pixels connexes. Les blobs restants sont remodelés en régions rectangulaires qui sont ensuite utilisés pour un traitement ultérieur. Afin de déterminer si une région rectangulaire sélectionnée est une personne, les auteurs font l'apprentissage d'un classificateur à l'aide des fonctionnalités en ondelettes et un schéma de mise en correspondance. En utilisant une base de données d'apprentissage d'images de piétons, des modèles sont construits, qui sont essentiellement une table normalisée de coefficients d'ondelettes. Le modèle final est constitué d'un élément de vecteur de 49 dimensions, qui est comparé à une fonction similaire construite pour chaque entrée : Si le nombre de coefficients similaires sont supérieures à un seuil, le rectangle est classé comme un piéton. De toute évidence, étant donné que les rectangles d'entrée peuvent être de taille différente, la mise en correspondance est réalisée à plusieurs échelles. La méthode proposée n'a pas prouvée ses performances en temps réel.

2.3.1.6 Toth et Aach [24]

La méthode présentée dans ce document effectue d'abord une soustraction de fond en utilisant différence inter frame, en se basant sur un balayage de fenêtre, la somme des différences absolues (SAD) de l'agrégation, et adaptation seuil. Les auteurs utilisent un champ aléatoire de Gibbs-Markov pour créer des seuils spatialement différents qui conduisent à lisser les formes. Les blobs de premier plan sont identifiés en utilisant des composants connectés, et la transformée de Fourier est appliquée à la forme périphérique.

2.3.1.7 Jiang et al [31]

Cette approche est basée sur la fusion de l'infrarouge (IR) des images avec des images à partir d'un appareil photo classique. Les êtres humains présentent une signature caractéristique dans les Images IR en raison de leur température de la

peau. Ils peuvent être fusionnés avec des images d'un type appareil photo pour obtenir des résultats de détection supérieures. La méthode proposée calcule en premier les pixels saillants dans les deux images (IR et visibles) à plusieurs échelles, et la fusion est réalisée sur la base de différence de contraste dans les deux images.

2.3.2 Détection directe

2.3.2.1 Cutler et Davis [32]

Les techniques présentées dans ce document mettent l'accent sur la détection de mouvements périodiques et les caractéristique biologiques (Ex.la marche). La vidéo de la caméra mobile une fois stabilisée, la différence inter frame et un seuillage sont appliqués pour détecter indépendamment les régions en mouvement. Les opérations morphologiques sont ensuite utilisées pour obtenir un ensemble d'objets suivis. Chaque objet segmenté est aligné le long de l'axe du temps (pour éliminer la translation, et sa taille est également constante dans le temps .Les matrices de similarités temporelles de l'objet sont calculées en utilisant des mesures de similarité (comme la corrélation). L'analyse de Temps de fréquence est basée sur la transformée de Fourier (STFT), l'autocorrélation est utilisée pour la détection de la périodicité. L'approche est non seulement capable de détecter un mouvement périodique des humains, mais elle est utile pour extraire plus d'informations sur la démarche telle que la longueur de la foulée. Le système est en temps réel.

2.3.2.2 Utsumi et Tetsutani [33]

Ce document utilise le fait que les positions relatives (distances géométriques) de diverses parties du corps sont communes à tous les êtres humains, bien que les valeurs de pixels peuvent varier en raison des vêtements ou l'illumination. La technique utilise une structure connue comme la carte de distance qui est construite en prenant une image d'un être humain et la subdiviser en MxN blocs. Une matrice de distance de taille MN x MN est ensuite calculée où chaque

élément exprime la distance entre les distributions de couleurs présentes dans une paire de blocs. Puis, en utilisant de telles cartes de distances pour une grande base de données d'images humains et non - humains , un modèle statistique est construit pour les cartes de distance de chaque type , qui se compose de la matrice moyenne et de covariance pour chaque bloc . Les deux distributions sont comparées en utilisant la distance de Mahalanobis et se sont révélées être très similaires, sauf pour quelques éléments. Ces quelques éléments indiquent une matrice de projection qui est le modèle utilisé pour la reconnaissance. Compte tenu d'une nouvelle image d'entrée, l'image des correctifs à plusieurs endroits et les échelles sont par rapport au modèle et un seuil est utilisé pour classer un patch comme humain ou non humain.

2.3.2.3 Viola et al. [35]

Ce document traite la détection directe de l'homme sur une image statique ainsi que sur une vidéo en utilisant un classificateur formé sur la forme humaine et les caractéristiques de mouvement. L'ensemble de données d'apprentissage se compose d'images et vidéos des exemples humains et non humains. Le document se limite pour le cas des piétons (où les humains sont toujours de bouts, pose de marche). Le détecteur statique utilise des images comme des entrées et extrait les caractéristiques rectangulaires en utilisant des images intégrées. Un classificateur en cascade est créé pour obtenir une bonne détection avec un faible taux de faux positifs.

Chaque étape du classificateur est apprise sur les vrais et les faux positifs de l'étape précédente à l'aide d'Adaboost pour sélectionner des classificateurs faibles.

Le détecteur dynamique est formé de façon similaire en utilisant une combinaison de caractéristiques rectangulaires statiques et animées. Les deux détecteurs sont rapides et donnent de bons résultats de détection sur une grande base de données des piétons.

2.3.2.4 Sidenbladh [36]

Ce document met l'accent sur les modèles de mouvements humains, cette détection est robuste car elle est relativement indépendante de l'apparence et de l'environnement. La technique est sur la base de la collecte des exemples de mouvement humain et non humain et le calcul de flux optique. Une machine à vecteurs de support (SVM) avec une fonction radiale de base (RBF) est formée sur les modèles de flux optique pour créer un classificateur humain.

Le classificateur obtenu peut être appliqué pour une nouvelle vidéo d'entrée à plusieurs positions. Le procédé n'est pas approprié pour détecter les humains partiellement occlus.

2.3.2.5 Dalal et Triggs [37]

Le point culminant de ce document est l'utilisation d'un histogramme de gradients comme l'espace des caractéristiques pour la construction d'un classificateur. Elle utilise le fait que la forme d'un objet peut être bien représentée par une distribution de gradients locaux d'intensité ou les directions de contour. Cela se fait en divisant l'image en petites parties spatiales (cellules) et de trouver les histogrammes des orientations de contour sur tous les pixels de la cellule.

Les entrées d'histogramme combinées forment la représentation de la fonction après la normalisation du contraste local en blocs. Pour la classification, un ensemble de données d'exemples humains et non-humains est créé, et un classificateur linéaire est formé en utilisant un SVM Linéaire sur les caractéristiques de l'histogramme de gradient à partir de ces deux classes. Ce classificateur peut ensuite être appliqué sur une nouvelle image d'entrée à plusieurs échelles pour détecter les humains.

Etant une méthode adoptée dans notre approche approchée cette méthode sera plus détaillée dans les prochains chapitres.

2.4 Estimation de la pose humaine

L'estimation de la pose du corps humain est un problème difficile, dû au nombre important du degré de liberté à estimer. En outre, apparence variée dû aux vêtements, la forme du corps humain (auto occlusion), ainsi que le problème de projection du 3D sur une image plane 2D.

Ces difficultés ont été abordées de plusieurs manières en fonction des données d'entrée fournies. Dans certain cas, l'information 3D peut être disponible en présence de plusieurs caméras sur la scène. A l'heure actuel, un certain nombre d'applications dans le domaine de l'estimation de la pose exploitant la profondeur dû l'apparition des caméras de profondeur à faible coût [9].

Les approches de l'estimation de la pose humaine peuvent être classées, dans une première étape, entre les méthodes basés ou non basés modèle [10]. D'une part, le modèle non basés modèle [11] , [12] sont ceux qui utilisent de l'apprentissage pour la mise en correspondance entre l'apparence et la pose du corps, conduisant à une performance rapide et des résultats précis pour certaines actions (ex. poses de la marche). Cependant, ces méthodes sont limitées par une première étape de soustraction de fond ou par la difficulté d'étendre les actions possibles. Par ailleurs, les méthodes basées modèle emploient la connaissance préalable sur la morphologie humaine.

Dans notre chapitre état de l'art pour la partie estimation de la pose, nous avons structuré notre présentation selon cinq modules « Apparence, prise de vue, les relations spatiales, les relations temporelles et comportement ».

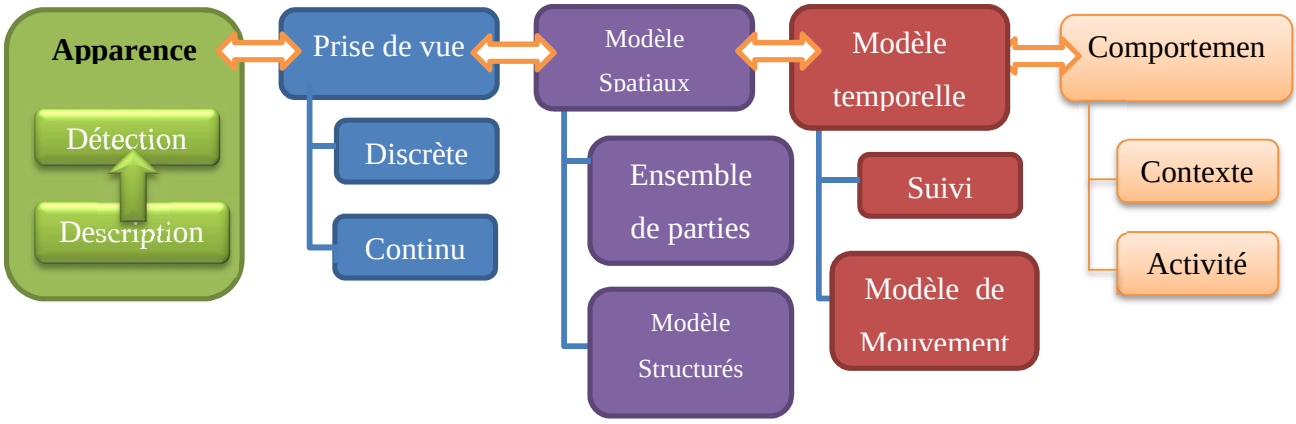


Figure 2-2 Taxonomie des approches de l'estimation de pose

2.4.1 Apparence

L'apparence peut être définie comme l'évidence d'image liée au corps humain et ses poses. Dans ces cas, les évidences sont non seulement les caractéristiques liées à l'image et les données d'entrée, mais également au processus d'étiquetage de pixels. Par conséquent, il peut être pris en compte à différents niveaux, du pixel à la région. L'apparence des personnes dans des images varie selon la pose humaine adoptée, l'éclairage et de l'habillement, et les conditions de changements de point de prise de vue, entre autres. Afin d'obtenir une détection précise et le suivi du corps humain, la connaissance préalable de la pose et de l'apparence sont nécessaire. Cette information déjà connue sur le corps humain peut être codifiée en deux étapes successives : description de l'image et la détection du corps humain (ou parties), appliquant généralement un processus d'apprentissage. Les procédures de description de l'image peuvent être effectuées à trois niveaux différents : pixel , local, et global (représenté sur la figure 1-13 (a) , (b) , (c)). Respectivement , ils conduisent à la segmentation d'images [38] , [39] , [20] , la détection de parties du corps [21] , [22] , [23] , [24] et localisation du corps entier [25] ,[26] . Il est largement reconnu que la description du corps humain comme un ensemble de

pièces améliore la reconnaissance du corps humain dans des poses complexes. En revanche, les descripteurs globaux sont largement exploités dans le domaine de la détection de personnes (telles que les piétons), comme ils servent dans l'étape d'initialisation pour le processus d'estimation de la pose humaine. La taxonomie pour les étapes de détection ainsi que la description sont décrites sur la figure 1-3 .

2.4.1.1 Détection :

La phase de la détection se réfère au processus spécifique de détection sur une image ou aux différents classificateurs de sortie qui codifient les informations sur le corps humain dans une image. Ce processus de synthèse peut être selon les quatre points ci-dessous :

2.4.1.1.1 Classifieurs discriminants :

Une technique couramment utilisée pour la détection de personnes dans des images consistant à décrire l'image régions à l'aide de descripteurs standards (ie l'Histogramme de Gradient Orienté (HOG) [25]) et l'apprentissage d'un classifieur discriminant (par exemple, Support Vector Machines) comme un descripteur de la posture entière du corps humain [25] ou comme une description multi-partie [27]. Certains auteurs ont étendu ce type d'approches, notamment les relations spatiales entre les objets tels que des *poselets* [26].

2.4.1.1.2 Classificateurs Génératifs :

Comme le cas des classifieurs discriminants, précédemment vus, les approches génératives ont été proposées pour le cas de la problématique de détection de personnes. Par exemple, l'approche de Rother, Kolmogorov et Blake [28] fait l'apprentissage d'un modèle de couleur pour l'arrière-plan , pour optimiser la probabilité fonctionnelle en utilisant la méthode Graph Cuts.

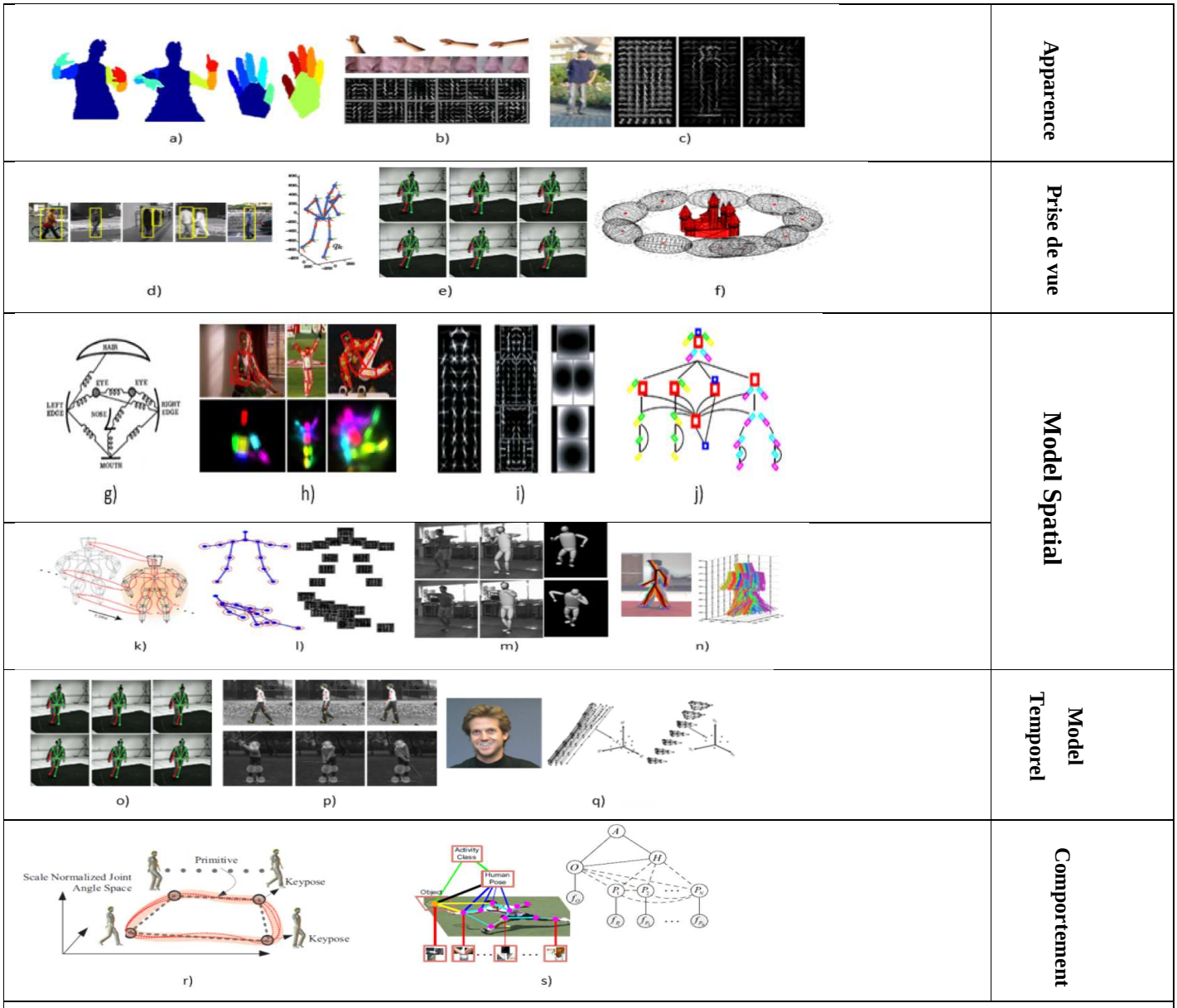


Figure 2-3 **Exemples de descripteurs appliqués au pixel, niveau local et global , respectivement** :

Approche Graph Cut [20]pour segmentation du corps et de la main;(b) Parties orientables (Steerable part) [24] ; (c) Méthode de HOG [25] Estimation de la prise e vue : (d) Discrète (e) Continue [10]; (f) Clustering de la pose de la caméra [46].Exemple des modèles de corps illustré comme un ensembles de parties : (g) initiale et (h) moderne [22] Structures picturales; (i) Model de personne basé grammaire[27] (j) composition hiérarchique du corps « pièces »[23];(k) Modèle spatio-temporel [49] (f) Des arbres différents obtenus à partir du mélange de pièces présentées dans [50]. Modèle structuré : (m) Deux exemples de 3D pose estimation [51];(n) 3D poses possibles (vers le bas) [45].Exemples de suivi: o) suivi 3D de l'ensemble du corps, à travers une approche d'hypothèses multiples [10], p) suivi 2D des parties du corps [52]. q) De gauche à droite: 3D dispose d'une bouche souriante et une comparaison de la forme et de l'espace trajectoire [53]. Pose humaine et comportement : r) exemples de marche différents (courbes), les modèles appris (lignes de morceaux) et ses poses clés [6]; s) modèle graphique proposé dans [54] pour la détection d'objet (O) estimation de pose (H) à partir de la détection des parties de corps (Pi).

2.4.1.1.3 Modèles:

Ces méthodes ont été proposées dans le domaine de l'estimation de la pose pour comparer l'image observée avec une base de données d'échantillons [10]. La limitation de ces approches est la restriction imposée aux poses utilisées dans la phase d'apprentissage, ce qui peut augmenter le nombre de faux positifs.

2.4.1.1.4 Les points d'intérêt:

Les points saillants ou des parties dans les images peuvent également être utilisées pour calculer la pose ou le comportement dans une séquence vidéo [29].

2.4.1.2 Description:

L'étape de détection analyse les informations extraites à partir des images dans la phase de description [31]. La plupart des méthodes courantes appliquées pour décrire l'image sont détaillés ci-dessous :

2.4.1.2.1 Silhouettes et contours :

Silhouettes et leurs contours (arêtes et contours) fournissent des descripteurs puissants invariants aux changements de couleur et de texture. Ils sont utilisés pour extraire le corps humain en images [32] , car en général, l'information sur la pose du corps humain réside dans la silhouette . Cependant, ces méthodes souffrent de mauvaises segmentations dans les scènes du monde réel, ainsi que la difficulté de récupérer certains degré de liberté (DOF) en raison de l'absence d'informations de profondeur .

2.4.1.2.1.1 Couleur et texture :

D'une part, les informations de la couleur et la texture sont eux-mêmes utilisés comme indices supplémentaires pour la description local des régions d'intérêt [10] . L'information de couleur est généralement codifiée au moyen d'histogrammes ou modèle d'espace de couleur (c.-à-d.- modèle de mélange gaussien) , quant à la

texture est décrite en utilisant la transformée de Fourier discrète (DFT) [33] ou ondelettes tels que des filtres de Gabor [34] .

D'autre part, les gradients sur les intensités de l'image sont les plus exploités pour la description de l'apparence d'une personne. En ce sens, HOG et SIFT , entre autres , sont largement exploités [25] .

2.4.1.2.2 Profondeur :

Récemment, la profondeur a été exploitée dans plusieurs systèmes de reconnaissance de la pose humaine en raison de la disponibilité des cartes de profondeur fournies par le **KinectTM** multi- capteur. Cette nouvelle représentation de la profondeur offre l'information 3D à partir d'un capteur synchronisé avec des données RVB. Exemple [35] utilisant les filtres de Gabor à travers une carte de profondeur pour la description de la main. Ces approches se caractérisent par un temps de calcul réduit, par contre, elle exploite des descripteurs spécifiques de l'image et une carte de profondeur qui ne peut être disponible dans la majorité des cas.

2.4.1.2.3 Mouvement :

Le flux optique [37] est l'approche la plus utilisée pour modéliser l'activité humaine [38]. En outre, d'autres travaux assurent le suivi visuel des descripteurs et codifient le mouvement par HOF (Histogram of Optical Flow) pour décrire les régions des parties du corps humain.

2.4.1.2.4 Logique :

Il est important de noter que de nouveaux descripteurs utilisant les relations logiques ont été récemment proposés. Ce c'est le cas de l'approche de Yao et Fei -Fei [39] , où les particularités locales sont codifiées à l'aide des opérateurs logiques ,permettant une description intuitive et discriminatoire de l'image (ou région) contexte .

2.4.1.3 Le point de vue « Viewpoint »

L'estimation du point de vue n'est pas uniquement nécessaire pour déterminer la relation entre la position et l'orientation entre le corps humain et la caméra (i.e. la pose de la caméra) mais offre aussi une possibilité de réduction de l'ambigüité 3D de la pose humaine [10]. Notons que, dans la littérature la pose de la caméra est désignée par « pose » mais dans nos travaux nous utilisons le terme **caméra** et réservant le terme **pose** pour la « **posture du corps humain ».**

Le point de vue n'est pas estimé directement dans le suivi des humains ou pour l'estimation de la pose humaine, mais il est indirectement considéré dans certain cas, car souvent on impose des contraintes sur la prise de vue. Par exemple, dans la littérature de la détection des parties supérieures du corps humain « Upper Body detection », dans la phase d'apprentissage, uniquement les images de face sont étudiées. Les recherches où le point de vue 3D est estimé, sont subdivisées en Classification discrète et continue.

Dans les approches discrètes, La géométrie 3D et l'apparence des objets sont capturées par un regroupement des descripteurs locaux des parties. Le point de vue discret est estimé pour les piétons par l'apprentissage de huit spécifiques prises de vue pour la détection des personnes.

L'approche continue pour l'estimation du point de vue se réfère à l'estimation des valeurs réelles des angles de point de vue du corps humain en 3D. Cette approche est très utilisée dans le domaine de l'estimation de point de vue en continu est largement étudié dans le domaine de l'enregistrement de la forme « Shape Registration », qui fait référence à la recherche des correspondances entre deux ensembles de points de récupération et la transformation qui fait correspondre un point fixé à l'autre.

Enregistrement de forme non-rigide monoculaire [44] peut être considéré comme un problème similaire à l'estimation de la pose humaine, puisque les points récupérés

ont la forme déformable qui peut être interprétée comme articulations du corps [45].Compte tenu des images fixes, la pose de la caméra en continue et l'estimation de la forme ont été étudiées pour l'estimation des surfaces rigides [46], ainsi que pour les formes déformables [47]. Dans les deux œuvres, l'information a priori sur la caméra a été communiquée par la modélisation des pose de caméra possible comme un modèle de mélange gaussien.

2.4.1.4 Modèles Spatiaux :

Ces modèles codifient la configuration du corps humain d'une manière dure (par exemple, squelette, a longueur des os) ou dans une manière douce (par exemple, les structures picturales, grammaires « **grammars** »). D'une part, les structures des modèles sont des squelettes 3D et des chaînes cinématiques précises. D'autre part, les projections dégénérées du corps humain dans l'image plane sont généralement modélisées par des ensembles de pièces.

2.4.1.4.1 Ensembles de pièces:

Les ensembles de pièces permettent de détecter les emplacements probables des différentes parties du corps dans une configuration compatible avec le corps humain, cette composition n'est pas définie par les contraintes physiques du corps humain mais elle est contraint par les emplacements des parties du corps dans l'image, qui peuvent faire face à la forte variabilité des poses de corps et points de vue.

Les structures picturales [48] sont des approches 2D génératrices d'assemblage de parties, où chaque partie est détectée avec son détecteur spécifique (représenté sur la figure1-3). Les structures picturales sont largement utilisées pour la détection des personnes et l'estimation de la pose humaine [55], [22]. Bien que la structure traditionnelle de la représentation soit un graphique [48], des approches plus récentes représentent le modèle de corps sous-jacent sous forme d'arbre, en raison de la facilité de l'inférence [55].

Les contraintes entre les parties sont modélisées par des modèles gaussiens. Ce qui semble ne pas correspondre à la réalité, Cependant, la distribution gaussienne ne correspond pas à une restriction dans le plan d'image en 2D, elle est appliquée dans un espace paramétrique, où chaque partie est représentée par sa position, l'orientation et l'échelle de représentation [55].

Modèles grammaires formalisés dans [56] fournissent une flexibilité et un cadre élégant pour détecter des objets [27], en particulier la détection des humains [27], [57]. Les règles de composition sont utilisées pour représenter des objets comme une combinaison d'autres objets. Dans ce cette façon, le corps humain peut être représenté sous la forme d'une composition du tronc, des membres et du visage, ainsi composée par les yeux, le nez et de la bouche. D'un point de vue théorique, les règles de déformation conduisent à des déformations hiérarchiques, permettant d'étudier des mouvements relatifs de chaque pièce à chaque niveau hiérarchique; Toutefois, les règles de déformation dans [27] sont traités comme des structures picturales (représenté sur la Figure 1-3). Suivant cette idée de composition, Les travaux de [23] se sont est basés sur **poselets** [26] représentant le corps comme une combinaison hiérarchique du corps "Pièces».

Des ensembles de pièces peuvent également être réalisées en 3D lorsque, par exemple, l'information 3D est disponible en utilisant un système multi-caméras [49], [58]. Une extension des structures picturales [58], où l'évolution temporelle est également prise en compte (représenté sur la Figure1-3). Les articulations sont modélisées suivant un mélange de gaussiennes, mais il est nommé modèle "loose-limbed" en raison du libre attachement entre les membres. Une puissante représentation graphique de la pose humaine en 2D, relativement inexploré, «**Graphe ET-OU »,** ce qui pourrait être vu comme une combinaison entre contexte stochastique et multi-niveaux des Champs de Markov.

De plus, leur structure permet une inférence probabiliste rapide avec contraintes logiques [60]. Beaucoup de recherches ont été faites dans le domaine d'inférence graphique, l'optimisation des algorithmes pour éviter les minima locaux. Les arbres Multi-vue représentent une alternative car l'optimum global peut être trouvé en utilisant la **programmation dynamique** [50]. De plus, dans [50] des paramètres du modèle de corps et l'aspect étaient appris en même temps [50] afin de faire face à la haute déformation du corps humain et des changements dans l'apparence.

2.4.1.4.2 Les modèles de structure:

En raison de l'efficacité des arbres et similitude entre le corps humain et les graphiques acycliques, la plupart d'entre les modèles de structure sont représentés comme des chaînes cinématiques suivant une configuration d'arbre. Contrairement aux arbres expliqués ci-dessus, dont les nœuds représentent les parties du corps, les nœuds de la structure d'arbres représentent généralement des joints, chacun paramétré avec ses degrés de liberté (DOF).

De la même manière que les ensembles de pièces sont plus souvent considérées comme en 2D. Les contraintes cinématiques de modèles de structure sont plus appropriées dans une représentation 3D. Cependant, l'utilisation de modèle 2D de structure est raisonnablement utile pour les **mouvements parallèles** par rapport au plan d'image (par exemple, l'analyse de la démarche [40]). La pose 2D est estimée dans [63] avec un modèle 2D dégénéré appris à partir des projections d'images. Dans ce cas, les mouvements parallèles sont autorisés, par conséquent, différents mouvements sont interprétés pour la même action, par exemple lors de la marche dans une direction opposée.

2.4.1.5 Les modèles temporels

Afin de réduire l'espace de recherche, la consistance temporelle est étudiée lorsque qu'une séquence vidéo est disponible. Les mouvements des parties du corps

peuvent être incorporés afin d'affiner la pose du corps ou d'analyser le comportement qui est en cours d'exécution.

2.4.1.5.1 Suivi:

Le suivi est appliqué pour assurer la cohérence parmi les poses sur la durée. Le suivi peut être appliqué séparément à toutes les parties du corps ou seulement une position présentative pour l'ensemble du corps peut être prise en compte. De plus, le suivi en 2D peut être réalisé à des positions de pixels ou être considéré quand la personne se déplace en 3D. Autre subdivision de suivi est le nombre d'hypothèses, qui peut être celui qui est maintenue sur la séquence ou multiple hypothèse qui peut être propagée dans le temps.

Le suivi unique est appliqué dans [40], où seule la partie centrale du corps est estimée au moyen des champs de Markov Cachés (HMM). Enfin la pose 2D est récupérée à partir de la position du corps. Toujours en2D, une seule hypothèse pour chaque articulation du corps(représentée sur la figure1-3)est propagée dans[52]. Dans la rubrique de récupération de forme, une formulation probabiliste est présentée dans [73], permettant de résoudre simultanément la pose de la caméra et la forme non-rigide d'un maillage dans le lot. Les positions possibles des points d'intérêt (c'est-à-parties du corps)et leurs covariances sont propagées le long de toute la séquence, en optimisant lecheminement3Dsimultanéede tous les points.

2.4.1.6 Modèle de mouvement :

Le corps humain peut effectuer une diversité de mouvements, cependant, des actions spécifiques pourraient être définies par de petits ensembles de mouvements (par exemple dans les actions cycliques tel que la marche). De cette façon, un ensemble de mouvement a priori peut décrire l'ensemble des mouvements du corps quand une seule action est effectuée. Cependant, des restrictions sur les mouvements récupérés sont ainsi établies. Un problème potentiel des mouvements a priori est que la variété des mouvements qui peuvent être décrits dépendant

fortement de la quantité et de la diversité des données d'apprentissage [63]. Les modèles de mouvement sont introduits dans [74], combinés avec les modèles de séquences de la marche et de course. Une réduction de dimensionnalité est effectuée en appliquant l'APC à travers les séquences d'angles de différents exemples, pour obtenir un suivi précis. Ce travail est étendu dans [75] pour le **golf** à partir d'images monoculaires dans un cadre semi-automatique. L'approche du processus gaussien des modèles de variables latentes échantillonnés « Scaled Gaussian Process Latent Variable Models » (SGPLVM) peut également représenter plus différemment les mouvements humains [76] cyclique (ex. marche) et actions acycliques (ex.golf), à partir de séquences d'images monoculaires.

Dans [77],par exemple, le problème d'initialisation de l'estimation de pose a été traité dans le domaine temporel. Les mouvements humains possibles ont été appris par une gaussienne.Pour réduire l'espace de recherche pour la récupération de la pose lors d'une activité telle que le ski, le golf ou le patinage.

Une point faible en utilisant l'information a priori est le sur-apprentissage sur les données d'apprentissage parce qu'ils ne peuvent être généralisées sur un petit ensemble de mouvements spécifiques.

D'autre part, la trajectoire générale basée sur la transformée en cosinus discrète (CDT) est introduite dans [78] pour reconstruire les différends. Par exemple, des visages et des jouets (représenté sur la Figure 1-3). Dans ce cas, le modèle de trajectoire est combiné avec des modèles spatiaux du suivi des objets. Les applications de ces modèles de mouvement liées à la pose humaine peuvent être trouvées dans [79].

2.4.1.7 Comportement

Le bloc de comportement dans notre taxonomie se réfère aux méthodes qui tiennent en compte des activités ou des informations particulières sur scène et le contexte, pour fournir une rétroaction à des modules de reconnaissance de la pose précédente, l'amélioration de la tâche finale de reconnaissance. La plupart des

approches décrites ci-dessus ne font pas inclure directement ce genre d'information. Cependant, les bases de données sont généralement organisées par des actions (par exemple, la marche, le jogging, la boxe [81].

En ce sens, l'élection d'un ensemble de données spécifiques pour l'apprentissage est une conséquence du choix directe ou indirecte de l'ensemble des actions que le système sera capable de détecter.

Il est à noter que le terme **comportement** « **behavior** » est utilisé pour inclure les actions et les gestes. Bien que l'analyse du comportement ne soit pas habituellement dans l'état de l'art de l'estimation de la pose, certains œuvres tiennent en compte le comportement l'activité pour estimer un ensemble précis de pièce formant la pose. Certains œuvres de la littérature récupèrent conjointement la pose et le comportement. Dans le travail de Yao etFei -Fei [54] , les auteurs incluent des informations de contexte sur l'activité humaine et de son interaction avec les objets (indiqué dans La figure 1-3 (s)) pour améliorer l'estimation de la pose finale de sujets et l'activité de reconnaissance. Il a été démontré que les ambiguïtés entre les classes sont mieux discriminées, et de meilleurs résultats sont obtenus.

De même, Andriluka et Sigal [82] étendent leurs travaux antérieurs en multi-personnes estimation de la pose 3D par la modélisation du contexte d'interaction humaine. Ils ont entrepris avec succès les résultats dans la concurrence de danse à travers les vidéos.

2.4.2 Outils de manipulation des paramètres du modèle

Le but de cette étape est fondamental puisqu'elle détermine la façon dont sont optimisés les paramètres du modèle après les avoir comparés aux observations sur l'image. Parmi les méthodes parcourues dans cette bibliographie, on distingue les approches déterministes qui consistent à optimiser une fonction objectif, ainsi que les solutions qui mettent en œuvre un apprentissage sous la forme d'une régression, de variétés, de clés de hachage ou de « shape context » . [68]

On trouve aussi les approches stochastiques qui introduisent des probabilités pour modéliser les incertitudes [69], les méthodes qui mettent en œuvre une série de règles pour constituer des assemblages cohérents entre les membres et les approches « template matching » qui consistent à rechercher les membres à partir de gabarits. [70]

2.4.2.1 Approches déterministes

La frontière entre déterministe et stochastique est parfois ténue et des choix ont dû être effectués pour constituer le classement proposé ici. Les approches qui modélisent la vraisemblance du modèle connaissant l'image ainsi que celles qui utilisent des outils probabilistiques (modèles graphiques probabilistes du corps, modèles de Markov cachés...) ont étés classées comme stochastiques.

En revanche, sont classées comme déterministes les méthodes qui consistent à adapter le modèle aux indices extraits de l'image à partir de l'optimisation d'une fonction de coût calculée selon une métrique définie. Même si cette métrique peut prendre la forme d'une probabilité ou qu'un apprentissage y soit utilisé, ces méthodes restent globalement déterministes. Dans ce cadre, l'optimisation du modèle utilise couramment au premier ordre : le gradient localselon Levenberg Marquardt [71] ,Newton Raphson [72] ou la projection du gradient de Rosen [73] et au second ordre le hessien . Alternativement, une méthode d'optimisation peut être gourmande en puissance de calcul. Les points générés par des acquisitions stéréoscopiques ou triloculaires peuvent êtres appariés avec un modèle 3D. L'algorithme d'ICP (Iterative closest point), consiste à trouver ces correspondances de manière à minimiser la distance euclidienne entre le modèle et ces points.

Une transformation géométrique est calculée d'après ces paires pour affiner les paramètres du modèle au cours des itérations. Cette étape peut précéder une optimisation dans un espace généré par une machine à vecteurs de support apprise sur les contraintes articulaires pour les prendre en compte.

Une autre approche [75] affecte le membre le plus proche à chaque point issu des données de disparité. L'optimisation est menée par la méthode des moindres carrés sur la distance globale entre le modèle et ces points. La cohérence temporelle du suivi est assurée en menant l'optimisation sur l'ensemble des images d'une séquence au lieu de le faire image par image. Une ACP sur des sujets marchant ou courant sélectionne un faible nombre des paramètres articulaires les plus représentatifs de ces deux comportements. Le mouvement est modélisé selon une combinaison linéaire des composantes principale d'après laquelle l'optimisation s'opère. Cette approche se borne au suivi hors ligne d'un sujet marchant ou courant. Le bruit contenu dans l'image des disparités peut aboutir à un débordement du modèle vis à vis de la réalité, surtout lorsque ce dernier est très précis [75]. Afin d'éviter cet artefact, les auteurs ont recours à l'ajout pondéré d'observations issues des points de contour de la silhouette.

Une approche mono-caméra consiste à comparer une silhouette extraite de l'image à la projection d'un modèle 3D . L'optimisation des paramètres s'appuie sur une vraisemblance comprenant deux termes. Le premier estime la surface en commun entre la silhouette cible et la projection du modèle au sens des moindres carrés. Le second est un terme d'attraction qui « pousse » le modèle à l'intérieur de la silhouette grâce à une distance estimée par la méthode du cheminement rapide (fast marching method). Cette méthode consiste à déplacer à vitesse égale tous les points d'un contour dans le sens centripète à la normale du contour.

Un formalisme mathématique élégant, les produits d'exponentiels de « twists » , est mis en œuvre dans le but de faire un suivi du corps . Comparé à la manipulation classique des matrices homogènes, cette nouvelle écriture simplifie la résolution des chaînes cinématiques.

Grâce à cette technique associée à un modèle de flot optique affine, l'estimation du mouvement dans le plan image génère un système d'équations permettant de trouver les paramètres du mouvement dans l'espace. La projection du modèle sur

l'image délimite des masques englobant chaque membre à l'intérieur desquels une sélection des pixels est opérée. Afin d'améliorer l'estimation du mouvement 3D, les pixels sélectionnés doivent valider le modèle de mouvement rigide local dans l'image. Les résultats expérimentaux montrent un suivi des membres rigoureux sur des courtes scènes de Muybridge. Appliquée en monoculaire, cette technique est aussi implémentable pour des captures multi-caméra.

2.4.2.2 Approches à base d'apprentissage

Ces méthodes utilisent un apprentissage préalable pour stocker des données inhérentes à la pose dans un but de comparaison avec les indices extraits de l'image test. Ces données peuvent prendre la forme de clés de hachage ou d'histogrammes log-polaires . Une seconde famille d'approches basées sur l'apprentissage consiste à établir une relation entre les indices tirés de l'image et les paramètres de la pose. Ces derniers peuvent être exprimés dans une base de fonctions vectorielles à partir d'une régression ou inclure une variable latente Gaussienne pour introduire des probabilités dans l'espace latent contenant les poses apprises.

2.4.2.3 Régressions et espaces latents

Classiquement, l'apprentissage d'une régression se fait en recherchant les coefficients a_k d'une combinaison linéaire de fonctions base $\phi_k(.)$ faisant le lien entre les indices extraits de l'image x et les paramètres de la pose y et minimisant l'erreur commise. Ce formalisme peut être utilisé avec de nombreux indices extraits de l'image, et en particulier des histogrammes d'orientations, les points de la silhouette ou des descripteurs de formes de type « shape context ». La recherche d'une fonction de transfert entre l'image et la pose peut s'accompagner d'une réduction de la dimension. L'analyse en composantes principales n'est pas adaptée aux occultations ou des tissus déformables qui introduisent des non-linéarités. Pourtant, l'idée de réduire la dimension de l'espace de recherche peut séduire vu le nombre important des paramètres à optimiser. Avec les variétés localement linéaires (LLE) un point de

l'espace de départ est exprimé comme une combinaison linéaire de ses N plus proches voisins. Les composantes principales sont calculées sur la matrice des poids de cette combinaison linéaire. Cette opération génère une variété de dimension réduite. Cette solution est utilisée pour suivre la silhouette d'un sujet animé du mouvement de la marche vu sous plusieurs angles .

Les variétés obtenues se distinguent clairement les unes des autres en fonction de l'angle choisi pour la prise de vue. Cette propriété permet de trouver l'angle de vue d'une séquence en comparant la variété obtenue à une base de variétés apprises. La fonction de transfert inverse donne alors la pose du personnage. Cependant, le nombre fini d'exemples appris crée une variété discrète procurant des sauts dans le suivi si la base n'est pas suffisamment dense.

L'ajout de probabilités par l'intermédiaire d'un bruit Gaussien engendre un modèle de processus à variable latente Gaussienne . Contrairement à l'analyse en composante principale probabiliste (PPCA) où l'apprentissage se fait en optimisant la vraisemblance de la base d'après la variable latente, on cherche à optimiser la vraisemblance d'après la fonction noyau qui exprime la similitude entre les vecteurs de la base (GPLVM et SGPLVM). Ces techniques génèrent un espace de modélisation continu et autorisent un suivi fluide du personnage. Le suivi peut être étendu à des gestes plus généraux comme les signes de guidage des avions au sol ou des gestes sportifs.

2.4.2.3.1 Comparaison à une base apprise

Quelques approches tentent de généraliser la reconnaissance aux poses quelconques en comparant les données extraites de l'image avec celles enregistrées dans une base d'apprentissage. Ces méthodes exigent des bases importantes ou utilisent des méthodes d'interpolation. Pour cette dernière approche, la base est traduite en histogrammes polaires grâce aux « shape context » pour faire de l'estimation de pose. Les histogrammes issus de l'image test sont comparés à ceux de la base d'apprentissage pour trouver la meilleure correspondance. La position 2D des

membres est déterminée en estimant, par la méthode des moindres carrés, la transformation géométrique locale entre l'image test et l'image d'apprentis- sage qui lui correspond le plus .

La position 3D des articulations est déduite grâce à l'algorithme de Taylor. Le fait d'utiliser des bases d'apprentissage de très grande taille exige de mettre en œuvre des stratégies adaptées afin d'accélérer la recherche des plus proches voisins d'une requête. Une solution est trouvée grâce au LSH, une technique de hachage qui conserve les distances entre les exemples . Cette méthode appliquée à l'estimation de poses quelconques et appelée « PSH » - Pose Sensitive Hashing_, permet de générer des clés binaires de faible dimension qui respectent les relations métriques entre les poses. Des indices extraits de l'image sont binarisés d'après un seuil optimal calculé sur les données apprises. Un sous-ensemble d'indices binaires tirés aléatoirement génère les fonctions de hachage et la comparaison d'une image test avec la base consiste à sélectionner un sous-ensemble des poses apprises d'après la similitude des clés. Celle-ci est efficacement mesurée grâce à la distance de Hamming et la pose résultat est extrapolée à partir d'une régression locale pondérée sur les plus proches voisins trouvés.

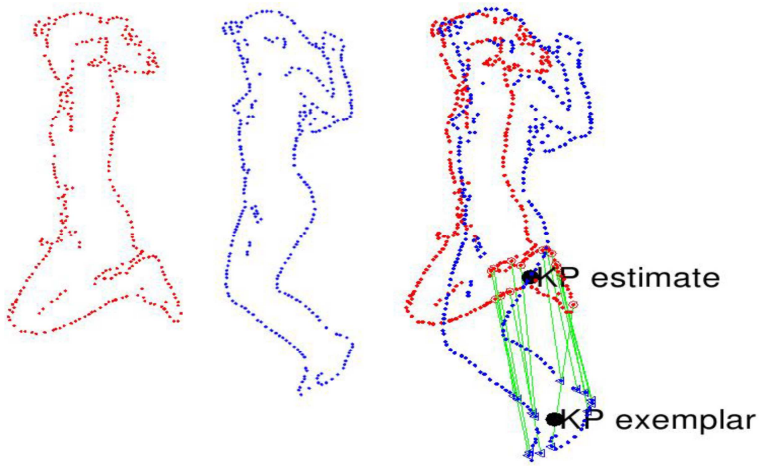


Figure 2-4 Shape context :
Localisation des articulations : Les points échantillonnés le long de la silhouette exemple (à gauche) et de test (au centre) sont mis en correspondance. Une transformation est estimée au niveau local pour retrouver le pied dans l'image test (lignes vertes dans l'image de droite).

2.4.2.4 Approches stochastiques

Le bruit induit par les caméras, les imprécisions du modèle ou les hypothèses simplificatrices génèrent des incertitudes dans les observations rendant pertinentes l'expression de la validité du modèle par une fonction de probabilité. Cette fonction étant multimodale dans un espace de grande dimension, les méthodes analytiques sont peu efficaces et on préférera discrétiser l'espace des poses en le divisant par zones ou grâce à un échantillonnage à base de MCMC (« Monte Carlo Markov Chain »), un échantillonnage d'importance (filtre à particules) ou une exploration méthodique dans l'image. Une représentation continue de la probabilité de la pose dans l'image sous la forme de noyaux Gaussiens peut être aussi envisagée.

Les approches stochastiques ont la capacité de fournir une approximation de la densité a posteriori et de propager les hypothèses pertinentes au cours des itérations. Ce principe paraît judicieux si on considère qu'une hypothèse non optimale à l'instant t peut être à l'origine de la bonne pose dans le futur. La propriété

multi-hypothèse de ces algorithmes leur permet donc de raccrocher la bonne solution après une erreur de suivi.

La résolution probabiliste d'un problème de vision par ordinateur consiste généralement à estimer la probabilité a posteriori p(x | y) des paramètres du modèle x en considérant les observations sur l'image y. À partir de cette estimation, il est possible d'en extraire le maximum a posteriori (MAP) :

$$\overset{*}{x} = \arg\max_{x} [p(x \setminus y)] \qquad \textbf{Eq2-1}$$

où l'espérance :

$$< x > = \int x\, p(x \setminus y)\, dx \qquad \textbf{Eq2-2}$$

L'écriture Bayésienne permet de décomposer l'estimation de la probabilité a posteriori :

$$p(x \setminus y) \propto p(y \setminus x)\, p(x) \qquad \textbf{Eq2-3}$$

La probabilité p(y | x) et notée L(x, y) est appelée vraisemblance. C'est elle qui est exprimée par un modèle génératif. La probabilité a priori sur le modèle p(x) représente l'ensemble des connaissances dont on dispose sur le modèle sans tenir compte des observations. Ces connaissances peuvent porter sur les liens physiques (les articulations et leur contraintes) et/ou temporels (cohérence temporelle d'un corps en mouvement), elles peuvent êtres apprises et être représentées par un mélange de Gaussiennes. Demirdjian et al [76] procèdent à une comparaison entre l'image test et une base apprise. Un processus d'optimisation est initialisé à partir des meilleures poses pour modéliser la vraisemblance avec un mélange de Gaussiennes.

2.4.2.5 Échantillonnage de la densité a posteriori

La densité a posteriori est généralement impossible à exprimer analytiquement et une approximation par échantillonnage est souvent utilisée dans la pratique. Lee et Cohen [77] font appel à un échantillonnage de type Métropolis Hastings pour estimer la densité a posteriori. La densité de proposition, classiquement un mouvement brownien, est remplacée par une fonction conditionnée par les observations. Cette technique d'échantillonnage par MCMC conduite d'après les données sur l'image (data driven Monte Carlo Markov Chain) permet de faire converger l'algorithme vers un optimum global plus efficacement.

Pour alimenter ce procédé, il est nécessaire d'extraire des cartes de probabilités pour chacun des membres recherchés. Ces « proposal maps » sont le fruit d'hypothèses pondérées par leur confiance et issues des caractéristiques extraites de l'image. Ces hypothèses sont modélisées sous la forme de Gaussiennes 2D sur l'image d'après lesquelles les échantillons sont tirés.

Dans le cadre du filtre à particules, l'approximation des distributions de probabilités a posteriori est conduite par un échantillonnage d'importance séquentiel suivant une distribution de proposition. Dans le cas de l'algorithme condensation, la cohérence temporelle est contrainte par une distribution de proposition Gaussienne centrée sur la pose trouvée à l'image précédente. Un ré-échantillonnage est nécessaire pour empêcher la dégénérescence des échantillons vers une solution unique. Le nombre de particules (les échantillons) doit être suffisant pour couvrir la plus grande partie des hypothèses plausibles. Or, la grande dimension de l'espace de recherche pour le suivi de personnes exige un nombre de particules souvent prohibitif.

Dans le but d'éviter le « street light effect », c'est-à-dire le fait de rechercher un optimum dans un espace trop restreint, une amélioration de l'étape de ré-échantillonnage peut offrir des solutions. Le fait de remarquer que la fonction de vraisemblance possède des maxima allongés sous la forme de vallées pousse à ré-échantillonner avec une covariance qui suit ces vallées grâce à la technique de « covariance » « scalled sampling ». Une seconde amélioration consiste à générer des échantillons vers les poses 2D qui présentent une ambiguïté avec la projection du modèle dans l'image. En exploitant la propriété multi-hypothèses du filtre à particules, le procédé de saut cinématique permet de raccrocher le suivi après une Désambiguïsation de la pose 2D grâce aux contraintes temporelles.

2.4.2.6 Modélisation du corps par un modèle graphique

Un modèle graphique est un graphe incluant des variables aléatoires représentés par des nœuds. Des liaisons forment un système de voisinage entre ces nœuds pour exprimer des couplages entre les variables aléatoires. L'intérêt d'un tel modèle se trouve dans sa capacité à mettre en évidence des relations d'indépendance entre les nœuds, lorsqu'il y a absence de liens, pour factoriser l'expression de la probabilité jointe sur l'ensemble des variables aléatoires du graphe.

Dans le contexte du suivi du corps humain, les nœuds modélisent souvent l'état des membres ou des articulations.

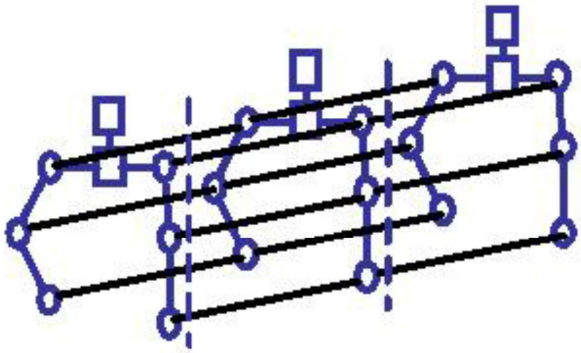


Figure 2-5Champ de Markov intégrant une fenêtre temporelle sur trois images

Le réseau Bayésien consiste en un graphe acyclique orienté. Il est commodément utilisé dans le cadre d'approches ascendantes pour modéliser les liens entre les membres du corps par une structure arborescente. Une interprétation Bayésienne des « structures picturales » adaptée aux problèmes d'estimation de la pose décrite précédemment. La recherche des membres consiste à calculer l'orientation des contours appartenant à la projection d'un modèle sur l'image. Pour chaque pixel de la projection, la distance qui le sépare du plus proche contour de l'image ayant une orientation similaire est mesurée.

Le résultat de la moyenne des distances par une exponentielle négative fourni une probabilité qui, mêlée à une probabilité d'apparence basée sur les couleurs et apprise sur les trente premières images, donne la probabilité de similitude du modèle. Plusieurs hypothèses sont retenues pour chaque image avant d'être départagées par l'algorithme de Viterbi. Pour une seconde approche, la recherche des membres est facilitée par le choix des photos de Muybridge où le corps apparaît généralement en clair sur un fond sombre. Le problème de reconnaissance suit une procédure classique inspirée du modèle de réseau Bayésien.

Si le corps comprends k membres et xi représente la configuration du membre i, la probabilité jointe P(x1, ..., xi, .., x_k)est factorisée en produit de probabilités conditionnelles du membre xi connaissant le membre parent xi−1. La probabilité de la racine xroot détermine la pose globale du corps et la probabilité jointe est :

$$p(x_1,...,x_i,...,x_k) = p(\mathrm{x}_{root}) \prod_{i \neq root} p(\mathrm{x}_i \setminus \mathrm{x}_{i-1}) \qquad \textbf{Eq2-4}$$

Dans le but d'établir un modèle plus général, des travaux font appel à une mixture d'arbres afin d'inclure des modèles d'arbres tronqués qui correspondent aux différents cas d'occultations.

Un champ de Markov aléatoire ou Markov Random Field (MRF) est un graphe non orienté. Il est utilisé dans [79] pour modéliser l'influence des contraintes articulaires ou temporelles grâce aux liaisons. Cette modélisation permet d'exprimer la probabilité jointe comme un produit de facteurs sur les cliques maximales (théorème de Hammersley-Cliford). L'utilisation d'un graphe de facteurs offre la possibilité de diviser les cliques pour ne garder que des facteurs entre paires de nœuds. Cette simplification est avantageusement exploitée pour parvenir à un algorithme de suivi temps réel.

2.4.2.7 Modèles de Markov cachés temporels

Les modèles de Markov cachés temporels (Hidden Markov Model - HMM) consistent en un graphe orienté dont chaque nœud représente l'état du système à un instant donné. Ces états sont raccordés par un réseau de liens qui détermine les probabilités de passage. Ce modèle est adopté par Lan et Huttenlocher [81] pour suivre un sujet animé de la marche. Il comprend les postures clés de la marche observées depuis huit angles de vue, décalés chacun de quarante-cinq (45°) degrés.

Les observations permettent d'associer chaque image avec un état du graphe d'après la distance de chanfrein entre le modèle « cardboard » et la silhouette. Le parcours dans la chaîne de Markov est retrouvé à l'aide de l'algorithme de Viterbi pour contraindre les règles articulaires et la cohérence temporelle du résultat.

Un réseau HMM est également adopté pour assurer la cohérence temporelle au cours d'un suivi de gestes quelconques [82]. Les hypothèses sont générées à partir d'un modèle hiérarchique en arbre qui représente le corps. Le torse est recherché en premier, suivi des membres qui s'y rattachent. Une autre approche consiste à utiliser des champs aléatoires conditionnels ou Conditional Random Fields (CRF) [83] pour estimer aisément la probabilité a posteriori. Cette approche à base d'apprentissage consiste à propager temporellement les hypothèses formulées dans un espace discrétisé sur les poses apprises. La vraisemblance est évaluée d'après un hachage de type LSH pour accélérer la comparaison des hypothèses avec l'image test. Cependant, pour considérer la fonction de probabilité constante dans les zones délimitées par les poses apprises, l'apprentissage doit être dense et exhaustif.

2.4.2.8 Propagation des croyances

Appliquée aux modèles graphiques, la propagation des croyances [84] permet d'estimer la probabilité marginale de chaque membre. Même dans le cas de graphes bouclés, il est montré que cet algorithme est capable de converger vers une bonne approximation probabiliste de la pose après un processus itératif [85]. Cette approche montante a l'avantage de limiter la dimension de l'espace d'investigation au nombre de *DDL* du membre recherché.

L'algorithme assure alors la cohérence de la solution fournie vis-à-vis des contraintes articulaires grâce à des facteurs de compatibilité entre les membres. Ces facteurs vont générer des messages qui sont propagés à travers le graphe pour exprimer l'influence de chaque membre sur le reste du corps. Cette méthode est adoptée pour assurer un suivi récursif sur une séquence d'images [86]ou en

intégrant au modèle graphique une fenêtre temporelle et des contraintes de non-collision entre les membres [86].

La propagation de croyance s'exprimant plus aisément dans les espaces discrets, les auteurs qui y font appel utilisent un espace discrétisé par une grille [79] ou des filtres à particules qui fournissent un ensemble d'échantillons pondérés pour explorer la vraisemblance du modèle.

Dans le cas d'un espace continu, les messages peuvent être modélisés par des mélanges de Gaussiennes [87] et leur mise à jour crée une explosion du nombre de gaussiennes du fait de la multiplication des mélanges entre eux. Pour empêcher cela, les auteurs font appel à un échantillonneur de Gibbs [88]. Une approche analogue à la propagation des croyances consiste à mêler l'algorithme du champ moyen aux techniques de Monte-Carlo [89]. Cette approche qui utilise les niveaux de gris et les contours est limitée aux poses en 2D du fait de son modèle.

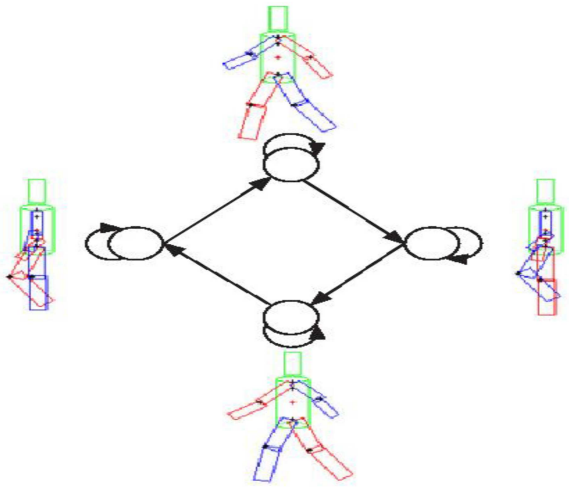


Figure 2-6Modèle de Markov caché comportant

2.4.2.9 Approches multicritères à base de règles

Ces approches se basent sur un ensemble de règles qui s'appuient sur plusieurs indices pour Constituer des assemblages cohérents en estimation de pose.

La fusion des observations sur l'amplitude des contours, la forme et le gradient de luminosité créé par l'ombre sur un membre comparé à une base d'apprentissage, permet de trouver de manière opportuniste un ensemble de membres candidats sur une image segmentée.

Le nombre de combinaisons est réduit en faisant des hypothèses sur la longueur moyenne des membres, leur emplacement (pas de membres isolés) ou la symétrie des vêtements.

Une approche analogue consiste à désigner les membres candidats à partir d'hypothèses sur des lignes parallèles issues d'une extraction de contours suivie d'une triangulation de Delaunay [90]. Similairement à l'algorithme précédent, dans [91] les critères de sélection sur les contours incluent l'orientation, la longueur, l'amplitude ou la distance entre les centres des segments. La réduction du nombre d'hypothèses peu probables se fait sur des critères anthropomorphiques et des contraintes de connexion apprises sur une base issue de séquences de patinage artistique. L'intégration d'un module d'intelligence artificielle par l'intermédiaire de tableaux noirs hiérarchisés permet de rationaliser le processus de désignation et d'élimination des membres candidats [92]. Ce processus est conduit par trois niveaux de sources de connaissance, chacun spécialisé dans une tâche précise. Au niveau le plus bas, les spécialistes sont chargés de détecter les membres de manière opportuniste : bras, torse, tête,etc. Le niveau intermédiaire a pour rôle de construire des membres complets : épaule + bras + avant-bras +main. Le plus haut niveau doit reconstituer un buste cohérent avec ces éléments. Les niveaux inférieurs sont sollicités par les niveaux supérieurs en fonction du contenu des tableaux noirs de manière à avoir assez d'hypothèses pour réussir la reconstruction. Un ensemble de règles d'ordre et d'adjacence sur les membres permet de mener à bien ces tâches.

2.4.2.10 Approches à base de « template matching »

Ces approches recherchent directement les membres [93] ou le corps entier [80] à partir de la scrutation de l'image avec des gabarits adaptés en utilisant des masques probabilistes adaptés aux membres recherchés ou des gabarits spatio-temporels pour détecter un sujet marchant [80].Les masques probabilistes [93] consistent à calculer, au centre et sur les bords du masque, deux histogrammes de couleurs. L'homogénéité chromatique de la vignette est testée en comparant le résultat à un modèle appris pour distinguer les cas où le masque est centré sur un membre (fortement homogène) ou non (peu homogène). Des contraintes sur les liaisons entre les membres et la symétrie des vêtements permettent d'affiner le processus de décision.

Dans le cas des gabarits spatio-temporels, la base apprise est composée des contours de la silhouette d'un sujet marchant sur un tapis roulant et vu de plusieurs angles. Une séquence de trois silhouettes consécutives autour d'une position clé pour laquelle les deux pieds touchent le sol constitue un élément de la base d'apprentissage et donc un gabarit. La comparaison entre la base apprise et la séquence test est opérée à partir du calcul d'une distance de chanfrein robuste qui pénalise les contours possédant une orientation trop différente.

2.5 Conclusion

Il existe un nombre important de travaux sur l'estimation de la pose humaine en 2-D et 3-D. Les articles du mémoire associés font référence à un certain nombre d'entre eux. Ici nous nous contenterons de nous rapporter brièvement aux travaux qui nous semblent être les plus pertinents, sans entrer dans les détails. On voit apparaître deux lignes de pensée qui s'appliquent à chaque problème: des approches « **articulées** » basées sur un **modèle géométrique explicite**; et des approches « **par apparence** » ou « **exemplaires** », basées en l'essentiel sur la sélection d'exemples similaires dans une base d'apprentissage, et sans modèle géométrique explicite.

Pour la détection de personne (corps entiers) sur des images, les approches « basées modèle » prônent les représentations basées sur les modèles 2-D articulés explicites – par exemple le « scaled prismatic model », pendant la phase de détection, le modèle doit être optimisé sur l'image, ce qui est souvent fait par un processus de programmation dynamique.

De l'autre côté, il y a les approches « détecteurs de piétons » qui prônent une modélisation directe de l'apparence et/ou une comparaison avec les exemples. Pour l'essentiel ce sont des méthodes à gabarits plus ou moins rigides, mais généralisées par une phase d'apprentissage, et avec en option une décomposition par fenêtres locales de l'objet à détecter afin de donner un peu plus de flexibilité au plan spatial. Certains auteurs apprennent à partir d'exemples des règles de décision globales telles que les machines à vecteur de support, d'autres développent des algorithmes efficaces pour comparer explicitement l'image avec une série d'« exemplaires », afin de décider si ou non il y a une personne présente.

Nous avons tenté l'expérience avec le domaine de l'estimation de pose, en exploitant une méthode sans utilisation de modèle explicit, en utilisant les informations de bas niveau de l'image tel que le contour, les coins, mouvement, etc. Ainsi que des données sur la morphologie humaine. Cette méthode nous a servi dans le cas où la localisation des bras n'est pas nécessaire tel que l'exploitation de l'application du domaine de détection de l'orientation des piétons.
Une fois nos perspectives élargis cette méthode s'avère limitée. Pour répondre à nos ambitions, nous avons proposé à une nouvelle approche **discriminante**, exploitant l'histogramme de gradient orienté HOG comme **descripteurs ,**ce dernier comme cité préalablement, a été manipulé introduit dans la première fois dans le domaine de la vision par ordinateur , pour la détection des piétons par Dala & Triggs ; Le choix

de ce descripteur est justifié par son efficacité à parer aux problème de l'apparence variée.

Notre approche est fondée sur le principe de « **Pictorial Structure** » : Une approche probabiliste exploitant un modèle 2D**déformable**,modélisant à la fois l'apparence de l'objet ainsi que la configuration spatiale des sous ensemble de l'objet, dans notre cas les parties du corps humain ,considéré également comme un réseau bayésien.

Le modèle définit une distribution de probabilité a posteriori sur les poses humaines dans une image d'entrée. Pour une image donnée, nous échantillons la configuration du corps humain en entier. Les configurations échantillonnées de la distribution a posteriori d'une manière efficace. Les configurations échantillonnées sont ensuite classées en fonction de leurs probabilités a posteriori et combinés par la suite par les configurations de la racine, dans notre cas le filtre représentant le corps en entier.

La notation finale de la pose du corps et la meilleure sélection de candidat est un problème difficile de lui-même.

Chapitre 2 : Motivation et Challenge

3.1 Introduction :

L'activité de l'humain est caractérisée par la pose adoptée pour accomplir une tâche, càd à la configuration de chaque partie de son corps. La question qui nous motive dans ce travail est comment assurer un moyen d'interaction homme-machine via un moyen naturel qui est la pose humaine. Étant donné une image, **comment détecter et estimer la pose 2 D d'une personne**, ce qui revient à récupérer la configuration spatiale des parties de son corps.

L'estimation de la pose humaine constitue un grand challenge dans le domaine de la vision par ordinateur. Ce challenge est relatif à plusieurs points liés à la structure humaine ainsi qu'aux problèmes inhérents au domaine de la vision par ordinateur.

Sachant que l'activité d'une personne est caractérisée par sa pose, cette dernière est caractérisée par la configuration (orientation, position) de chaque membre du corps ; Nous tenons à relever ce défi et d'implémenter une solution efficace capable d'extraire la silhouette d'une personne sur image et raffiner l'extraction par la suite en la localisant les différentes parties du corps humain.

Le système proposé exploite une approche discriminatoire basée modèle à structure déformable permettant de codifier à la fois l'apparence et la configuration de la partie du corps passant par deux processus , un d'entre eux détecte le corps humain en sa globalité faisant recours à l'approche de Dalal et triggs et le deuxième processus recherche les parties du corps.

L'algorithme de l'Histogramme de Gradient orienté a été introduit pour modéliser l'apparence des objets, l'idée est que l'apparence locale et la forme d'objet dans une image peut être décrite par la distribution d'intensité des gradients ou de direction des contours.

Dans ce qui suit nous allons présenter, notre intérêt d'entreprendre un tel challenge face aux problèmes rencontrés, ainsi qu'une description générale de l'approche proposée.

3.2 Motivation

La détection automatique des humains et la localisation de leurs parties du corps constituent un grand challenge dans le domaine de la vision par ordinateur. La solution à cette problématique est exploitée dans de domaines diverses tel que : Navigation robot, vidéo surveillance, Machine-humain interaction, mesure des performances des athlètes et patients, réalité virtuelle.

Dans ce travail, nous nous intéressons plus spécialement au domaine de l'interaction homme-machine.

La perception du mouvement humain, non verbale, un aspect important de l'interaction homme-Robot. Pour permettre aux robots de devenir des collaborateurs fonctionnels dans la société, ils doivent être en mesure de prendre des décisions en fonction de leur perception de l'état de l'humain. De plus, les connaissances sur l'état de l'humain est crucial pour les robots d'apprendre des stratégies de contrôle et de l'observation directe de l'homme.

L'activité humaine est caractérisée par la pose adoptée par l'humain, cependant l'état de l'humain, englobe une grande et une diversifié de variables, y compris cinématique, affectif, et de l'information axée sur les buts , s'avérant difficiles à être modéliser et à inférer.

Notre conviction principale est de considérer que l'interaction sociale avec les robots doit être mesurée par des distributions probabilistes. Un humain, dans ce processus d'interaction, fait une multitude de geste affectant plusieurs membres or que le robot ne peut percevoir qu'une' information limitée sur l'apparence et la

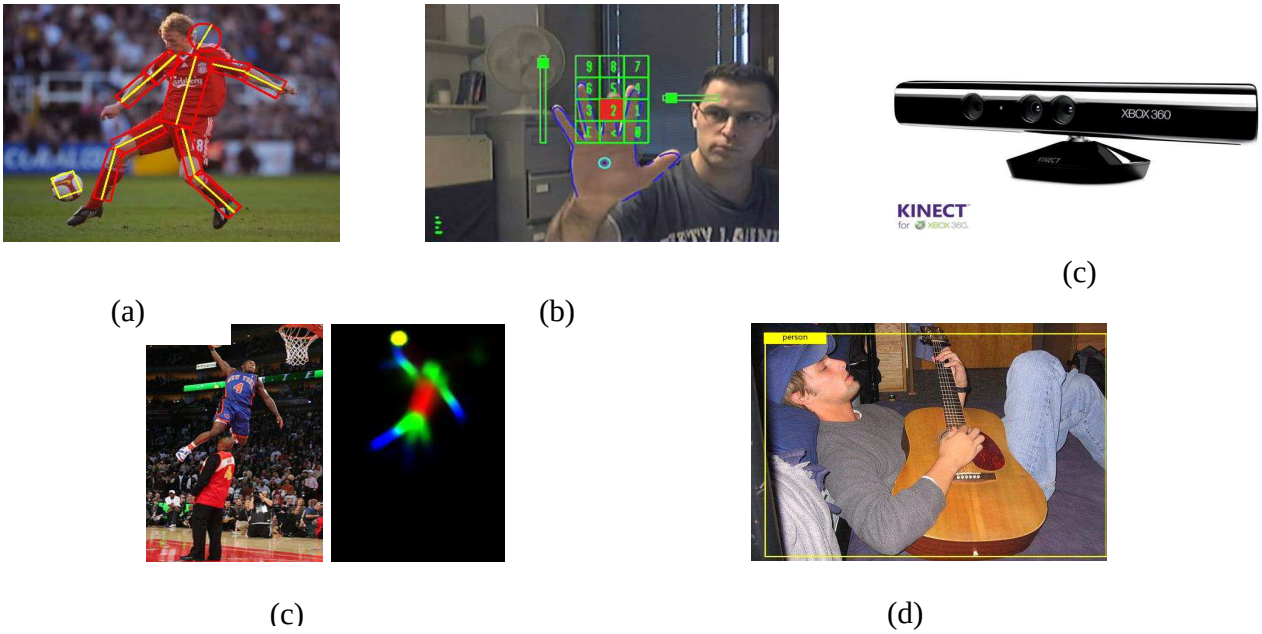
cinématique de l'humain provenant des images. Un robot doit inverser les informations partielles provenant de l'humain à travers son modèle en injectant des croyances préalables sur la cinématique et la morphologie de l'humain pour prendre de décision sur la pose à estimer pour l'humain.

Dans ce qui suit, nous étalons nos objectifs ainsi que les différentes lacunes liées à un tel challenge, en répondant aux questions ci-dessous :

- Quel intérêt d'utiliser la vision par ordinateur pour le domaine de l'estimation de la pose?
- Pourquoi la difficulté d'analyser une pose ?
- Qu'est ce qui est voulu dire par pose et quel niveau de détail est approprié?

Figure 3-1Détermination de l'activité par la pose

Figure 3-2interaction homme machine



(a) (b) (c)

(c) (d)

(a) Reconnaissance d'actions, (b) Interface homme-machine, (c) jeux et loisir, (d) segmentation, (e) reconnaissance d'objet.

Figure 3-3 Domaine d'exploitation de l'estimation de la pose humaine

3.3 Description générale du système

Cette thèse a pour objectif d'exploiter le domaine de la vision par ordinateur dans le domaine de l'estimation de la pose humaine en implémentant un système qui accepte comme entrée une séquence d'images et en sortie un vecteur de pose des parties du corps prédéfinie par un modèle.

L'estimation de la pose humaine se concrétise par la localisation de la personne sur l'image, et la détermination de la position ainsi que l'orientation des différentes parties de corps.

Dans nos propos, nous nous intéressons aux parties de corps suivantes : Tête, bras supérieurs et inférieurs, torse, cuisses et jambes ; Nous exploitons une approche basée modèle.

Nous adoptons les directives suivantes :

- Exploitation d'un système d'acquisition d'image simple : Une webcam
- Le sujet étudié doit être soumis au minimum de contraintes possibles : Pas de contrainte sur la pose , pas de contrainte ou spécification spéciale sur l'habillage du sujet , pas de contrainte sur la scène, etc.
- Le système doit être automatique sur le choix du résultat, sans autant faire recours à l'introduction de l'humain pour le choix de la meilleure pose à partir de plusieurs propositions.
- Pour approche basée modèle, on spécifie explicitement un modèle qui rapporte les contraintes entre les parties du corps. Ces contraintes ne sont pas limitées à des contraintes cinématiques des articulations, ou leurs structures, mais inclut aussi des contraintes sur l'apparence, d'échelle et ainsi des contraintes angulaires entre eux.

3.3.1 Schéma synoptique de l'interface de la pose humaine

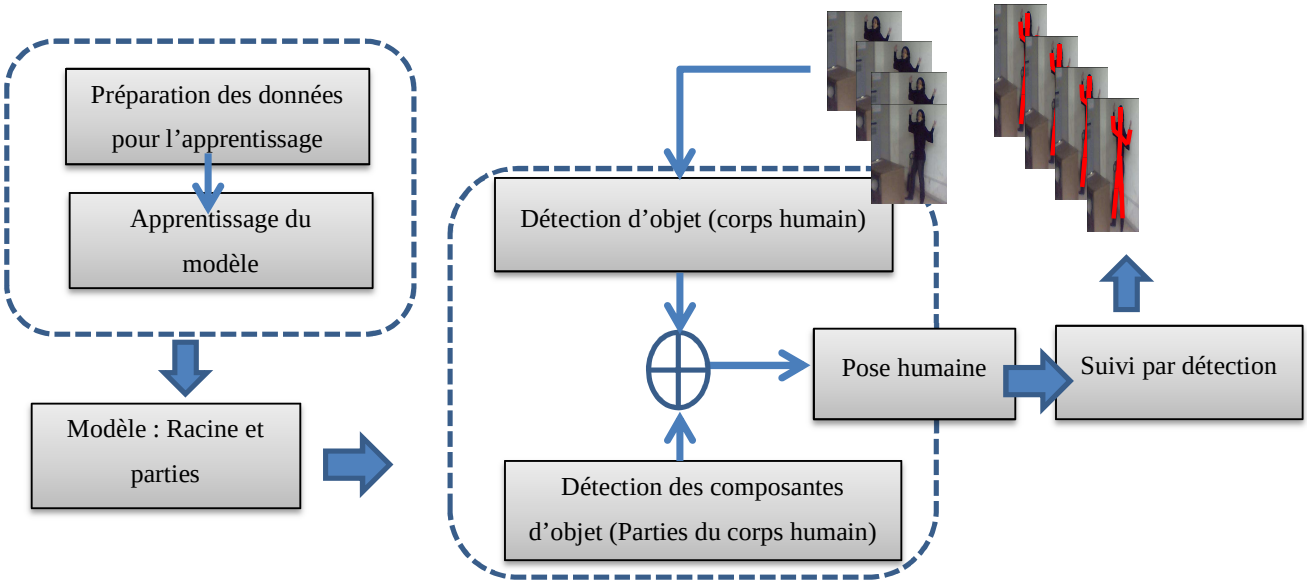


Figure 3-4 Schéma synoptique de l'approche proposée

3.3.2 Formulation de problème :

Afin de formaliser notre problématique, nous adoptons en premier lieu l'équation ci-dessous :

$\Phi : W_c \rightarrow I_c$

Φ : C'est le processus de projection une image le monde réel W_c à une image I_c, contient de nombreux processus qui contribuent à la façon d'une image est formée en prenant en compte le facteur d'éclairage, la perspective, le modèle de la caméra, à la lumière, configuration 3D de la scène .

Pour cette raison de complexité, l'estimation de pose automatique du corps humain utilise certains composants d'extraction de caractéristiques qui tentent de filtrer

l'image d'entrée pour ne sauvegarder que les paramètres 3D de la configuration humaine.

Supposons que nous nommons tous les paramètres spécifiques à une pose humaine θ et le processus de filtrage dans l'espace de caractéristique χ.

Le mappage de l'espace de configuration à l'espace de caractéristique est spécifié comme suit :

$$M\ (\theta_m) \rightarrow\ x_n \qquad \textbf{Eq3-1}$$

Où $\theta_m \in \theta$ est une instance dans l'espace de pose et $x_{n\times 1} \in \chi$ est une instance dans l'espace de caractéristiques.

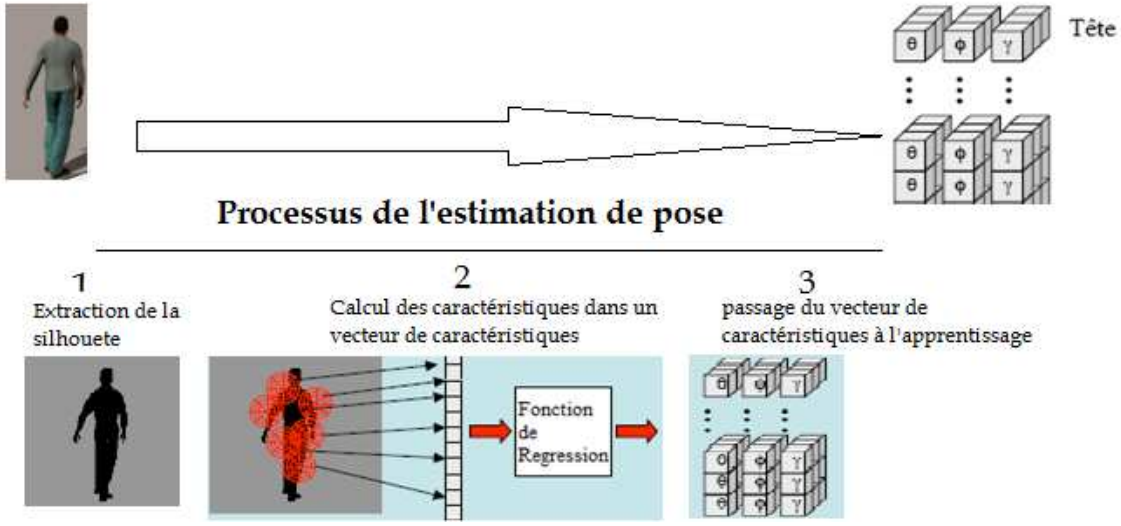


Figure 3-5 illustration du la formulation du problème

3.4 Contraintes :

L'estimation de la pose humaine constitue un grand challenge, ce dernier est alimenté par plusieurs contraintes :

3.4.1 Choix du niveau de représentation de la pose humaine

Beaucoup d'applications sont à envisager, et cela dépend du degré de l'interprétation de la pose humaine. Un système d'estimation de pose parfait doit être **moins couteux, non intrusif et précis** Malheureusement de tel systèmes sont rares, par exemple pour atteindre un haut niveau de précision il faut faire recours à des systèmes très développés ce qui rend le système très couteux.

Nous distinguons trois niveaux de représentations soient :

- **Echelle basse :**

Un exemple des systèmes de détails bas, les systèmes de télésurveillances et les systèmes intelligents ; Ces derniers exigent le temps réel, la robustesse et opèrent dans un intervalle de temps très grand. Ces systèmes ne sont pas intrusifs et n'exigent pas que l'environnement soit structuré. Une personne est un blob couvrant un nombre réduit de pixel.

De telles contraintes engendrent un détail bas de la représentation de la pose humaine, tel que fournir la position de l'humain dans la salle.

- **Échelle moyenne :**

A ce niveau les parties du corps humain couvrent un grand nombre de pixel pour être détectées**.** L'information en temps réel dans de tels systèmes est très importante.

- **Echelle haute** :

L'exemple des applications dans ce domaine : L'analyse de la performance des sportifs, le domaine médical. La représentation de l'humain est généralement en 3D , il est à noter que tel système requière des environnements spécialisés, un temps de calcul considérable ce qui rend de tels systèmes en offline.

3.4.2 Dimension de l'espace :

On exige généralement un vecteur d'état dimensionnel élevé pour décrire la configuration du corps dans un tel cadre. Par exemple, supposons que nous décrivons une personne en utilisant une représentation en 2D. Chacun de dix segments du corps (torse, tête, bras supérieur et inférieur et les jambes) sera représentée par un rectangle de taille fixe . Cette représentation aura recours à un minimum absolu de 12 variables d'état (position et l'orientation pour un rectangle, et l'orientation relative de chaque autre). Un plus dans la pratique chaque rectangle représentant une partie du corps doit être capable de glisser en fonction de son parent, ce qui nous oblige à gérer 27 variables d'état.

3.4.3 La dynamique:

Il y' a de bonnes preuves que des mouvements tels que la marche ont prévisibles, à faible structure tridimensionnelle. Cependant, le corps peut se déplacer extrêmement rapide (Figure 2-6), avec de fortes accélérations. Ces accélérations importantes signifie que l'on peut arrêter de bouger très vite prévisible - par exemple, sauter en l'air lors d'une promenade. Pour des raisons simples mécaniques, les parties du corps qui se déplacent le plus rapide ont tendance à être faible et à une extrémité d'un long levier qui a de gros muscles à l'autre extrémité (avant-bras, les doigts et les pieds, par exemple). Cela signifie que les segments du corps que le modèle dynamique ne parvient pas à prédire vont être difficiles à

trouver parce qu'ils sont petits. Par conséquent, un suivi précis de l'avant-bras peut être très difficile.

Figure 3-6 Humain qui court rapidement

3.4.4 Des phénomènes d'apparence complexes:

Dans la plupart des applications on est le suivi des personnes vêtues. Vêtements pouvez modifier l'apparence de façon spectaculaire comme il se déplace, parce que les forces du corps s'applique à la modification de vêtements, de sorte que le motif de plis, causée par flambage, les changements. Il ya deux résultats importants. Tout d'abord, le modèle des occlusions de changements de texture, ce qui signifie que la texture apparente du segment du corps peut changer. Deuxièmement, chaque pli aura un motif typique ombrage attaché, et ces motifs se déplacer dans l'image que les plis se déplacer sur la surface. Encore une fois, le résultat est que la texture apparente des changements des segments corporels.

Figure 3-7 Des exemples de diverses apparences dues aux vêtements

3.4.5 Association de données :

Il n'y a généralement pas de couleur ou de texture distinctive qui identifie une personne. Un repère possible est que de nombreux segments du corps apparaissent à l'échelle distinctive des régions étendues avec des côtés plutôt à peu près parallèles. Ce n'est pas trop utile, car il y'a beaucoup d'autres sources de ces régions. Des fonds texturés sont une source particulièrement riche de structures de fausses cartes de pointe. Une grande partie de ce qui suit porte sur des méthodes pour traiter les données des problèmes d'association pour le suivi des personnes.

3.4.6 Variation de mouvement :

Le corps humain contient 230 d'articulations mobiles ou légèrement mobiles, et si on se contente uniquement de six degrés de liberté pour chacun d'eux il donne un total de 1380 DOF

Figure 3-8 Variation de mouvement humain

Figure 3-9 Variation par rapport à la rotation

Figure 3-10 Variation par rapport à la perspective

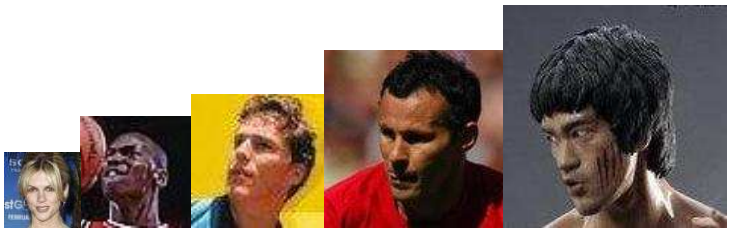


Figure 3-11 Variation par rapport à l'échelle

Figure 3-12En dehors du plan de rotation

Figure 3-13 Variation par rapport au format

3.4.7 Une très grande variation Intra-classe :

Nous distinguons une très grandes variété intra-classe , par exemple , si nous prenons l'exemple du visage , nous avons des image de profile, face, avec moustache ou sans, etc.

Figure 3-14 Variété Intra-classe

3.4.8 Occlusion, problème de luminosité, etc. :

Additivement aux problèmes liés précédemment, nous distinguant d'autre problème liée à la nature non rigide du corps humain ainsi que d'autre liée au domaine de traitement d'images.

Figure 3-15 Problème d'ombre

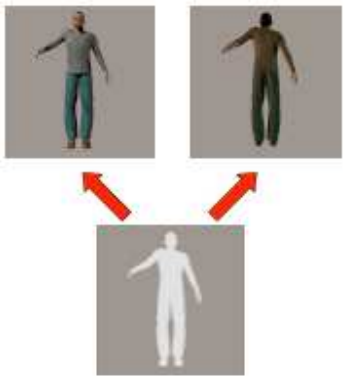


Figure 3-16 Ambiguïté (de face ou de dos

Figure 3-17Problème d'occlusion

3.5 Conclusion

Exploitée dans plusieurs applications dans le domaine de la vision par ordinateur, l'estimation de la pose humaine constitue un grand challenge vue ses innombrables contraintes liées à la fois au traitement d'images ainsi qu'à la morphologie de l'humain.

Intéressé par l'exploitation de la pose humaine dans le domaine de l'interaction homme-machine, nous avons lancé ce défi, en fournissant une contribution dans tel domaine.

Notre approche est fondée sur deux concepts dans la vision par ordinateur : Histogramme de gradient orienté pour la détection des piétons ainsi que les structures picturales, en manipulant ainsi un modèle déformable, modélisant à la fois l'apparence et la configuration spatiale des parties du corps.

Dans le chapitre suivant, nous allons aborder l'histogramme de gradient orienté , choisi dans nos travaux comme descripteur de l'apparence.

Chapitre 3 : Histogramme du gradient orienté (HOG)

4.1 Introduction

Le choix d'un descripteur est très crucial pour la performance d'un détecteur d'objet : il doit être le plus représentatif, le plus discriminatoire et le plus rapide en termes de temps de calcul.

L'approche proposée dans ce travail est basée sur deux grands concept de la vision par ordinateur : Détection de personne par le descripteur de **l'histogramme de gradient orienté (HOG)** et l'approche **Structure Picturale**, s'inspirant ainsi par le concept : Chercher un objet articulé par la recherche de ses composantes, dans notre cas les parties du corps humain.

Balayant l'image par l'approche : Fenêtre coulissante « Sliding Window » , à la recherche d'une personne sur l'image en exploitant l'algorithme proposé par Dalal & Triggs [37], une fois l'objet détecté nous entamons la localisation des composantes du corps humain , dans notre cas : Tête, Bras, Jambes, Torse.

Nous nous inspirons de l'approche de Dalal &Triggs pour l'utilisation de l'Histogramme de gradient orienté comme descripteur. Permettant ainsi une classification plus fiable qui doit capturer les similitudes essentielles entre les objets de la même classe et les différences avec des objets de classes concurrentes en se basant sur l'apparence de l'objet. Ces descripteurs ont l'exclusivité de mieux représenter la structure interne d'un objet via l'information du gradient, permettant ainsi de surmonter les problèmes liés à l'apparence de l'objet : pose, éclairage, occlusion, texture de fond, etc.

Ayant comme entrée du système des images du monde réel avec l'information de l'emplacement des boites englobantes des parties du corps et de la personne. Ces

fenêtres sont représentées par la suite, en utilisant le descripteur de HOG, finalement les fenêtres représentatives sont échantillonnées à des échelles différentes pour obtenir les modèles définitifs des parties composantes du corps humain.

4.2 Définition

Les descripteurs HoG sont introduits par Navneet Dalal et Bill Triggs, chercheurs à l'INRIA de Grenoble, à la conférence CVPR de juin 2005dans leurs travaux de détection des piétons. L'idée essentielle derrière l'histogramme de gradient orienté c'est que l'apparence locale et la forme d'objet dans une image peut être décrite par la distribution d'intensité des gradients ou de direction des contours. La mise en œuvre de ces descripteurs peut être obtenue en divisant l'image en petites régions connectées, appelées cellules, et pour chaque cellule on calcule un histogramme des directions de gradient ou des orientations de contour pour les pixels dans la cellule. La combinaison de ces histogrammes représente alors le descripteur.

Le descripteur HoG maintient quelques avantages clés par rapport aux autres méthodes. Puisque le descripteur histogramme de gradient orienté « HoG » opère sur les cellules localisées, la méthode maintient l'invariance à des transformations géométriques et photométriques, ces changements ne feront leur apparition que dans les larges régions d'espaces.

4.3 Notions de base

Dans le but de mieux comprendre le principe de l'histogramme de gradient orienté, nous faisons un tour sur le principe de l'application de la fonction de gradient.

Soit $f(x_1,...,x_n)$ une fonction arbitraire avec n paramètres. Un gradient ∇ de cette fonction est comme suit :

$$\nabla f = \begin{pmatrix} \frac{\partial}{\partial x_1} f \\ . \\ . \\ . \\ \frac{\partial}{\partial x_n} f \end{pmatrix} \qquad \textbf{Eq4-1}$$

Ainsi, un gradient est un vecteur avec les dérivées partielles de la fonction f comme entrées. Nous tenons à souligner que le terme de gradient en général ne décrit pas seulement un vecteur, mais un Champ de vecteurs, un pour chaque combinaison de paramètres d'entrée dans la fonction d'origine. Nous ne discuterons uniquement des fonctions de 2 dimensions, comme une image numérique est essentiellement une fonction de deux paramètres x et y d'entrée indiquant une position de pixel sur une grille discrète et une valeur d'intensité / triple en tant que sortie.

Un vecteur de gradient à 2 dimensions, dans un plan a deux propriétés importantes, à savoir une **amplitude** ou la **longueur** et une **direction**. La grandeur d'une intensité à 2 dimensions d'un gradient G (x, y) est donnée par

$$r_{x,y} = \sqrt{G_x(x,y)^2 + G_y(x,y)^2} \qquad \textbf{Eq4-2}$$

Et sa direction θ est calculée comme suit :

$$\theta(x,y) = \arctan\left(\frac{G_y(x,y)}{G_x(x,y)}\right) \qquad \textbf{Eq 4-3}$$

Comprendre le principe du gradient est fondamental pour comprendre notre approche, nous discuterons des gradients dans ce contexte plus vivement : Une image peut être considérée comme une fonction discrète à 2 dimensions. Les deux paramètres d'entrée x et y décrivent les coordonnées de chaque pixel de cette fonction, et la valeur d'intensité à une position de pixel est le résultat de la fonction.

La taille d'une image est finie, ce qui résulte qu'on peut calculer un nombre fini de vecteurs de gradients pour cette image. Le gradient est défini par les dérivées

partielles d'une fonction, de sorte que la question se pose de la façon dont une fonction discrète comme une image peut être obtenue. Dans les algorithmes de traitement d'image, la dérivée d'une image peut être approchée par le biais d'un filtrage différentiel.

Citons à titre d'exemple, l'algorithme le plus populaire de Sobel , adopté pour notre approche, défini comme ci-dessous :

$$S_x = \begin{pmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{pmatrix} \qquad \textbf{Eq4-4}$$

Et

$$S_y = S_x^T \qquad \textbf{Eq4-5}$$

Pour la dérivée de l'image en direction x et y, respectivement. Ces filtres Sont appliqués à une image via une convolution discrète, donnée par

$$(I * F)(x, y) = \sum_{j=-1}^{1} \sum_{i=-1}^{1} I(i, j).F(x - i, y - j) \qquad \textbf{Eq4-6}$$

Tandis que I (x, y) représente l'image, F (x, y) est un filtre arbitraire. Notez que la réponse au filtre peut être plus petite que l'image originale, comme les pixels du bord n'ont pas assez de pixels voisins pour calculer une réponse correcte. Ceci dépend de la nature de l'application si les pixels voisins manquants sont interpolés ou le calcul du gradient de pixels de bordure est complètement ignoré.

Afin de calculer le gradient d'une image, deux filtres gradient sont appliqués à l'image, un pour chaque dimension. Un vecteur de gradient pour un pixel (x, y) est alors formé en combinant les entrées des deux cartes de réponse du filtre à la position (x; y) dans un seul vecteur.

L'idée de base d'introduire la notion d'intensité de gradient dans nos travaux est d'exploiter l'information de base liée à l'image qui le contour : L'amplitude d'un vecteur de gradient d'une région homogène d'une image est égale à zéro. Un vecteur de gradient sur un contour n'a pas seulement une valeur élevée indiquant l'intensité de ce contour, mais il représente également la direction du contour car il a toujours une orientation perpendiculaire à ce contour.

La prise en compte de tous les vecteurs gradients d'une image pour la détection de l'objet n'est pas utile pour deux raisons: (1) Temps de calcul trop élevé, (2) Un modèle qui tient en compte toutes les informations du gradient d'un objet ne serait pas efficace contre les transformations liées à la rotation, mise à l'échelle, le déplacement et la déformation de l'objet dans différentes images. C'est pourquoi les groupes de vecteurs gradients sont combinés à une fonctionnalité appelée histogrammes de gradients orientés.

Le terme **histogramme** signifie la représentation d'une distribution des données. Il permet de lier des variables continues ensembles dans des classes non chevauchées. Dans le but de donner l'estimation de l'intensité de ses données. Les applications les plus répondues exploitants cette approche statistique dans le domaine de traitement d'image est l'histogramme de couleurs ou de l'intensité.

Un histogramme de l'intensité représente la distribution de valeurs d'intensité dans une image en comptant les occurrences des intensités et les visualiser dans un diagramme à barres.

Habituellement, la taille de la classe est augmentée ce qui conduit à une résolution plus grossière de l'histogramme. Augmentation de la taille de la classe, a par exemple 10, signifie que tous les pixels ayant une valeur comprise entre 0 et 9 sont réunies dans la première classe, des valeurs de 10 à 19 sont réunies dans le deuxième classe et ainsi de suite.

La notion de base d'histogrammes est applicable à un gradient de même. Rappeler que l'orientation est $\theta(x, y)$ et l'amplitude r (x, y) d'un vecteur de gradient à une position (x, y) dans une image. L'orientation du gradient peut être discrétisée dans l'une des p classes. Bien que la taille d'une classe dans un histogramme de couleur ou d'intensité est un intervalle de valeurs d'intensité, dans un histogramme de gradient orienté c'est un intervalle de valeurs d'angles.

Il existe deux formules de discrétisation de gradient :

$$B_1(x, y) = round\left(\frac{p.\theta(x, y)}{2\pi}\right) \bmod p \qquad \textbf{Eq4-7}$$

$$B_2(x, y) = round\left(\frac{p.\theta(x, y)}{\pi}\right) \bmod p \qquad \textbf{Eq4-8}$$

où B1 désigne un descripteur sensible au contraste et B2 représente un descripteur insensible au contraste. Notons que la seule différence entre ces deux formules est que l'orientation est divisée par 2π ou π respectivement. La division par π conduit à des vecteurs pointant dans des directions opposées (plus précisément: classes opposées) étant réunis dans une seule et même classe dans le cas insensible. Ainsi, la définition de l'insensibilité ne considère pas que le gradient est calculé dans une zone où l'intensité de pixel le long de ses augmentations ou diminutions d'orientation; comme une droite g qui tourne avec π à n'importe quel points $P \in g$, mappage de la ligne sur elle-même, il capture seulement l'orientation d'une arête,

dans le domaine, nous considérons deux termes « signé » et « non signée » dans le cas de la division par π ou 2π respectivement.

Avec une des deux formules citées préalablement, nous pourrons définir un vecteur de caractéristiques à (x,y) :

$$F(x,y)_b = \begin{cases} r(x,y), si & b = B_{1/2}(x,y) \\ 0, & sinon \end{cases} \quad \textbf{Eq4-9}$$

Avec $b \in \{0,...,p-1\}$ Pour chaque pixel (x,y), le vecteur de gradient est discrétisé à une des p classes.

4.4 Construction du vecteur de descripteur de l'histogramme du gradient orienté « HOG »

L'histogramme de gradient orienté (HOG) est conçu par l'agrégation spatiale des histogrammes de contour dans une cellule $C(i,j)$, pour $0 \le i \le [(w-1)/k]$et $0 \le j \le [(h-1)/k]$. Cette agrégation offre l'invariance par rapport aux petites déformations et réduit la taille de la carte de descripteurs. L'approche, la plus simple pour agréger les descripteurs est de mapper chaque pixel *(x,y)* dans une cellule $([x/k],[y/k])$ et définir le vecteur de descripteurs pour chaque cellule comme étant la somme de tous les descripteurs –niveau pixel- des pixels appartenant à cette cellule.

En pratique, le plus commun, on définit le nombre de classe $p = 9$ et la taille de la cellule $k = 8$. Ce qui entraine un vecteur de descripteur à p-dimension pour chaque HOG-cellule . Dalal and Triggs propose une étape de normalisation et de troncature pour les descripteurs de HOG pour obtenir l'invariance aux changements de partialité.

La normalisation d'un descripteur $C(i, j)$ est accomplie à travers la normalisation de facteurs $N_{\delta,\gamma}(i, j)$ avec $\delta, \gamma \in \{-1,1\}$, donnée par :

$$N_{\delta,\gamma} = (\|C(i, j)\|^2 + \|C(i+\delta, j)\|^2 + \|C(i, j+\gamma)\|^2 + \|C(i+\delta, j+\gamma)\|^2)^{0.5} \quad \textbf{Eq4-10}$$

Ainsi, nous obtenons quatre facteurs de normalisations pour toutes les lignes de (δ, γ). La normalisation prend quatre lignes des HOG-cellules, selon l'équation ci-dessus, et prend la somme de leurs énergies globales.

La troncature $T_\alpha(C(i, j))$ signifie le processus de remplacement un élément c_b dans le vecteur de descripteur *C(i,j)* avec le $\min(\alpha, c_b)$. Le descripteur HOG est ensuite défini comme une matrice 9x4 :

$$H(i, j) = \begin{pmatrix} (T_\alpha(C(i, j)) / N_{-1,-1}(i, j)) \\ (T_\alpha(C(i, j)) / N_{+1,-1}(i, j)) \\ (T_\alpha(C(i, j)) / N_{+1,+1}(i, j)) \\ (T_\alpha(C(i, j)) / N_{-1,+1}(i, j)) \end{pmatrix} \quad \textbf{Eq4-11}$$

Comme résultat de la normalisation, les descripteurs des bordures de la carte des descripteurs à base de cellules ne devraient pas être pris en considération pour le calcul en outre par le fait que la normalisation complète prend l'énergie du gradient de tous les huit voisins d'une cellule en compte, de sorte que les cellules de bordure ne peut pas être convenablement normalisé. En général, ces caractéristiques sont jetées.

4.4.1 Diagramme de calcul de descripteur de HOG

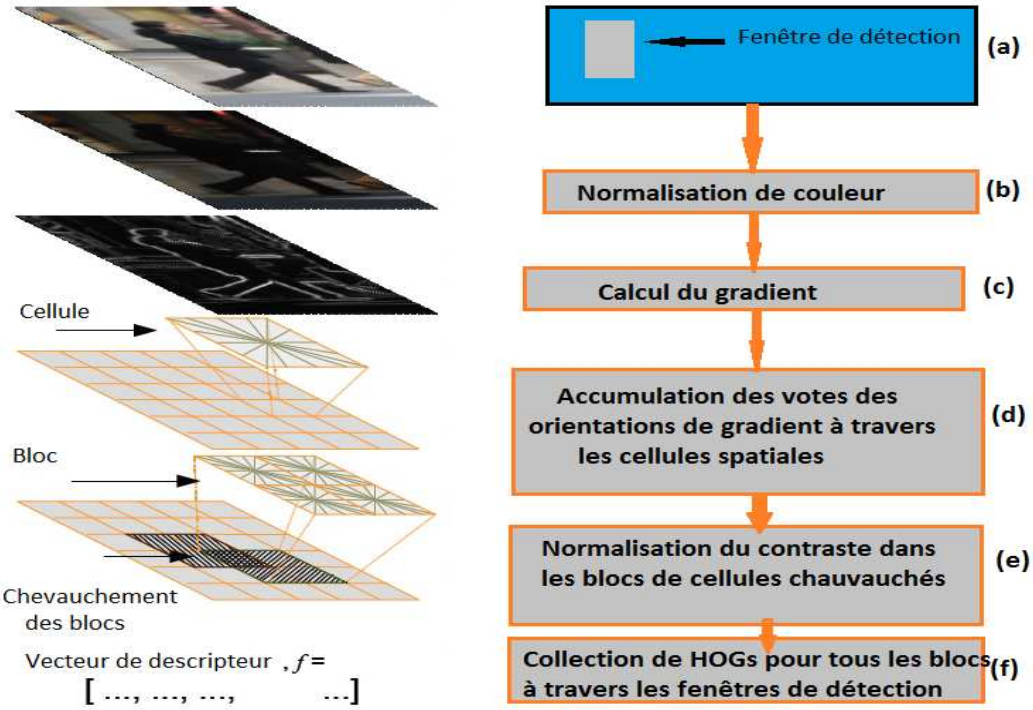


Figure 4-1Histogramme de gradient orienté tel que proposé par Dalal & Triggs [37]

4.4.1.1 Normalisation Gamma-couleur de l'image

Le système d'extraction de caractéristiques est visualisé dans la figure 3.1. Descripteurs du HOG sont extraits pour chaque fenêtre séparément - voir la figure 3.1 l'étape (a). Dans le schéma initial de Dalal et Triggs [23] l'image est d'abord normalisée (b). Dans la plupart des approches exploitant HOG comme descripteur , cette étape est sautée.

4.4.1.2 Cumul des réponses Gradient

Pour chaque pixel à l'intérieur de la fenêtre de détection du gradient de l'image est calculée - cela est visualisé à l'étape (c) de la figure 3.1.Dans la prochaine étape (d) les réponses de gradient sont accumulés dans des classes à travers (i)l'orientation, et (ii) les régions spatiales (« cellules »). Une cellule individuelle et une grille de cellules est montrés dans la figure 3.1 (d). Au sein de chaque cellule, un histogramme d'orientation de gradient est calculé par la quantification de l'orientation de gradient et votant de l'amplitude dans les classes d'histogramme. Ces votes sont généralement une bi- interpolation linéaire c-à-dire l'amplitude est partagée entre les classes voisines.

L'idée derrière la quantification des orientations et les positions spatiales est d'introduire une invariance à la déformation locale d'une image- les parties d'objets peuvent se déplacer autour de chaque cellule dans une certaine mesure sans modifier le descripteur HOG.

4.4.1.3 Blocs et normalisation de contraste

La deuxième étape de l'accumulation spatiale (étape (e)) regroupe les rangs de cellules contiguës des cellules dans des blocs , voir figure 3-2. Le descripteur de chaque bloc est ensuite normalisé indépendamment pour avoir une norme constante par la division de chaque classe d'histogramme dans le bloc par la norme ℓ_2. Les blocs se chevauchent généralement par une ou plusieurs cellules de sorte que chaque cellule est représentée plusieurs fois (avec différentes normalisations) dans le descripteur final formé par la concaténation de tous les blocs. Le schéma bloc de normalisation donne une invariance de contraste sur une échelle plus grande que celle des cellules individuelles.

4.5 Dimension d'un vecteur de descripteur HOG

Felzenszwalb et al. [47] proposent d'utiliser des descripteurs plus petits afin de minimiser le nombre de paramètres dans leurs modèles et accélérer les processus de détection et d'apprentissage. Ils ont analysé les composantes principales d'une collection de descripteurs HOG- dérivant des descripteurs de 13 dimensions en obtenant le même rendement que celui de 36 dimensions d'origine, tel que démontré dans le tableau 3-1 , représentant la précision moyenne pour les différentes dimensions du vecteur de descripteur de HOG.

La nouvelle fonction est calculée en additionnant à travers les quatre normalisations et en additionnant sur les neuf classes d'orientation dans la matrice de l'équation 3.11. Cela conduit à un 9 + 4 – dimensions, nouveau descripteur insensible au contraste. En outre, leurs travaux ont indiqué que certaines classes d'objets bénéficient des caractéristiques de contraste insensibles alors que les performances de détection s'améliorent pour les autres classes en utilisant des éléments sensibles de contraste.

Le système final implémente des descripteurs qui combinent à la fois les descripteurs sensibles et insensibles au contraste. Rappelons que les descripteurs insensibles originaux, sans normalisation et troncature, sont à p = 9 dimensions. Nous pouvons facilement définir un descripteur sensible au contraste dans les équations (3.9) avec (3.7) et p = 9 et concaténer les deux à un (9 + 18)- dimension.

L'équation (3.11) indique que la version normalisée est une matrice (9 + 18) . 4 = 108 entrées . La version améliorée de ce vecteur de descripteurs est alors obtenue en additionnant les 9 +18 = 27 colonnes de la matrice et la concaténation avec 4 sommes sur l'orientation des 9 contrastes orientations insensibles. Cela conduit à vecteur de descripteur de 27 + 4 = 31 dimensions.

Felzenszwalb et al. ajoutent une autre entrée à la fin de chaque vecteur, appelée une fonction de troncature , qui est généralement mis à 0 sauf pour les fonctions générés artificiellement pour le rembourrage , où il est mis à 1 pour indiquer la différence entre une cellule de frontière rembourré et un «vrai "caractéristique.

Le secret derrière la recherche de compression du vecteur de descripteurs est dans la quantité importante de produit scalaire, une opération de compression réduit énormément le temps de calcul tout en préservant l'information nécessaire.

En dehors de la normalisation et de la troncature, « **Soft-binning »** est appliqué comme une autre optimisation des descripteurs de HOG au niveau cellulaire qui améliore la robustesse de l'ensemble du système. L'idée derrière « **soft-binning** » est que le gradient à un pixel (x, y) doit non seulement donner toute son énergie, à savoir l'amplitude, à une seule cellule, mais également à des cellules adjacentes ainsi. Nous examinons le gradient dans un des quatre quadrants d'une cellule de bordure. Chaque quadrant est voisin de trois cellules du HOG-adjacents. Soft-binning calcule la distance de la position du gradient au centre de sa propre cellule et les bordures de cellule de son quadrant. Si sa position est très proche du centre de sa propre cellule, la presque totalité de l'amplitude de gradient est ajouté à son propre histogramme. Plus il est proche des cellules voisines si, plus de son énergie est ajoutée aux classes correspondantes des histogrammes adjacentes. Ce qui fait une carte- HOG complète beaucoup plus robuste contre petit déplacement d'objets dans les images.

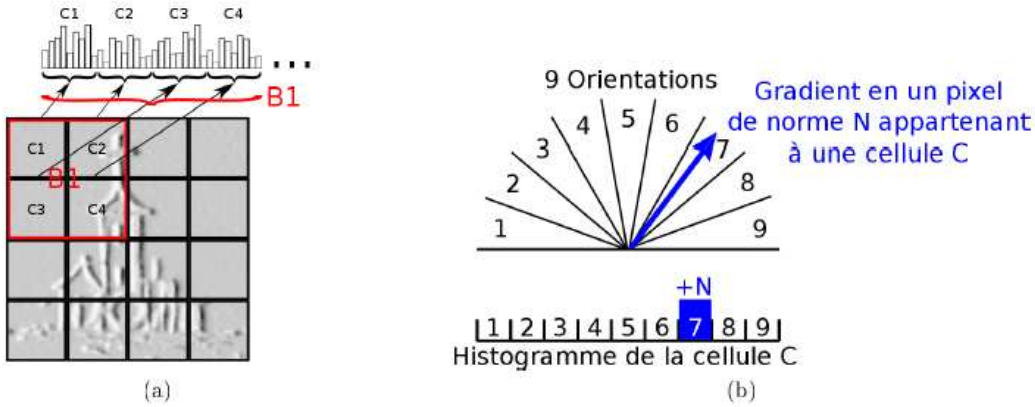


Figure 4-2 [96] Formulation des descripteurs de HOG

Type HOG	INRIA PERSONNE	VOC2006 VOITURE	 PERSONNE
HOG_{36}	79.0	51.7	22.2
HOG_{13}	79.3	50.6	21.0
HOG_{31}	78.7	55.5	23.2

Tableau 4-1 AP (Précision Moyenne pour les détecteurs linéaires (SVM Latent)

4.6 Pyramide de descripteur

Fondamentalement, La fonction qui calcul une pyramide de descripteurs prend comme entrée l'image d'origine et crée une pyramide d'images d'elle-même en utilisant un facteur de redimensionnement. Cela signifie que l'image est à échelle réduite à chaque fois jusqu'à ce qu'il atteigne une taille qui est plus petite que le modèle (ce qui serait impossible à réaliser), et agrandit également l'image jusqu'à un niveau défini. Le grand avantage de cette méthode , est de pouvoir détecter des personnes à partir de modèle unique, tout au long de différentes échelles d'images.

En utilisant cette technique, on peut détecter des personnes aux différentes échelles avec un modèle unique, ce qui est moins coûteux en termes de calcul que de faire l'apprentissage de modèle pour chaque échelle d'images et de les utiliser pour détecter les objets.

Une pyramide de descripteurs est un ensemble de carte de descripteurs pour un nombre fini d'échelles. Pratiquement, une pyramide de descripteur est calculée par une itération d'opération de lissage et de sous-échantillonnage de l'image originale et le calcul des descripteurs à chaque échelle. Par conséquent, le paramètre λ définit le nombre de niveau dans une **octave** comme la distance entre deux niveaux dans la pyramide de descripteurs, où le niveau le plus haut à une double résolution par rapport au plus bas. La Figure 3-1 démontre le principe de la construction d'une pyramide de descripteurs.

4.6.1 Interpolation d'une image

Pour redimensionner les images d'entrée à une échelle donnée, nous faisons recours à une **interpolation par le plus proche voisin » (Nearest-neighbor interpolation)** .L'idée principale est que pour chaque emplacement (i,j) dans l'image de sortie , nous faisons un mappage vers l'emplacement le plus proche des coordonnées dans l'image d'entrée tel que illustré sur la figure 3-3 :

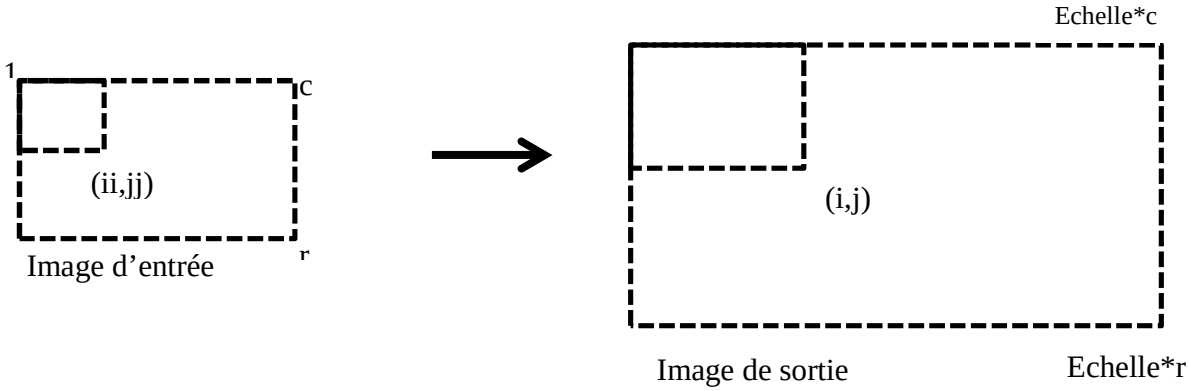


Figure 4-3 Principe d'interpolation d'une image

Pour les deux premiers niveaux de la pyramide ; l'image est interpolée une seule fois avec l'échelle appropriée, l'astuce est de calculer la carte de descripteurs du niveau le plus haut en utilisant la moitié de la taille de la cellule utilisée, (utiliser 4x4 pour la taille d'une cellule au lieu de 8x8). Ce qui implique à avoir une carte de descripteurs avec une double résolution sans à recourir à une seconde interpolation.

Le reste de la pyramide est calculé par une opération consécutive d'interpolation d'image avec le facteur 0.5 et le calcul de la carte de descripteurs.

4.6.2 Approche de la fenêtre coulissante

Supposons que nous avons affaire à des objets qui ont une apparence relativement bien comportés, et ne se déforment pas beaucoup. Alors nous pouvons les détecter avec une méthode très simple : Nous construisons un ensemble de fenêtres étiquetées d'image de taille fixe (par exemple, n**X**m). Les exemples étiquetés positif doivent contenir de grandes instances, centrées sur l'objet, contrairement aux exemples négatifs labélisés. Nous construisons alors un classifieur pour ces exemples. Nous passons maintenant chaque fenêtre n**X**m dan l'image au classifieur.

il s'agit d'une recherche sur l'emplacement, que nous pourrions représenter avec le coin supérieur gauche de la fenêtre.

Il y'a deux subtilités dans l'application de cette méthode. Premièrement, toutes les instances d'objets doivent être de la même taille dans l'image : Contrainte qui n'est pas toujours vrai, ce qui signifie qu'il faut opter pour une recherche par échelle. La façon la plus facile est de préparer une pyramide gaussienne des images , et ensuite appliquer une recherche par fenêtre de nXm pour chaque niveau de pyramide. La recherche dans une image dont la longueur du contour a été réduite par s pour n × m des fenêtres est un peu comme rechercher dans l'image d'origine pour des fenêtres (sn) × (sm) (les différences sont dans la résolution, dans la facilité de l'apprentissage, et en temps de calcul).

La deuxième subtilité est que certaines fenêtres d'images se chevauchent très fortement. Chacun d'un ensemble de fenêtres qui se chevauchent peut contenir la totalité (ou une fraction substantielle de) l'objet. Cela signifie que chacun pourrait être étiqueté positif par le classificateur, ce qui signifie que nous serions comptés le même objet plusieurs fois. Cet effet ne peut pas être durci par passage dans un ensemble d'apprentissage d'ensemble et à produire un classificateur qui est si étroitement réglé qu'il répond uniquement lorsque l'objet est exactement centrée dans la fenêtre. C'est parce qu'il est difficile de produire des classificateurs bien réglés, et parce que nous ne serons jamais en mesure de placer une fenêtre exactement autour d'un objet, de sorte qu'un classificateur bien réglé aura tendance à mal se comporter.

La stratégie habituelle pour la gestion de ce problème **« non maximum Suppression ».** Dans cette stratégie, les fenêtres avec un maximum local de la réponse de classificateur suppriment les fenêtres voisines. Nous résumons l'approche globale de l'algorithme.

L'approche de détection par fenêtre coulissante est totalement générique et offre de bon résultat dans la pratique. Différentes applications nécessitent des choix différents de fonctionnalité et parfois bénéficient de différents choix de fonction.

Notez qu'il existe une interaction subtile entre la taille de la fenêtre, et les pas Δx, Δy , et le classificateur . Par exemple, si nous travaillons avec des fenêtres qui entourent étroitement l'objet , alors nous pourrions être en mesure d'utiliser un classificateur qui est plus étroitement ajusté , mais nous aurons à utiliser de plus petits pas et avoir par la suite plusieurs fenêtres à gérer . Si nous utilisons les fenêtres qui sont un peu plus grande que l'objet, alors nous pouvons avoir moins de fenêtres , mais notre capacité à détecter des objets près de l'autre pourraient être affectés , comme pourraient notre capacité à localiser les objets .

La validation croisée « **Cross-Validation** » est une façon de faire des choix appropriés. En conséquence, il existe une certaine variation de l'apparence de la fenêtre provoqué par le fait que notre recherche est quantifiée en translation et par échelle.

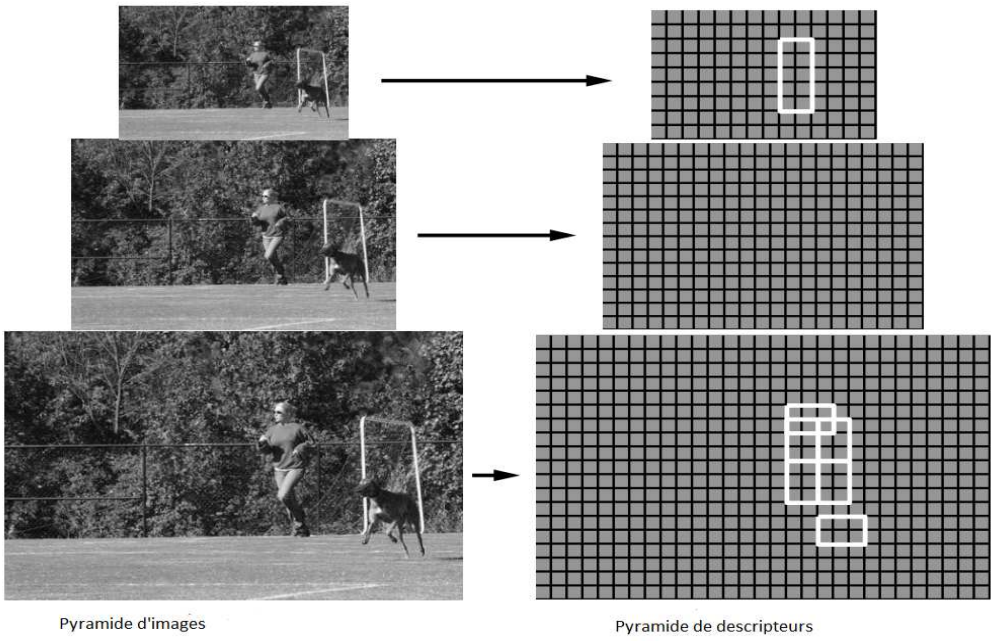


Figure 4-4 Construction de pyramide de descripteurs

4.7 Conclusion

Le corps humain est considéré comme étant un objet articulé, soumis à des variations d'apparence, de forme. La détection de ses composantes, partie du corps, dans le domaine de la vision par ordinateur est très contraignante. Dans ce contexte, nous proposons une approche basée sur un modèle déformable combinant à la fois deux algorithmes : Détection de personne par le descripteur de HOG , proposé par Dallal& Triggs [37] et les structures picturales[45].

L'élément clé dans telle recherche est le choix de descripteur, dans notre recherche, nous utilisons , tel que [37] l'Histogramme de Gradient Orienté (HOG).

Le descripteur basé histogramme de gradient orienté a montré son performance dans de nombreux travaux de recherches, proposé par Dalal & Triggs pour la détection de piétons , nous l'exploitons pour capter l'apparence d'une partie du corps, il a l'exclusivité de mieux représenter la structure interne d'un objet via

l'information du gradient, permettant ainsi de surmonter les problèmes liés à l'apparence de l'objet : pose, l'éclairage, l'occlusion, texture de fond, etc. Un autre problème se pose, autre que l'apparence, est l'échelle de représentation de l'objet ce qui nous pousse à faire l'apprentissage de modèle pour chaque échelle d'images , ce qui demande un énorme temps de calcul, pour cela nous introduisons un nouveau concept de représentation combinant à la fois le concept de l'histogramme de gradient orienté avec la pyramide d'image.

Chapitre 4 : Champs aléatoires conditionnels (CRF)

5.1 Introduction :

Le processus de l'estimation de la pose humaine est considéré comme étant un problème de labellisation plus précisément de classification, comment assigner un emplacement d'une image observé $x \in X$ à une partie du corps humain quelconque $y \in Y$.

Cette tâche peut être abordée avec la théorie des probabilités par la spécification d'une distribution de probabilité pour définir la classe la plus probable "y" pour une observation donnée "x".

La modélisation de toutes les dépendances dans une distribution de probabilité est généralement très complexe en raison de l'interdépendance entre les éléments. L'hypothèse de Bayes natif (**Native Bayes)** de toutes les fonctions conditionnellement indépendantes est une approche pour résoudre ce problème.

Dans le scénario d'apprentissage structuré, les variables d'observations sont considérées comme ce qui implique une complexité encore plus élevé dans la distribution de probabilité. C'est le cas pour les données d'image ou de la musique ainsi que pour le texte en langage naturel. Comme pour les images, les pixels près les uns aux autres sont très susceptibles d'avoir une couleur similaire.

Une approche pour la modélisation des **séquences de structures linéaires**, tel que le texte en langage naturel, sont les **modèles de Markov cachés**[102]. Par souci de réduction de la complexité, des hypothèses d'indépendance fortes entre les variables d'observation sont faites ce qui affecte la précision du modèle. Les **modèle de champs aléatoires conditionnels** (**Conditional Random Fields)** (CRF [103]), sont développées exactement pour combler ces lacunes.

Les modèle CRF ont été exploités dans de nombreuses applications traitant avec des données structurées. Malgré l'application fréquente de chaîne linéaire. Les CRF sont actuellement une technique state-of-the-art pour un grand nombre de ses sous-tâches, y compris la segmentation du texte[104] , les processus d'annotation , marquage de discours[102], l'analyse profonde [106], la résolution de syntagmes nominaux elliptiques[107].
Les CRF ont été révélés très utiles dans la reconnaissance d'entités nommées, en particulier sur les documents du domaine biomédical[108], [109], [110], [111]. En outre, ils ont été appliqués à la prédiction des gènes[112] l'étiquetage d'image[113] la reconnaissance d'objet set de[114], et aussi dans la télématique pour la détection d'intrusion[115] et le capteur de la gestion des données[116].

Dans ce chapitre, nous visons à donner un aperçu de la théorie fondamentale des **modèles des champs aléatoires conditionnels** « Conditional Random Fields » et nous illustrons comment celles-ci sont liés à d'autres modèles probabiliste présentant un bref aperçu des trois modèles probabilistes classiques : **Bayes**, **Hidden Markov**, et l'**entropie maximale**. Les relations entre eux être des présentations graphiques de ces différentes approches, ainsi que les concepts de base de CRF.

5.2 Modèles Probabilistes

5.2.1 Naïve bayes

Une approche de classification d'une seule classe où les valeurs d'entrées sont considérées à être **conditionnellement indépendantes.**

- **Le modèle de « Naive bayesian » sont appelés aussi « Modèles Génératifs »**

Une probabilité conditionnelle est une distribution de probabilité $p(y/\vec{x})$ avec un vecteur d'entrée $\vec{x} = (x_1, ..., x_m)$ où $x_i\,(1 \le i \le m)$ sont les caractéristiques et y la

classe de variable à prédire. Cette probabilité peut être formulée avec la loi de Bayes :

$$p(y \setminus \vec{x}) = \frac{p(y)\, p(\vec{x} \setminus y)}{p(\vec{x})} \qquad \textbf{Eq5-1}$$

Le dénominateur $p(\vec{x})$ n'est pas important pour la classification car elle peut être comprise comme une constante de normalisation qui peut être calculé en tenant compte de toutes les valeurs possibles pour y.

Le numérateur peut aussi être écrit comme une probabilité conjointe :

$$p(y)\, p(\vec{x} \setminus y) = p(y, \vec{x}) \qquad \textbf{Eq5-2}$$

Qui peut être trop complexe pour être calculés directement(surtout quand le nombre de composants $\vec{x}$ est élevé). Une décomposition générale de cette probabilité peut être formulée l'application de la règle de la chaîne :

$$p(x_1, \ldots\ldots x_m) = \prod_{i=2}^{m} p(x_i \setminus x_{i-1}, \ldots\ldots x_1) \qquad \textbf{Eq5-3}$$

$$p(y, \vec{x}) = p(y) \prod_{i=2}^{m} p(x_i \setminus x_{i-1}, \ldots\ldots x_1, y) \qquad \textbf{Eq5-4}$$

Dans la pratique, il est souvent supposé, que toutes les variables d'entrée x_i sont conditionnellement indépendantes les uns des autres, connu par **l'hypothèse natif de Bayes**. Cela signifie que $p(x_i \setminus y, x_j) = p(x_i \setminus y)$ détient pour tous $i \neq j$.Sur la base de cette simplification; Un modèle connu sous le nom Naive classificateur de Bayes est formulée comme suit:

$$P(y \setminus \vec{x}) \propto p(y, \vec{x}) = p(y) \prod_{i=1}^{m} p(x_i \setminus y) \qquad \textbf{Eq5-5}$$

Cette distribution de probabilité est moins complexe que celle formulée dans l'équation 4-2. Les dépendances entre les variables d'entrée ne sont pas modélisées, ce qui conduira probablement une représentation imparfaite du monde réel. Néanmoins, le modèle de Bayes est bien répondu dans de nombreuses applications du monde réel, comme la classification e-mail.

5.2.2 Modèle de Markovcaché (Hidden Markov Models)

Dans le modèle Naive Bayes, seules les variables de sorties ont été prises en considération. Pour prédire une suite de variables de classe $\vec{y} = (y_1, ...y_n)$ pour une séquence d'observation $\vec{x} = (x_1, ...x_n)$, un simple modèle de séquence peut être formulé comme un produit plus modèles de Bayes autochtones. Les dépendances entre les positions des séquences simples ne sont pas pris en compte. Notez que contrairement au Modèle de Bayes il n'y a qu'une seule fonction à chaque position de la séquence, à savoir l'identité de l'observation respective:

$$p(\vec{y}, \vec{x}) = \prod_{i=1}^{n} p(y_i).p(x_i \setminus y_i) \qquad \textbf{Eq5-6}$$

Chaque observation x_i ne dépend que de classe de la variable y_i à la position de la séquence respective. En raison de cette hypothèse d'indépendance, les probabilités de transition d'une étape à l'autre ne sont pas inclus dans ce modèle. En fait, cette hypothèse est rarement rencontrée dans la pratique résultant en une performance limitée de ces modèles. Ainsi, il est raisonnable de supposer qu'il existe des dépendances entre les observations à des positions de séquences consécutives.

Pour modéliser cela, probabilités de transition d'état sont ajoutés:

$$p(\vec{y},\vec{x}) = \prod_{i=0}^{n} p(y_i \setminus y_{i-1})\, p(x_i \setminus y_i) \qquad \textbf{Eq5-7}$$

Cela conduit au modèle de Markov caché :

$$p(\vec{x}) = \sum_{y \in Y} \prod_{i=0}^{n} p(y_i \setminus y_{i-1}) p(x_i \setminus y_i) \qquad \textbf{Eq5-8}$$

Où Y est la série de toutes les séquence de labels $\vec{y}$.

Les dépendances entre les variables de sortie $\vec{y}$ sont modélisées. Un problème réside dans l'hypothèse d'indépendance conditionnelle (voir équation **Eq 4-6**) entre les variables d'entrée $\vec{x}$ en raison de problèmes de complexité. Comme nous le verrons plus tard, CRF répond exactement à ce problème.

5.2.3 Modèle d'Entropie Maximale (Maximum Entropy Model)

Les deux modèles présentés ci-dessous sont formés afin de maximiser la **probabilité conjointe**. Dans ce qui suit, le modèle entropie maximale est examinée plus en détail, car iles fondamentalement lié à l'CRF. Le modèle d'entropie maximale est un modèle de **probabilité conditionnelle**.

Il est basé sur le principe de l'entropie maximale [117] qui stipule que si des informations incomplètes sur une distribution de probabilité est disponible, la seule hypothèse biaisée qui peut être faite est une distribution qui est aussi uniforme que possible, compte tenu des informations disponibles. Dans cette hypothèse, la distribution de probabilité appropriée est celle qui maximise l'entropie étant donné que les contraintes liées à la phase d'apprentissage.

Pour le modèle conditionnel $p(y \setminus x)$, l'entropie conditionnelle $H(y \setminus x)$ [118] est définie comme:

$$H(y \setminus x) = -\sum_{(x,y)\in\square} p(y,x)\log p(y \setminus x) \qquad \textbf{Eq5-9}$$

La série $\square = X \times Y$, X consiste en la série de toutes les entrées x de variables possibles , et Y est la série de toutes les sorties possibles . Notons que $\square$ contient non seulement les combinaisons de x et Y survenant durant la phase d'apprentissage mais aussi d'autres combinaisons possibles.

L'idée de base derrière les **modèles d'entropie maximale** est de trouver le modèle $p^*(y \setminus x)$qui d'une part possède la **plus grande entropie conditionnelle possible**, mais est d'autre part encore consistant avec l'information fournie par le matériel d'apprentissage.
La fonction objective, plus tard appelé problème primal, est donc :

$$p^*(y \setminus x) = \arg\max_{p(y\setminus x)\in p} H(y \setminus x) \qquad \textbf{Eq5-10}$$

Où P est la série de tous les modèles consistants durant l'apprentissage.
Le matériel d'apprentissage (training materiel) est représenté par des caractéristiques. Ici , sont définis comme des valeurs binaires .

La fonction $f_i(x,y) \in \{0,1\}(1 \le i \le m)$ dont dépend à la fois la variable d'entrée x et l classe de la variable y .

Un exemple de cette fonction est :

$$f_i(x,y)=\begin{cases}1 & if\ y=name\ and\ x=Mister\\ 0 & otherwise\end{cases} \qquad \textbf{Eq5-11}$$

La valeur attendue de chaque caractéristique f_i est estimée à partir de la distribution empirique $\tilde{p}(x,y)$.

La distribution empirique est obtenu en comptant simplement combien de fois les différentes valeurs des variables de se produire dans les données d'apprentissage:

$$\tilde{E}(f_i)=\sum_{(x,y)\in\mathbb{Z}}\tilde{p}(x,y)f_i(x,y) \qquad \textbf{Eq5-12}$$

Toutes les paires possibles (x,y) sont pris en compte ici. Comme la probabilité empirique pour une paire (x,y) qui n'est pas contenue dans le matériel d'apprentissage est égal à 0, $\tilde{E}(f_i)$ peut être reformulée comme suit:

$$\tilde{E}(f_i)=\frac{1}{N}\sum_{(x,y)\in T}f_i(x,y) \qquad \textbf{Eq5-13}$$

La taille Del' ensemble d'apprentissages N= $|T|$. Ainsi, $\tilde{E}(f_i)$ peut être calculée en comptant combien de fois une caractéristique f_i se trouve avec la valeur 1dans les données d'apprentissage $T\subseteq Z$ et en divisant ce nombre parla tailleN de l'ensemble de données d'apprentissage.

De manière analogue à l'équation 4-10, la valeur attendue d'une fonction sur le modèle de distribution est défini comme suit:

$$E(f_i) = \sum_{(x,y)\in\mathbb{Z}} p(x,y) f_i(x,y) \qquad \textbf{Eq5-14}$$

Contrairement à l'équation 4-12 (la valeur attendue sur la distribution empirique), le modèle la distribution est pris en compte ici. Bien sûr, $p(x,y)$ ne peut être calculé de manière générale parce que le nombre de tous les possibles $x \in X$ peuvent être énormes.

Cela peut être adressée par la réécriture $E(f_i)$ par :

$$E(f_i) = \sum_{(x,y)\in\mathbb{Z}} p(x)p(y \setminus x)\, f_i(x,y) \qquad \textbf{Eq5-15}$$

et la substitution $p(x)$ avec la distribution empirique $\tilde{p}(x)$. Il s'agit d'un rapprochement à faire réduire le calcul de $E(f_i)$.Ce résultat en :

$$E(f_i) \approx \sum_{(x,y)\in\mathbb{Z}} \tilde{p}(x)p(y \setminus x)\, f_i(x,y) \qquad \textbf{Eq5-16}$$

Qui peut(de manière analogue à l'équation4-14)se transforme en:

$$E(f_i) = \frac{1}{N}\sum_{x\in T}\sum_{y\in Y} p(y \setminus x) fi(x,y) \qquad \textbf{Eq5-17}$$

Seules les valeurs x apparaissant dans les données d'apprentissage $(x \in T)$ sont pris en compte lors de toutes les valeurs y possibles sont prises en compte $(y \in Y)$. Dans de nombreuses applications, l'ensemble Y contient généralement seulement un petit

nombre de variables. Ainsi, en sommant sur y est possible ici et $E(f_i)$ peut être calculé de manière efficace. Équation 4-10postule que le modèle $p^*(y \setminus x)$ est consistante avec les éléments de preuve trouvés dans le matériel d'apprentissage. Cela signifie que f_i, pour chaque fonction de sa valeur attendue sur la distribution empirique doit être égale à sa valeur attendue sur le modèle de distribution en particulier, ce sont les contraintes m premières.

$$E(f_i) = \tilde{E}(f_i) \qquad \textbf{Eq5-18}$$

Une autre contrainte est d'avoir une probabilité conditionnelle assurée par $p(y \setminus x) \geq 0 \; for\; all\; x, y$

$$\sum_{y \in Y} p(y \setminus x) = 1 \; for\; all\; x \qquad \textbf{Eq5-19}$$

Chercher $p^*(y \setminus x)$ sous les contraintes peut être formulée comme une optimisation de contraintes . Pour chaque contrainte un multiplicateur de Lagrange λ_i est introduit. Cela conduit à la fonction suivante Lagrange $\Lambda(p, \vec{\lambda})$

$$\Lambda(p, \vec{\lambda}) = \underset{1}{H(y \setminus x)} + \sum_{i=1}^{m} \lambda_i \underbrace{\left(E(f_i) - \tilde{E}(f_i) \right)}_{2} + \lambda_{m+1} \underbrace{\left(\sum_{y \in Y} p(y \setminus x) - 1 \right)}_{3} \qquad \textbf{Eq5-20}$$

1- Problème Intitial À partir de l'équation 11
2- Contrainte !=0 À partir de l'équation 19
3- Contrainte !=0 À partir de l'équation 20

Ceci est maximisé pour obtenir la formulation du modèle $p_{\vec{\lambda}}^{*}(y \setminus x)$ de l'équation4-29.Dans ce qui suit, un calcul détaillé est donné. De la même manière que pour faire des valeurs moyennes de l'équation4-16, $H(y \setminus x)$ est approchée:

$$H(y \setminus x) \approx -\sum_{(x,y)\in\mathbb{Z}} \tilde{p}(x) p(y \setminus x) \log p(y \setminus x) \qquad \textbf{Eq5-21}$$

La dérivation de l'équation 4-21 est donnée par :

$$\begin{aligned} \frac{\delta}{\delta p(x,y)} H(y \setminus x) &= -\tilde{p}(x)\left(\log p(y \setminus x) + \frac{p(y \setminus x)}{p(y \setminus x)}\right) \\ &= -\tilde{p}(x)\left(\log p(y \setminus x) + 1\right) \end{aligned} \qquad \textbf{Eq5-22}$$

La dérivation des m première contrainte dans l'équation (4-20) est donné par :

$$\frac{\delta}{\delta p(x,y)} \sum_{i=1}^{m} \lambda_i \left(\sum_{(x,y)\in Z} \tilde{p}(x) p(y \setminus x) f_i(x,y) - \left(\sum_{(x,y)\in Z} \tilde{p}(x,y) f_i(x,y) \right) \right) = \sum_{i=1}^{m} \lambda_i \tilde{p}(x) f_i(x,y) \qquad \textbf{Eq5-23}$$

La dérivation complète de la fonction de Lagrange de l'équation4-20 est alors:

$$\frac{\delta}{\delta p(y \setminus x)} \Lambda\left(p, \vec{\lambda}\right) = -\tilde{p}(x)\left(1 + \log p(y \setminus x)\right) + \sum_{i=1}^{m} \lambda_i \tilde{p}(x) f_i(x,y) + \lambda_{m+1} \qquad \textbf{Eq5-24}$$

Assimiler ce terme à 0 et en résolvant par $p(x, y)$ conduit à

$$0 = -\tilde{p}(x)\left(1+\log\left(y \setminus x\right)\right)+\sum_{i=1}^{m}\lambda_i \tilde{p}(x)\, f_i(x,y)+\lambda_{m+1}$$

$$1+\log p\left(y \setminus x\right)=\sum_{i=1}^{m}\lambda_i f_i(x,y)+\frac{\lambda_{m+1}}{\tilde{p}(x)}$$

$$\tilde{p}(x)\left(1+\log p\left(y \setminus x\right)\right)=\sum_{i=1}^{m}\lambda_i \tilde{p}(x)\, f_i(x,y)+\lambda_{m+1}\ \log p\left(y \setminus x\right)=\sum_{i=1}^{m}\lambda_i f_i(x,y)+\frac{\lambda_{m+1}}{\tilde{p}(x)}-1$$

$$p\left(y \setminus x\right)=\exp\left(\sum_{i=1}^{m}\lambda_i f_i(x,y)\right)+\exp\left(\frac{\lambda_{m+1}}{\tilde{p}(x)}-1\right) \qquad \textbf{Eq5-25}$$

La seconde contrainte est donnés par la contrainte de l'équation 4-17 comme :

$$\sum_{y\in Y} p\left(y \setminus x\right)=1 \qquad \textbf{Eq5-26}$$

En remplaçons l'équation 4- 25 dans 4- 26 , nous aurons :

$$\sum_{y\in Y}\exp\left(\sum_{i=1}^{m}\lambda_i f_i(x,y)\right)+\exp\left(\frac{\lambda_{m+1}}{\tilde{p}(x)}-1\right)=1$$

$$\exp\left(\frac{\lambda_{m+1}}{\tilde{p}(x)}-1\right)=\frac{1}{\sum_{y\in Y}\exp\left(\sum_{i=1}^{m}\lambda_i f_i(x,y)\right)} \qquad \textbf{Eq5-27}$$

En remplaçons l'équation 4-27 dans l'équation 4-25 cela résulte en :

$$p(y \setminus x) = \exp\left(\sum_{i=1}^{m} \lambda_i f_i(x, y)\right) \frac{1}{\sum_{y \in Y} \exp\left(\sum_{i=1}^{m} \lambda_i f_i(x, y)\right)} \quad \textbf{Eq5-28}$$

Il s'agit de la forme générale du modèle pour répondre aux contraintes. Le maximum Modèle entropie peut alors être formulé comme :

$$p^*_{\bar{\lambda}}(y \setminus x) = \frac{1}{Z_{\bar{\lambda}}(x)} \exp\left(\sum_{i=1}^{m} \lambda_i f_i(x, y)\right) \quad \textbf{Eq5-29}$$

Où $Z_{\bar{\lambda}}(x) = \sum_{y \in Y} \exp\left(\sum_{i=1}^{m} \lambda_i f_i(x, y)\right)$

Discussion :

Deux types de modèles probabiliste sont été introduits. D'une part, les **modèles génératifs,** comme **Naive Bayes** et **modèles de Markov cachés**, qui sont basées sur des distributions de **probabilités conjointes**. Comme on le voit dans la formule6 et 8, dans de tels modèles les variables d'observation génèrent les variables d'entrée. Cette caractéristique peut être vu ans la représentation graphique; la section suivante, D'autre part, les **modèles discriminants,** tels que modèles **d'entropie maximale**, sont basées sur des **distributions de probabilités conditionnelles.** Dans la section suivante, les deux groupes de modèles sont passés en revue à partir d'un point de vue différent: **Leurs représentations graphiques.**

5.3 Representation graphique

Les distributions de probabilité sous-jacentes de modèles probabilistes peuvent être représentées sous une forme graphique, c'est pourquoi ils sont souvent appelés des modèles graphiques probabilistes. **Un modèle probabiliste graphique** est une représentation schématique d'une distribution de probabilité. Dans un tel graphe existe un nœud pour chaque variable aléatoire. L'absence d'une arête entredeux variables représente l'indépendance conditionnelle entre ces variables.

L'indépendance conditionnelle signifie que deux variables aléatoires A et B sont indépendants étant donné tiers variable aléatoire si elles sont indépendantes dans leur distribution de probabilité conditionnelle, formellement:

$$p(a,b \setminus c) = p(a \setminus b)\, p(b \setminus c)$$

De tels graphes, également appelé **graphe in dépendant,** on peut lire les propriétés d'indépendance conditionnelle de la distribution sous-jacente. Notez que l'indépendance d'un graphique entièrement connecté ne contient aucune information sur la distribution de probabilité, que l'absence d'arêtes est informative: l'indépendance conditionnelle dans la distribution de probabilité à ne pas provoquer l'absence de l'arête dans le graphe.

5.3.1 Independence conditionnelle :

L'indépendance conditionnelle est un concept important car il peut être utilisé pour décomposer les distributions de probabilité complexes en un produit de facteurs, dont chacune est constituée du sous-ensemble de variables aléatoires correspondantes. Ce concept permet des calculs complexes (qui sont par exemple nécessaires pour l'apprentissage ou l'inférence) beaucoup plus efficaces.

En général, la décomposition, en fait, une factorisation de la distribution de probabilité, est écrit comme le produit de ses facteurs ψ_s, avec $\vec{v}_s$ le sous-ensemble des variables aléatoires respectives constituant un tel facteur :

$$p(\vec{v}) = \prod_s \psi_s(\vec{v}_s) \qquad \textbf{Eq5-30}$$

Soit $G = (V, E)$ un graphe avec sommets V et bords E. Dans un graphe indépendance(par exemple celui représenté sur la figure 7 (a)), les sommets, avec X et ensembles Y de variables aléatoires, sont représentées par des cercles. Il sera généralement considérée comme l'ensemble des entrées ou des variables d'observations (cercles grisés), et comme l'ensemble des variables de sorties(les nœuds vides). Un graphe indépendant peut être dirigé ou non, en fonction du type de modèle graphique adopté. Dans un graphe facteur ,tel que celui illustré à la figure4-1 (b), les cercles représentent, comme dans un graphe indépendance, les variables aléatoires de la distribution sous-jacente, représentées par des cercles. En outre, un graphe facteur contient des nœuds de facteurs, représentées par de petits carrés noirs, qui représentent les facteurs ψ_s (comparer avec l'équation4-30).

Dans un graph facteur, les arrêtes sont toujours non orienté, reliant les variables aléatoires aux nœuds des facteurs. Un facteur ψ_s comprend toutes les variables aléatoires à laquelle le noeud facteur respectif est directement connecté par un bord.

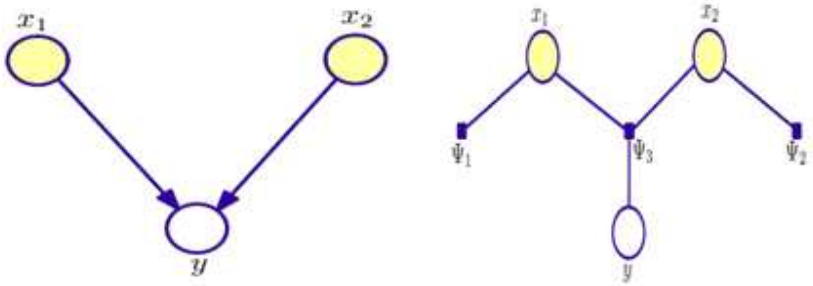
Ainsi, un graphique facteur représente de façon plus explicite la factorisation de la distribution de probabilité sous-jacente. Graphiques indépendance de à la fois dirigé et non orientés modèles graphiques peuvent être transformées en graphiques facteur. A titre d'exemple, supposons une distribution de probabilité $p(x_1, x_2, y)$ à factoriser comme :

$p(\vec{x}) = p(x_1)\, p(x_2)\, p(y \setminus x_1, x_2)$.

Elle aura comme facteur :

$\Psi_1(x_1) = p(x_1)$, $\Psi_2(x_2) = p(x_2)$, $\Psi_3(y) = p(y \setminus x_1, x_2)$.

Ici, x_1 et x_2 sont conditionnellement indépendants y donné. La figure 4-1 montre un graphe Indépendance et un graphique facteur représentant cette distribution



(a) Graphe Indépendant (b) Graphe facteur

Figure 5-1 Un modèle de graphe directe

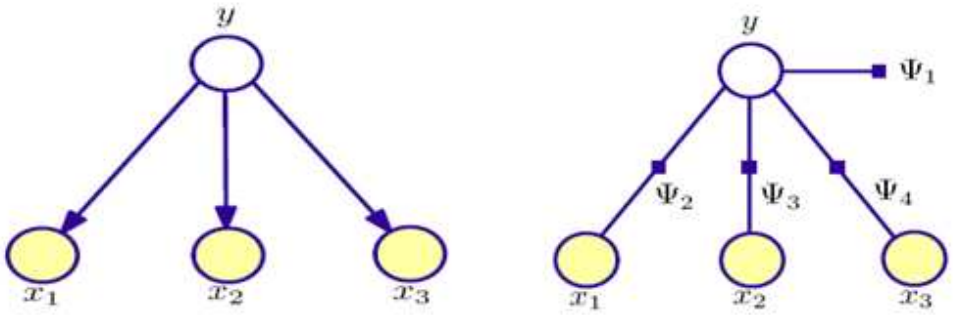
5.3.2 Modèle de graphe Graphique

Une distribution **conjointe** $p(\vec{v})$ peut être factorisée comme un produit des distributions conditionnelles pour chaque noeud v_k de sorte que chaque distribution conditionnelle est subordonnée à l'ensemble de nœuds parents $v_k{}^p$

$$p(\vec{v}) = \prod_{k=1}^{K} p(v_k \setminus v_k^p) \qquad \textbf{Eq5-31}$$

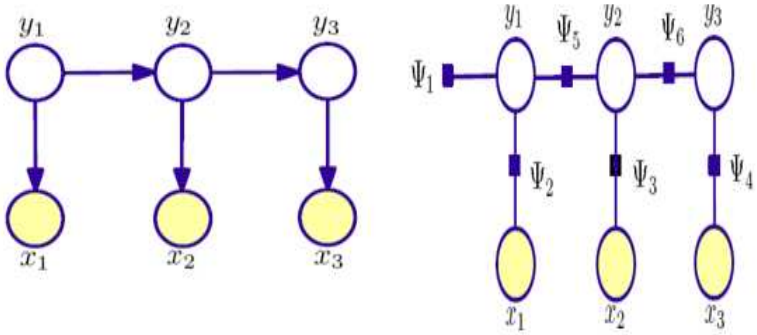
C'est le même genre de factorisation comme le montre la figure 4-1 pour la distribution par exemple $p(x_1, x_2, y)$. Comme autre exemple, prenez le classificateur Naive Bayes qui est discuté dans les sections précédentes. Figure 4-2représente graphiquement un tel modèle à trois variables d'observation. La distribution de probabilité correspondante se factorise en $p(y, x_1, x_2, x_3) = p(y).p(x_1 \setminus y).p(x_2 \setminus y).p(x_3 \setminus y)$, À la suite de l'hypothèse Naive Bayes De façon analogue, la figure 4-3 montre un classificateur MMC pour une séquence de trois variables x_1, x_2, x_3 d'entrée. La factorisation est

$p(x_1, x_2, x_3, y_1, y_2, y_3) = \Psi_1(y_1).\Psi_2(x_1, y_1).\Psi_3(x_2, y_2).\Psi_4(x_3, y_3).\Psi_5(y_1, y_2).\Psi_6(y_2, y_3)$ qui correspond à la HMM .



(a) Graphe Indépendant (b) Graphe Facteur

Figure 5-2 Classifieur Bayesien Naive



(a) Graphe Indépendant (b) Graphe Facteur

Figure 5-3 Graphe Indépendance et le facteur pour le modèle de Markov caché.

5.3.3 Modèle de graphe directe

Une distribution de probabilité peut être représentée par un modèle non orienté graphique en utilisant un produit de non-négatifs fonctions des cliques maximales de G. La factorisation est effectuée avec des nœuds conditionnellement indépendants ne semblent pas dans le même facteur, ce qui signifie qu'ils appartiennent aux cliques différents:

$$p(\vec{v}) = \frac{1}{Z}\prod_{c\in C}\Psi_c(\vec{v}_c) \qquad \textbf{Eq5-32}$$

Les facteurs $\Psi_c \geq 0$ nommés comme des fonctions potentielles des variables aléatoires $\vec{v}_c$ dans les cliques $c \in C$.

Les fonctions potentielles peuvent être n'importe quelle fonction arbitraire. En raison de cette généralité des fonctions potentielles , elles ne doivent pas nécessairement être des fonctions de probabilité. Ceci est en contraste aux graphes orientés, où la distribution conjointe est factorisé en un produit de distributions conditionnelles. Ainsi, la normalisation du produit de fonctions potentielles est nécessaire pour atteindre une mesure de probabilité correspondant. Ceci est donné par un facteur de normalisation Z. Calcul Z est l'un des principaux défis liés aux paramètre d'apprentissage comme en sommant sur toutes les variables possibles est nécessaire:

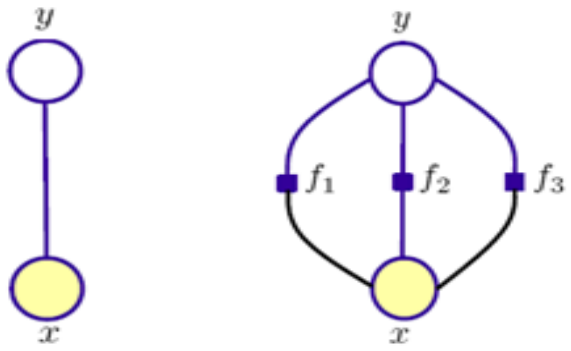
$$Z = \sum_{\vec{v}}\prod_{c\in C}\Psi_c(\vec{v}_c) \qquad \textbf{Eq5-33}$$

Le modèle entropie maximale a été discutée, qui peuvent être formulées par un tel produit de la non-négatifs potentiels des fonctions(à comparer l'équation4-29)

$$p_{\vec{\lambda}}(y \setminus x) = \frac{1}{Z_{\vec{\lambda}}}\prod_{i=1}^{m}\exp(\lambda_i f_i(x,y)) \qquad \textbf{Eq5-34}$$

Dans ces modèles log-linéaires, fonctions potentielles sont formulés comme la fonction exponentielle de caractéristiques pondérées. Une telle formulation est souvent utilisé car il répond à l'exigence de positivité stricte disjonctions potentielles. Figure4-4 (a) montre le graphique indépendance pour un classificateur

d'entropie maximale avec une variable x d'observation, un graphe facteur correspondant à trois fonctions est illustré à la figure4-4 (b).



(a) Graphe Independant (b) Graphe facteur

Figure 5-4 Classifieur de Maximum d'entropie

Graphe non orienté, ils diffèrent dans la manière dont la distribution de probabilité d'origine est factorisée. La factorisation en un produit de distributions de probabilités conditionnelles comme cela se fait dans un modèle orienté graphique est simple. Dans les graphes non orientés la factorisation des fonctions arbitraires est faite par la spécification explicite de la façon dont les variables sont liées. Mais il se fait au détriment d'avoir à calculer le facteur de normalisation.

5.4 Champs aléatoires conditionnels (CRF)

Dans la section précédente, certains modèles probabilistes bien connus ont été discutées à partir d'un point de vue mathématique. En outre, la représentation graphique, qui caractérise la distribution de probabilité du modèle sous-jacent, est représentée. Un **modèle de Markov caché** peut être compris comme la version séquence d'un modèle de Naive Bayes: au lieu de simples décisions indépendantes, un modèle de Markov caché modélise une séquence linéaire de décisions. En conséquence, Champs aléatoires conditionnels peut être comprise comme la version séquence de modèles d'entropie maximale, ce qui signifie qu'ils sont également des modèles discriminants.

En outre, contrairement à modèles de Markov cachés, Les modèle de champs aléatoires conditionnels « Conditionel Random Field »ne sont pas liés à la structure linéaire de séquence, mais peut être arbitrairement structuré. Dans ce qui suit, l'idée et le fondement théorique de champs aléatoires conditionnels est illustré. Tout d'abord, une formulation générale des champs aléatoires conditionnels est donnée suivie d'une discussion en profondeur de la forme la plus populaire des CRF, ceux qui ont une structure séquence linéaire. Un objectif principal sont les aspects de la formation et l'inférence. Cette section se termine par une brève discussion de la CRF arbitrairement structurés.

5.4.1 Principes de bases :

Introduit par [118], les champs aléatoires conditionnels (CRF) sont des modèles probabilistes pour le calcul de la probabilité $p(\vec{y} \setminus \vec{x})$ d'une sortie possible $\vec{y} = (y_1, ...y_n)$, compte tenu de l'entrée $\vec{x} = (x_1, ...x_n) \in X^n$ qui est aussi appelée l'observation. Une CRF en général peuvent être dérivée de la formule33:

$$p(\vec{v}) = \frac{1}{Z}\prod_{c\in C}\Psi_c(\vec{v}_c) \qquad \textbf{Eq5-35}$$

La probabilité conditionnelle $p(\vec{y} \setminus \vec{x})$ peut être écrite comme :

$$\begin{aligned}(\vec{y} \setminus \vec{x}) &= \frac{p(\vec{x},\vec{y})}{p(\vec{x})} \\ &= \frac{p(\vec{x},\vec{y})}{\sum_{\vec{y}'} p(\vec{y}',\vec{x})} \\ &= \frac{\frac{1}{Z}\prod_{c\in C}\psi_c(\vec{x}_c,\vec{y}_c)}{\frac{1}{Z}\sum_{\vec{y}'}\prod_{c\in C}\psi_c(\vec{x}_c,\vec{y}'_c)}\end{aligned} \qquad \textbf{Eq5-36}$$

De là, la formulation du modèle général de CRF est dérivé:

$$p(\vec{y} \setminus \vec{x}) = \frac{1}{Z(\vec{x})}\prod_{c\in C}\psi_c(\vec{x}_c,\vec{y}_c) \qquad \textbf{Eq5-37}$$

Nous notons ψ_C que sont les différents facteurs correspondant à cliques maximales du graphe indépendance. Voir Figure 4-5 pour un exemple d'un linéaire Trésor. Chaque facteur correspond à une fonction de potentiel qui combine des caractéristiques f_i différentes de la partie considérée de l'observation et la sortie. La normalisation suit à partir du dénominateur de l'équation 4-36:

$$Z(\vec{x}) = \sum_{c\in C}\psi(\vec{x}_c,\vec{y}') \qquad \textbf{Eq5-38}$$

En fait, pendant l'apprentissage et l'inférence, chaque instance d'un graphique distinct est construit à partir de modèles cliques soi-disant. Modèles Cliques fondés sur des hypothèses sur la structure des données sous-jacentes, en définissant la composition des cliques. Chaque clique est un ensemble de variables interdépendantes, à savoir celles contenues dans la fonction potentielle correspondant. Exemples de modèles de cliques sont donnés dans les sections suivantes.

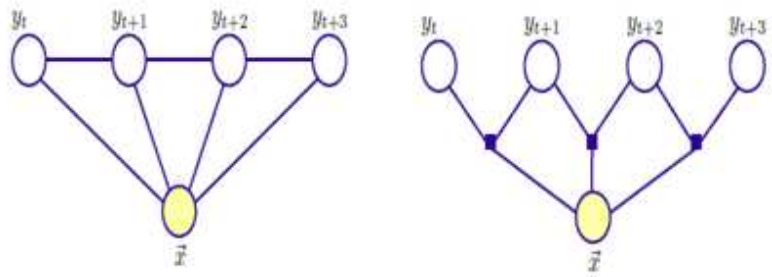
5.4.2 Chaines linéaires « Linear chain CRFs »

Une forme particulière de CRF , qui est structuré comme une chaîne linéaire, les modèles des variables de sortie sont modélisées par séquence. La figure 4-5 montre l'indépendance respective et graphique des facteurs. Le CRF a introduit dans l'équation4-37 peut être formulée comme.

$$p(\vec{y} \setminus \vec{x}) = \frac{1}{z(\vec{x})} \prod_{j=1}^{n} \psi_j(\vec{x}, \vec{y}) \qquad \textbf{Eq5-39}$$

Avec

$$Z(\vec{x}) = \sum_{\vec{y}'} \prod_{j=1}^{n} \psi_j(\vec{x}, \vec{y}') \qquad \textbf{Eq5-40}$$



(a) Graphe Independant (b) Graphe facteur

Figure 5-5 Chaine linéaire : Champs aléatoires conditionnels

Ayant les facteurs $\psi_j(\vec{x},\vec{y})$ dans la forme :

$$\psi_j(\vec{x},\vec{y}) = \exp(\sum_{i=1}^{m} \lambda_i f_i(y_{j-1}, y_j, \vec{x}, j)) \qquad \textbf{Eq 5-41}$$

Avec la supposition que n+1 est la longueur de la séquence d'observation , une chaine linéaire de CRF peut être écrite sous :

$$p_{\vec{\lambda}}(\vec{y} \setminus \vec{x}) = \frac{1}{Z_{\vec{\lambda}}(\vec{x})}.\exp(\sum_{j=1}^{n}\sum_{i=1}^{m} \lambda_i f_i(y_{j-1}, y_j, \vec{x}, j)) \qquad \textbf{Eq 5-42}$$

L'indice j est nécessaire par rapport au modèle de l'entropie maximale, car une étiquette de séquence est considérée au lieu d'une seule étiquette d'être prédite. Dans l'équation4-42, spécifie la position dans la séquence d'entrée $\vec{x}$. Notez que les poids λ_i ne sont pas dépendant de la position j. Cette technique, connue sous le nom de paramètre clef, est appliquée afin d'assurer un ensemble spécifié de variables d'avoir la même valeur.

La Sommation sur y, l'ensemble de toutes les séquences possibles d'étiquettes, est effectuée pour obtenir une probabilité possible.normalisation de [0; 1] est donnée par :

$$Z_{\vec{\lambda}}(\vec{x}) = \sum_{\vec{y} \in Y} \exp\left(\sum_{j=1}^{n} \sum_{i=1}^{m} \lambda_i f_i(y_{j-1}, y_j, \vec{x}, j)\right) \quad \textbf{Eq5-43}$$

Dans l'équation4-41 fournie la formulation d'une chaîne linéaire CRF est donné. Le déplacement de la somme sur les positions des séquences sous une fonction exponentielle, généralement la factorisation réelle pour une CRF devient plus évident:

$$p_{\vec{\lambda}}(\vec{y} \setminus \vec{x}) = \frac{1}{Z_{\vec{\lambda}}(\vec{x})} \cdot \prod_{j=1}^{n} \exp\left(\sum_{i=1}^{m} \lambda_i f_i(y_{j-1}, y_j, \vec{x}, j)\right) \quad \textbf{Eq5-44}$$

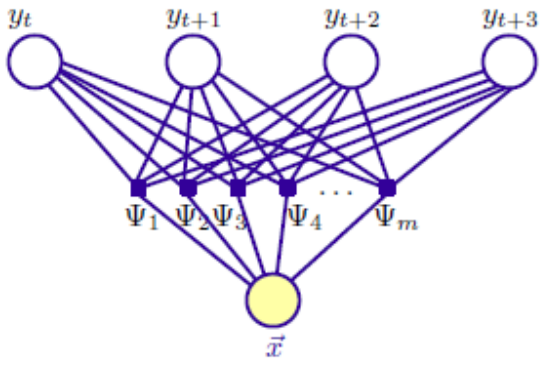


Figure 5-6 Interprétation alternative de chaine linéaire CRF.

Le facteur graphe dans la figure 4-5 (b) correspond à cette factorisation. On pourrait aussi déplacer la somme sur les différentes caractéristiques à travers fonction exponentielle:

$$p_{\vec{\lambda}}(\vec{y} \setminus \vec{x}) = \frac{1}{Z_{\vec{\lambda}}(\vec{x})} \cdot \prod_{i=1}^{m} \exp\left(\sum_{j=1}^{n} \lambda_i f_i(y_{j-1}, y_j, \vec{x}, j)\right) \quad \textbf{Eq 5-45}$$

Dans cette interprétation, les facteurs ne sont pas \ en cours d'exécution "au cours de la séquence, mais sur les caractéristiques.
Le facteur graphe avec $\psi_i = \exp(\sum_{j=1}^{n} \lambda_i f_i(y_{j-1}, y_j, \vec{x}, j))$ des facteurs correspondant aux caractéristiques f_i est donné dans la figure 4-6. Cette interprétation est moins intuitive, mais montre la relation avec le modèle entropie maximale (figure 4-4). Le modèle peut être interprété avec bien plus de facteurs en déplaçant les deux sommes à l'avant de la fonction exponentielle.

$$p_{\vec{\lambda}}(\vec{y} \setminus \vec{x}) = \frac{1}{Z_{\vec{\lambda}}(\vec{x})} \cdot \prod_{i=1}^{m} \prod_{j=1}^{n} \exp(\lambda_i f_i(y_{j-1}, y_j, \vec{x}, j)) \quad \textbf{Eq 5-46}$$

Le facteur graphe facteur correspondant n'est pas représenté ici en raison du nombre important de facteurs dans le graphe. La factorisation sur la base de cliques maximales(voir l'équation 4-44)est celui habituellement appliqué pour une CRF à chaîne linéaire. Les deux autres factorisations (voir les équations 4-45et 4-46)ne sont pas conforme à cette maximalité.
En général, factoriser selon cliques composées de moins de nœuds variables que le plomb clique maximale à des inexactitudes, car toutes les dépendances sont correctement pris en compte.

Dans ce cas, cependant, il conduit à des calculs redondants que l'on peut voir dans l'équation4-46.Le reste du chapitre est basé sur l'idée de la factorisation première.

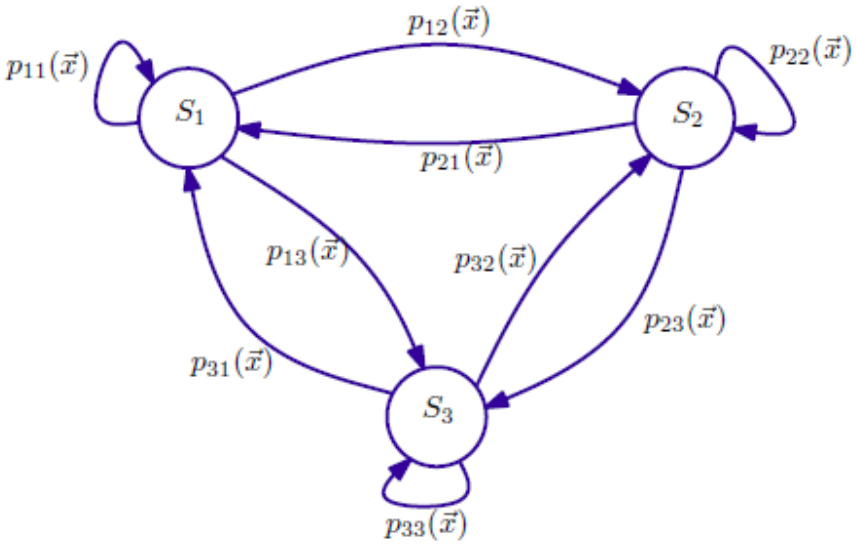


Figure 5-7 Exemple d'un automate à états finis stochastique

CRF à chaîne linéaire sont exactement un modèle une seule clique $c \in C$: Il spécifie l'indépendance graphique pour se composent de connexions entre y_j et y_{j-1} et $\vec{x}$:

$C = \{\psi_j(y_j, y_{j-1}, \vec{x}) \mid \forall j \in \{1, ..., n\}\}$.

En raison de cette structure particulière, il est possible de représenter une chaîne linéaire CRF par un automate d'état finis stochastique(SFSA) similaire à modèles de Markov cachés. Pour des fins de mise en œuvre. Dans cet automate les probabilités de transition dépendent de la séquence d'entrée $\vec{x}$. Sa structure est en général arbitraire, mais l'approche (straight-forward) est d'utiliser un automate entièrement connecté avec les États S_l où $l \in L$ un État est utilisé pour chaque symbole dans l'alphabet d'étiquetage. Cet automate est représenté par $|L| = 3$ dans la figure 4-7.

Comme indiqué dans l'équation 4-42, les fonctions dépendent de la séquence étiquette et la présente sur les transitions d'état dans l'automate à états finis. Donc, il est important de souligner que seul un sous-ensemble de toutes les caractéristiques f_i est utilisée dans chaque transition dans le graphique :

- En raison de cette structure particulière, il est possible de représenter une chaîne linéaire CRF par une Construire un SFSA $S = (S, T)$ sur l'ensemble des états S (avec des transitions $T = (s, \dot{s}) \in S^2$). Il peut être entièrement connecté, mais il est également possible d'interdire certaines transitions.
- Indique un ensemble de modèles de caractéristiques $F = \{g_1(\vec{x}, j), ..., g_h(\vec{x}, j)\}$ sur la séquence d'entrée. Ceux-ci ne sont pas utilisés directement mais pour la génération des caractéristiques f_i
- Générer un ensemble de caractéristiques $F = \{\forall s, \dot{s} \in S . \forall g_0 \in F : f_k(s, \dot{s}, g_0)\}$. Jusqu'à présent, seuls de premier ordre à chaîne linéaire CRF ont été pris en considération .Pour définir la chaîne linéaire de CRF avec un ordre plus élevé, les caractéristiques doivent avoir la forme :

$$f_i(\vec{y}, \vec{x}, j) = f_i(h_j(\vec{y}), \vec{x}, j) \qquad \textbf{Eq 5-47}$$

- Generate set of features $F = \{\forall s, \dot{s} \in S . \forall g_0 \in F : f_k(s, \dot{s}, g_0)\}$. Until now, only first-order linear-chain CRFs have been considered. To define linear chain CRFs with a higher order (see McDonald and Pereira, 2005), the features need to have the form

Avec:

$$h_j(\vec{y}) = (y_{j-k+1}, ...y_j) \qquad \textbf{Eq5-48}$$

L'ordre est donné par k. Pour ordres supérieurs $(k > 2)$ le même automate probabiliste d'état est utilisé parla combinaison de différentes valeurs de sortie précédentes y_i dans l'état spécial. Par exemple, pour $k = 3$ l'ensemble des états serait $S' = \{(S_i, S_j)\}$ *for all* $i, j \in \{1, ..., |S|\}$ (selon le premier ordre SFSA à la figure 4-7).

Pour cela spéciale à chaîne linéaire la structure de la CRF, l'apprentissage et l'inférence sont formulées d'une manière similaire à celle de modèles de Markov cachés comme des problèmes de base :

- Compte tenu de l'observation $\vec{x}$ et un CRF M Comment trouver la séquence d'étiquettes $\vec{y}$ la plus probable?
- Compte tenu des séquences d'étiquettes Y et des séquences d'observation X : Comment trouver les paramètres d'unCRF M afin de maximiser la probabilité $p(\vec{y} \setminus \vec{x}, M)$?

Problème I est l'application la plus courante d'un champ aléatoire conditionnel, de trouver une séquence d'étiquettes pour une observation.

Problème II estla question sur la façon de former, d'ajuster les paramètres M qui sont en particulier les poids λ_i.

Dans les approches discriminantes, la probabilité $p(\vec{x} \setminus M)$ n'est pas modélisée. Son estimation est une autre question fondamentale dans le contexte de modèles de Markov cachés et n'est pas considérée ici.

5.4.2.1 Apprentissage

Pour tous les types de CRF, ainsi que pour les modèles d'entropie maximale, la méthode du maximum de vraisemblance peut être appliquée pour l'estimation des paramètres. Cela signifie, que la formation du modèle se fait en maximisant la log-vraisemblance l sur les données d'apprentissage τ :

$$\bar{l}(\tau) = \sum_{(\vec{x},\vec{y})\in\tau} \log p(\vec{y} \setminus \vec{x}) \qquad \textbf{Eq5-49}$$

$$= \sum_{(\vec{x},\vec{y})\in\tau} \left[\log\left(\frac{\exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i f_i(y_{j-1}, y_j, \vec{x}, j)\right)}{\sum_{\vec{y}'\in Y}\exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i f_i(y'_{j-1}, y'_j, \vec{x}, j)\right)}\right) \right]$$

Pour éviter le débordement de la probabilité, on emploie le terme $-\sum_{i=1}^{n}\frac{\lambda_i^2}{2\sigma^2}$.

Cette technique est établi pour une utilisation dans des modèles d'entropie maximale et peut également être appliquée ici .le paramètre σ^2 modélise un compromis entre la fréquence d'observation d'une caractéristique et la norme carré du vecteur des poids.

Les poids les plus petits sont forcés à être de telle sorte que la chance que peu de poids élevés dominent est réduite. Pour la dérivation, la notation de la fonction de vraisemblance $l(\tau)$ est réorganisée:

$$l(\tau) = \sum_{(\vec{x},\vec{y})\in\tau}\left[\log\left(\frac{\exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i f_i\left(y_{j-1},y_j,\vec{x},j\right)\right)}{\sum_{\vec{y}'\in Y}\exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i f_i\left(y'_{j-1},y'_j,\vec{x},j\right)\right)}\right)\right]-\sum_{i=1}^{m}\frac{\lambda_i^2}{2\sigma^2} \quad \textbf{Eq5-50}$$

$$= \sum_{(\vec{x},\vec{y})\in\tau}\left[\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i f_i\left(y_{j-1},y_j,\vec{x},j\right)\right)-\log\left(\sum_{\vec{y}'\in Y}\exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i f_i\left(y'_{j-1},y'_j,\vec{x},j\right)\right)\right)\right]-\sum_{i=1}^{m}\frac{\lambda_i^2}{2\sigma^2}$$

$$= \underbrace{\sum_{(\vec{x},\vec{y})\in T}\sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i f_i\left(y_{j-1},y_j,\vec{x},j\right)}_{A} - \underbrace{\sum_{(\vec{x},\vec{y})\in T}\log\underbrace{\left(\sum_{\vec{y}'\in Y}\exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i f_i\left(y'_{j-1},y'_j,\vec{x},j\right)\right)\right)}_{Z_{\vec{\lambda}}(\vec{x})}}_{B} - \underbrace{\sum_{i=1}^{m}\frac{\lambda_i^2}{2\sigma^2}}_{C}$$

Les dérivées partielles de $l(\tau)$ par les coefficients de pondération λ_k sont calculée séparément pour les parties A, B et C. Le calcul de la partie A est donnée par :

$$\frac{\partial}{\partial\lambda_k}\sum_{(\vec{x},\vec{y})\in T}\sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i f_i\left(y_{j-1},y_j,\vec{x},j\right) = \sum_{(\vec{x},\vec{y})\in T}\sum_{j=1}^{n} f_i\left(y_{j-1},y_j,\vec{x},j\right) \quad \textbf{Eq5-51}$$

La dérivation de la partie B, qui correspond à la normalisation, est donnée par :

$$\frac{\partial}{\partial\lambda_k}\sum_{(\vec{x},\vec{y})\in T}\log Z_{\vec{\lambda}}(\vec{x}) = \sum_{(\vec{x},\vec{y})\in T}\frac{1}{Z_{\vec{\lambda}}(\vec{x})}\frac{\partial Z_{\vec{\lambda}}(\vec{x})}{\partial\lambda_k} \quad \textbf{Eq5-52}$$

$$= \sum_{(\vec{x},\vec{y})\in T}\frac{1}{Z_{\vec{\lambda}}(\vec{x})}\sum_{\vec{y}'\in Y}\exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i f_i(y'_{j-1},y'_j,\vec{x},j)\right).\sum_{j=1}^{n} f_k(y'_{j-1},y'_j,\vec{x},j)$$

$$= \sum_{(\vec{x},\vec{y})\in T}\sum_{\vec{y}'\in Y}\frac{1}{Z_{\vec{\lambda}}(\vec{x})}\underbrace{\exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i f_i(y'_{j-1},y'_j,\vec{x},j)\right)}_{=p_{\vec{\lambda}}(\vec{y}\setminus\vec{x})\ \textit{see equation 42}}.\sum_{j=1}^{n} f_k(y'_{j-1},y'_j,\vec{x},j)$$

$$= \sum_{(\vec{x},\vec{y})\in T}\sum_{\vec{y}'\in Y} p_{\vec{\lambda}}(\vec{y}'\setminus\vec{x}).\sum_{j=1}^{n} f_k(y'_{j-1},y'_j,\vec{x},j).$$

Partie C, la dérivation du terme sanction, est donnée par :

$$\frac{\partial}{\partial \lambda_k}(-\sum_{i=1}^{m}\frac{\lambda_i^2}{2\partial^2}) = -\frac{2\lambda_k}{2\sigma^2} = -\frac{\lambda_k}{\sigma^2} \qquad \textbf{Eq5-53}$$

La fonction log-vraisemblance dans l'équation4-52 est concave: Le premier terme est linéaire(voir l'équation 4-51), le second terme appartient à la normalisation. Par conséquent, il ne change pas la concavité de la fonction et le dernier terme est concave(voir l'équation 4-53), est si toute la fonction. L'équation4-51,la dérivation d'une partie A ,est la valeur attendue dans la distribution empirique d'une caractéristique f_i.

$$\tilde{E}(f_i) = \sum_{(\vec{x},\vec{y})\in T}\sum_{j=1}^{n} f_i(y_{j-1}, y_j, \vec{x}, j)$$

Par conséquent, l'équation 4-52 ,la dérivation de la partie B, est l'espérance dans le cadre du modèle de distribution:

$$E(fi) = \sum_{(\vec{x},\vec{y})\in T}\sum_{\vec{y}'\in Y} p_{\vec{\lambda}}(\vec{y}' \setminus \vec{x})\sum_{j=1}^{n} f_i(y'_{j-1}, y'_j, \vec{x}, j) \qquad \textbf{Eq5-54}$$

Les dérivées partielles de $l(\tau)$ peut aussi être interprété comme:

$$\frac{\partial l(\tau)}{\partial \lambda_k} = \tilde{E}(f_k) - E(f_k) - \frac{\lambda_k}{\sigma^2} \qquad \textbf{Eq5-55}$$

Notez la relation d'équations 4-54 et 4-55 et les équations 4-13 et 4-17 qui ont été formulées

pour le modèle entropie maximale. Outre le fait que, pour le CRF possèdent plusieurs variables. La différence est l'absence du facteur $\frac{1}{N}$, qui n'est pas pertinent

pour trouver le maximum par le rapprochement de la première dérivation $\tilde{E}(f_k) - E(f_k) - \frac{\lambda_k}{\sigma^2} = 0$

Le calcul du $\tilde{E}(f_i)$ se fait facilement en comptant combien de fois chaque caractéristique se produit dans le les données d'apprentissage. Le calcul $E(f_i)$ est directement impraticable en raison du nombre élevé de séquences de balises possibles ($|Y|$). Rappelons que, pour les modèles entropie maximale, peut être calculée efficacement en raison du petit nombre de variables de sortie y différente dans la plupart des applications.

Dans un CRF, des séquences de variables de sortie conduisent à une complexité combinatoire énorme. Ainsi, une approche de programmation dynamique est appliqué, connu sous le nom de l'algorithme avant-arrière (Forward-Backward)décrit à l'origine pour les modèles de Markov cachés. Cet algorithme peut aussi être utilisé pour chaîne linéaire aléatoire dans une forme légèrement modifiée.

Selon[121] une fonction $T_j(s)$ est définie, qui associe un état s à un seul j position d'entrée à un ensemble de permis états suivants à la positionj+1 et la fonction inverse $T_j^{-1}(s)$, qui associe l'ensemble des états de ses prédécesseurs s possibles. États spéciaux $\perp$, T sont définis pour le début et la fin de la séquence. Un exemple pour les Etats de la figure 4-1 est $T_j(S_1) = \{S_1, S_2, S_3\}$, avant (α) et en arrière scores (β) seront utilisés, qui peut être comprise en général sous forme de messages envoyés sur le réseau, dans ce qui suit supposé être une chaîne linéaire:

$$\alpha_j(s \setminus \vec{x}) = \sum_{s' \in T_j^{-1}(s)} \alpha_{j-1}(s' \setminus \vec{x}).\Psi_j(\vec{x}, s', s) \qquad \textbf{Eq5-56}$$

$$\beta_j(s \setminus \vec{x}) = \sum_{s' \in T_j^{-1}(s)} \beta_{j-1}(s' \setminus \vec{x}).\Psi_j(\vec{x}, s, s') \qquad \textbf{Eq5-57}$$

En ce qui concerne la définition des potentiels dans l'équation4-41, les caractéristiques sont définies sur les états spéciaux: $\Psi_j(\vec{x}, s, s') = \exp(\sum_{i=1}^{m} \lambda_i f_i(y_{j-1} = s, y_j = s', \vec{x}, j))$.

Les α sont les messages envoyés depuis le début de la chaîne à la fin. Les β fonctions sont des messages envoyés à partir de la fin de la chaîne au début. Ils sont initialisés par :

$$\alpha_0(\perp | \vec{x}) = 1 \qquad \textbf{Eq5-58}$$

$$\beta_{|\vec{x}|+1}(T \setminus \vec{x}) = 1 \qquad \textbf{Eq5-59}$$

Avec ces messages, il est possible de calculer l'espérance dans le cadre du modèle de distribution efficacement par:

$$E(f_i) = \sum_{(\vec{x},\vec{y}) \in T} \frac{1}{Z_{\vec{\lambda}}(\vec{x})} \sum_{j=1}^{n} \sum_{s \in S} \sum_{s' \in T_j(s)} f_i(s, s', \vec{x}, j).\alpha_j(s \setminus \vec{x})\Psi_j(\vec{x}, s, s')\beta_{j+1}(s' \setminus \vec{x}) \qquad \textbf{Eq5-60}$$

La partie soulignée de la formule4-61 peut être comprise comme le calcul des potentiels dans toutes les combinaisons de séquences d'états dans les données d'entraînement. Une visualisation agréable est le diagramme en treillis que l'on appelle dans la figure4-7, dans le quelles chemins possibles de messages envoyés sont représentés. Les valeurs de α et β sont stockés après une itération de sorte qu'ils doivent être calculés qu'une seule fois. Le facteur de normalisation est calculée par :

$$Z_{\vec{\lambda}}(\vec{x}) = \beta_0(\perp \setminus \vec{x}) \qquad \textbf{Eq5-61}$$

L'algorithme d'avant-arrière (Forward-Backward) a une complexité d'exécution de l'ordre de $O(|S|^2 n)$, il est donc linéaire en la longueur de la séquence et quadratique en le nombre d'états.

5.4.2.2 Inférence

Le problème de l'inférence est de trouver la séquence $\vec{y}$ la plus probable pour la note des observations donnée $\vec{x}$, que ce n'est pas sur le point de choisir une séquence d'états, qui sont individuellement plus probable. Ce serait la maximisation du nombre d'états corrects dans la séquence. En revanche, pour trouver la séquence la plus probable de l'algorithme de Viterbi est appliqué. L'algorithme de Viterbi est similaire à l'algorithme avant-arrière (Forward-Backward). La différence principale est, qu'au lieu de sommation, **une maximisation** est appliquée.
La quantité $\delta_j(s \setminus \vec{x})$ qui est le meilleur score long d'un trajet, à la position j qui se termine dans l'état s, est défini comme :

$$\delta_j(s \setminus \vec{x}) = \max_{y_1, y_2, \ldots, y_{j-1}} p(y_1, y_2, \ldots, y_j = s \setminus \vec{x}) \qquad \textbf{Eq5-62}$$

L'étape induction est :

$$\delta_{j+1}(s \setminus \vec{x}) = \max_{s' \in S} \delta_j(s').\Psi_{j+1}(\vec{x}, s, s') \qquad \textbf{Eq5-63}$$

Le vecteur $\Psi_j(s)$ permet de suivre les valeur sj et s. L'algorithme fonctionne alors comme suit:

1. **Initialization:**

Les valeurs pour toutes les étapes de l'état de départ $\perp$ à tous les premiers états possibles sont mis à la valeur du facteur correspondant.

$$\forall s \in S: \quad \delta_1 = \Psi_1(\vec{x}, \perp, s) \qquad \textbf{Eq 5-64}$$
$$\psi(s) = \perp$$

2. **Recursion:**

Les valeurs pour les prochaines étapes sont calculées à partir de la valeur actuelle et les valeurs maximales en ce qui concerne tous les états possibles suivants s'

$$\forall s \in S: 1 \le j \le n: \quad \delta_j(s) = \max_{s' \in S} \delta_{j-1}(s') \Psi(\vec{x}, s', s) \qquad \textbf{Eq 5-65}$$
$$\psi_j(s) = \arg\max_{s' \in S} \delta_{j-1}(s') \Psi(\vec{x}, s', s)$$

3. **Terminaison:**

$$p^* = \max_{s' \in S} \delta_n(s') \qquad \textbf{Eq 5-66}$$

$$\vec{y}_n^* = \arg\max_{s' \in S} \delta_n(s') \qquad \textbf{Eq5-67}$$

5.4.2.3 Chemin (séquence d'état) retour en arrière:

Recalculer le chemin optimal à partir du réseau en utilisant la piste en gardant les valeurs ψ_t

$\vec{y}_t^* = \psi_{t+1}(\vec{y}_{t+1}^*) \qquad t = n-1, n-2, \ldots, 1$

Étapes1-3sont très similaires à l'algorithme avant-arrière (Forward-Backward) .Une lattice est rempli avec les meilleurs valeurs. Étape 4 lit le meilleur chemin à partir de ce réseau.

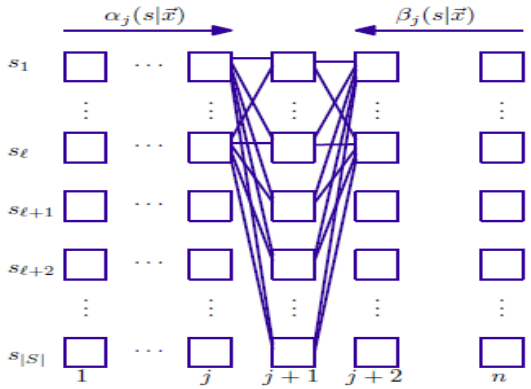


Figure 5-8 Passage de message sur l'algorithme d'avant_arrière (Forward-Backward)

5.5 Conclusion:

Dans les sections ci-dessous, nous avons vu que le **modèle Naive Bayes** est une approche pour classer les variables de classe unique en fonction des valeurs de plusieurs caractéristiques. Dans tel modèle, les valeurs d'entrée sont supposés être conditionnellement indépendantes. Il s'agit d'une approche dite **générative**, la modélisation de la probabilité conjointe $p(y, \vec{x})$ des valeurs d'entrée $\vec{x}$ et la variable y de classe. Le **modèle de Markov caché** est une extension au modèle **Naive Bayes** pour les données structurées de façon séquentielle représentant également les dépendances des variables comme une distribution de probabilité conjointe.

La modélisation de la probabilité conjointe présente des inconvénients en raison de la complexité des calculs. Le **Modèle entropie maximale**, en revanche, est basé surla modélisation de la **probabilité conditionnelle** $p(y \setminus x)$ .Comme le modèle **Naive Bayes,** il s'agit d'une approche pour classer une variable de **classe unique** en fonction des valeurs de plusieurs caractéristiques. La **différence** est la prise en compte de la **probabilité conditionnelle au lieu de la probabilité conjointe**. Même si un modèle de Markov caché est une extension séquentielle du modèle Naive Bayes.

Le modèle de champs conditionnels aléatoires (CRF) peut être considéré comme une extension séquentielle au modèle **d'entropie maximale,** les deux modèles : l'entropie maximale ainsi que champs conditionnel aléatoire sont connus comme **des approches discriminatoires**.

Une comparaison graphique de ces modèles est donnée dans la figure 4-9.

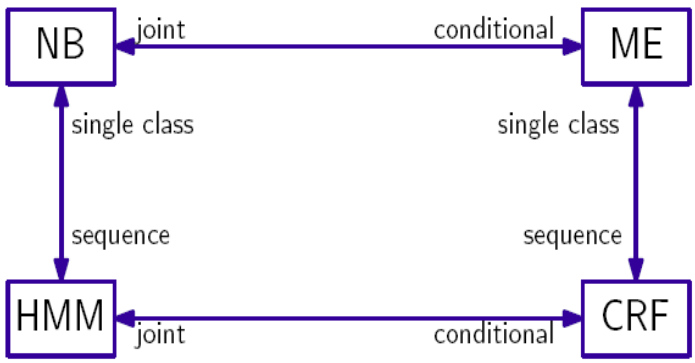


Figure 5-9Vue d'ensemble des modèles probabilistes

Chapitre 5Détection des parties du corps humain

6.1 Introduction

Le problème de détection et localisation des objets, tel que les personnes, dans des images arbitraires constitue un grand challenge dans le domaine de la vision par ordinateur. Les objets sont souvent alignés selon des poses arbitraires, soumises aux problèmes d'occlusion et variation de taille (échelle). Pour surmonter ces problème ; et relever le défi, les nouvelles recherches font recours à des approches basé-modèle déformable subdivisé capables à surmonter le problème de l'apparence variée liée aux objets articulés.

Les modèles à base de structure déformable exploitent l'histogramme de gradient orienté « HOG » comme descripteur d'apparence vu son efficacité à décrire la structure locale d'un objet grâce au gradient.

Les structures picturales exploitent un modèle composé de racine et des pièces composantes, comme par exemple dans le cas de la détection de visage, généralement le nez est considéré comme étant la racine du modèle, cette considération est nécessaire durant le processus de l'inférence.

Dans notre approche nous considérant la racine comme étant l'objet en sa globalité , càd le corps humain et les autres parties comme étant les parties du corps humain , dans notre cas nous considérons : La tête, le torse, les membres inférieures et supérieures.

Les parties sont apprises d'une façon non supervisés en les modélisant comme des variables cachées « Latent Variables », ceci pour surmonter le problème d'occlusion.

Le modèle est modélisé sous une forme d'arbre à structure d'étoile pour codifier les liaisons entre les parties, nécessaire pour la phase inférence.

Dans ce travail, nous proposons une approche basée modèle à structure déformable, utilisant un modèle spatial subdivisé en parties pour décrire les fenêtres d'objet

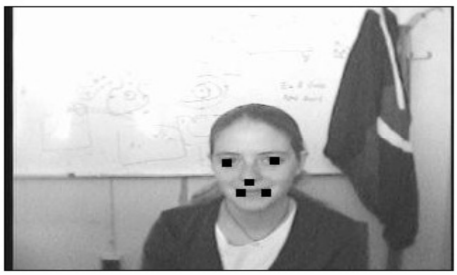
permettant de surmonter le problème liée à l'occlusion et la variation de la pose. Nous utilisons l'histogramme de gradient orienté pour la description de l'apparence des objets, une pyramide de descripteurs est exploitée afin de détecter l'objet à toutes les échelles possibles.

6.2 Cadre probabiliste

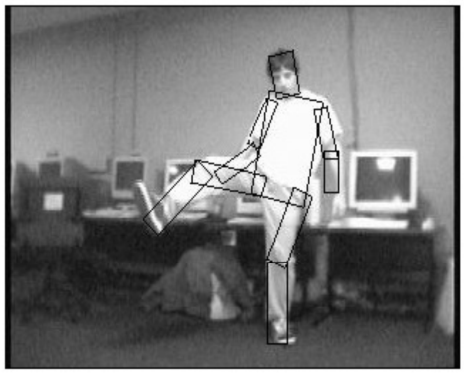
6.2.1 Définition

Le modèle basé structure picturale est une collection de pièce d'un objet disposé dans une configuration déformable. Chaque partie du modèle codifie des propriétés visuelles locales de l'objet, et la configuration déformable peut être comparé à un ressort en forme de connexions entre certaines paires de pièces. Le modèle d'apparence de chaque partie est donné par le biais d'une fonction, qui mesure la ressemblance d'un emplacement dans une image à la partie correspondante. Le meilleur résultat d'un tel modèle sur l'image donnée se trouve en minimisant la fonction d'énergie qui mesure à la fois le coût d'un résultat pour chaque partie et un coût de déformation pour chaque paire de parties reliées. Voir la figure5.1 pour un exemple de deux modèles appariés contre les images.
Le modèle d'apparence pour chaque partie peut être assez générique. C'est parce que nous ne reconnaissent pas les parties de façon indépendante, mais avec d'autres parties de l'objet.
Ceci est différent de la plupart des méthodes qui utilisent des approches basées-parties « part-based », où, en une phase initiale, les pièces sont reconnues individuellement, et dans la phase suivante, ils sont assemblés en groupes pour former des objets.

Reconnaissance individuelle des parties exige des modèles de parties complexes, alors que si nous nous appuyons sur leur configuration en distinguant que nous pouvons utiliser les modèles d'apparence assez simples pour chaque partie.



(a)



(b)

Figure 6-1 Exemple de deux structures picturales mises en correspondance avec des images. (a) détection du visage ; (b) détection du corps humain.

6.2.2 Formulation du problème

Comme nous avons précisé dans notre chapitre « Motivation et Problématique » , que notre but est de trouver une approximation grossière de la réalité . La formulation probabiliste fournit un moyen naturel pour trouver plusieurs bonnes mises en correspondances d'un modèle à une image. Nous pouvons y parvenir par échantillonnage des configurations d'objets à partir de leur distribution de probabilité postérieure pour une image observée.

En outre, le cadre probabiliste peut être utilisé pour apprendre les paramètres du modèle. La façon la plus standard de rapprocher un problème de reconnaissance d'objet aux paramétrages probabilistes est comme suit :

a. Soit θ une série de paramètres du modèle, I représente une image, et un vecteur aléatoire $L=(l_0,...,l_{n-1})$ représente la configuration d'un objet.
b. La distribution $p(I/L,\theta)$ est la probabilité d'observer une image particulière en ayant un emplacement d'objet.
c. La distribution $p(L\setminus\theta)$ mesure la probabilité à priori d'une configuration particulière.
d. La distribution à postériori $P(L\setminus I,\theta)$ spécifie la probabilité qu'une partie est à un emplacement particulier, en ayant une image observée.

En utilisant la loi de Bayes, la distribution postérieures , peut être écrite comme :

$$p(L \setminus I,\theta) \propto p(L \setminus \theta)\, p(I \setminus L,\theta) \qquad \textbf{Eq 6-1}$$

Une difficulté inhérente l'approche bayésiennes de déterminer à priori la distribution $p(L)$. Souvent, le priori est considéré comme uniforme, c'est une connaissance préalable est complètement rejetée. Cependant, seulement avec une information préalable nous obtenons une véritable approche bayésienne [173].
Pour les structures basées modèle déformables l'apriori à travers les configurations de l'objet encode l'information sur les emplacements relatifs des parties. Dans le cas d'un modèle humain, il spécifie une distribution de la pose humaine. L'apriori est un élément important dans le modèle.
Afin de réaliser l'inférence, nous avons besoin de décomposer $p(L\setminus I)$:

1- On considère $p(I\setminus L)$, la probabilité d'apercevoir une image , en ayant une configuration d'un objet particulier.
2- On suppose que la configuration $L=1$, nous pouvons décomposer l'image I en une série de sous images $I_B,I_0,...,I_{n-1}$, de telle sorte qu'une sous-image

I_i, $i = 0,...,n-1$ contient uniquement les parties v_i et les pixels qui l'entourent.

3- I_B : est l'arrière-plan, pour illustration référer vous à l'image….

En outre, nous supposons que ce que nous voyons dans la sous-image est indépendant de l'emplacement et le contenu des autres, compte tenu des paramètres du modèle.

Cette supposition nous permet une abstraction de certaines informations telles que la symétrie dans l'habillement. Cette décomposition est nécessaire pour notre approche.

Nous pouvons écrire :

$$p(I/L) = p(I_B, I_0, ..., I_{n-1}/L) = p(I_B/L)\,p(I_0/L)...p(I_{n-1}/L) = p(I_B)\,p(I_0/L_0)...p(I_{n-1}/L_{n-1}) \propto p(I_0/L_0)...p(I_{n-1}/L_{n-1})$$

Où $p(I_B)$ est uniforme, nous obtenons :

$$p(I/L) \propto \prod_{i=0}^{n-1} p(I_i/L_i) \qquad \textbf{Eq6-2}$$

De la discussion qui précède, nous voyons que cette approximation est bonne si les paires ne se chevauchent pas et ne sont pas obstrués. Cependant, Ceci est rarement vrai. En conséquence, pour obtenir une configuration postérieur multimodale, où la configuration correcte correspond à la somme des maximums locaux. C'est l'une des raisons en faveur d'échantillonnage par rapport à la solution MAP.

Figure 6-2 Découpage de l'image en ayant une configuration l=(l1,l2) et le paramètre du modèle θ

Considérons la distribution a priori sur les configurations d'objets p(L). Un graphe non orienté G pour qu'il soit un graphe d'indépendance du modèle graphique représentant la distribution a priori $p(L_0, \ldots, L_{n-1})$.

Des algorithmes d'inférences efficaces telles que la propagation de croyance (Belief propagation) existent si le graphe d'indépendance est un **arbre**. On peut utiliser ces algorithmes d'inférence pour calculer le meilleur correspond (MAP), ainsi que pour l'échantillon à travers le postériori. Appliquer une restriction sur le graphe de l'indépendance pour qu'il soit un arbre peut être naturel. Par exemple, on peut prendre un arbre correspondant à une structure de squelette d'un objet articulé. Un modèle d'arbre introduit, cependant, une forte hypothèses concernant les indépendances entre les parties du corps de la structure picturale. Ainsi, il manque des informations importantes sur l'objet articulé comme les coordonnées des articulations.

Pour plus de simplicité dans la partie restante de la thèse, nous supposons que le graphe est un arbre indépendant.

Pour nos travaux, nous utilisons un **graphe indépendant direct** $G=(A,V)$. Tel que spécifié précédemment $V=\{v_0,...,v_{n-1}\}$ est une série de sommets , où un sommet v_i correspond à un emplacement d'une partie l_i.

Soit $A=\{(v_i \rightarrow v_j)\}$ une série d'arc , où v_i est un nœud parent et v_j est le nœud enfant.

Nous considérons v_0 comme la racine de l'arbre direct.

L'arbre direct G représente un réseau Bayesien, où les probabilités de distribution de jointure sont factorisées comme :

$$p(L)=p(L_0)\prod_{(v_i \rightarrow v_j)\in A} p(L_j / L_i) \qquad \textbf{Eq 6-3}$$

Dans les paragraphes précédents , nous avons obtenu une décomposition factorielle (5.3) de $p(I/L)$, la probabilité d'apercevoir une image dans la mesure où l'objet se trouve à une certaine configuration ; Ainsi que décomposition (5.4) de $p(L)$, la probabilité a priori que l'objet assumera une configuration particulière. Ceux-ci peuvent être remplacés dans (4.2) ce qui donne :

$$p(L/I) \propto p(L_0)\prod_{(v_i \rightarrow v_j)\in A} p(L_j / L_i)\prod_{v_i \in V} p(I_i / L_i) \qquad \textbf{Eq6-4}$$

Ensuite, nous allons avoir une relation entre la solution MAP et la solution optimale du problème de minimisation de l'énergie(5,1). Pour nous supposons le priori $p(L_i)$ est une distribution uniforme. On obtient alors de l'équation (5-4)

Supposons maintenant, que nous voulons trouver la solution MAP pour le problème cité précédemment. Donc nous avons besoin de la configuration des parties du corps l^{**} maximisant $p(l/I)$, ce qui revient à dire :

$$l^* = \arg\max_L p(L/I) = \arg\max_L \prod_{(v_i,v_j)\in E} p(L_j, L_i) \prod_{v_i \in V} p(I_i / L_i) \qquad \textbf{Eq6-5}$$

Supposons maintenant, que nous voulons trouver la solution MAP pour le problème cité précédemment. Il est à noter que le problème de la mise en correspondance des parties à une image peut être défini comme un problème de **minimisation d'énergie**.

Soit $m_i(l_i)$ une fonction de mesure du degré de non-correspondance, lorsque la partie v_i est placée sur l'emplacement l_i sur l'image. Pour une paire de parties interconnectées, soit $d_{ij}(l_i, l_j)$ la fonction qui mesure le degré de déformation du ressort en forme de liaison entre les parties v_i et v_j, quand elles sont placées aux emplacements l_i et l_j sur l'image.

Donc nous avons besoin de la configuration des parties du corps l^{**} maximisant $p(l/I)$, qui revient à la minimisation $-\log p(l/I)$. En utilisant l'équation (5.5), nous auront besoin de minimiser l'équation $\sum_{v_i \in v} -\log p(I_i/l_i) + \sum_{(v_i,v_j)\in E} -\log p(l_i, l_j)$.

Maintenant, soit $m_i(l_i) = -\log p(I_i/l_i)$ la mesure du coût de mise en correspondance quand la partie v_i est mise en correspondance sur l'image à l'emplacement l_i, et $d_{ij}(l_i, l_j) = -\log p(l_i, l_j)$ la mesure du coût de déformation, à quel point les emplacements des pièces v_i et v_j s'accordent avec le modèle a priori. Remarquez maintenant que la solution MAP l^{**} correspond exactement à la solution optimale l^* obtenant en résolvant le problème de minimisation d'énergie (5.1). Ainsi le cadre probabiliste permet de formuler le problème d'optimisation de la section

précédente. De plus, il nous permet d'appliquer les techniques d'inférence approximatives, tel que l'échantillonnage.
Maintenant, l'adaptation optimale du modèle à l'image est naturellement définie comme :

$$l^* = \arg\min_I \left(\sum_{i=1}^{n} m_i(l_i) + \sum_{(v_i,v_j)\in E} d_{ij}(l_i,l_j)\right) \quad \textbf{Eq 6-6}$$

Ce qui est une configuration qui minimise la somme des coûts de la m_i correspondantes et la somme des coûts de déformation d_{ij} pour des paires de parties liées.

6.2.3 Estimation des paramètres du modèle

Supposons que nous avons une série d'images $\{\mathrm{I}^1,...,\mathrm{I}^m\}$ et $\{\mathrm{L}^1,...,\mathrm{L}^m\}$ représentent les configurations correspondantes à l'objet pour chaque image. Nous apprenons les paramètres du modèle $\theta=(u,\mathrm{E},\mathrm{c})$, où $u=\{u_1,...,u_n\}$ sont les paramètres de l'apparence pour chaque partie. *E* est la série des connexions entre les parties et $c=\{\mathrm{c}_{ij}\setminus(\mathrm{v}_i,\mathrm{v}_j)\in\mathrm{E}\}$ sont les paramètres de connexion. Nous aurons donc :

$$p(\mathrm{I}^1,...,\mathrm{I}^m,\mathrm{L}^1,...,\mathrm{L}^m\setminus\theta)=\prod_{k=1}^{m} p(\mathrm{I}^k,\mathrm{L}^k\setminus\theta)$$

$$\theta^* = \arg\max_\theta \prod_{k=1}^{m} p(\mathrm{I}^k\setminus\mathrm{L}^k,\theta)\prod_{k=1}^{m} p(\mathrm{L}^k\setminus\theta) \quad \textbf{Eq 6-7}$$

Le premier terme de cette équation ne dépend que de l'aspect des pièces, tandis que le second terme dépend uniquement de l'ensemble des connexions et des paramètres de connexion. Ci-dessous, on montre que l'on peut régler de façon indépendante pour les modèles d'aspect des parties individuelles et le modèle structurel proposée par les connexions et de leurs paramètres. En conséquence, tout

type de modèles de pièces peut être utilisé dans ce cadre tant qu'il s'agit d'une procédure d'évaluation de probabilité maximale pour l'apprentissage les paramètres du modèle pour une seule pièce à partir d'exemples. Nous utilisons des modèles de pièces très simples dans cet article parce que notre objectif est de développer un cadre général et de fournir des algorithmes efficaces qui peuvent être utilisés avec de nombreux systèmes de modélisation différents

6.2.3.1 Estimation de l'apparence

A partir de l'équation (5-7) nous aurons :

$$\mu^* = \arg\max_{\mu} \prod_{k=1}^{m} p(\mathrm{I}^k \setminus \mathrm{L}^k, \mu)$$

La probabilité d'observer une image I^k , avec une configuration L^k donnée pour un objet ; est donnée par l'équation (5-4) donc ,

$$u^* = \arg\max_{u} \prod_{k=1}^{m} \prod_{i=1}^{n} p(\mathrm{I}^k \setminus l_i^k, u_i) = \arg\max_{u} \prod_{i=1}^{n} \prod_{k=1}^{m} p(\mathrm{I}^k \setminus l_i^k, u_i) \qquad \textbf{Eq 6-8}$$

Nous observons que pour résoudre u^* , nous pouvons résoudre u_i^* indépendamment :

$$u_i^* = \arg\max_{u_i} \prod_{k=1}^{m} p(\mathrm{I}^k \setminus l_i^k, u_i)$$

Ceci est exactement l'estimation de la probabilité des paramètres d'apparence pour une partie v_i , en ayant des exemples indépendants $\{(\mathrm{I}^1, l_i^1), ..., (\mathrm{I}^m, l_i^m)\}$. Résoudre u_i^* dépend du schéma de modélisation spécifique choisi.

6.2.3.2 Estimation des dépendances

A partir de l'équation (5-7) , nous aurons :

$$E^*, c^* = \arg\max_{E,c} \prod_{k=1}^{m} p(\mathrm{L}^k \setminus \mathrm{E}, \mathrm{c}) \qquad \textbf{Eq6-9}$$

Nous avons besoin de choisir un ensemble d'arêtes qui forment un arbre et les paramètres de connexion pour chaque contour. Cela peut se faire d'une manière similaire à l'algorithme de Chow et Liu décrit dans [11], qui estime une distribution d'arbre pour variables aléatoires.

La formule de probabilité d'un objet adoptant une configuration L^k , comme :

$$p(\mathrm{L}^k \setminus \mathrm{E}, \mathrm{c}) = \prod_{(\mathrm{v}_i, \mathrm{v}_j) \in \mathrm{E}} p(l_i^k, l_j^k \setminus c_{ij})$$

Remplaçant la formule ci-dessous dans l'équation (5-8) , nous aurons :

$$E^*, c^* = \arg\max_{E,c} \prod_{(\mathrm{v}_i, \mathrm{v}_j) \in \mathrm{E}} \prod_{k=1}^{m} p(l_i^k, l_i^k \setminus c_{ij}) \qquad \textbf{Eq 6-10}$$

Nous pouvons estimer les paramètres pour chaque connexion possible indépendamment, avant même que nous sachions quelles connexions seront effectivement dans E tel que,

$$c_{ij}^* = \arg\max_{c_{ij}} \prod_{k=1}^{m} p(l_i^k, l_j^k \setminus \mathrm{c}_{ij}) \qquad \textbf{Eq 6-11}$$

C'est l'estimation ML pour la distribution conjointe de $l_i$et l_j, pour des exemples indépendants $\{(l_i^1, l_j^1), ..., (l_i^m, l_j^m)\}$. La résolution de c_{ij}^* dépend du choix de la représentation spécifique pour les distributions conjointes. Indépendamment de la forme exacte de $p(l_i, l_i \setminus c_{ij})$ et la façon de calculer c_{ij}^* (que nous considérons plus tard, car il varie en fonction du système de modélisation), nous pouvons caractériser

la « qualité » d'un lien entre deux parties que la probabilité des exemples dans l'estimation ML pour leur distribution commune,

$$q(v_i, v_j) = \prod_{k=1}^{m} p\left(l_i^k, l_i^k \setminus c_{ij}^*\right) \qquad \textbf{Eq6-12}$$

Intuitivement, la qualité d'une liaison entre deux parties mesure le degré de liaison entre leurs emplacements. Ces quantités peuvent être utilisées pour estimer l'ensemble de connexion E^*.

Nous savons E^* que devrait former un arbre, et selon l'équation 5-9), nous aurons,

$$E^* = \arg\max_{E} \prod_{(v_i, v_j) \in E} q(v_i, v_j) = \arg\min_{E} \sum_{(v_i, v_j) \in E} -\log q(v_i, v_j) \qquad \textbf{Eq6-13}$$

Le côté droit est obtenu en prenant le logarithme négatif de la quantité étant maximisé (et de trouver ainsi l'argument minimisant la valeur, au lieu de maximiser elle). Le calcul de E^* est équivalent au problème du calcul « Minimum Spanning Tree » (MST) d'un graphe.

Nous construisons un graphe complet sur le sommets V, et associons le poids $-\log q(v_i, v_j)$ à chaque arête (v_i, v_j).

Le MST de ce graphe est l'arbre avec le poids total minimal, ce qui est exactement l'ensemble des contours définis par l'équation (5-10). Le problème MST est bien connu et peut être résolu efficacement. L'algorithme de Kruskal peut être utilisé pour calculer le MST en temps de O (n2 log n), puisque nous avons un graphe complet à n nœuds.

6.3 Exploration des partie du corps humain

L'approche de la fenêtre coulissante est extrêmement puissante. Elle ne suppose pas que les fenêtres sont indépendantes, mais nous avons montré des façons de gérer le coût de cette hypothèse. Cependant, l'approche doit échouer lorsque le classificateur échoue. Il ya deux effets importants qui pousse le classificateur à l'échec: L'objet peut changer de forme, généralement appelée **déformation**; et nous pourrions voir l'objet de points de vue différents, généralement appelés **aspect**. Des travaux récents ont montré que ces effets peuvent être atténués de façon très significative par les changements naturels au classificateur.

Pour faire face à l'**aspect**, nous pourrions construire plus d'un classificateur pour le même objet. Chaque classificateur répond à différents points de vue de cet objet. La réponse de ce système des classificateurs à une fenêtre donnée est obtenue en prenant le maximum de chacune des réponses de classificateurs séparés. La procédure d'apprentissage devra en tenir compte, afin de s'assurer que les classificateurs sont calibrés à l'autre. En particulier, la procédure d'apprentissage sera considérablement simplifiée si elle sait **qui** de multiples classificateurs devraient avoir la réponse la plus forte pour chaque exemple d'apprentissage.

Nous ne nous attendons pas que cette information fasse partie de l'ensemble d'apprentissage, et il forme ainsi une variable d'une variable latente qui simplifie la modélisation, mais est lui-même inconnu. Nous aurons à estimer lors de l'apprentissage.

Cette notion de **variable latente** donne des méthodes extrêmement puissantes pour faire face à la déformation, aussi. Déformation est un concept plutôt fluide, car il doit couvrir de tels effets aussi variés que les personnes se déplaçant leurs bras et les jambes autour, la tendance que certains automobiles d'être plus longs que d'autres, et la tendance de (par exemple) les amibes ou des méduses à avoir forme peu fiable du tout. Le sens le plus utile à ce jour est l'observation que de nombreux objets ont

des points en commun la tête d'une personne, ou le capot d'une voiture, mais peut être trouvé dans un peu différents endroits dans différentes instances de l'ensemble objet. Par exemple, un break et une berline sont comme un autre parce que chacun a des portes , des roues et des phares - fiables taches qui ressemblent un de l'autre , mais des photos de breaks pourraient avoir les phares et les roues plutôt plus loin des portes de photos de berlines volonté . Cela suggère la modélisation d'un objet comme une racine d'un modèle approximatif qui donne la situation générale de l'objet et un ensemble de pièces objet composants qui ont l'apparence tout à fait fiables. Les pièces sont généralement beaucoup plus petites que les racines. Chaque partie dispose d'un modèle d'apparence et d'un emplacement naturel. Trouver une fenêtre qui ressemble beaucoup à la partie proche de l'emplacement naturel de cette partie, à l'égard de la preuve de la racine prouvant que l'objet est présent sur l'image.

Nous construisons maintenant classificateurs qui utilisent ce modèle d'un objet, et les appliquons dans notre fenêtre coulissante. Le score global pour une fenêtre sera la somme de plusieurs scores distincts. On compare la racine de la fenêtre. Chaque partie a son propre score séparé, constitué d'un terme d'apparence et un terme de localisation . Le terme d'aspect compare l'aspect de la partie à l'image, et le terme de localisation pénalise la partie si elle se trouve trop loin de son emplacement naturel. Le modèle d'apparence pour la racine et pour chaque partie sera une fonction linéaire de HOG , de sorte que le modèle résultant aura de fortes analogies avec un linéaire SVM- en fait, un SVM linéaire est un cas particulier du modèle que nous construisons , dans ce qu 'il a une racine , mais pas de pièces .

Nous pouvons maintenant introduire quelques notations pour décrire un modèle de composant unique.

Le modèle de la racine sera constitué d'un ensemble de poids linéaires, $\beta^{(r)}$, à appliquer au vecteur de caractéristiques décrivant la fenêtre racine. Le modèle de la $i^{éme}$ partie sera composé d'un ensemble de poids linéaires, $\beta^{(p_i)}$ à être appliquée sur le vecteur de caractéristique décrivant la fenêtre de la pièce ; une compensation naturelle par rapport à la racine , $v^{(p_i)} = \left(u^{(p_i)}, v^{(p_i)}\right)$; et un ensemble de poids de distance $d^{(p_i)} = \left(d_1^{(p_i)}, d_2^{(p_i)}, d_3^{(p_i)}, d_4^{(p_i)}\right)$.

Maintenant, nous écrivons φ (x , y) pour le vecteur de caractéristiques décrivant la fenêtre de la pièce à la position (x , y) par rapport à la racine du système de coordonnées .

Ecrire $(dx, dy) = \left(u^{(p_i)}, v^{(p_i)}\right) - (x, y)$ pour le système de décalage de l'emplacement idéal de la partie à la racine de coordonnées. Le score de la $i^{éme}$ partie à cet endroit (x, y) par rapport à la racine est donnée par :

$$\begin{aligned} Score\_partie\,à\,(x,y) &= Score\_Apparence - Coût\_Déformation \\ &= S^{(p_i)}\left(x, y, \beta^{(p_i)}, d^{(p_i)}, v^{(p_i)}\right) \\ &= \beta^{(p_i)}.\Phi(x,y) - \left(d_1^{(p_i)}dx + d_2^{(p_i)}dy + d_3^{(p_i)}dx^2 + d_4^{(p_i)}dy^2\right) \end{aligned} \qquad \textbf{Eq6-14}$$

Aussi, nous définissons le score de la $i^{ème}$ partie d'être le meilleure score obtenu à travers tous les déplacements possibles, soit :

$$Score\_partie\_i = \max_{(x,y)} S^{(p_i)}\left(x, y; \beta^{(p_i)}, d^{(p_i)}, v^{(p_i)}\right) = D_i \qquad \textbf{Eq 6-15}$$

Maintenant, le score pour le model d'objet à une fenêtre d'une racine particulière est :

$$\begin{aligned} Score\_Model &= Score\_apparence\_racine + \sum score\_partie\_i \\ &= \beta^{(r)} \varphi(\mathrm{x}, y) + \sum D_i(x, y) \end{aligned} \qquad \textbf{Eq6-16}$$

Supposons que nous avons un modèle d'objet articulé (composé), la détection consiste en :

Pour chaque fenêtre, on calcule le score du modèle pour chaque composant, prenons le maximum sur tous les composants, et utilisons ce maximum dans notre fenêtre coulissante. Pour ce faire, nous devrons maximiser le score pour chaque partie en fonction de (x, y).

En revanche, pour la phase d'apprentissage, nous devrions faire face à deux types de variables latentes. Tout d'abord, nous ne savons pas quel composant doit répondre pour chaque exemple positif, les exemples négatifs sont plus en moins facile à traiter vu qu'ils doivent avoir systématiquement un sore négatif ;

Deuxièmement, nous ne connaissons pas les emplacements exacts des parties dans la base d'apprentissage. Notez que, si on avait l'information sur le composant et l'emplacement de la partie pour chaque exemple, l'apprentissage aurait pu être fait avec un SVM Linéaire. Néanmoins, une stratégie basée sur l'estimation répétée. Nous supposons que les composants ainsi que leurs emplacements sont connus, ensuite nous calculons l'apparence de chaque partie et le déplacement du modèle pour chaque composant. En ayant cette information, nous pouvons estimer les emplacements et les composants de nouveau.

L'approche de la fenêtre coulissante doit traiter un nombre immense de fenêtre d'images dont la plupart sont négatives. En conséquence, les taux de faux positif peuvent être un problème majeur. Il est extrêmement important de faire de l'apprentissage du modèle avec une très large base de données d'images, de les exposer à autant d'exemples négatifs que possible.

Une stratégie intéressante, introduite par Felzenszwalb et al. Connue par « Exploration des données négatives durs », Comme nous apprenons le classificateur, nous l'appliquons à des exemples négatifs, à la recherche de ceux qui obtiennent une réponse forte; ceux-ci sont mis en cache, et utilisés dans le prochain cycle d'apprentissage. Si cela est bien fait, on peut garantir que le classificateur a les mêmes vecteurs supports, qu'il aurait pu avoir si on l'aurait appliqué sur tous les exemples négatifs.

6.3.1 Apprentissage

6.3.1.1 Machine à vecteurs de support

6.3.1.2 Machine à vecteurs de support Latent

Considérons un classificateur qui comme entrée un exemple x avec une fonction de la forme,

$$f_\beta(x) = \max_{z \in Z(x)} \beta.\Phi(x,z) \qquad \textbf{Eq6-17}$$

Dans cette équation :

e. β est un vecteur des paramètres du modèle,

f. z sont les valeurs latentes.

L'ensemble $Z(x)$ définit les valeurs latentes possibles pour l'exemple x. Une étiquette binaire de x peut être obtenue par seuillage de son score. Comparativement avec l'approche classique du SVM nous appliquons un apprentissage à partir d'un ensemble d'exemples labélisés $D = (\langle x_1, y_1 \rangle, ..., \langle x_n, y_n \rangle)$ $où\ y_i \in \{-1,1\}$ en minimisant la fonction objectif.

$$L_D(\beta) = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i f_\beta(x_i)) \qquad \textbf{Eq 6-18}$$

Où $\max(0,1-y_i f_\beta(x_i)$ est :

Notons que le SVM linéaire est un cas particulier du vecteur de support latent (Latent SVM) , en ayant un seule valeur cachée possible pour chaque exemple $(|Z(x_i)|)=1$, et avec f_β est linéaire dans β .

6.3.1.2.1 Semi-Convexité

Un SVM latent conduit à un problème d'optimisation non convexe. Cependant, un SVM latente est semi-convexe dans le sens décrit ci-dessous, et le problème de l'apprentissage devient convexe une fois que l'information latente est spécifiée pour les exemples d'apprentissages positifs.

Rappelons que le maximum d'un ensemble de fonctions convexes est convexe. Dans un SVM linéaire nous avons $f_\beta(x)=\beta.\Phi(x)$ linéaire. Dans ce cas, la fonction perte charnière est convexe, pour chaque exemple, car il est toujours le maximum de deux fonctions convexes.

On notera que $f_\beta(x)$ tel que défini dans (18) est un maximum de fonctions dont chacune est linéaire en. Ainsi $f_\beta(x)$ est convexe et donc la perte de la charnière, $\max(0,1-y_i f_\beta(x_i))$, est convexe lorsque $y_i=-1$. Autrement dit, la fonction de perte est convexe dans les exemples négatifs. Nous appelons cette propriété de la fonction de perte semi-convexité. Dans une SVM latente général la perte de la charnière est non convexe pour un exemple positif, car il est le maximum d'une fonction convexe (zéro) et une fonction concave $1-y_i f_\beta(x_i)$.

Considérons maintenant un SVM latente où il ya une valeur latente possible unique pour chaque exemple positif. Dans ce cas $f_\beta(x_i)$, est linéaire pour un exemple positif et la perte due à chaque positif est convexe. Combiné avec la propriété semi - convexité, (Eq 5-19) devient convexe.

6.3.1.2.2 Optimisation

Soit Z_p spécifiant la latente pour chaque exemple positif dans un ensemble d'apprentissage D. Nous pouvons définir une fonction objectif $L_D(\beta, Z_p) = L_{D(Z_p)}(\beta)$, où $D(Z_p)$ est dérivé à partir de D en limitant les valeurs latentes pour les exemples positifs selon Zp. C'est, pour un exemple positif nous avons mis en Z (xi) = fzig où zi est la valeur latente spécifié pour xi par Zp. Notez que

$$L_D(\beta) = \min_{Z_p} L_D(\beta, Z_p) \qquad \textbf{Eq 6-19}$$

En particulier $L_D(\beta) \le L_D(\beta, Z_p)$; La fonction objectif auxiliaire délimite l'objectif de LSVM. Cela justifie l'apprentissage d'un SVM latente en minimisant $L_D(\beta, Z_p)$.

En pratique, nous minimisons $L_D(\beta, Z_p)$ en utilisant l'approche «Les coordonnées descentes":

1- Re-labéliser les exemples positives : Optimiser $L_D(\beta, Z_p)$ à travers Z_p en sélectionnant les plus haut score de la valeur latente pour chaque exemple positif.
$z_i = \arg\max_{z \in Z(x_i)} \beta.\Phi(x_i, z)$

2- Optimiser bêta: Optimiser $L_D(\beta, Z_p)$ à travers β en charge par la résolution du problème d'optimisation convexe définie par $L_D(Z_p)(\beta)$.

Les deux étapes améliorent ou maintiennent toujours la valeur de $L_D(\beta, Z_p)$. Après convergence, nous avons relativement un optimum local fort dans le sens que l'étape 1 recherche sur un espace de façon exponentielle les valeurs latentes des exemples positifs tandis que l'étape 2 recherches sur tous les modèles possibles, considérant implicitement un espace de valeurs latentes exponentiellement grand

pour tous négatifs exemples. Nous notons, toutefois, que l'initialisation de β peut être nécessaire parce que sinon on peut sélectionner des valeurs latentes déraisonnables pour les exemples positifs de l'étape 1, et cela pourrait conduire à un mauvais model. La propriété semi-convexité est importante car elle conduit à un problème d'optimisation convexe à l'étape 2, même si les valeurs latentes pour les exemples négatifs ne sont pas fixes. Une procédure similaire qui fixe les valeurs latentes pour tous les exemples pour chaque itération ne donnera pas forcément de bons résultats. Soit Z spécifiant des valeurs latentes pour tous les exemples de D. Au moment où $L_D(\beta)$ est maximisée efficacement sur toutes les valeurs négatives latentes, $L_D(\beta)$ pourrait être beaucoup plus grande que $L_D(\beta,Z)$ et nous ne devrions pas attendre que la maximisation de $L_D(\beta,Z)$ conduirait à un bon modèle.

6.3.1.2.3 *Descente de gradient stochastique*

Etape 2 (Beta Optimisée) de la méthode de coordonnées descentes peut être résolue via une programmation quadratique [3]. Il peut également être résolu par gradient descent stochastique. Nous décrivons ici une approche de descente de gradient pour optimisation de β à partir d'un ensemble arbitraire d'apprentissage D. En pratique, nous utilisons une version modifiée de cette procédure qui fonctionne avec un de cache de vecteurs de caractéristiques pour $D(Z_p)$.

Soit : $z_i(\beta) = \arg\max_{z \in Z(x_i)} \beta.\Phi(x_i, z)$

Alors : $f_\beta(x_i) = \beta.\Phi(x_i, z_i(\beta))$

Nous pouvons calculer un sous-gradient de la fonction objectif LSVM, comme suit,

$$\nabla L_D(\beta) = \beta + C\sum_{i=1}^{n} h(\beta, x_i, y_i) \qquad \textbf{Eq 6-20}$$

$$h(\beta, x_i, y_i) = \begin{cases} 0 & Si\ y_i f_\beta(x_i) \geq 1 \\ -y_i \Phi(x_i, z_i(\beta)) & Sinon \end{cases} \qquad \textbf{Eq 6-21}$$

En descente de gradient stochastique nous approchons ∇L_D en utilisant un sous-ensemble des exemples et faire un pas dans sa direction négative. L'utilisation d'un seul exemple, $\langle x_i, y_i \rangle$, nous approchons $\sum_{i=1}^{n} h(\beta, x_i, y_i)$ avec $nh(\beta, x_i, y_i)$. L'algorithme qui en résulte mis à jour β comme ci-dessous :

1) Soit α_t le taux d'apprentissage pour l'itération t.
2) Soit i un exemple aléatoire.
3) Soit $z_i = \arg\max_{z \in Z(x_i)} \beta.\Phi(x_i, z)$
4) Si $y_i f_\beta(x_i) = y_i(\beta.\Phi(x_i, z_i)) \geq 1$ alors $\beta := \beta - \alpha_t \beta$.
5) Sinon $\beta := \beta - \alpha_t(\beta - Cny_i\Phi(x_i, z_i))$

Comme dans les méthodes de descente de gradient pour SVM linéaires, nous obtenons une procédure qui est assez similaire à l'algorithme du perceptron. Si f_β est classe correctement l'exemple aléatoire xi , nous diminuons β tout simplement. Sinon, nous reculons et ajoutons un scalaire multiple $\Phi(x_i, z_i)$.

Pour un SVM linéaire le taux d'apprentissage $\alpha_t = 1/t$ a donné de bons résultats. Cependant, le temps de convergence dépend du nombre d'exemples d'apprentissage.

6.3.1.2.4 Exploration des données (Data_mining), la version SVM

Lorsque l'apprentissage d'un modèle pour la détection de l'objet, nous avons souvent un très grand nombre d'exemples négatifs (une seule image peut donner 105 exemples pour un classificateur de la fenêtre de balayage). Cela peut rendre impossible d'examiner tous les exemples négatifs simultanément. Au lieu de cela, il est courant de construire des données d'apprentissage, comprenant des cas positifs et les cas "dur négatives» (Hard negative).

Méthodes de Boostrap [97] forme un modèle avec un sous-ensemble initial d'exemples négatifs, puis recueille des exemples négatifs qui sont mal classés par ce modèle initial pour former un ensemble de négatifs dur. Un nouveau modèle est formé avec les exemples négatifs durs et le processus peut être répété plusieurs fois.

Ici, nous décrivons un algorithme d'exploration de données motivée par l'idée de boostrap pour la formation d'un (non latent) SVM classique. La méthode permet de résoudre une séquence d'apprentissage :

Ici, nous décrivons un algorithme d'exploration de données motivée par l'idée du boostrap pour l'apprentissage d'un (non latent) SVM classique. Le procédé permet de résoudre une série de problèmes d'apprentissage à l'aide d'un relativement nombre d'exemples durs petit et converge vers la solution exacte du problème d'apprentissage définie par un grand ensemble d'apprentissage. Cela exige une définition des exemples durs.

On définit les cas duret simples d'un ensemble d'apprentissage D pour β , comme suit :

$$H(\beta, D) = \{\langle x, y\rangle \in D \setminus y f_\beta(x) \prec 1\} \qquad \textbf{Eq6-22}$$

$$E(\beta, D) = \{\langle x, y\rangle \in D \setminus y f_\beta(x) \succ 1\} \qquad \textbf{Eq6-23}$$

Autrement dit $H(\beta,D)$ sont les exemples de D qui sont mal classés ou à l'intérieur de la marge du classifieur définie par β. De même $E(\beta,D)$ sont les exemples de D qui sont correctement classés et en dehors de la marge.

Soit $\beta^*(D)=\arg\min_\beta L_D(\beta)$

Tant que L_D est strictement convexe $\beta^*(D)$ est unique.

Compte tenu d'un grand ensemble d'apprentissage D nous aimerions trouver une petite série d'exemples $C\subseteq D$ tel que $\beta^*(C)=\beta^*(D)$.

La méthode commence par un "cache" initiale d'exemples et alterne entre la formation d'un modèle et la mise à jour du cache. Dans chaque itération, nous enlevons des exemples faciles à partir du cache et ajoutons de nouveaux exemples durs. Un cas particulier consiste à garder tous les exemples positifs dans le cache et l'extraction de données sur les négatifs.

Soit $C_1\subseteq D$ un premier cache d'exemples. L'algorithme apprend le modèle d'une façon répétitive et mis à jour le cache comme suit:

1) Soit $\beta_t:=\beta^*(C_t)$ (apprendre le modèle, en utilisant C_t).
2) Si $H(\beta_t,D)\subseteq C_t$ stop et retourner.
3) Soit $C_t^{'}:=C_t\setminus X$ pour chaque X tel que $X\subseteq E(\beta_t,C_t)$.(diminution du cache).
4) Soit $C_{t+1}:=C_t^{'}\cup X$ pour chaque X tel que $X\subseteq D$ et $X\cap H(\beta_t,D)\setminus C_t\neq\varnothing$ (Augmenter le cache).

Dans l'étape 3, nous reculons le cache en supprimant des exemples de C_t qui sont en dehors de la marge définie par β_t. Dans l'étape 4, nous augmentons le cache en ajoutant des exemples de D, y compris au moins un nouvel exemple qui est à

l'intérieur de la marge définie par β_t. Tel exemple doit exister, sinon nous aurions retourné à l'étape 2.

Le théorème suivant montre que lorsque nous nous arrêtons, nous avons trouvé $\beta^*(D)$.

- **Théorème 1 :** Soit $C \subseteq D$ et $\beta = \beta^*(C)$. Si $H(\beta, D) \subseteq C$ alors $\beta = \beta^*(D)$.
- **Preuve :** $C \subseteq D$ implique $L_D(\beta^*(D)) \geq L_C(\beta^*(C)) = L_C(\beta)$. Tant que $H(\beta, D) \subseteq C$ tous les exemples dans $D \setminus C$ ont une perte nulle dans β. Ceci implique $L_C(\beta) = L_D(\beta)$. Nous incluons $L_D(\beta^*(D)) \geq L_D(\beta)$, parce que L_D a un unique minimum $\beta = \beta^*(D)$.
- Le résultat suivant démontre que l'algorithme s'arrête après un nombre fini d'itérations. Intuitivement cela découle du fait que $L_{C_t}(\beta^*(C_t))$ croît à chaque itération, mais elle est délimitée par $L_D(\beta^*(D))$

- **Théorème 2 L'algorithme d'exploration de données se termine.**
- **Preuve** : Lorsque nous reculons le cache $C_t^{'}$ contenant tous les exemples de C_t avec une perte non nulle autour de β_t. Cela implique $L_{C_t^{'}}$ est identique à L_{C_t} autour de β_t, et quand β_t est le minimum de L_{C_t}, il doit aussi être le minimum de $L_{C_t^{'}}$. Donc $L_{C_t^{'}}(\beta^*(C_t^{'})) = L_{C_t}(\beta^*(C_t))$.
- Lorsque nous augmentons le cache $C_{t+1} \setminus C_t^{'}$ contenant au moins un exemple $\langle x, y \rangle$ avec une perte non nulle à β_t. Tant que $C_t^{'} \subseteq C_{t+1}$, nous avons $L_{C_{t+1}}(\beta) \geq L_{C_t^{'}}(\beta)$ pour tous les β.
- Si $\beta^*(C_{t+1}) \neq \beta(C_t^{'})$ alors $L_{C_{t+1}}(\beta^*(C_{t+1})) \geq L_{C_t^{'}}(\beta^*(C_t^{'}))$ parce que $L_{C_t^{'}}$ a un minimum unique. Si $\beta^*(C_{t+1}) = \beta^*(C_t^{'})$ alors $L_{C_{t+1}}(\beta^*(C_{t+1})) \geq L_{C_t^{'}}(\beta^*(C_t^{'}))$ dû à $\langle x, y \rangle$.

Nous concluons $L_{C_{t+1}}\left(\beta^*\left(C_{t+1}\right)\right) \geq L_{C_t'}\left(\beta^*\left(C_t'\right)\right)$ .Comme il existe un nombre fini de caches la perte du cache ne peut se développer qu'à un nombre fini de fois.

6.3.1.2.5 *Exploration des données (Data_mining), la version LSVM*

Maintenant, nous décrivons un algorithme d'exploration de données pour l'apprentissage d'un SVM latente lorsque les valeurs latentes des exemples positifs sont fixés. Nous optimisons $L_{D(z_p)}(\beta)$, et non $L_D(\beta)$. Comme discuté ci-dessus cette restriction assure que le problème d'optimisation est convexe.

Pour un SVM latente au lieu de garder un cache d'exemples x, nous gardons un cache de paires (x,z) où $z \in Z(x)$.Ceci permet d'éviter de faire inférence sur la totalité de Z (x) dans la boucle intérieure d'un algorithme d'optimisation telle que la descente de gradient. En outre, dans la pratique, nous pouvons garder un cache de vecteurs de caractéristiques, $\Phi(x,z)$, Au lieu des paires (x,z). Cette représentation est plus simple (son application indépendante) et peut être beaucoup plus compact.

Un cache de vecteur de caractéristique F est un ensemble de paires (i,v) où $1 \leq i \leq n$ est l'indice d'un exemple et $v = \Phi(x_i, z)$ pour un certain $z \in Z(x_i)$. Notez que nous pouvons avoir plusieurs paires $(i,v) \in F$ pour chaque exemple x_i. Si l'ensemble d'apprentissage a nombre fixé de labels pour des exemples positifs ce qui peut encore être vrai pour les exemples négatifs.

Soit $I(F)$ un ensemble d'exemples indexés par *F*. Le vecteur de caractéristiques dans *F* définit la fonction d'objective pour β, où nous considérons uniquement les exemples indexés par $I(F)$, et pour chaque exemple , nous considérons le vecteur de caractéristiques dans le cache ,

$$L_F(\beta) = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1 \in (F)} \max(0, 1 - y_i\left(\max_{(i,v)\in F} \beta.v\right)) \qquad \textbf{Eq 6-24}$$

Nous pouvons optimiser L_F via la descente du gradient par la modification de la méthode (vu dans la section descente du gradient). Soit $V(i)$ l'ensemble de vecteurs de caractéristiques v tel que $(i,v)\in F$. Ensuite chaque itération de la descente du gradient est simplifiée comme suit :

1) Soit α_t le taux d'apprentissage pour l'itération t.
2) Soit $i\in I(F)$ un exemple aléatoire indexé par F .
3) Soit $v_i = \arg\max_{v\in V(i)} \beta.v$.
4) Si $y_i(\beta.v_i)\geq 1$ alors $\beta := \beta - \alpha_t\beta$.
5) Sinon $\beta := \beta - \alpha_t(\beta - Cny_iv_i)$

Maintenant, la taille de $I(F)$ contrôle le nombre d'itération nécessaires pour la convergence, tandis que la taille de $V(i)$ contrôle le temps d'exécution de l'étape 3. Dans l'étape 5 , $n = |I(F)|$.

Soit $\beta^*(F) = \arg\min_\beta L_F(\beta)$. Nous voudrions de trouver un petit cache pour $D(Z_p)$ avec $\beta^*(F) = \beta^*(D(Z_p))$.

Nous définissons les vecteurs de caractéristiques dures pour l'ensemble d'apprentissage D , relatif à β comme,

$$H(\beta,D) = \left\{ (i,\Phi(x_i,z_i)) \mid z_i = \arg\max_{z\in Z(x_i)} \beta.\Phi(x_i,z) \; et \; y_i(\beta.\Phi(x_i,z_i)) \right\} \qquad \textbf{Eq6-25}$$

C'est-à-dire , $H(\beta,D)$ sont des paires (i,v) où v est le vecteur de caractéristique qui a le score le plus élevé à partir d'un exemple x_i , qui est à l'intérieur de la marge définie par le classificateur.

Nous définissons les vecteurs de caractéristiques faciles dans un cache F comme,

$$E(\beta,F)=\{(i,v)\in F \setminus y_i(\beta.v)\succ 1\} \qquad \textbf{Eq 6-26}$$

Ce sont les vecteurs de caractéristiques de F qui sont en dehors de la marge définie par β. Notez que si $y_i(\beta.v)\leq 1$ alors (i,v) ne sont pas considérés comme **facile** , toute foisil existe un autre vecteur de caractéristique pour le $i^{ème}$ exemple dans le cache avec le score le plus élevé que v sous β.

Maintenant, nous décrivons l'algorithme d'exploration de données pour le calcul $\beta^*(D(Z_p))$. L'algorithme fonctionne avec un cache de vecteurs de caractéristiques pour $D(Z_p)$. Il alterne entre l'apprentissage du modèle et mise à jour du cache.

Soit F_1 , un cache initial de vecteur de caractéristique maintenant, nous considérons l'algorithme itératif suivant :

1) Soit $\beta_t := \beta^*(F_t)$ (apprentissage du model).
2) Si $H(\beta, D(Z_p)) \subseteq F_t$ stop et retourner β_t.
3) Soit $F_t^{'} := F_t \setminus X$ pour chaque X tel que $X \subseteq E(\beta_t, F_t)$ (Diminuer le cache).
4) Soit $F_{t+1} := F_t^{'} \cup X$ pour chaque X , tel que $X \cap H(\beta_t, D(Z_p)) \setminus F_t \neq \varnothing$ (Augmenter le cache).

L'étape 3 diminue le cache par la suppression les vecteurs de caractéristiques faciles. Étape 4 augmente le cache en ajoutant de "nouveaux" vecteurs de caractéristiques, y compris au moins un de $H(\beta_t, D(Z_p))$. Notez qu'à travers le temps nous allons accumuler plusieurs vecteurs de caractéristiques à partir du même exemple négatif dans le cache. Nous pouvons montrer que cet algorithme finira par s'arrêter et

retourner $\beta^*\left(D\left(Z_p\right)\right)$. Cela découle des arguments analogues à ceux utilisés dans la section précédente.

6.3.1.3 Apprentissage du modèle

Maintenant, nous considérons le problème d'apprentissage de modèle à partir des images labélisées avec des boîtes englobantes autour des objets d'intérêt. Notez que ceci est une labélisation faible « **weakly labeled** » , vue que les boîtes englobantes ne précisent pas les étiquettes des composants ou les emplacement des pièces.

Nous décrivons une procédure d'initialisation de la structure du modèle et l'apprentissage de tous les paramètres. L'apprentissage des paramètres se fait par SVM latente. Nous faisons l'apprentissage du SVM latente en utilisant l'approche de coordonnées descentes ainsi que les algorithmes de data mining qui travaillent avec un cache de vecteurs caractéristiques de la section précédente. Puisque la méthode de descente de coordonnées est susceptible aux minima locaux, nous devons veiller à assurer une bonne initialisation du modèle.

6.3.1.3.1 Apprentissage des paramètres :

Soit C une classe d'objet. Nous supposons que les exemples d'apprentissage pour les C sont donnés par des boîtes englobantes positives P et un ensemble d'images de fond N. P est un ensemble de paires (I, B) où I est une image et B est un cadre de sélection d'un objet de classe C en I .

Soit M un (Mélange) modèle avec une structure fixe. Rappelons que les paramètres d'un modèle sont définis par un vecteur. Pour en savoir, nous définissons un problème d'apprentissage SVM latente avec un ensemble défini implicitement d'apprentissage D, avec des exemples positifs de P, et des exemples négatifs de N.

Chaque exemple $\langle x, y \rangle \in D$ a une image associée et une pyramide de caractéristiques $H(x)$. Les valeurs latentes $z \in Z(x)$ spécifiant les une instanciation de M dans la pyramide de caractéristiques $H(x)$. Maintenant, nous définissons $\Phi(x,z) = \psi(H(x), z)$. Alors $\beta.\Phi(x,z)$ est exactement le score de l'hypothèse z pour M dans $H(x)$. Une boîte de sélection positive $(I,B) \in P$ spécifie que le détecteur d'objet doit donner un grand score dans l'emplacement défini par la boite B.

Pour chaque $(I,B) \in P$, nous définissons un exemple positif x pour le problème d'apprentissage du LSVM. Nous définissons $Z(x)$ de sorte que la fenêtre de détection du filtre racine spécifié par une hypothèse $z \in Z(x)$ se chevauche avec B d'au moins 50%.

Il y'a généralement, de nombreux emplacement concurrent pour le filtre racine à des échelles différents qui définissent des fenêtres de détection avec 50% de recouvrement. Nous avons trouvé que le traitement de l'emplacement du filtre racine comme une variable latente est utile pour compenser les erreurs dans les délimitations des boites englobantes.

Considérons maintenant une image de fond $I \in N$ .Nous ne voulons pas que le détecteur d'objet se coïncide avec n'importe quel endroit de la pyramide de caractéristiques pour I. Cela signifie que le score global (7) de chaque emplacement racine devrait être faible. Soit ς un ensemble dense d'emplacement dans la pyramide de caractéristiques. Nous définissons un autre exemple négatif x pour chaque emplacement $(i, j, l) \in \varsigma$.

Nous définissons Z (x) de sorte que le niveau du filtre racine spécifié par $z \in Z(x)$ est l , et le centre de sa fenêtre de détection est (i, j). Notez qu'il existe un très grand nombre d'exemples négatifs obtenus à partir de chaque image. Ceci est cohérent avec l'exigence d'un classificateur à base de fenêtre de balayage qui doit avoir un faible taux de faux positif.

La procédure **apprentissage** est décrite ci-dessous. La boucle la plus externe implémente un nombre fixe d'itérations de coordonnées descentes dans $L_D(\beta, Z_p)$. Les lignes 3-6 implémentent l'**étape de labélisation**. Le vecteur de caractéristiques résultant, un pour chaque exemple positif, sont stockés dans F_p. Lignes 7-14 implémentent l'étape beta **optimisation.**

Vu que le nombre d'exemples négatifs implicitement définis par N est très grand, nous utilisons l'algorithme d'exploration de données LSVM(Data mining). Nous réitérons l'extraction de données d'un nombre fixe d'itération que de plutôt attendre la convergence pour des raisons pratiques. A chaque itération, nous recueillons les exemples négatifs durs dans F_n, apprendre un nouveau modèle utilisant une descente de gradient, puis réduisant F_n en enlevant les vecteurs de caractéristiques faciles. Pendant le processus d'extraction de données , nous augmentons le cache en itérant à travers les images dans N séquentiellement jusqu'à la limite de la mémoire.

Algorithme : Apprentissage	
Données	
Exemples positifs $P=\{(I_1,B_1),...,(I_n,B_n)\}$	
Images négatives $N=\{J_1,...,J_m\}$	
Modèle initial β	
Résultat : Nouveau modèle β	
1	$F_n := \varnothing$
2	Pour $relabl := 1$ *à* num_relab **fait**
3	$F_p := \varnothing$
4	**Pour** $i := 1$ à n **fait**
5	**Ajout** Meilleur_detect(β, I, B_i) à F_p
6	**Fin**
7	**Pour** $dataMine := 1$ à $num\_dataMine$ **fait**
8	**Pour** $j := 1$ à m fait
9	**Si** $\lvert F_n \rvert \geq memoire\_Limit$ **alors stop**
10	**Ajout detect_tous**$(\beta, J_j, -(1+\delta))$ à F_n
11	**Fin**
12	$\beta := descente\_gradient(F_p \cup F_n)$
13	**Supprime** (i,v) avec $\beta.v \prec -(1+\delta)$ de F_n
14	**Fin**
15	**Fin**

La fonction $\text{Meilleur\_detect}(\beta, I, B_i)$ **cherche le meilleur score avec un filtre racine qui significativement couvre** B dans I. La fonction detect_tous (β, I, t) calcul la meilleure hypothèse d'objet et sélectionner celui quia le score supérieur à t. Ces deux fonctions sont implémentées par l'algorithme de mise en correspondance vu précédemment.

La fonction $descente\_gradient(F)$ apprend β en utilsant le vecteur de caractéristique dans le cache tel décrit précédemment. Dans la pratique, nous modifiant l'algorithme par fixer les valeurs des coefficients quadratiques par des valeurs supérieures à 0.01. Ceci assure que les coûts de déformations seront convexes .Aussi , le modèle doit être symétrique tout au long de l'axe vertical.

Les filtres qui sont positionnés le long de l'axe vertical central du modèle sont contraints à être auto-symétrique. Les filtres des parties qui sont hors-centre ont un rôle symétrique de l'autre côté du modèle. Cela réduit efficacement le nombre de paramètres à tirer dans la moitié.

- ***initialisation***

L'algorithme LSVM « coordonnées descentes » est sensible aux minima locaux et donc sensible à l'initialisation. Il s'agit d'une limitation commune à d'autres méthodes qui utilisent les informations latentes , ainsi. On initialise et on apprend le mélange de modèles dans les trois phases suivantes.

Phase 1. Initialisation des filtres racines: Pour l'apprentissage de mélange de modèle à m composantes nous trions les boîtes englobantes dans P par leur rapport d'aspect et les diviser en m groupes de taille égale $P_1, ..., P_m$. L'aspect ration est utilisé comme un simple indicateur des interclasses variation extrêmes. Nous apprenons m filtres différents $F_1, ..., F_m$, un pour chaque groupe de boites englobantes positives.

Pour définir les dimensions de F_i, nous sélectionnons la moyenne des ratios des boites dans P_i et la plus large et la zone l plus grande ne dépassant pas les 80% des boite. Ceci garantit que pour la plupart des paires $(I,B) \in P_i$, nous pouvons placé F_i dans la pyramide de caractéristiques de I, donc elles se chevauchent significativement avec B.

L'apprentissage des filtres F_i est fait en utilisant SVM standard , sans aucune information latente. Pour $(I,B) \in P_i$, nous déformons la région d'image dans B pour que sa carte de caractéristique soit de même dimension que F_i . Ceci conduit à un exemple positif. Nous sélectionnons des sous-fenêtres aléatoires de dimension appropriée à partir d'images en N pour définir des exemples négatifs. Les figures 5 (a) et 5 (b) montrent le résultat de cette phase lors de l'apprentissage d'un modèle de voiture à deux composants.

Phase 2. Fusion des Composants: Nous combinons les filtres racines initiaux dans un modèle de mélange sans pièces et nous ré- apprenons les paramètres du modèle combiné en utilisant la méthode **apprentissage** dans l'ensemble des données P et N (sans décomposition ni déformation). Dans ce cas, les composants labélisés et l'emplacement de la racine sont les seules variables latentes pour chaque exemple. L'algorithme d'apprentissage des coordonnées descentes peut être considéré comme une méthode de **classification discriminatoire** qui alterne entre l'affectation des labels aux clusters (mélange) pour chaque exemple positif et estimation du cluster «moyens» (filtres racines).

Phase 3. Initialisation des filtres partis (pièces): On initialise les parties de chaque composant à l'aide d'une heuristique simple. Nous fixons le nombre de pièces à neuf (06) par composant, et en utilisant un pool de pièces rectangulaires, nous plaçons les parties pour couvrir les régions à haute énergie dans le filtre racine .Une partie est soit ancrée le long de l'axe vertical central du filtre de racine, ou elle est non centrée ,et ayant une partie symétrique de l'autre côté du filtre racine . Une fois qu'une pièce est placée, l'énergie de la partie couverte du filtre est mise à zéro, et on cherche la région suivante avec une énergie élevée, jusqu'à ce que les neuf pièces soient choisies.

Les filtres des parties sont initialisés par interpolation du filtre racine à deux fois la résolution spatiale. Les paramètres de déformation pour chaque partie sont initialisés à di = (0, 0, 1, 1).

6.4 Conclusion

L'estimation de la pose humaine constitue un grand challenge dans le domaine de la vision par ordinateurs, notre approche proposée constitue une combinaison entre deux grandes recherches dans le monde de détection d'objet.

Le corps humain est un objet articulé soumis à des variations d'apparence et de forme, composé de plusieurs membres : La façon la plus intuitive est de le décomposer en des parties lors de la détection, or les mouvement d'un membre de corps humain dépendent fortement des autres, ce qui a orienté notre choix vers l'approche Structure picturale. Cette dernière est une combinaison de plusieurs parties dont une d'entre elle est considérée comme une racine , codifiant à la fois l'apparence de la partie ainsi que les contraintes spatiales entre elles.

Par ailleurs, et en considérant la contrainte de traitement en un temps réduit, la phase de mise en correspondance du modèle sur l'image devient très couteuse en terme de temps de calcul , à cet effet , dans le but d'isoler la cible , personne dans

notre cas, nous faisons recours à l'approche proposée par le laboratoire INRIA , Dalal & Triggs , dans leurs travaux de détection de piétons. De ce fait, la racine dans la structure picturale deviendra l'objet dans sa globalité.

Aussi, Le choix d'un descripteur est très crucial pour la performance d'un détecteur d'objet : il doit être le plus représentatif, le plus discriminatoire et le plus rapide en termes de temps de calcul. Dans nos propose nous exploitant le descripteur de HOG « Histogramme de gradient orienté », il a l'exclusivité de mieux représenter la structure interne d'un objet via l'information du gradient, permettant ainsi de surmonter les problèmes liés à l'apparence de l'objet : pose, l'éclairage, l'occlusion, texture de fond, etc.

Dans le chapitre suivant, nous décrivons la mise en œuvre de cette approche et les métriques d'évaluation permettant de valider nos propos.

Chapitre 6 Implémentation et évaluation

7.1 Introduction

Dans le cadre de notre projet, nous proposons une interface pour la détection des parties du corps humain dans deux scénarios : Statique et Dynamique, en faisant recours à une combinaison de deux grandes approches dans le domaine de la reconnaissance des objets : Approche de structure picturale proposée par Fischler [45] et l'approche de l'histogramme de gradient orienté pour la détection des piétons proposée par N.Dalal and B.Triggs [37]. Le but de cette combinaison est de mettre en place un système robuste, fiable, capable à surmonter les problèmes liés à la reconnaissance d'objet tant cités dans la littérature de la vision par ordinateur.
Ain de valider nos travaux, nous proposons l'application IDHPE « Interface Design for Human Pose Estimation » consistant en une implémentation de notre approche proposée , décrite dans le chapitre précèdent. Dans ce qui suit, nous discuterons de la stratégie adoptée durant toutes les phases d'apprentissage et de détection.

Il serait judicieux de noter qu'afin de satisfaire notre cahier des charge relatif à l'estimation de la pose humaine, nous avons débuté nos travaux par la proposition d'un système non basé modèle, exploitant les caractéristiques d'image et les opérations morphologiques et nous avons injecté des données relatives à la morphologie humaine. Ce système à montrer ces limites, ce que nous allons voir plus tard. Nous avons jugé nécessaire de donner un aperçu de ces travaux précédents dans ce chapitre afin de justifier le choix de notre approche.

7.1.1 Travaux Précédents

Afin de satisfaire notre cahier des charge relatif à l'estimation de la pose humaine, nous avons débuté nos travaux par la proposition d'un système non basé modèle , exploitant les caractéristiques d'image et les opérations morphologiques et nous avons injecté des données relatives à la morphologie humaine. Ce système à montrer ces limites, ce que nous allons voir plus tard, Le schéma synoptique détaillé du système précédemment proposé et comme suit :

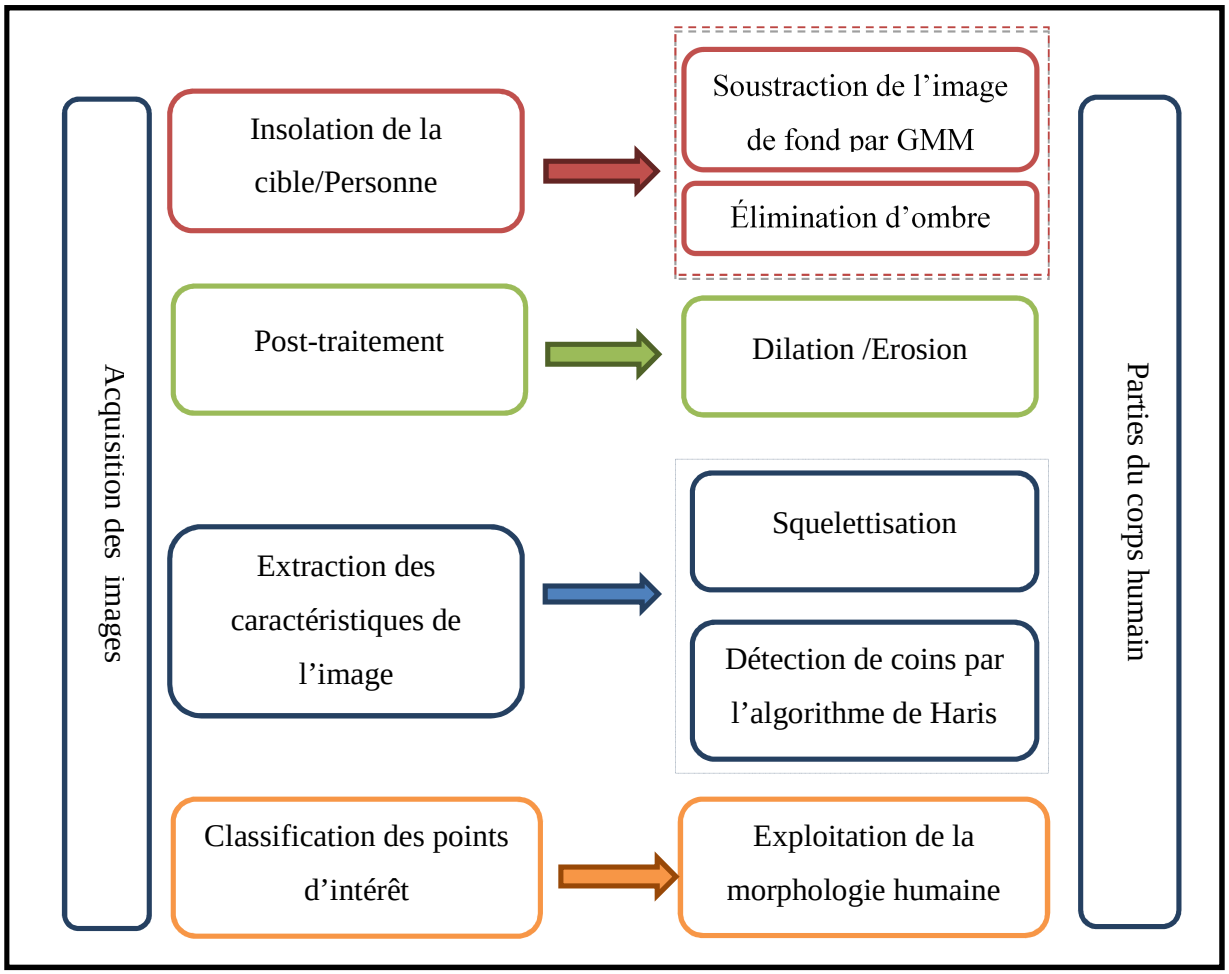


Figure 7-1 Schéma synoptique de notre approche précédemment proposée

7.1.1.1 Détection

Pour le processus de détection, nous avons opté par la méthode de soustraction de l'image de fond, en mettant comme contrainte : une seule personne sur la scène d'acquisition des images. Pour le processus de soustraction de l'image de fond, nous avons fait recours à l'algorithme de « Modèle de Mélange de gaussienne » _ *Gaussian Mixture Model*_ pour la construction d'un arrière-plan adaptatif.

7.1.1.1.1 Model de Mélange de Gaussiennes

Un **modèle de mélange gaussien** (usuellement abrégé par l'acronyme anglais **GMM** pour *Gaussian Mixture Model*) est un modèle statistique exprimé selon une densité de mélange. Elle sert usuellement à estimer paramétriquement la distribution de variables aléatoires en les modélisant comme une somme de plusieurs gaussiennes (appelées *noyaux*). Il s'agit alors de déterminer la variance, la moyenne et l'amplitude de chaque gaussienne. Ces paramètres sont optimisés selon un critère de maximum de vraisemblance pour approcher le plus possible la distribution recherchée. Cette procédure se fait le plus souvent itérativement via l'algorithme espérance-maximisation (*EM*) [94].

7.1.1.1.2 Algorithme :

Nous considérons les valeurs d'un pixel particulier au fil du temps comme un «processus de pixel ». Ce dernier est une série temporelle des valeurs de pixels, par exemple, scalaires pour les images en gris et des vecteurs pour les images couleurs. A tout moment, t, ce qui est connu au sujet d'un pixel particulier à la position (x_0, y_0) est son historique.

$\{X_1,...,X_t\} = \{I(x_0, y_0, i) : 1 \le i \le t\}$ Où I est la séquence d'images. La valeur de chaque pixel représente une mesure de la luminance dans la direction du capteur du premier objet intersecté par le rayon optique du pixel.

Avec un fond statique et de l'éclairage statique, cette valeur serait relativement constante. Si nous supposons que l'indépendance, un bruit gaussien est engagé dans le processus d'échantillonnage, sa densité peut être décrite par une distribution gaussienne centrée unique à la valeur moyenne des pixels. Malheureusement, les séquences vidéo les plus intéressantes impliquent des changements d'éclairage, les changements de scène et les objets en mouvement. Si des changements d'éclairage ont eu lieu dans une scène statique, il serait nécessaire pour la gaussienne de suivre ces changements. Si un objet statique a été ajouté à la scène et n'a pas été incorporé dans l'arrière-plan jusqu'à ce qu'il ait été là plus longtemps que l'objet précédent, les pixels correspondants pourraient être considérés comme de premier plan pour des périodes arbitrairement longues. Cela conduirait à des erreurs accumulées dans l'estimation de premier plan, qui entraîne un mauvais suivi. Ces facteurs donnent à penser que les observations les plus récentes peuvent être plus importantes dans la détermination des estimations des paramètres gaussiens. Un autre aspect de la variation se produit si des objets en mouvement sont présents dans la scène. Même sans cela, un objet avec une apparence variée peut altérer les estimations.

La probabilité d'observer la valeur du pixel courant pour une gaussienne "G" est la suivante:

$$p(X_t / G) = \sum_{j=1}^{K} p(G).p(G / X_t) \qquad \textbf{Eq 7-1}$$

$$p(X_t / G) = \sum_{i=1}^{K} p(G).g(\mu_i, \Sigma_i) \qquad \textbf{Eq7-2}$$

Où K est le nombre de distributions, μi est la valeur moyenne de la i-ème gaussienne dans le mélange à l'instant t, Σi, t est la matrice de covariance de la

gaussienne i dans le mélange à l'instant t, et où g une fonction gaussienne de densité de probabilité:

$$g(\mu_i, \sum_i) = \frac{1}{(2\pi)^{1/2} |\sum|^{1/2}} \exp(-\frac{1}{2}(X_t - \mu)^T \sum^{-1}(X_t - \mu)) \quad \textbf{Eq 7-3}$$

K est déterminé par la mémoire disponible et la puissance de calcul. Actuellement, 3 à 5 sont utilisés. En outre, pour des raisons de calcul, nous supposons que les valeurs de pixel rouge, vert et bleu sont indépendantes et ont les mêmes écarts. Alors que ce n'est certainement pas le cas, l'hypothèse nous permet d'éviter une inversion de matrice, coûteuse en temps de calcul , mais au détriment d'une certaine précision. Ainsi, la matrice de covariance est supposée être de la forme:

$$\sum_j = \begin{pmatrix} {}_rV_i & 0 & 0 \\ 0 & {}_gV_i & 0 \\ 0 & 0 & {}_bV_i \end{pmatrix} = \begin{pmatrix} {}_r\sigma^2{}_i & 0 & 0 \\ 0 & {}_g\sigma^2{}_i & 0 \\ 0 & 0 & {}_b\sigma^2{}_i \end{pmatrix} \quad \textbf{Eq7-4}$$

$$\sum_{k,t} = \sigma^2_k . I \quad \textbf{Eq7-5}$$

Ainsi, la distribution des valeurs observées récemment de chaque pixel de la scène se caractérise par un mélange de gaussiennes. Une nouvelle valeur de pixel, en général, peut être représentée par l'un des composantes principales du modèle de mélange pour mettre à jour le modèle. Si le processus de pixel pourrait être considéré comme un processus stationnaire, une méthode standard pour maximiser la probabilité des données observées est "EM". Parce qu'il y'a un modèle de mélange pour chaque pixel de l'image, mettre en œuvre un algorithme exact EM sur une

fenêtre de données récentes serait coûteux. Au lieu de cela, nous mettons en œuvre l'algorithme de K-means , rapprochant chaque nouvelle valeur de pixel ; Xt, aux distributions K gaussiennes existantes, jusqu'à ce qu'une correspondance soit trouvée. Une marge est défini comme une valeur de pixel dans les 2,5 'écarts types de la distribution, est choisi 2,5 à 95%. Ce seuil peut être perturbé avec peu d'effet sur la performance. Il s'agit effectivement d'un seuil de distribution par pixel. Cela est extrêmement utile lorsque les différentes régions disposent d'un éclairage différent, parce que les objets qui apparaissent dans les régions ombragées ne présentent pas généralement autant de bruit que les objets dans les régions éclairées. Un seuil uniforme se traduit souvent par des objets qui disparaissent quand ils entrent dans les régions ombragées.
Si aucunes des distributions K correspondent à la valeur de pixel actuelle, la distribution moins probable est remplacée avec une distribution avec la valeur actuelle de la valeur moyenne, un écart initialement élevée, et un faible poids.

Les poids sont ajustés comme suit:

$$\omega_t(t) = (1 - \alpha).\omega_t(t-1) + \alpha \ .q_t$$

Eq7-6

où α est le taux d'apprentissage et vaut 1 pour le modèle correspondant et à 0 pour les modèles restants. Après cette approximation, les coefficients de pondération sont renormalisés. 1 / α définit la constante de temps qui détermine la vitesse à laquelle le changement de paramètres de la distribution. ωk, est effectivement un lien de causalité passe-bas moyenne filtrée de la probabilité (seuillée) postérieure que les valeurs de pixels avons identifié le modèle k .

Les paramètres σ et μ pour les distributions inégalées restent les mêmes. Les paramètres de la distribution qui correspond à la nouvelle observation sont en place datée comme suit:

$$\mu_t = (1 - \rho)\mu_{t-1} + \rho X_t \quad \textbf{Eq7-7}$$

$$\sigma_t^2 = (1-\rho)\sigma_{t-1}^2 + \rho(X_t - \mu_t)^T (X_t - \mu_t) \quad \textbf{Eq7-8}$$

Où :

$$\rho = \alpha . g(X_t / \mu_t, \Sigma_t) / \omega_t \quad \textbf{Eq7-9}$$

Afin d'assurer une convergence rapide, nous avons choisi dans nos implémentations :

$$\rho = \alpha / \omega_t \quad \textbf{Eq7-10}$$

Ceci est effectivement le même type de causalité filtre passe-bas comme mentionné ci-dessus, sauf que seules données qui correspondent au modèle sont incluses dans l'estimation. Un des avantages importants de cette méthode est que quand un pixel est autorisé à faire partie de l'arrière-plan, il ne détruit pas le modèle actuel de l'arrière-plan. La couleur de fond d'origine reste dans le mélange jusqu'à ce qu'il devienne l'Kème la plus probable et une nouvelle couleur est observée. Par conséquent, si un objet est à l'arrêt juste assez longtemps pour faire partie de l'arrière-plan, puis il se déplace, la distribution décrivant le contexte précédent existe toujours avec le même μ et σ, mais avec le plus faible ω et sera rapidement réintégrés dans l'arrière-plan.

7.1.1.1.3 *Estimation du modèle de l'arrière-plan:*

Comme les paramètres du modèle de mélange de chaque pixel changent, nous tenons à déterminer lequel des gaussiennes du mélange sont les plus susceptibles aux changements.

La variance de l'objet mobile devrait rester supérieure à un pixel de fond jusqu'à ce que l'objet en mouvement s'arrête. Pour modéliser cela, nous avons besoin d'une méthode pour décider quelle partie du modèle de mélange représente le mieux les processus d'arrière-plan. Tout d'abord, la gaussienne sont classés par la valeur de ω / ∂. Cette valeur augmente à la fois comme une distribution. Après la réestimation des paramètres du mélange, il suffit de trier de la distribution correspond à la distribution d'arrière-plan le plus probable, parce que seuls les modèles appariés valeur relative qui aura changé.

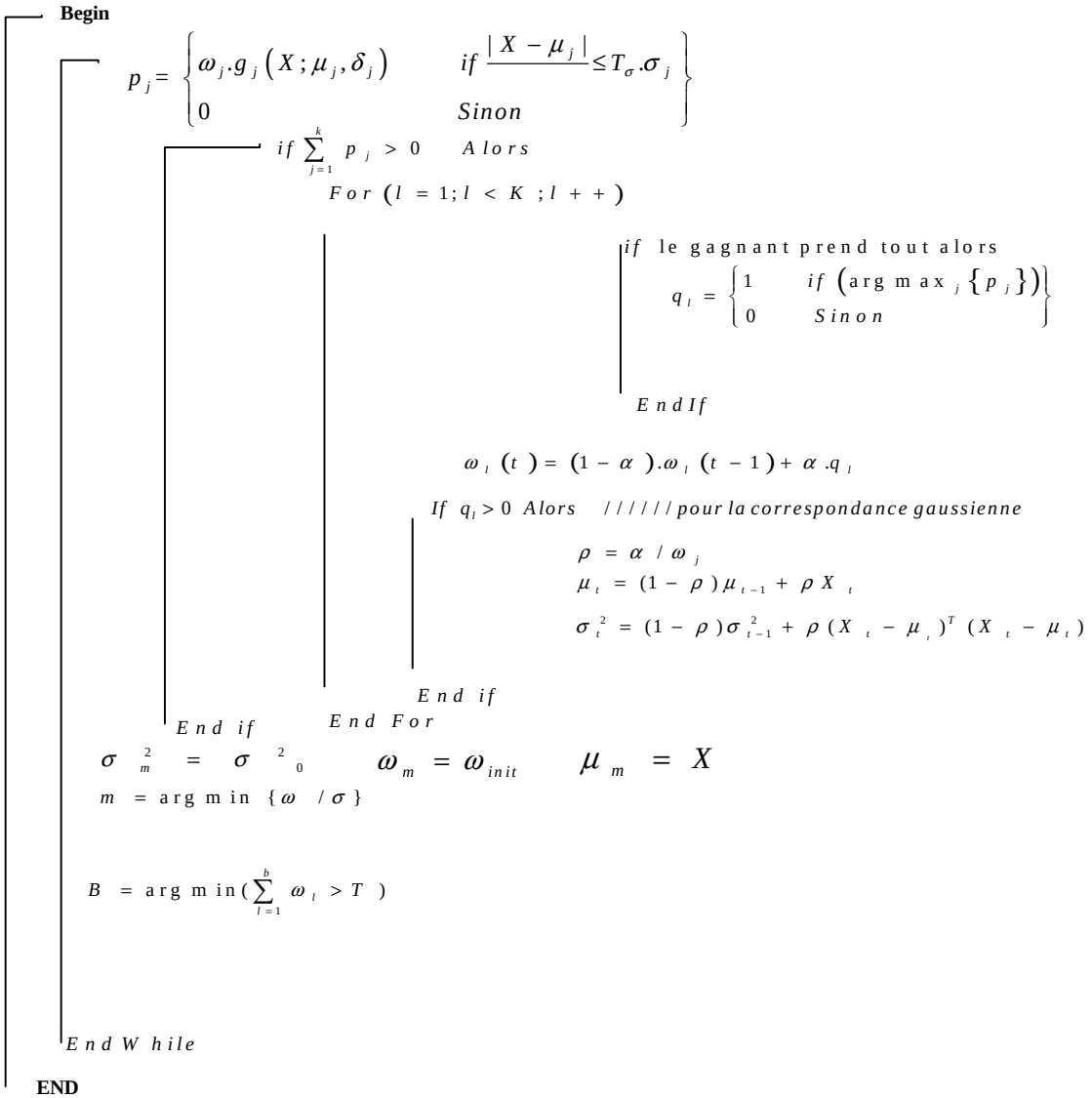
Puis les distributions B premiers sont choisis comme un modèle d'arrière-plan, où:

$$B = \arg\min\left(\sum_{l=1}^{b} \omega_l > T\right) \qquad \textbf{Eq7-11}$$

où T est une mesure des données qui doivent être pris en compte par le fond. Cela prend le «meilleur» des distributions jusqu'à ce qu'une certaine portion, T, des données récentes ont été prises en compte. Si une petite valeur pour T est choisie, le modèle d'arrière-plan est généralement uni-modal. Si tel est le cas, en utilisant uniquement la distribution la plus probable permettra d'économiser le traitement.

Si T est plus élevé, une distribution multi-modale causée par un mouvement de fond répétitif (par exemple les feuilles sur un arbre, un drapeau dans le vent, un clignotant de la construction, etc) peut entraîner plus d'une couleur étant incluses dans le modèle de fond. Il en résulte un effet de transparence qui permet à l'arrière-plan d'accepter deux ou plusieurs couleurs distinctes.

Algorithme 7-1 Mélange de Gaussiennes (GMM)



Initialisation :

- $\underline{X}_t$: *Valeur à l'instant*
- μ, σ_0 : *Moyenne et Variance*
- T: Seuil du modèle de l'arrière-plan
- K : Nombre de gaussiennes
- α, ρ: Taux

7.1.1.2 Traitement de l'image

7.1.1.2.1 Filtrage

Le filtrage consiste à appliquer une transformation à toute ou une partie de l'image en appliquant un opérateur. Cette transformation va avoir un impact sur l'image en la modifiant. L'utilisation courante des filtres est de détecter les contours dans une image, la rendre plus floue pour éliminer le bruit, la rendre plus nette pour intensifier les contours.

Nous utilisons des **filtres** pour **éliminer** le **bruit** de l'image. Le **brui**t caractérise les parasites ou interférences d'un signal, c'est-à-dire les parties du signal déformées localement. Ainsi le bruit d'une image désigne les pixels de l'image dont l'intensité est très différente de celles des pixels voisins. Le bruit peut provenir de différentes causes : Environnement lors de l'acquisition, défaut du capteur, qualité de l'échantillonnage.

7.1.1.2.1.1 Filtrage Moyenneur :

La suppression du bruit est un des prétraitements essentiels en vision par ordinateur : Il s'agit d'éliminer ce qui est dû aux aléas des mesures tout en essayant de ne pas altérer l'information utile contenue dans l'image. Ici, nous nous limitons à la suppression du bruit additif gaussien. Les filtres linéaires, dont les plus utilisés en pratique sont le filtre moyenneur et le filtre gaussien, en général, sur de petites fenêtres de taille 3*3 ou 5*5, permettant de réduire le bruit gaussien additifs. Leur inconvénient est qu'ils émoussent les contours et font disparaître les lignes trop fines. Dans notre étude, ceci est considéré plutôt comme un avantage. En effet, nous cherchons l'information globale, la précision des contours n'est donc pas une priorité. De plus, l'utilisation d'un tel filtre permet de ne pas désentrelacer la vidéo rendant les mouvements géométriquement plus gros.

-Suite à plusieurs acquisitions, nous avons remarqué que certaines parties de corps se séparent, Donc il fallait trouver un moyen pour les reconnecter entre eux. Pour

remédier à ce problème nous avons appliqué aux images une opération de dilatation suivie par une opération d'érosion .

7.1.1.2.2 Fermeture de contour :

Afin de parer aux problèmes liés à l'ouverture des contours chose qui faussera l'estimation des paramètres du corps nous avons opté par une fermeture de contour qui consiste en une opération de dilatation et d'érosion. Avec l'utilisation d'un élément structurant sous forme de disque. Pour notre projet, nous avons utilisé un élément structurant sous forme de disque avec un rayon égale à 3.

7.1.1.2.2.1 Dilatation :

Pour chaque position de B (B est l'élément structurant) sur l'image X, si un, au moins, des pixels de B fait partie de X, alors l'origine de B appartient à l'image générée.

Notation :

$$Y = X \oplus B = \{a + b : a \in X, b \in B\} \qquad \textbf{Eq7-12}$$

7.1.1.2.2.2 Erosion :

Pour chaque position de B sur l'image X, si tous les pixels de B font partie de X, alors l'origine de B appartient à l'image générée.

Notation

Eq7-13

$$Y = X \ominus B = (X^c + B)^c$$

7.1.1.3 Extraction des primitives

Le centre d'intérêt de ce travail est le corps humain. Le mouvement d'une personne est réalisé garce à ces différentes partie du corps d'où l'intérêt de les estimer. Mais cette dernière constitue un vrai challenge, comme cité auparavant, vu la nature du corps humain, sa forme variée selon les vêtements.

L'approche proposée consiste en la création d'une représentation du corps humain. Cette modélisation à la particularité d'être capable de fournir une description hiérarchique de la posture prise par le sujet a un instant donné. En effet, ce modèle du corps humain dont la mise en œuvre permet l'obtention d'une information sur la localisation des différents membres du sujet dans une image.

Cette représentation est sous forme de squelette. Le choix d'une telle configuration était pour les points suivants :

- Le squelette est une représentation naturelle du corps humain.
- Une représentation hiérarchique.
- Facilite le processus de la labellisation lors du suivi.

Afin de parer aux problèmes liés aux vêtements, il serait judicieux d'exploiter un algorithme permettant de réduire l'information utile pour pouvoir appliquer une analyse efficace, d'où l'utilisation de l'algorithme de la squelettisation. Dans ce qui suit nous allons présenter quelques concepts liés à la squelettisation.

7.1.1.3.1 Algorithme de la squelettisation :

7.1.1.3.1.1 Historique :

Le concept de squelette a été introduit pour la première fois par H. Blum en 1964, en vue de créer un nouveau descripteur de formes. Il utilise le concept de feu de prairie, c'est-à-dire, des feux provenant des points de contour de l'objet et qui se propagent vers l'intérieur à vitesse constante. Le squelette est alors formé par les points où les fronts de ces feux créent une intersection. Ces points sont aussi appelés points d'extinction. Une autre définition donnée par L.Calabi en 1965

considère le problème d'un point de vue topologique. Cette définition est basée sur le concept de boules maximales. Il définit le squelette d'un objet comme étant l'ensemble des centres de ses boules maximales. Une boule incluse dans un objet est dite maximale s'il n'existe pas d'autres boules incluses dans l'objet la contenant entièrement.

7.1.1.3.1.2 Définitions et propriétés :

Formellement, un squelette est une représentation géométrique d'un objet dans une dimension inférieure. Il permet de décrire d'une manière compacte les propriétés d'un objet, en particulier sa forme. Dans le plan, le squelette d'un objet est un ensemble de lignes passant en son milieu appelé axe médian (« medial axis »).
Les squelettes présentent quelques propriétés intéressantes telles que :

- **Invariance par translation et rotation** : Le squelette est invariant par translation et rotation. Étant donné une translation ou une rotation **g** et un objet X. Soit S(X) le squelette de l'objet X. Nous avons S(g(X)) = g(S(X)).

- **Réversibilité :** A partir des points du squelette et des rayons des boules maximales, il est possible de reconstruire la forme. Ainsi la squelettisation est réversible à condition d'avoir mémorisé en chaque point p du squelette, le rayon r(p) de la boule maximale centrée en p. La fonction r est appelée fonction d'étanchéité.
- **Structure de graphe** : Sous certaines hypothèses de régularité, il est possible de montrer que le squelette a une structure de graphe, où les nœuds sont considérés comme des articulations et les arêtes comme des os. Ainsi, les techniques issues de la théorie de graphes peuvent être appliquées directement aux objets.
- **Homotopie :** La notion mathématique permettant de formaliser le concept d'objets topologiques équivalents est le type d'homotopie. Deux objets ont le

même type d'homotopie s'ils sont typologiquement équivalents. C'est à dire, si nous pouvons passer de l'un à l'autre par une déformation continue, à condition que les points de l'objet, qui étaient proches les uns des autres avant transformation, demeurent proches les uns des autres dans l'objet transformé. Dans le plan, deux objets homotopes ont le même aspect et justifie l'utilisation du squelette comme descripteur de formes.

- **Minceur :** Le squelette est typologiquement mince, c'est-à-dire qu'il a un pixel d'épaisseur, sauf aux jonctions pour lesquelles un pixel ne suffit pas à garantir l'homotopie.
- **Localisation :** Le squelette est situé au centre de l'objet.

7.1.1.3.1.3 Algorithme d'amincissement topologique :

A l'heure actuelle, il existe plusieurs manières de calculer un squelette.

Voici les principales :

- Par simulation des déplacements des fronts d'onde d'un feu de prairie
- Par extraction des lignes de crêtes dans une carte de distance. Une carte de distance est une image, ou chaque point est associé à la distance entre ce point et le bord le plus proche. Les lignes de crêtes représentent les maxima locaux.
- Par amincissement topologique. Nous allons détailler cette méthode dans ce qui suit.
- Par calcul analytique des axes médians. Cette technique consiste à modéliser le contour de l'objet par des objets dont le squelette est connu (des polygones) puis à rassembler les squelettes pour obtenir le squelette global.

Dans le cadre de notre projet, nous utilisons l'algorithme d'amincissement topologique dont voici une description.

7.1.1.3.1.3.1 Principe :

Consiste à retirer au fur et à mesure les points du contour de la forme, tout en préservant ses caractéristiques topologiques. Pour ce faire il part du contour initial de l'objet, étudie la connexité de chaque point du contour dans un voisinage, et enlève ceux dont la suppression n'affecte pas la topologie de l'objet. Le squelette est obtenu en érodant itérativement les couches frontières de l'objet (dans notre cas du corps humain).

Les points supprimables sont enlevés soit successivement, soit en parallèle, ou encore à l'aide d'opérations morphologiques. Ces méthodes conduisent à un squelette homotope à l'objet (ou au corps humain) par construction, mince, géométriquement représentatif mais pas forcément centré.

7.1.1.3.1.3.2 Algorithme :

Comme cité auparavant, la squelettisation consiste à réduire des couches qui entourent le pixel. Cette réduction s'opère par la répétition d'une procédure d'amincissement, qui examine les pixels du bord de la figure, et enlève ceux qui vérifient un critère :

Algorithme : Réduction

Répéter

Amincissement

Jusqu'à ce qu'à stabilité

Le processus d'amincissement est réalisé par l'algorithme de Hilditch, réalisé en plusieurs passes. Afin de définir la procédure d'amincissement il faut définir deux fonctions :

NZ(p1) : Nombre de pixels voisins de p, différents de zéro.

C(p1) : Nombre de pixels connexes.

Ainsi qu'une notion de voisinage tel que (8 pixel voisins):

Algorithme Amincissement

Répéter pour chaque pixel faire

Changer la valeur du pixel du 1 à0 s'il satisfait les 4 conditions :

- $2 \leq NZ(P1) \leq 6$.
- C(P1)==1.
- P2*P4*P8=0 ou C(P2) !=1.
- P2.P4.P6=0 Ou C(P4) !=1.

Jusqu'à ce qu'il n'aura aucun changement.

Résultats :

Figure 7-3 Personne sur scène

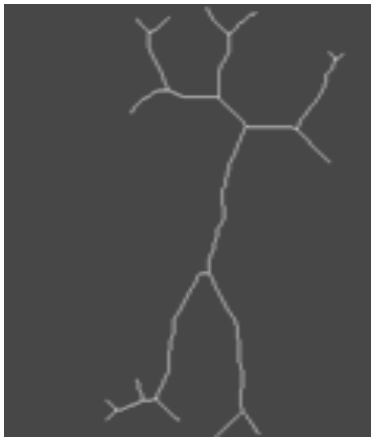


Figure 7-2 Résultat de la squelettisation

7.1.1.3.2 Extraction des points d'intérêts :

Afin de réaliser un suivi, il faut faire recours à des points d'intérêts, en analogies avec les systèmes de capture : « Extraction de primitives ». Vue que notre but est de se baser sur les algorithmes de traitements d'images nous avons remarqué que le squelette obtenu est constitué en des intersections ces intersection peuvent être considérés comme étant les articulations du corps humain.

Pour cela, nous avons fait recours à l'algorithme de détection de coin Harris.

7.1.1.3.2.1 Algorithme d'Harris :

Cette méthode différentielle se fonde sur l'analyse de la variation de la luminance au voisinage d'un point. Dans le cas d'un coin ou d'une autre configuration complexe qui n'est pas soumise au problème d'ouverture, la variation de l'image est grande quelle que soit la direction dans laquelle on effectue un décalage de l'image, par opposition à un contour ou à une zone homogène.

7.1.1.3.2.1.1 Etape de calcul des points d'intérêt par Harris (Détection de coins) :

1- Calcul des images des gradients X et Y de l'image
2- Filtrer ces images résultantes par un filtre Gaussien de taille 3x3
3- Calcul des gradients XX et YY de l'image (en utilisant pour l'image 2 fois de suite l'opérateur dérivé)
4- Calcul des coins de Harris pour chaque pixel de l'image de la façon suivante :
5- Calcul de la matrice de Harris :

$$M = \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \qquad \textbf{Eq7-14}$$

Avec

Ix est le gradient suivant l'axe X et *Iy* le gradient suivant l'axe Y

Ix^2 et Iy^2 sont les images convoluées 2 fois avec le gradient X et Y

- Calcul de la trace et du déterminant de la matrice

$$Trace(M) = \lambda_1 + \lambda_2 = M_{1,1} + M_{2,2}$$
$$Det(M) = \lambda_1 \lambda_2 = M_{1,1} M_{2,2} - M_{1,2} M_{2,1}$$

- Calcul de la réponse *R* du détecteur :

$R = Det(M) - k(Trace(M))^2$, *avec k* est un paramètre à régler –

Typiquement k = 0.04

6- Extraction des maxima locaux positifs dans un voisinage 3x3 de la réponse *R* (c'est-à-dire mettre à zéro tous les points négatifs ou dont la valeur n'est pas supérieure à celle des 8 voisins)
7- Extraction des *n* meilleurs points de Harris (par tri par insertion dans un tableau de taille *n*) ,*n* étant un paramètre à configurer, avec par exemple *n*=50

La matrice M dite du gradient carré moyen décrit la manière dont la variation de luminance se comporte autour du point. Une configuration où l'image présente des variations dans toutes les directions correspond au cas où les deux valeurs propres de cette matrice sont de même importance et suffisamment grandes pour ne pas être causées par du bruit.

Dans le cas d'un contour, une valeur propre est prépondérante, et la variation de luminance ne dépend que d'une seule direction, et reste faible selon la direction du contour.
Pour éviter le calcul des valeurs propres, prend comme "valeur d'intérêt" la grandeur suivante:

$\det(M)-K*\text{trace}^2(M)=(\lambda_1 * \lambda_2)-K-(\lambda_1+\lambda_2)^2$ où λ_1 etλ_2 sont les valeurs propres de M. Cette grandeur favorise en effet les configurations où deux valeurs propres sont grandes et d'égale importance. En dehors du seuil sur la "valeur d'intérêt", nous avons trois paramètres : k, qui influence le nombre de points détectés, la taille de la fenêtre de lissage, ainsi que le choix d'une méthode d'estimation de la dérivée. Concernant k, l'expérimentation montre selon plusieurs auteurs qu'un nombre optimal de points est obtenu pour une valeur de l'ordre de 0,04 (on considère qu'un maximum local de la valeur d'intérêt définit un point d'intérêt s'il est positif).

Résultats :

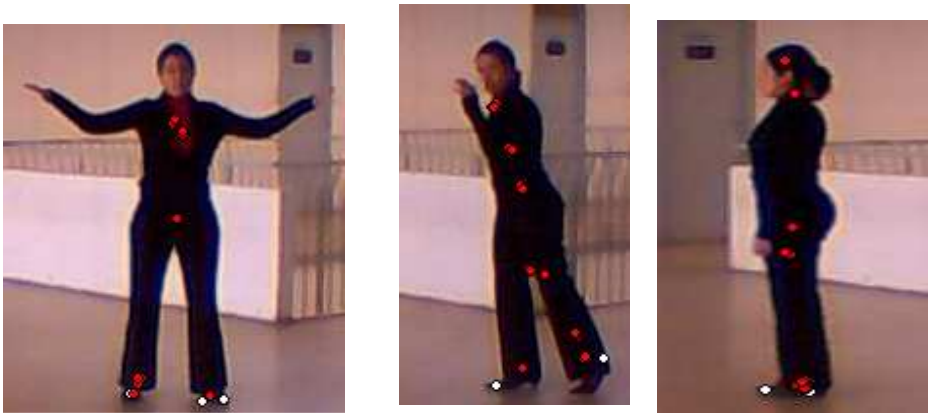


Figure 7-4 Résultat du détecteur de coins Harris

7.1.1.3.3 *Classification des points d'intérêts :*

Après l'utilisation du détecteur de harris, nous avons constaté un regroupement de point d'intérêt cela est dû à la sensibilité de l'algorithme de la squelettisation au bruit.

Pour parer à ce problème, pour avoir une représentation raffinée, nous avons opté par une classification basée sur la distance euclidienne tel que :

$$\forall\, Pi, Pj \in Coin, Si\ distance\ (Pi, Pj) < Seuil\ alors\ Pi, Pj \in Ck$$
$$Sinon\ Pi \in Ck, Pi \in Ck + 1 \qquad \textbf{Eq7-15}$$

Le seuil sera fixé en fonction des expérimentations De chaque classe, il ne sera retenu que le représentant de chaque classe.

Résultats

P.S : Les points représentants sont colorés en jaune.


Figure 7-5 Classification des points d'Harris

Il est remarquable, que ces points d'intérêts offre une analyse globale sur le mouvement de la personne sans prendre en considération les mouvements locaux(mouvement relatif à une partie précise du corps). Pour parer à ce problème nous avons renforcé cette approche faisant recours à la morphologie humaine.

7.1.1.4 Extraction du squelette par la biologie humaine :

Après avoir extrait les points d'intérêt par le biais de l'algorithme de Harris sur le squelette extrait via la méthode de la squelettisation. Ceci n'a pas suffi de donner un squelette modélisant la morphologie humaine. Dans ce but, nous avons fait recours à la biologie afin de créer un squelette qui nous permettra par la suite une analyse et un suivi des parties du corps.

7.1.1.4.1 Description du modèle du squelette :

Le squelette est un modèle linéaire qui représente la pose de la personne sur l'image. La pose idéale est lorsque la personne se trouve face à la caméra contrairement au cas où la personne est de profile par rapport à celle-ci.

Notre modèle de squelette peut être représenté par un vecteur de 6 parties de corps : Tête, torse , cuisses et jambes.

7.1.1.4.2 Représentation mathématique :

La représentation du squelette humain est réalisé par le vecteur B qui englobe six partes du corps, tel que :

$$B = \{bp1, bp2 \dots\dots\dots\dots\dots bp6\}$$

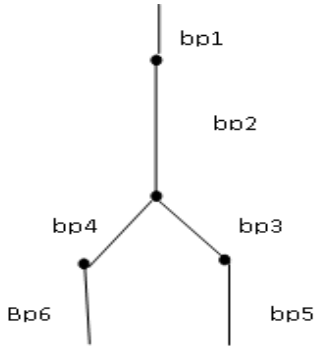


Figure 7-6 Squelette biométrique

Chaque partie du corps à sa propre proportion et deux extrémités :

bpi={ex $_{i;1}$, ex $_{i;2}$} où ex $_{i;j}$ ={x_{ij} , y_{ij} }

-x_{ij} : Coordonnées x de l'extrémité

- y_{ij} : Coordonnées y de l'extrémité.

7.1.1.4.2.1 Extraction du torse:

Pour extraire le torse , il faut extraire ces différentes extrémités : ex $_{2.1}$ et ex$_{2.2}$.

-ex $_{1.1}$ = ex $_{2.1}$: représente le cou.

- ex $_{2.2}$: Représente le point entre les jambes.

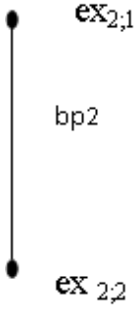


Figure 7-7 Extrémités du torse

7.1.1.4.2.2 Estimation du cou :

T={bp2}

Représente la colonne vertébrale de la personne, le positionnement des extrémité s est basé ur les proportions morphologiques du corps humain.

$ex_{2;1} = 2/15\ T$

tel que T, représente la taille de la personne. Estimation du point entre les jambes : En faisant référence à la morphologie humaine, la longueur en moyenne des jambes est de 4 têtes.

$ex_{2;2} = 4*(ex_{1;2} - ex_{1;1})$

Donc le point « $ex_{2;2}$ » ce trouve à une distance de 4 têtes par rapport aux pieds.

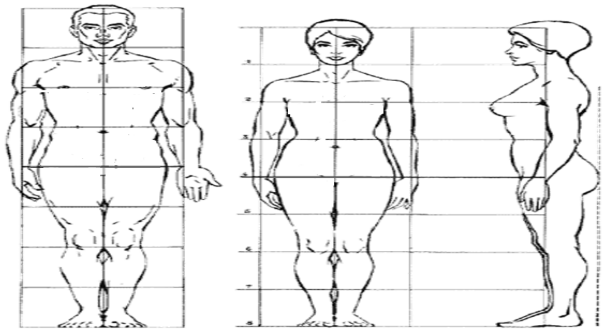


Figure 7-8 Mesure du corps humain par la tête

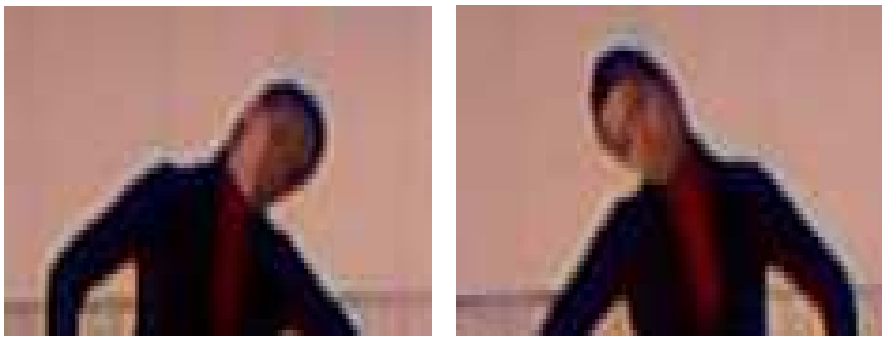


Figure 7-9Différentes orientations de la tête

7.1.1.4.2.3 Extraction de la tête :

Le point constituant l'extrémité de la tête, est le point extrême de tout le corps humain. Afin de définir l'orientation de la tête on construit une région d'intérêt, qui contiendra les extrémités possibles de la tête :RI={E1,E2,E3}.

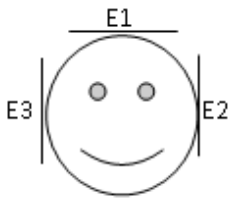


Figure 7-10 Définition des extrémités de la tête

Le point qui constituera extrémité de la tête, est celui qui respectera la règle suivante :

$ex_{1;1}=\text{Max}\left(\sqrt{((Ex\ 2.1-\ Exi)2\ +(\ Ex\ 2.2-\ Eyi)2)}\right)$ **Eq7-17**

Exi : Coordonnée x du point Ei., i=1…3

Eyi : Coordonnée y du point Ei. i=1…3

7.1.1.4.2.4 Estimation des jambes :

Connaissant deux points extrêmes : A et B tel que :

A= $ex_{2;2}$ .Il représente l'intersection des jambes.

B : Le point extrême d'une jambe . Tel que :

Le choix de B, est détaillé dans l'organigramme suivant :

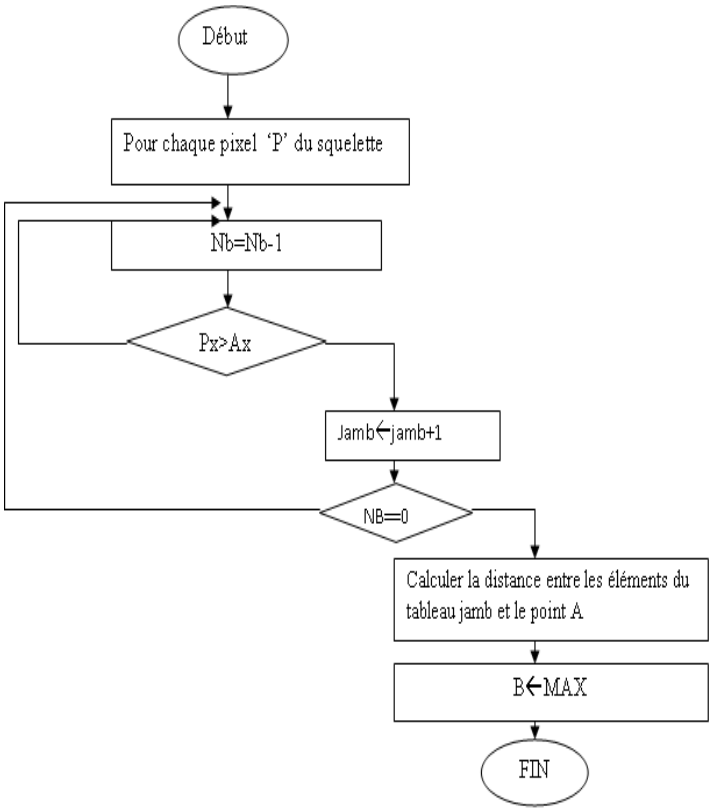


Figure 7-11 Organigramme de l'extraction de l'extrémité basse de la jambe

Ax,Px : Cordonnée représentant la ligne du pixel.

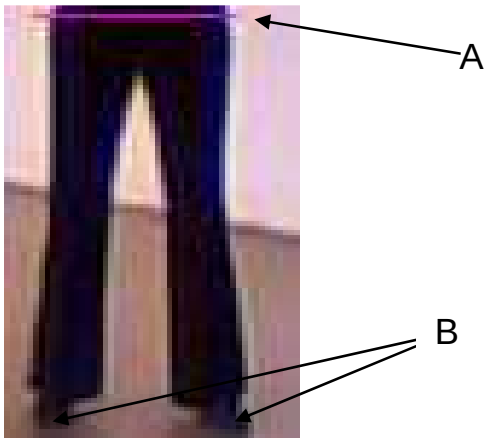


Figure 7-12 Extrémité de la partie inférieure du corps

Une fois que les deux points extrêmes des jambes trouvés, il faut passer une représentation hiérarchique : La recherche des genoux.

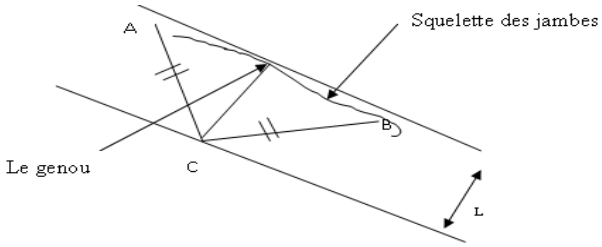


Figure 7-13 Estimation du genou

Nous cherchons à estimer le point « C », tel que : $AC \cong AB$. Nous effectuons par la suite une recherche récursive le long de « L », afin de trouver le point du squelette correspondant.

7.1.1.4.3 Combinaison des résultats :

Afin de combiner les points extraits par l'algorithme de harris et ceux par la morphologie humaine nous avons subdivisé le corps en t régions d'intérêts tels que :

- Tête : C1.
- Torse : C2.
- Cuisse gauche C3.
- Jambe gauche C4.
- Cuisse droite C5.
- Jambe droite C6.

On définit l'organigramme, qui offre une classification de tous pixel provenant des deux méthodes :

Px coordonnée x du pixel candidats.

Py coordonnée y du pixel candidats.

GG : genou gauche.

GD : Genou droit

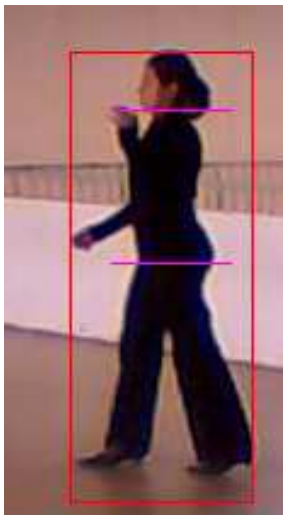


Figure 7-14 Segmentation du corps en trois parties

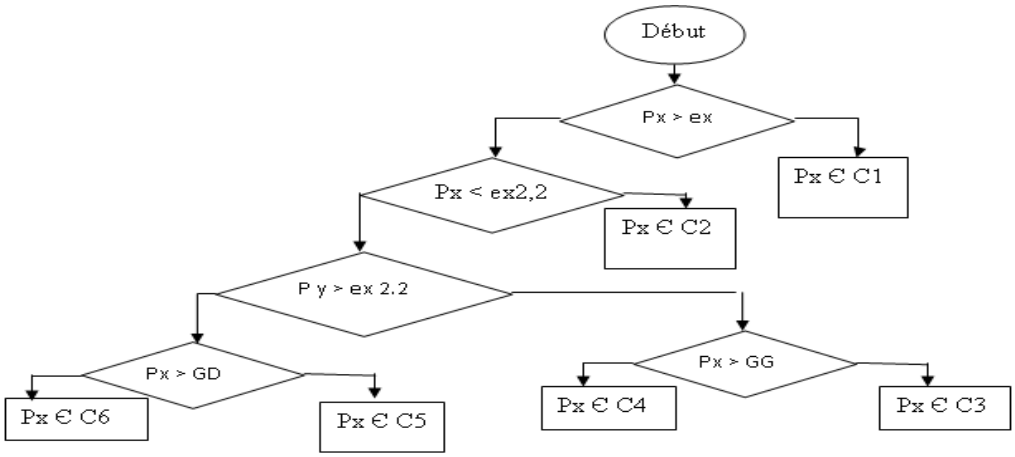


Figure 7-15Classification des points d'intérêts

7.1.1.4.4 *Connexion :*

La connexion des pixels afin de former les segments du corps set à la base d'une classification intra classe à l'aide de la distance euclidienne. Un pixel au sein de la même classe est connecté au pixel le plus proche à lui.

7.1.1.4.4.1 Algorithme connexion

Pour tout pixel « Pi » de la même classe

Min←Rechercher_min(Pi,pj) /j= ! i, j=1…….N=nombre de pixel dans la classe

Relier (Min,Pi).

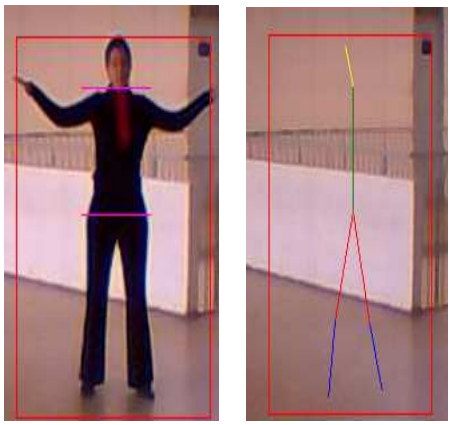


Figure 7-16 Squelette de la personne

7.1.1.4.5 Discussion

Dans cette phase de projet, nous nous sommes servis de l'information de base de l'image (contour, coins, mouvement etc) ainsi que des informations liées à la morphologie de l'être humain. Ce qui nous a permet de détecter les membres moins soumis aux contraintes d'auto-occlusions, en fixant la posture à adopter, tel que tête, torse et jambes.

En généralisant cette méthode sur d'autres postures cette approche montre ses limites, principalement dans la détection des bras.

Aussi, dans la phase de détection en adaptant l'approche de soustraction de l'image de fond pour la détection de la personne, nous avons remarqué que tous les membres statiques peuvent être considérés comme appartenant à l'image de fond ce qui pourrait fausser la phase de l'estimation de pose.

A cet effet, nous avons élargi notre champ de travail, en adoptant d'autres méthodes pour chaque phase de projet :

- Algorithme basé histogramme de gradient orienté pour la détection de la personne.
- Algorithme basé modèle pour l'estimation de la pose , en adoptant la méthode de « Pictural structure ».

7.2 Stratégie adoptée

Afin d'avoir une version rapide de notre algorithme, nous avons mis en avant lors de la phase de l'implémentation deux principaux objectifs :

1- Minimisation de la consommation de la mémoire : En évitant les opérations inutiles d'allocation\dés-allocation de la mémoire ainsi que la conservation des données en mémoire. L'approche générale est d'essayer de traiter les données aussi vite que possible et les supprimer si elles ne sont plus utilisées.
2- Optimisation de l'opération de l'inférence, par le choix pertinent d'un algorithme de convolution.

7.3 Environnement de développement

L'implémentation de notre interface a été réalisée sur le logiciel Matlab 7.11.0 (R2010b), le choix de cette version de logiciel n'est pas anodin, pour des raisons de réduction de temps de traitement cette version de Matlab est adoptée pour sa compatibilité avec Visual studio C++2010 Express pour l'environnement Mex.
Le développement a été réalisé sur une machine Intel ® Core ™ i5-2410 CPU @2.30 GHZ , RAM 4.00 GO .

7.4 Détection de personnes

Dans cette partie nous nous focalisons sur la localisation de la cible qui est dans notre cas le corps humain en entier, pour nos propos nous avons exploité deux approches : Extraction de l'image de fond par le Mélange de gaussiennes , le diagramme 6-1 illustre le processus, et l'autre détection de personne par histogramme de gradient orienté, où respectivement la première utilise le mouvement comme primitive et la deuxième l'information de gradient, sachant que la détection de la personne par histogramme de gradient orienté représentera une phase dans le processus de l'estimation de la pose humaine.

7.4.1 Extraction de fond par mélange de gaussiennes

Nous exploitons l'information de mouvement pour la détection de la personne, et afin de parer à tous changement dans l'image de fond pouvant altérer le processus de détection, nous exploitons le mélange de gaussienne pour la construction de l'image de fond , ce qu'on appelle un arrière-plan adaptatif.

L'approche de mélange de gaussiennes « GMM » est implémentée avec les paramètres initiaux suivants :

$k = 4$: Le nombre de composante de mélange.

$\alpha = 0.002$ s : Le temps nécessaire pour adapter les poids des composants

$\rho = 0.002$ s : Le temps nécessaire pour adapter les moyennes et les covariances des composantes.

DT : Le seuil utilisé pour la mise en correspondance = 44.6976

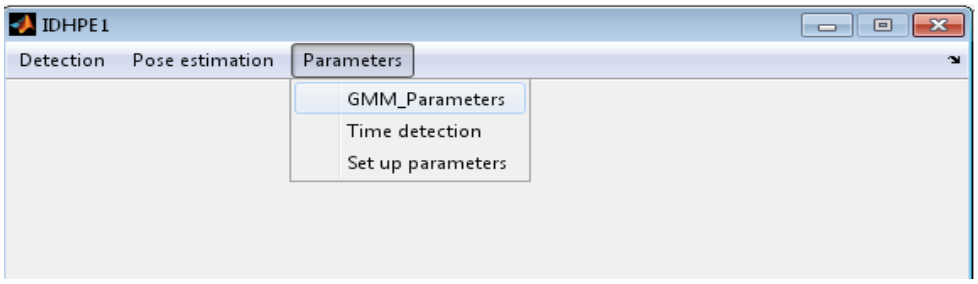
IV : Variance initiale = 7.99693

IM : La probabilité initiale apriori pour les composants nouvellement placées = 0.00868042

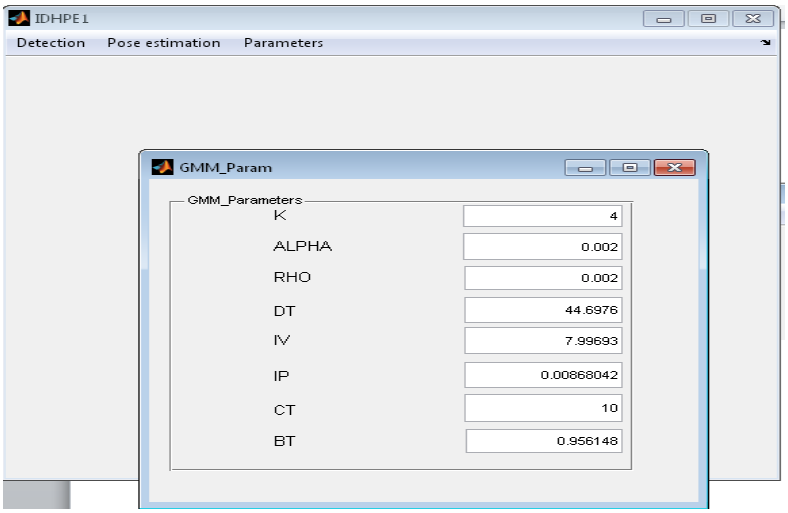
CT : Seuil pour filtrer les composantes connexes de petite taille =10

BT : Pourcentage de poids qui doit être représenté par des modèles d'arrière-plan = 0.956148

Dans notre interface, nous prévoyons un accès aux paramètres par l'utilisateur



(a)



(b)

Figure 7-17 Accès aux Paramètre GMM sur l'interface IDHPE (a,b)

Une fois les paramètres initialisés, l'utilisateur lancera la détection via l'approche de construction de l'image de fond par GMM comme illustré sur les images ci-dessous :

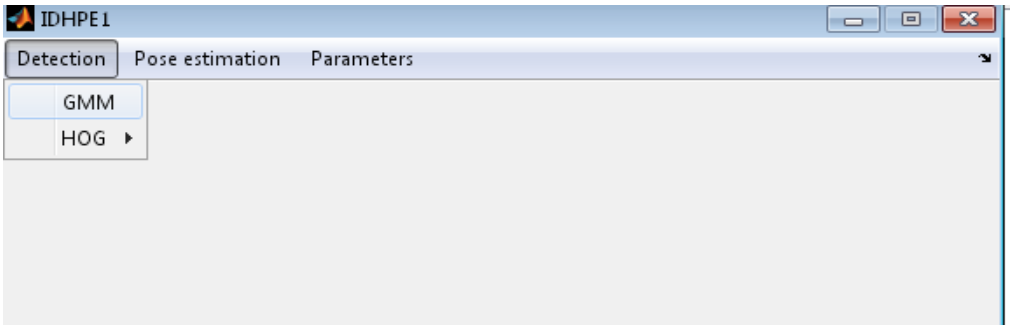


Figure 7-18 Sélectionner Detection-> GMM

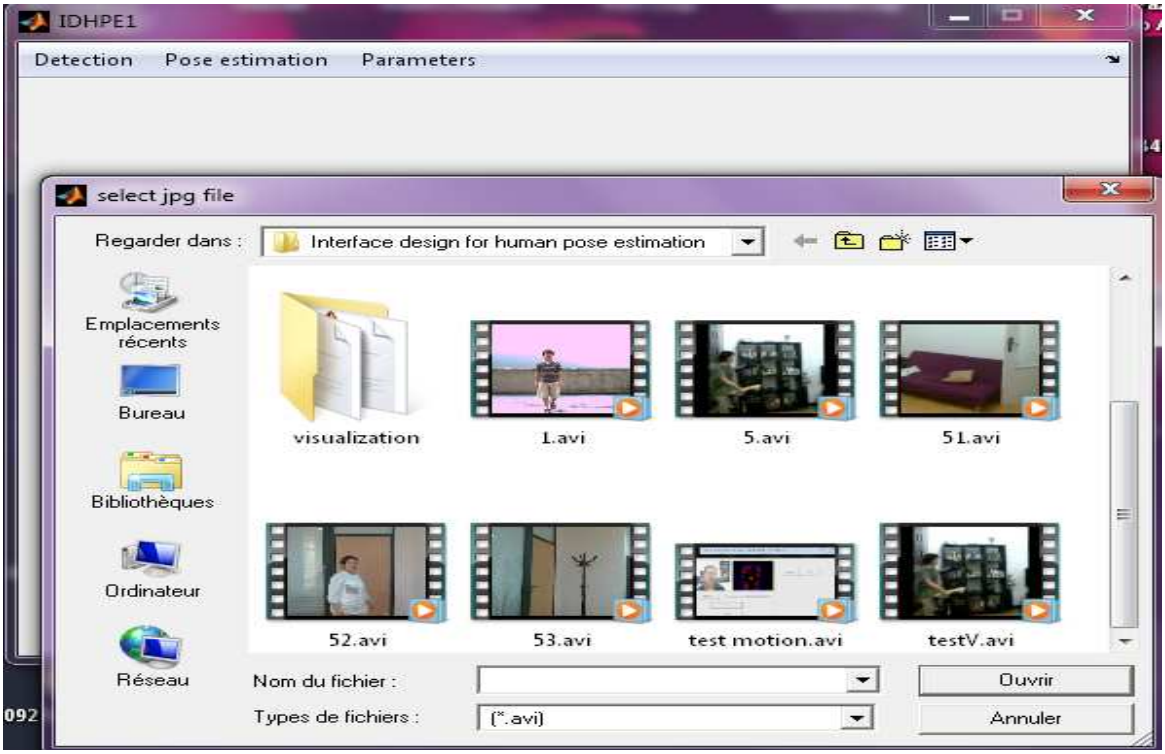


Figure 7-19 Sélectionner une vidéo, format AVI

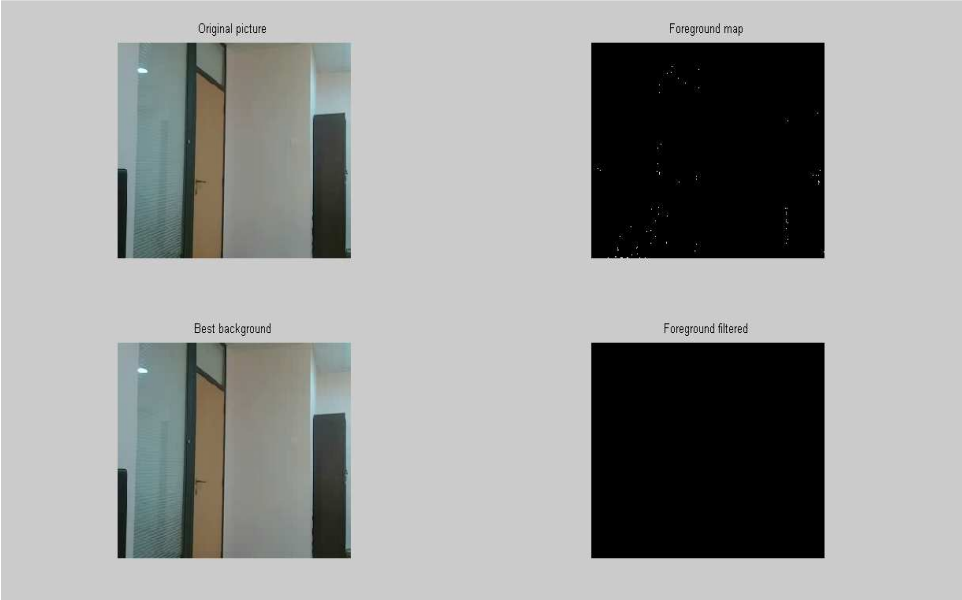


Figure 7-20 Construction d'un arrière-plan adaptative

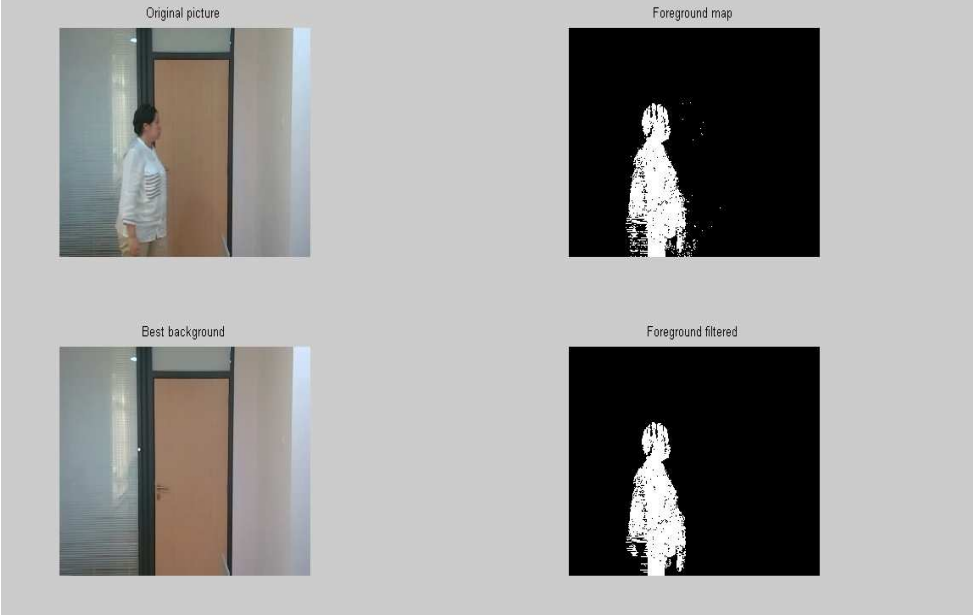


Figure 7-21 Détection par soustraction de l'image de fond

Le temps de détection est estimé à deux images par seconde.

La détection d'objet en mouvement par extraction de fond adaptatif est soumise à un problème lié à la nature de mouvement lui-même, càd , si l'objet exhibe un mouvement lent où qu'il soit statique , après la mise à jour de l'image de fond les pixel qui appartenaient à l'image de fond appartiendront à l'image de l'arrière-plan tel que illustré dans la figure

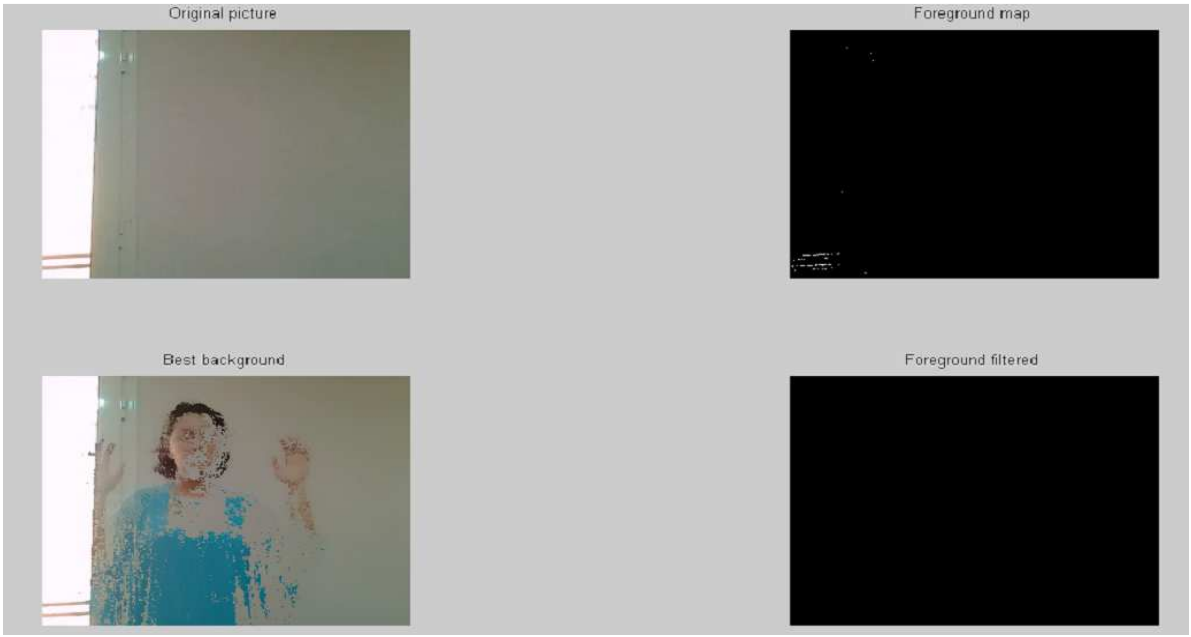


Figure 7-22 Mouvement lents et détection par extraction de fond adaptative

7.4.2 Détection par Histogramme de Gradient Orienté

7.4.2.1 Apprentissage du modèle

7.4.2.1.1 Données d'apprentissage

Dans la littérature, l'histogramme de gradient orienté pour la détection de personne a été proposé par la première fois dans les recherches de Dalal& Triggs [37] pourla détection de piétons. La phase d'apprentissage a été réalisé sur la base de données INRIA [100], devenue un benchmark dans le domaine de la vison par ordinateur ; La base de données INRIA, contient 2416 images positives, 1218 images négatives et 288 images de test.

Ayant pour but la détection de piétons, les personnes prises dans les images ont une contrainte d'avoir une position droite (debout).

Rappelons notre objectif primaire qui est l'estimation de la pose humaine , quel que soit la pose humaine adoptée, nous devrions avoir à un jeu de données d'images sur des poses épousant des poses variées.

Pour nos propos, nous réalisons la phase d'apprentissage sur la base de données : Pattern Analysis , Statistical Modelling and Computational Learning Visual Object Classes (**PASCAL VOC**) ; Réseau d'excellence financé par l'UE au titre du programme IST de l'Union européenne.Relative à l'année 2007.

Le "Pattern Analysis , Statistical Modelling and Computational Learning Visual Object Classes " est considéré comme un des plus grand défi et référence dans le domaine de l'étude de la reconnaissance d'objet pour la communauté des chercheurs dans le domaine de la vision par ordinateur. Toutes les images sont collectées du site web **flickr partage de photos**,Ce qui rend la base de données **impartiale** (unbiased) , dans le sens où les images n'ont été choisies pour un objectif

particulier . Qualitativement, les images contiennent un grand nombre de variabilité de conditions (pose, éclairage, etc).

Pour la mise en place du processus d'apprentissage pour la détection de la personne , la racine par rapport à notre approche d'estimation de la pose humaine, nous exploitons le package SVM Light .

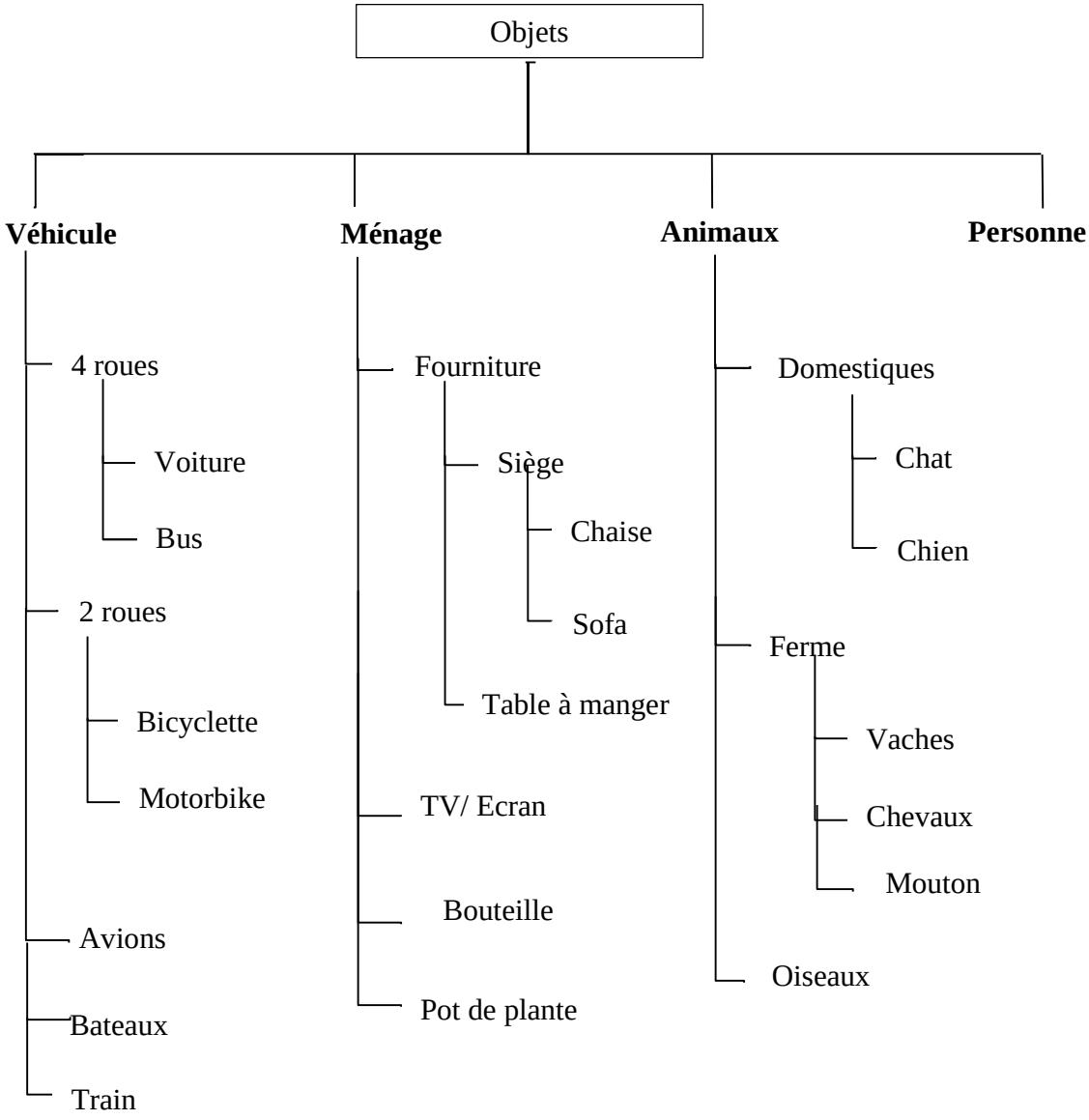


Figure 7-23 Classes Pascal VOC

7.4.2.1.2 Processus d'apprentissage

Nous exploitons une métrique très importante lors de la phase de l'apprentissage, nous voudrions obtenir un modèle du corps humain universel, à cet effet, nous nous basons sur les statistiques dans notre base d'apprentissage. Ci-dessous, nous décrivons les différentes phases d'apprentissage :

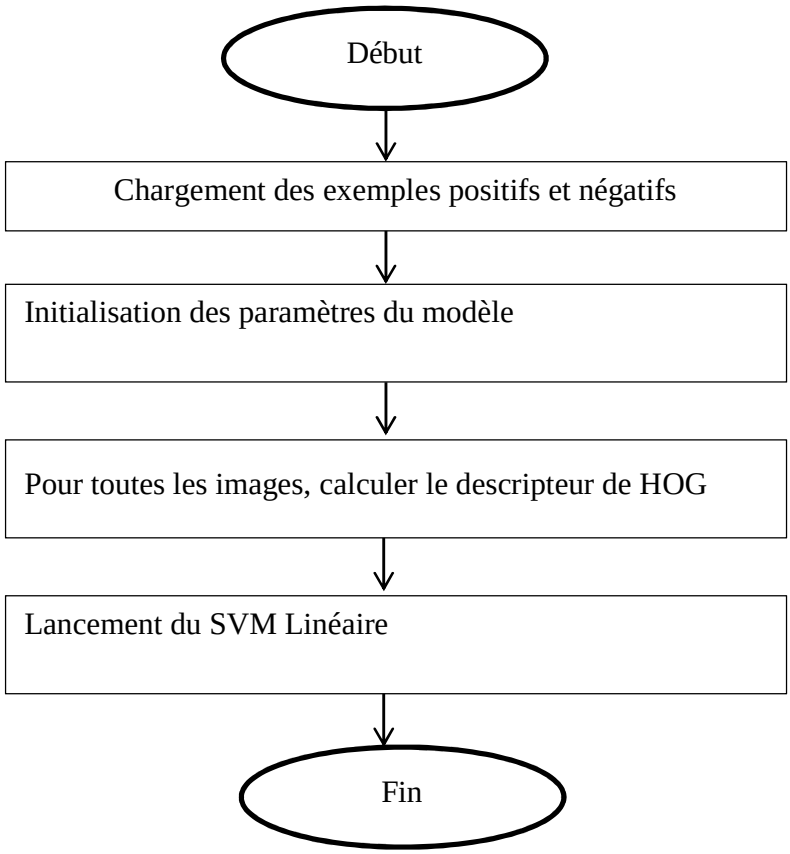


Diagramme 7-1 Processus d'apprentissage

a. Chargement des exemples positifs et négatifs :

Les exemples positifs sont construits à partir des exemples de la base d'apprentissage **sans occlusion** (tel que étiquetés dans les données de PASCAL). Nous utilisons des sous-fenêtres aléatoires à partir d'images négatives pour générer des exemples négatifs. Pour toutes les images, calculer le descripteur de hog.

b. **Initialisation des paramètres du modèle**

L'initialisation du modèle est fondée sur les statistiques sur les boites englobantes de la base d'apprentissage :

1- Calcul de l'aspect ratio des images :

Aspect_ratio_image = hauteur de l'image/largeur de l'image.

L'aspect ratio pris en compte pour l'apprentissage du modèle est le plus commun dans le jeux de données.

La taille du filtre est initialisée à la plus grande taille ne dépassant pas 80% des données.

2- Paramètres d'initialisation

-Nombre de classe pour la construction de l'histogramme de gradient orienté : 9.

-w=sqrt (plus grande suface image/aspect)

-h=w*aspect

-Taille initiale du filtre racine : h/nbr de classe w/nbr classe

-interval (pour la pyramide de descripteurs) = 10

-Block=2

-Calcul du seuil de détection

Le seuil de détection est liée à la base d'apprentissage, il est choisi par rapport aux paramètres : Précision et rappel caractérisant la base d'apprentissage [101].

A partir de la courbe PR (précision/rappel), sélectionner l'index correspondant au le plus petit seuil tel que précision >= rappel.

A partir des score confidence des boites englobantes de la base d'apprentissage, classe personne, choisir le score relatif à l'index.

Calcul du vecteur de descripteur HOG

a. Calcul de vecteur de descripteur HOG pour chaque image positive
b. Calcul de vecteur de descripteur HOG pour chaque image positive renversée horizontalement
c. Calcul de vecteur de descripteur HOG pour chaque image négative
d. Calcul de vecteur de descripteur HOG pour chaque image négative renversée horizontalement.

Pour le processus d'apprentissage, nous exploitons le logiciel SVM Light, avec c= 0.02 comme paramètre pour la fonction objective.

7.4.2.2 Détection

Nous illustrons dans le diagramme2 ci-dessous le processus de détection par fenêtre coulissante, le principe est basé sur une mise à l'échelle de l'image, voir diagramme 6-3, à chaque niveau un vecteur de descripteur de HOG est extrait pour être classifié par la suite.

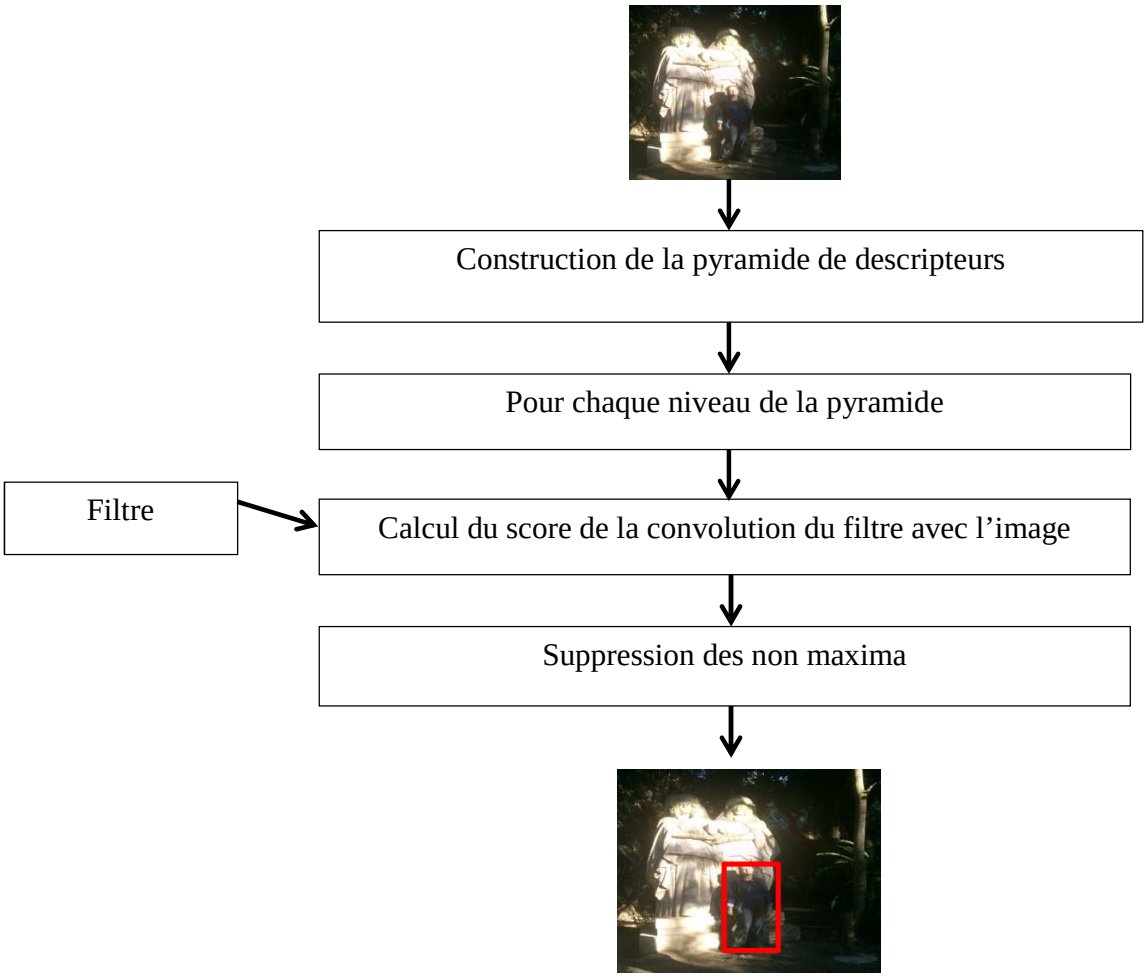


Diagramme 7-2 Processus de détection

7.4.2.2.1 Suppression de nom-maxima

Pour l'algorithme de la suppression de non-maxima est basé sur la notion de chevauchement entre les boites englobantes, le chevauchement est défi par la l'équation **Eq** 5-18 :

$$Chevauchement = \frac{\sup(BEi \cap BEj)}{\sup(BEi \cup BEj)}$$ **Eq** 7-18

Avec BEi et BEj deux boites englobantes.

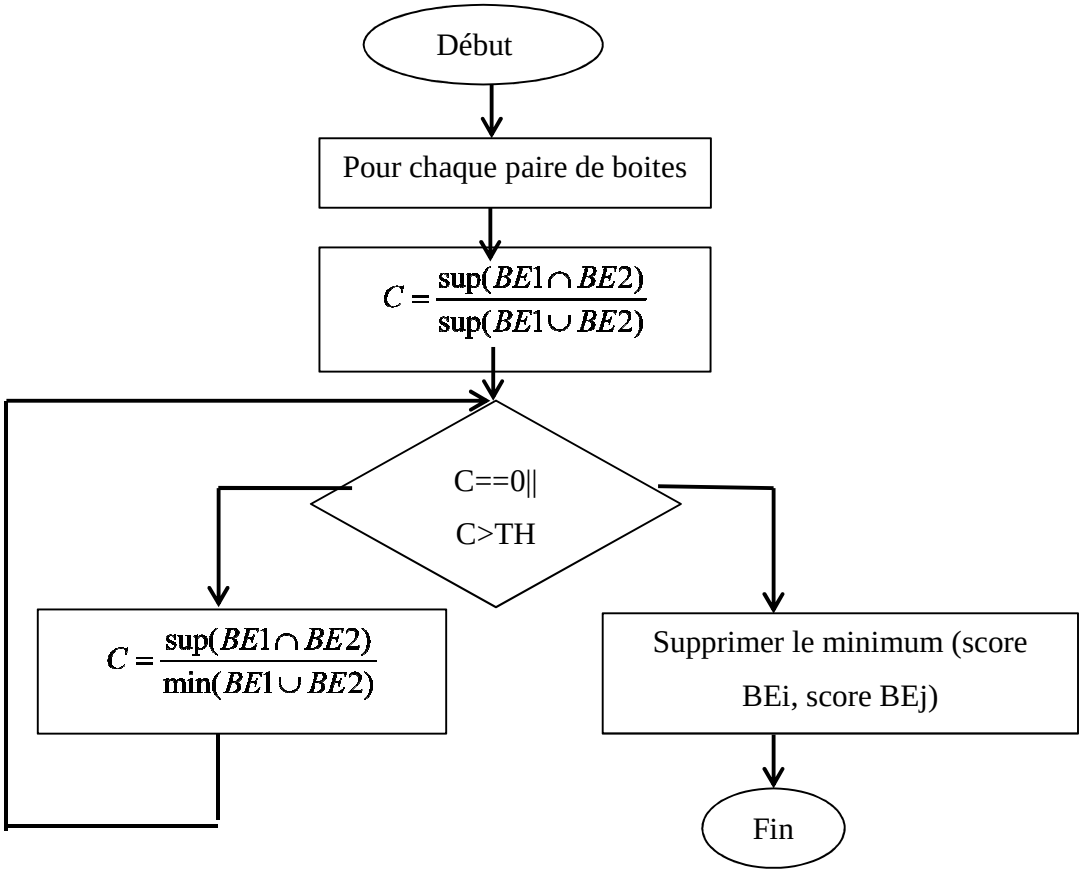


Diagramme 7-3 Suppression de non maxima

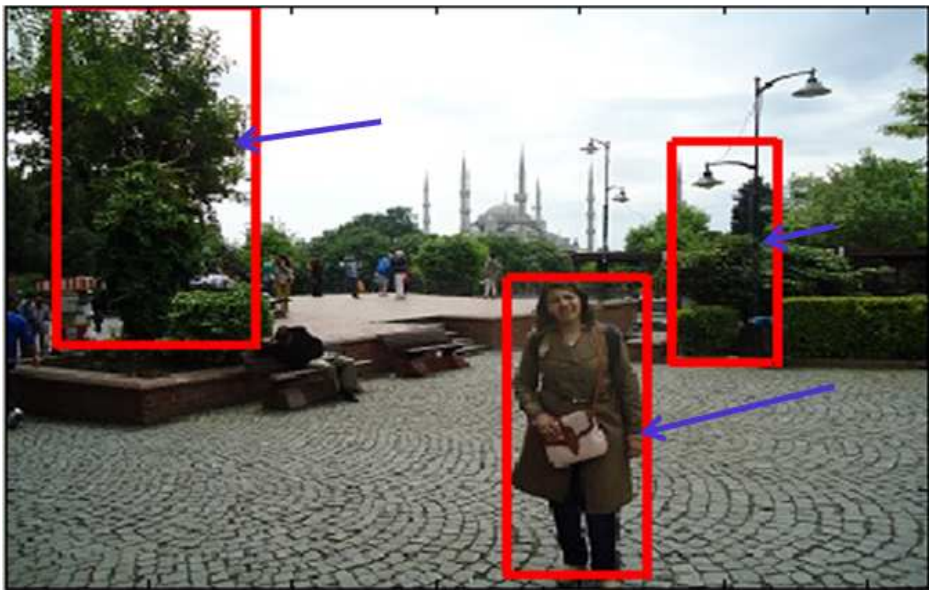


Figure 7-24 Résultats concurrents pour la détection de la personne

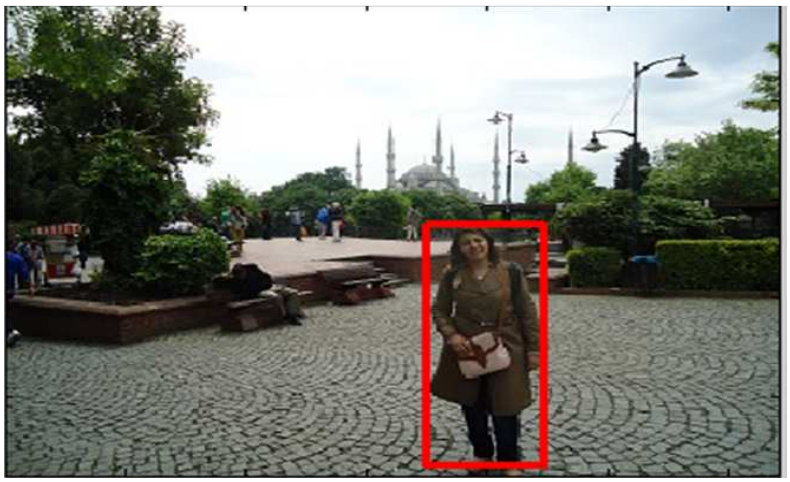


Figure 7-25 Application de l'algorithme de suppression du non-maxima

Un autre cas de figure, est celui quand l'image d'entrée ne contient pas d'objet cible,càd , une personne, notre logiciel est capable de déceler ceci et d'avertir l'utilisateur de l'interface , tel que illustré sur l'image 6-11.

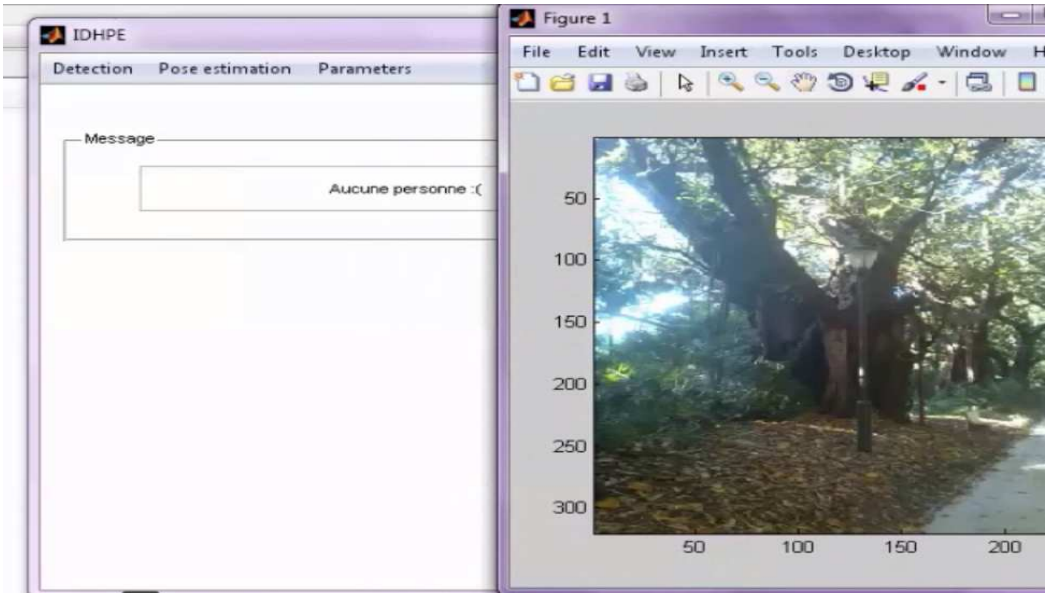


Figure 7-26Absence de personne sur une image

7.5 Estimation de la pose humaine

Pour l'estimation de la pose humaine, nous avons opté pour deux approches différentes, la première , comme décrit précédemment , non basée modèle exploitant principalement les algorithmes de traitement d'images tel que l'algorithme de squelettisation, Harris et les données sur la morphologie humaine La deuxième basée modèle, une approche discriminatoire.

Pour la première approche, nous avons limité nos propos pour la localisation des membres suivants : Tête; torse et jambes (ne considérant qu'une seule configuration possible , tel que les jambes écartées) et ceci est motivé par le choix des membres moins soumis au problème de l'auto occlusion.

Généralisant cette approche sur l'ensemble des parties du corps cette méthode montre ses limites, d'où notre recours à une approche discriminatoire basé modèle
Dans cette section, nous décrivons principalement notre deuxième approche .
Nous distinguons deux phases pour l'implémentation de notre approche, peuvent être dissociées : Apprentissage et détection.
Dans ce qui suit, nous allons décrire les différentes étapes et démarches entreprises pour chacune des phases.

7.6 Apprentissage

La phase d'apprentissage dépend fortement de l'espace des données d'entrée pour la définition des paramètres du modèle, dans ce sens, nous décrivons ci-dessous, notre base de données ainsi que l'étape de préparation de données.

7.6.1 Description de la base de données d'apprentissage

Nous avons testé nos propos sur la base de données Pascal VOC 2007, ce choix est justifié par le fait qu'elle contient des images du mondes réel, ce choix est justifié par le challenge lancé par nos propos: Surmonter les problèmes rencontrés dans le domaine de la reconnaissance d'objet.
A partir de cette base de données, nous avons composé une base de données personnalisée composée de 305 images positives dont 100 sont exploitées durant la phase d'apprentissage et les autres pour la phase de test ainsi que 100 images négatives, prise aléatoirement, ne contenant pas d'humains.
Aucune contrainte n'a été imposée sur le choix des images, ces dernières ont été recueillies à partir des bases de données antérieures des personnalités sportives et des photos personnelles.



.

Figure 7-27Exemples des images positives exploitées durant la phase d'apprentissage


Figure 7-28 Exemples des images négatives exploitées durant la phase d'apprentissage.

Les images exploitées ont été mises à l'échelle pour contenir des personnes couvrant d'environ 150 pixels de hauteur, de taille de 320x240 pixel environ.

L'apprentissage de modèle pour les parties du corps humain, nécessite comme entrées des images labélisées, où les articulations de l'objet personne doivent être localisées. A cet effet, nous avons fait recours à un processus de labellisation manuelle visant à extraire les articulations nécessaires pour la phase de test , telle que décrit dans l'image ci-dessous. Les autres labelles 12 représentent le point

couvrant le milieu du segment reliant une articulation avec la suivante, comme décrit précédemment le recours à ces points intermédiaires est dans le but de parer aux problèmes de variation de pose, se retrouvant ainsi avec 26 parties, au total.

Pour le processus de labélisation, nous avons fait recours au logiciel Gimp [98], les emplacements des articulations pour chaque image sont enregistrés dans un fichier Excel, converti par la suite sous format (.mat) pour être exploité par Matlab.
Notons que la base d'apprentissage Pascal VOC manipule des boites englobantes, ce qui revient à dire aucune information sur l'emplacement exactes des parties du corps , ce qui justifie le choix de l'utilisation du LSVM.

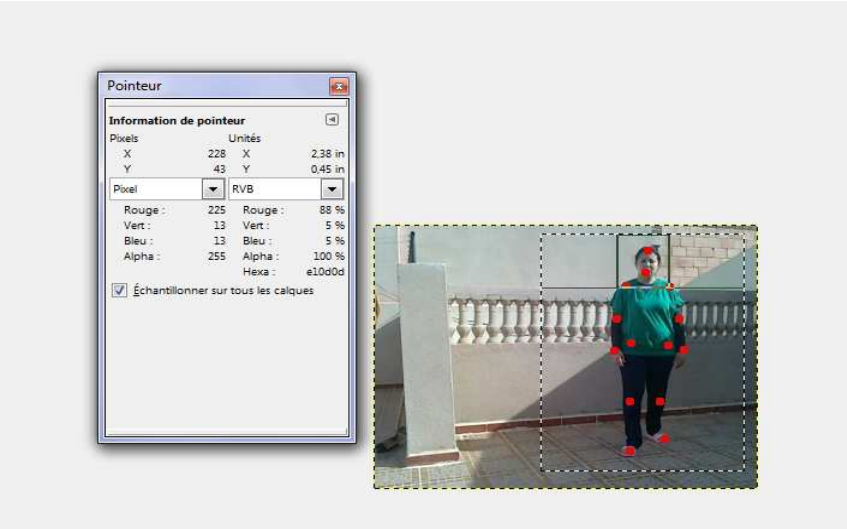


Figure 7-29 L'abellisation des articulations sur une image

N°	Désignation de la partie
1	Haut de la tête
2	Bas de la tête
3	Épaule droit
4	Coude droit
5	Poignet droit
6	Épaule gauche
7	Coude gauche
8	poignet gauche
9	Hanche droite
10	Genou doit
11	Cheville droite
12	Hanche gauche
13	Genou gauche
14	Cheville gauche

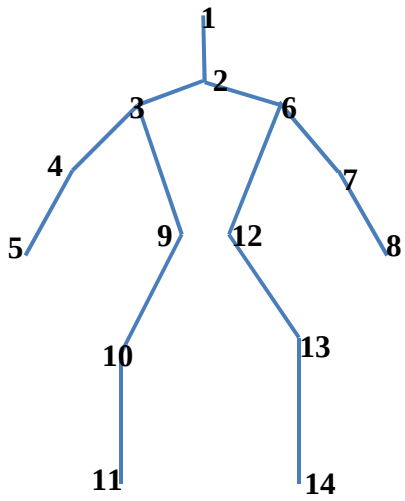


Tableau 7-1Tableau récapitulant les parties du corps humain prises en compte durant la phase d'annotation

(a) (b) (c) (d)

Figure 7-30 Exemple des images avec rotation : (a) rotation droite 15°, (b) rotation gauche 15°, (c) rotation gauche 7°, (d) rotation droite 7°

7.6.2 Conversion format point en format boites englobantes

Notre approche adoptées exploite le concept des boites englobantes, ce concept est largement exploités pour parer aux problèmes d'occlusion, nous aurons à faire à une région plutôt qu'un point particulier. Ci-dessous l'organigramme de mise en place de cette approche, permettant de convertir les points relatifs aux articulations en une région représentant des boites englobantes.

St : Vecteur définissant la structure arborescente du modèle (père_fils) : St={1,2,3,4,5,6,7,8,9,10,11,12,13,14}

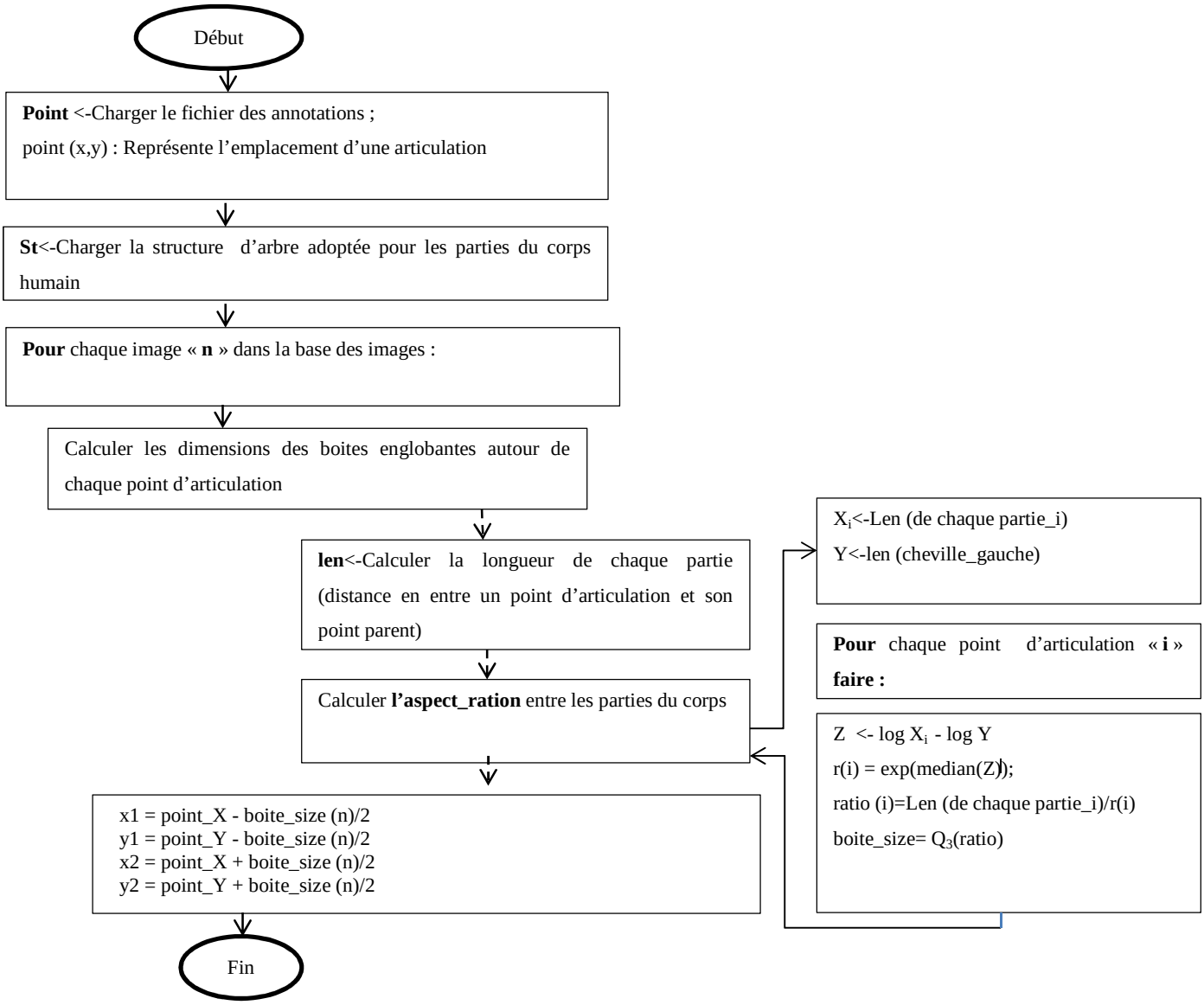


Diagramme 7-4Conversion de format

Figure 7-31 Exemples de découpages en boite englobantes

7.6.3 Paramètres du modèle

Paramètres	**Désignation**
K	Nombre de mélange pour chaque partie
Cellule	Taille de la cellule
Inter	Nombre d'octave pour chaque niveau
Taille_max	Taille maximale d'un filtre HOG
Def	Les coefficients de déformation
Filtres	Chaque filtre contient deux paramètres : e. Le poids : Contenant les poids des filtres appris. f. Taille : taille du filtre g. Le filtre parent
St	Structure arborescente du modèle, relation père –fils
T	Seuil de détection

7.6.4 Processus d'apprentissage

Notre approche d'apprentissage est fondée sur le concept des variables latentes, une étape clé pour le succès de l'approche des sous-catégories latentes est de générer une bonne initialisation des sous-catégories. Notre méthode d'initialisation est de rassembler tous les cas positifs à un espace commun $\phi(.)$, et d'effectuer une classification non supervisée dans cet espace. Dans notre expérience, nous faisons recours à l'algorithme de clustering par Kmeans exploitant ainsi la fonction de la distance euclidienne afin d'assurer une bonne initialisation.

La phase d'apprentissage est procédée selon trois phases.

- Phase 1 : Permet d'apprendre les paramètres du filtre racine pour chaque composant séparément.
- Phase 2 : Concaténer les paramètres de tous les composants et apprend le modèle de mélange des filtre de racine.
- Phase 3 : combine les parties et les paramètres de déplacement pour chaque composant et apprend le modèle de mélange final.

Les phases 2 et 3 exploitent l'algorithme des coordonnées descentes, tel que décrit dans le chapitre 5,

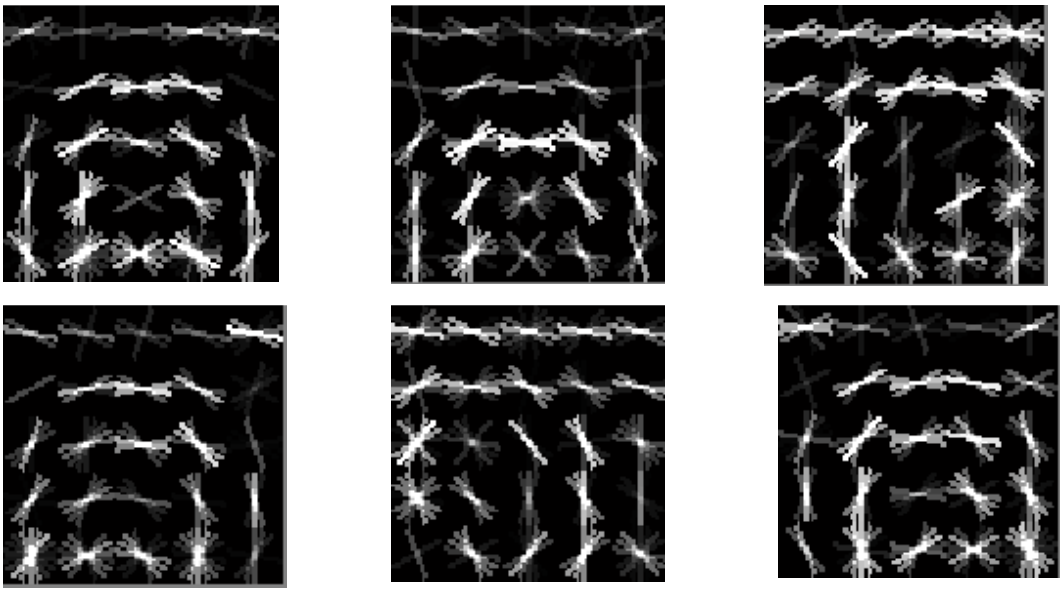


Figure 7-32 Exemple de modèle de tête appris (un pour les six clusters)

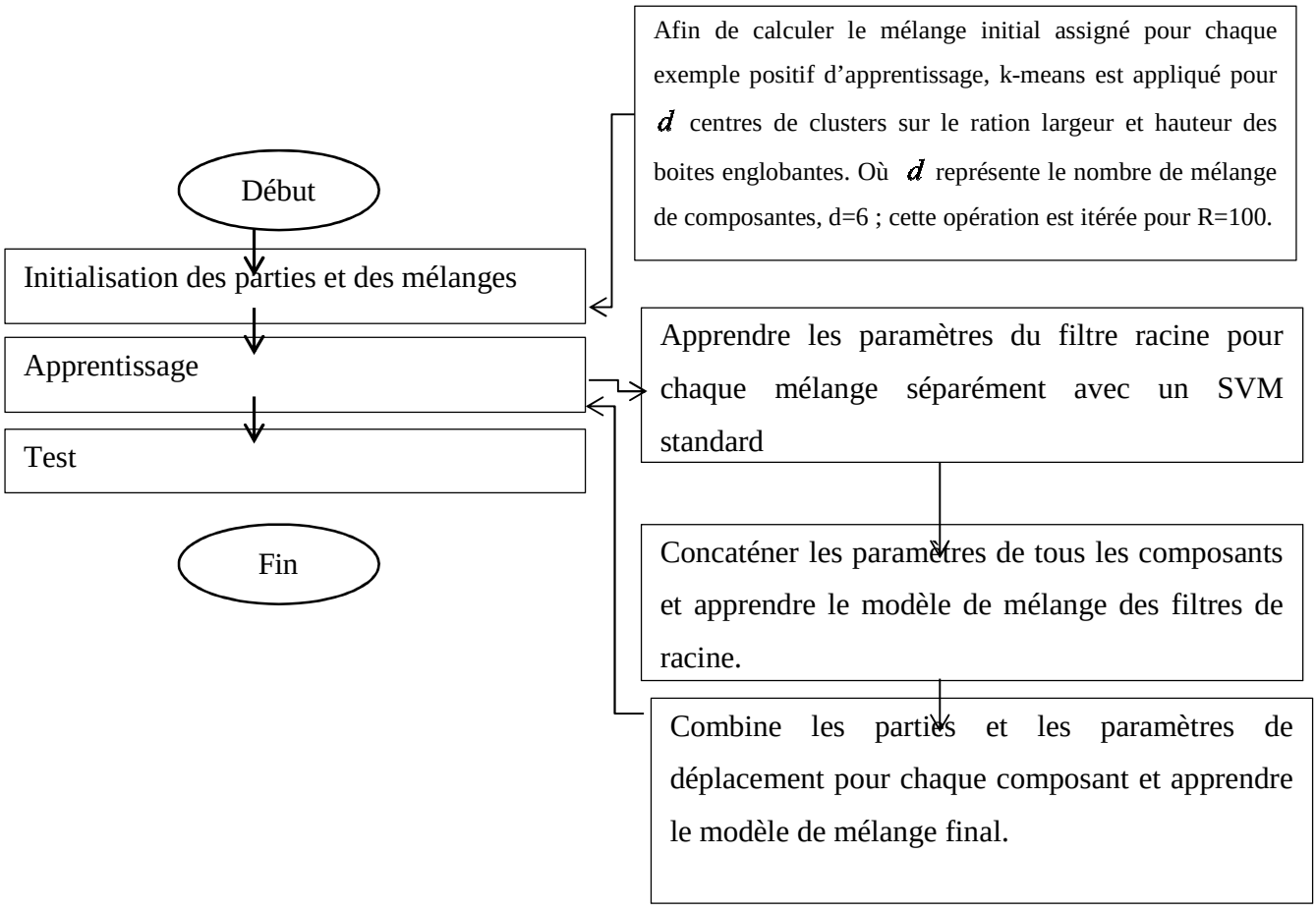


Diagramme 7-5Phases d'apprentissage de données

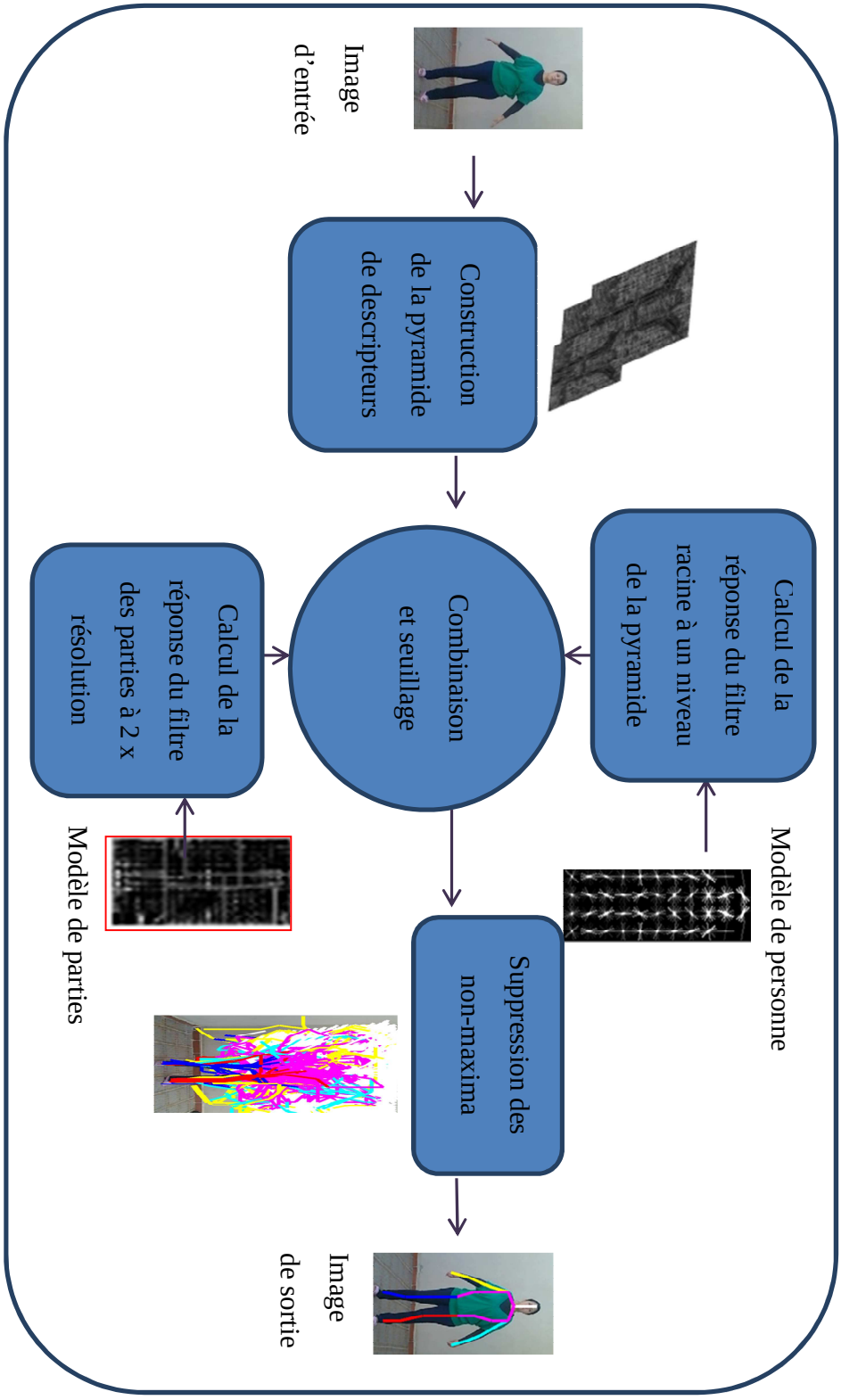


Diagramme 7-6Processus de détection

7.7 Détection

7.7.1 Calcul de pyramide de caractéristiques

Dans un souci de minimiser l'espace mémoire consommé. Nous ne construisons pas tous les niveaux de la pyramide à l'avance. Nous exploitons l'équation ci-dessous, afin de fixer le nombre de niveau nécessaire pour la construction de la pyramide :

$Echelle\_Max = 1 + \left\lfloor \log\left(\frac{\min(taille(image))}{5.\text{model.sbin}} \right) \right\rfloor / \log(sc)$	

Sachant que $sc = 2^{o^{-1}}$, $o = 10$

En outre, on peut calculer la façon dont la dimension d'un niveau de pyramide se développe à travers l'ensemble des calculs en tant que multiples étapes de remplissage sont appliqués à des résultats intermédiaires : Cela permet de créer un score de pyramide pour chaque sous-modèle à l'avance, créer un seul niveau de pyramide de descripteurs et traiter ce niveau au point où les réponses de transformées de filtre sont ajoutées sur les scores de pyramides correspondantes.

Une fois terminé, le niveau de la pyramide en question pour être libéré.

Les scores pyramides pour chaque sous-modèle sont de petites unités que nous devons garder en mémoire tout le temps. Nous pouvons supprimer les emplacements et les tailles des instances de l'objet de leur part. Après le calcul de la hauteur des pyramides et la taille optimale de chaque niveau.

Nous générons ensuite de manière itérative un niveau de la pyramide de descripteurs par un sous-échantillonnage de l'image et de calculer les caractéristiques.

7.7.2 Localisation de l'objet

Pour trouver la position et la dimension d'une instance d'objet dans une image, les scores de tous les sous-modèles sont combinés à une pyramide de score unique. C'est assez simple: en balayant sur chaque niveau de chaque pyramide de score de sous-modèle, nous avons une pyramide avec des scores de détection de l'ensemble du modèle de différents tailles d'objets. Nous pouvons maintenant analyser cette pyramide et enregistrer uniquement des résultats de haut score. Un résultat avec un score élevé défini comme ayant un score supérieur à un certain seuil. Tous les résultats de score élevés sont écrits dans un tableau spécial avec le numéro du niveau.

La position et le niveau permettent de retrouver l'objet en question dans l'image d'origine en ayant l'échelle du niveau. Par contre, si nous définissons la position d'un objet par sa boite englobante, nous obtenant uniquement le coin haut-gauche de sa boite englobante.

7.7.3 Post Traitement : Suppression des non-maxima

La dernière étape dans notre démarche est l'étape « non maxima suppression », c'est un post traitement pour améliorer les détections pour supprimer les doubles détections des instances de l'objet. Toutes les détections sont classifiées selon leur score. Les détections qui se chevauchent avec plus de 50% avec une ultérieure boîte de score sont rejetées.

Le critère de chevauchement mis en œuvre divise l'intersection des deux boîtes à travers la dimension de la boîte avec le score mineur. Si le coefficient est supérieur à 0,5, la boîte est supprimée.

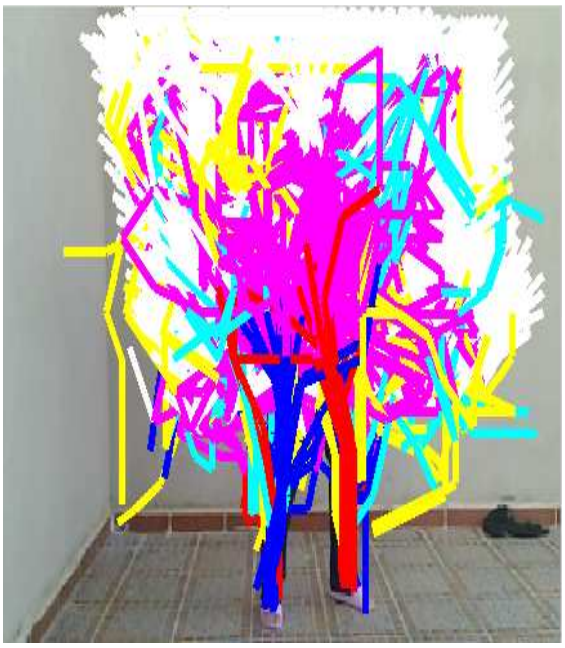


Figure 7-33 Exemple de détection sans l'étpe de non maxima supression

7.8 Interface du logiciel

Comme décrit précédemment, notre interface développée sous Matlab 7.10 donne accès à deux modules : détection ; Estimation de la pose humaine, est ceci su deux cas de figures : Le statique pour une seules image, ainsi que pour le cas dynamique, dans le cas d'une séquence d'images.

La figure 6-22 illustre notre interf*ace* de logiciel

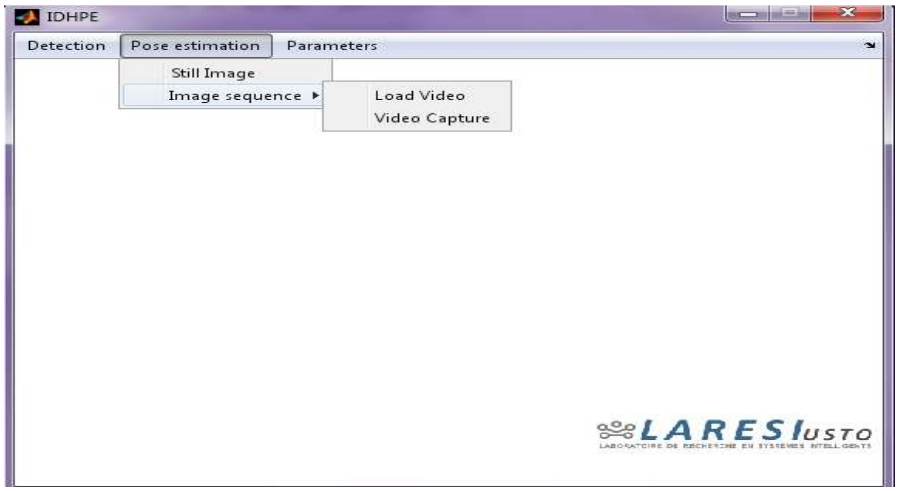


Figure 7-34 Module d'estimation de la pose humaine

7.8.1 Évaluation

Pour l'évaluation de nos travaux, nous avons fait recours à la base d'apprentissage de Parse-set [119], cette dernière est fournie avec les coordonnées des quatorze (14) articulations qui représentent les valeurs relatives à la réalité du terrain (ground truth) .

Après avoir appliqué notre algorithme de détection, nous calculons le PCK qui représente le pourcentage des points correctement labélisés, tel que spécifié sur le diagramme 6-7 où **IS**: Membres inférieurs parties supérieures, **II**: Membres inférieurs parties inférieures, **SS**: Membres supérieures parties supérieures, **SI**: Membres supérieures parties inférieures.

Ainsi la figure 7-35 illustre une étude comparative entre nos résultats obtenus et les résultats des travaux d'Andriluka et Johnson, à travers cette comparaison on remarque bel et bien que notre approche offre de meilleures résultats aussi, le pourcentage le plus faible pour les trois approches est celui des membres supérieures et ceci est justifié par le faite que se sont les membres les plus soumis aux problèmes d'occlusion.

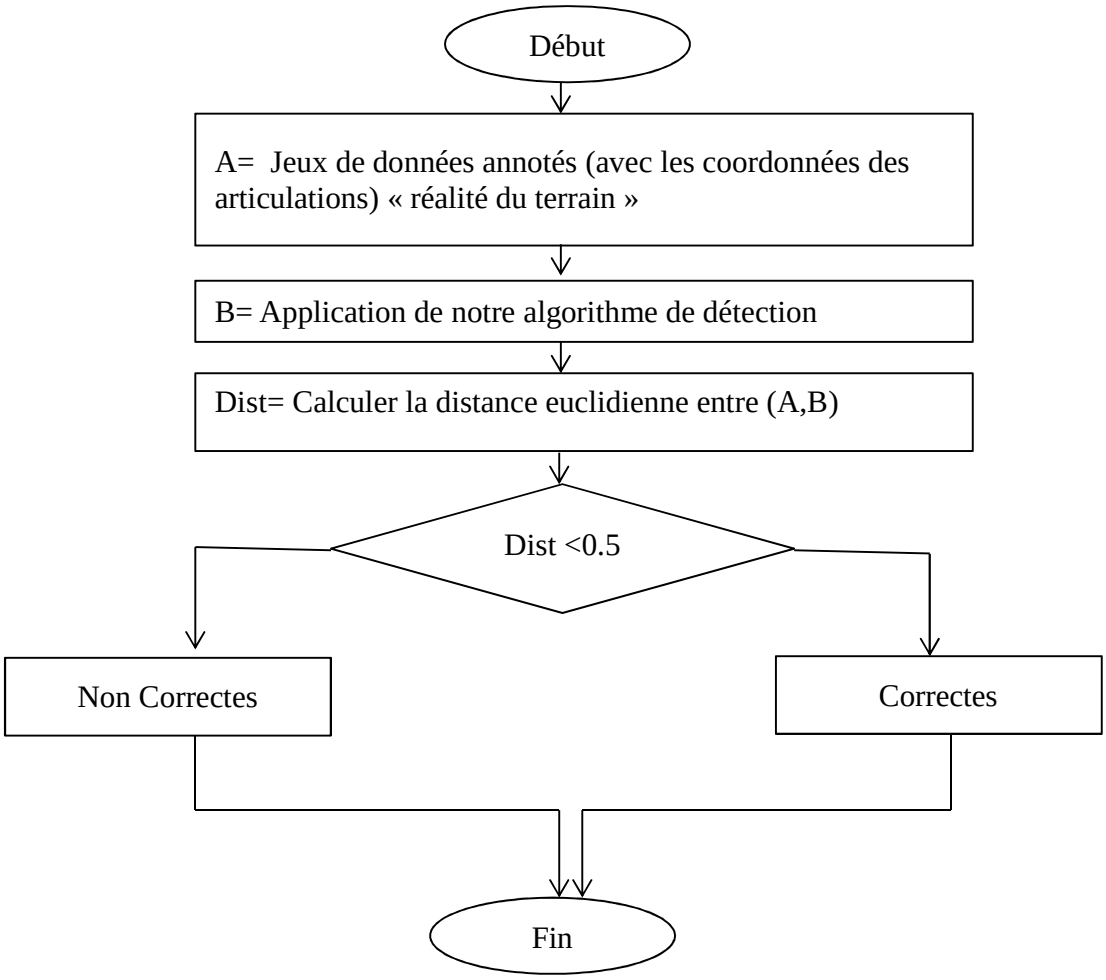


Diagramme 7-7 Calcul des points correctement labellisés

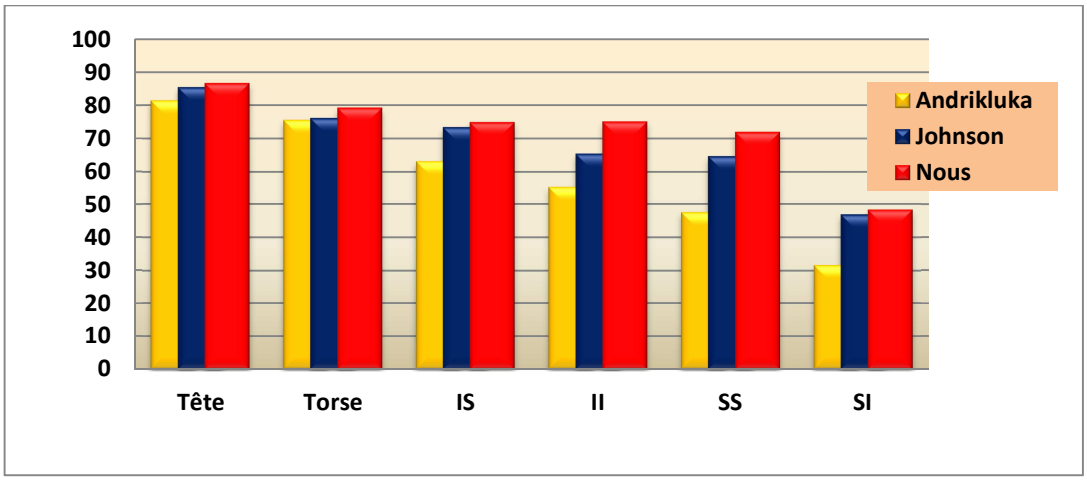


Figure 7-36 Pourcentage des points corrects

Aussi nous faisons recours à une autre métrique qui est le temps de détection moyens , en considérons les points cités ci-dessous , nous avons obtenu un temps de calcul moyen de trois (03) secondes :

- Taille de l'image : 320x240
- Taille de cellule pour HOG : 8
- Nombre d'intervalle : 10.

Étant un temps de calcul considérable, nous avons fait recours à la fonction « **Profiler** » de Matlab pour analyser le temps d'exécution de chaque fonction, nous avons remarqué que le temps d'exécution le plus élevé est relatif au balayage de la pyramide de descripteurs (~1.79s). A cet effet, nous avons modifié nos contraintes initiales comme suit :

- Taille de l'image : 212x130
- Taille de cellule pour HOG : 8
- Nombre d'intervalle : 2.

Les critères cités ci-dessus, nous ont permis de réduire le temps d'exécution moyen à 0.5s.

7.8.2 Résultats et expérimentations

La figure 6-37 illustre quelques résultats sur nos images de test, nous avons choisi des images avec un arrière-plan simple et complexe, des poses variées, et des apparences variées.

Figure 7-37 Quelques résultats (différentes poses)

7.8.3 Conclusion

Ce chapitre consiste en une concrétisation de notre approche d'estimation de la pose humaine, les étapes de réalisation ainsi l'évaluation de nos propos.

Notre travail a été sanctionné par la mise en œuvre du logiciel IDHPE pour « Interface Design for Human Pose estimation » : une application permettant de localiser les différentes parties du corps humain , nous considérons dans notre travail les parties suivantes : Tête, bras gauche (parties supérieure et inférieure),, bras droit (parties supérieure et inférieure), torse, jambes gauches (parties supérieure et inférieure), et jambes droites (parties supérieure et inférieure),

Nous avons fait recours à un modèle déformable pour épouser la forme articulé du corps humain, codifiant l'apparence via l'algorithme de l'histogramme de gradient orienté.
Suite aux différents tests, sur des images choisies aléatoirement, nous approuvons que l'approche proposée offre de bons résultats par rapport à l'estimation de la pose humaine sous différentes contraintes imposées lors du test : des contraintes liées à l'apparence, pose et scène.Néanmoins dans nos travaux futures , nous espérons améliorer le temps d'execution de notre approche par le recours aux nouvelles technologies de calcul , nous avons remarqué que l'opération la plus couteûse en terme de temps de calcul est le nombre important

Conclusion générale

Dans cette thèse , nous avons lancé une mise œuvre d'un système d'estimation de la pose humaine , à travers nos travaux précedents , nous avons remarqué que les approches basées seulement sur l'information de mouvement pour l'estimation de la pose huamine sont limitées par le mouvement d'objet lui-même, les parties statiques de l'objet sont omises lors de la détéction par mouvement, confusion dûe à la dynamique de l'objet. Aussi, les algorithmes de traitement d'images sont limités par la nature du corps humain et les condition d'acquisition d'images.
Par ailleurs, nous avons fixés les objectives suivants : Estimer la pose humaine quelle que soit la configuration spatialle des parties du corps humain, les conditions d'acquisition de la scène (indoor,outdoor) ainsi que la taille de la personne.

Pour répondre à ce défi , nous avons opté pour une approche descriminatoire basé modèle déformable épousant ainsi la forme articulé du corps humain , des modèles codifiant à la fois l'information de l'apparence par l'utilisation de l'algorithme de l'histogramme de gradient orienté et l'information de la configuration des parties par l'adoption d'un modèle à structure d'arbre codifiant ainsi une relation hiérarechqiue entre les parties du corps humain.

Aussi , un autre problème qui surgi lors d'un processus d'estimation de la pose humaine est celui de la mise à l'échelle de l'objet ainsi que ses composantes, ceci a été réglé par la construction d'une pyramide de descripteurs, le processus de localisation est réalisé via l'approche de la fenêtre coulissante.

Nous avons révélé dans notre chapitre 2 « Motivation et Challenge » le problème de la variation intra-classe, ceci a été résolu par une classification via l'approche de kmeans , par la construction de sous catégories pour chaque classe de parties et

l'apprentissage d'un filtre racine , représenant l'hallure général des personnes via un SVM Linéaire.
Pour l'apprentissage des parties du corps humain , nous considérons le problème d'apprentissage de modèle à partir des images labélisées avec des boîtes englobantes autour des objets d'intérêt. Notez que ceci est une labélisation faible « **weakly labeled** » , vue que les boîtes englobantes ne précisent pas les étiquettes des composants ou les emplacement des pièces.

L'apprentissage des paramètres se fait par SVM latente. Une stratégie intéressante, introduite par Felzenszwalb et al **[47]**. Connue par « Exploration des données négatives durs », Comme nous apprenons le classificateur, nous l'appliquons à des exemples négatifs, à la recherche de ceux qui obtiennent une réponse forte; ceux-ci sont mis en cache, et utilisés dans le prochain cycle d'apprentissage. Si cela est bien fait, on peut garantir que le classificateur a les mêmes vecteurs supports, qu'il aurait pu avoir si on l'aurait appliqué sur tous les exemples négatifs.

Nous exploitons pour nos travaux l'histogramme de gradient orienté comme descripteur d'apparence pour les parties du corps humain et le filtre racine ; Ces descripteurs ont l'exclusivité de mieux représenter la structure interne d'un objet via l'information du gradient, permettant ainsi de surmonter les problèmes liés à l'apparence de l'objet : pose, éclairage, occlusion, texture de fond, etc.

Nous validons nos propos par la mise en place d'une application « IDHPE : **I**nterface **D**esign for **H**uman **P**ose **E**stimation » : Une application permettant de localiser les différentes parties du corps humain , nous considérons dans notre travail les parties suivantes : Tête, bras gauche (parties supérieure et inférieure),, bras droit (parties supérieure et inférieure), torse, jambes gauches (parties supérieure et inférieure), et jambes droites (parties supérieure et inférieure).

Bibliographie

[1] T. Moeslund, A. Hilton, and V. Kr¨uger, "A survey of advances in vision-based human motion capture and analysis," *CVIU*, vol. 104, no. 2, pp. 90–126, 2006.

[2] D. Marr and L. Vaina, "Representation and recognition of the movements of shapes," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 214, no. 1197, pp.501–524, 1982.

[3] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari,"Articulated human pose estimation and search in (almost) unconstrained still images," *Technical Report No 272*, no.272, 2010.

[4] D. Gowsikhaa, S. Abirami, and R. Baskaran, "Automated human behavior analysis from surveillance videos: a survey,"*Artificial Intelligence Review*, pp. 1–19, 2012.

[5] P. Doll´ar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *PAMI*, no. 99, pp. 1–1, 2011.

[6] V. Singh and R. Nevatia, "Action recognition in cluttered dynamic scenes using pose-specific part models," in *ICCV*. IEEE, pp. 113–120.2011.

[7] E. Seemann, K. Nickel, and R. Stiefelhagen, "Head pose estimation using stereo vision for human-robot interaction," in *Automatic Face and Gesture Recognition*. IEEE, pp. 626–631.2004.

[8] K. Nickel and R. Stiefelhagen, "Visual recognition of pointing gestures for human-robot interaction," *Image and Vision Computing*, vol. 25, no. 12, pp. 1875–1884, 2007.

[9] S. Escalera, "Human behavior analysis from depth maps," in *AMDO*, 2012.

[10] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3d pose estimation and tracking by detection," in *CVPR*. IEEE, pp. 623–630.2010.

[11] A. Agarwal and B. Triggs, "Recovering 3d human pose from monocular images," *PAMI*, vol. 28, no. 1, pp. 44–58, 2006.

[12] G. Rogez, C. Orrite-Uruuela, and J. Martnez-del Rincn, "A spatio-temporal 2d-models framework for human pose recovery in monocular sequences," *PR*, vol. 41, no. 9, pp.2926 – 2944, 2008.

[13] T. Acharya and A. K. Ray *Image processing*: *principles and applications*, New Jersey: Wiley-Interscience.2005

[14] M. Shah, O. Javed and K. Shafique .Automated visual surveillance in realistic scenarios.*IEEE MultiMedia*, vol.14, no.1, pp.30-39.2007

[15] W. M. Hu, T. N. Tan, L. Wang and S. Maybank .A survey on visual surveillance of object motion and behaviors.*IEEE Transactions on Systems, Man and Cybernetics*, vol.34, no.3, pp. 334-352. 2004

[16] D. A. Forsyth and J. Ponce. *Computer vision: a modern approach*, Prentice Hall. 2003

[17] P. Perez, C. Hue, J. Vermaak and M. Gagnet .Color-Based Probabilistic Tracking.*In Proceedings of ECCV*, pp. 661–675.2002.

[18] K. Pahlavan and J.O. Eklundh .A head-eye system- analysis and design.*CVGIP: Image Understanding*, vol. 56, pp. 41–56.1992

[19] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfinder: real-time tracking of the human body. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7):780–785, 1997.

[20] C. Beleznai, B. Fruhstuck, and H. Bischof.Human detection in groups using a fast mean shift procedure.International Conference on Image Processing, 1:349–352, 2004.

[21] T. Haga, K. Sumi, and Y. Yagi.Human detection in outdoor scene using spatio-temporal motion analysis. International Conference on Pattern Recognition, 4:331–334, 2004.

[22] H. Eng, J. Wang, A. Kam, and W. Yau.A bayesian framework for robust human detection and occlusion handling using a human shape model.International Conference on Pattern Recognition, 2004.

[23] H. Elzein, S. Lakshmanan, and P. Watta.A motion and shapebased pedestrian detection algorithm. IEEE Intelligent Vehicles Symposium, pages 500–504, 2003.

[24] D. Toth and T. Aach. Detection and recognition of moving objects using statistical motion detection and fourier descriptors. International Conference on Image Analysis and Processing, pages 430–435, 2003.

[25] D. J. Lee, P. Zhan, A. Thomas, and R. Schoenberger. Shape-based human intrusion detection. SPIE International Symposium on Defense and Security, Visual Information Processing XIII, 5438:81–91, 2004.

[26] J. Zhou and J.Hoang.Real time robust human detection and tracking system.IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 3:149 – 149, 2005.

[27] S.Yoon and H. Kim. Real-time multiple people detection using skin color, motion and appearance information. International Workshop on Robot and Human Interactive Communication,pages 331–334, 2004.

[28] F.Xu and K. Fujimura. Human detection using depth and gray images. EEE Conference on Advanced Video and Signal Based Surveillance, pages 115–121, 2003.

[29] L. Li, S.Ge, T. Sim, Y.Koh, and X. Hunag. Object-oriented scale-adaptive filtering for human detection from stereo images. IEEE Conference on Cybernetics and Intelligent Systems, 1:135–140, 2004.

[30] J.Han and B. Bhanu. Detecting moving humans using color and infrared video. IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, 30:228–233, 2003.

[31] L.Jiang, F.Tian, L.Shen, Shiqian Wu, S.Yao, Z. Lu, and L.Xu. Perceptual-based fusion of ir and visual images for human detection. International Symposium on Intelligent Multimedia, Video and Speech Processing, pages 514– 517, 2004.

[32] R. Cutler and L. S. Davis. Robust real-time periodic motion detection,
analysis, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8):781–796, 2000.

[33] A.Utsumi and N.Tetsutani. Human detection using geometrical pixel value structures. Fifth IEEE International Conference on Automatic Face and Gesture Recognition, page 39, 2002.

[34] D. M. Gavrila and J. Giebel. Shape-based pedestrian detection and tracking. IEEE Intelligent Vehicle Symposium, 1:8–14, 2002.

[35] P. Viola, M. J. Jones, and D. Snow.Detecting pedestrians using patterns of motion and appearance. IEEE International Conference on Computer Vision, 2:734–741, 2003.

[36] H. Sidenbladh. Detecting human motion with support vector machines. Proceedings of the 17th International Conference on Pattern Recognition, 2:188–191, 2004.

[37] N.Dalal and B.Triggs. Histograms of oriented gradients for human detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 1063–6919, 2005.

[38] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, and M. Finocchio, "Real-time human pose recognition in parts from single depth images," 2011.

[39] A. Hern´andez, M. Reyes, S. Escalera, and P. Radeva, "Spatiotemporal grabcut human segmentation for face and pose recovery," in *CCVPR Workshops*. IEEE, pp. 33–40.2010.

[40] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3d pose estimation and tracking by detection," in *CVPR*. IEEE,pp. 623–630.2010.

[41] M. Salzmann, F. Moreno-Noguer, V. Lepetit, and P. Fua, "Closed-form solution to non-rigid 3d surface registration,"in *ECCV*, pp. 581–594.2008.

[42] E. Simo-Serra, A. Ramisa, G. Alenya, C. Torras, and F. Moreno-Noguer, "Single image 3d human pose estimation from noisy observations," in *CVPR*. IEEE, 2012.

[43] F. Moreno-Noguer, V. Lepetit, and P. Fua, "Pose priors for simultaneously solving alignment and correspondence," in *ECCV*. Springer-Verlag, pp. 405–418.2008

[44] J. S´anchez-Riera, J. Ostlund, P. Fua, and F. Moreno-Noguer, "Simultaneous pose, correspondence and non-rigid shape," in *CVPR*. IEEE, 2010, pp. 1189–1196.

[45] M. Fischler and R. Elschlager, "The representation and matching of pictorial structures," *Computers, Transactions on*, vol. 100, no. 1, pp. 67–92, 1973.

[46] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade, "Trajectory space: A dual representation for nonrigid structure from motion," *PAMI*, vol. 33, no. 7, pp. 1442–1456, 2011.

[47] P. Felzenszwalb and D. McAllester, "Object detection grammars,"University of Chicago, Tech. Rep., computer Science TR.2010

[48] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained partbased models," *PAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[49] Y. Wang, D. Tran, and Z. Liao, "Learning hierarchical poselets for human parsing," in *CVPR*. IEEE, pp. 1705–1712.2011.

[50] R. Girshick, P. Felzenszwalb, and D. McAllester, "Object detection with grammar models," *PAMI*, vol. 33, no. 12, 2011.

[51] L. D. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *ICCV*, pp. 1365–1372.2009

[52] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard, "Tracking loose-limbed people," in *CVPR*, vol. 1.IEEE, pp.I–421.2004.

[53] L. Sigal, M. Isard, H. Haussecker, and M. Black, "Looselimbed people: Estimating 3d human pose and motion using non-parametric belief propagation," *International journal of computer vision*, pp. 1–34, 2012.

[54] Y. Chen, L. Zhu, C. Lin, A. Yuille, and H. Zhang, "Rapid inference on a novel and/or graph for object detection, segmentation and parsing," *NIPS*, vol. 20, pp. 289–296, 2007.

[55] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *CVPR*. IEEE, pp. 1385–1392.2011

[56] I. Karaulova, P. Hall, and A. Marshall, "A hierarchical model of dynamics for tracking people with a single video camera,"in *British Machine Vision Conference*, vol. 1, pp. 352–361.2000.

[57] A. Agarwal and B. Triggs, "Tracking articulated motion with piecewise learned dynamical models," in *ECCV*, vol. 3, pp. 54–65.2004

[58] R. Urtasun, D. Fleet, and P. Fua, "Temporal motion models for monocular and multiview 3d human body tracking,"*CVIU*, vol. 104, no. 2, pp. 157–177, 2006.

[59] F. Moreno-Noguer and J. Porta, "Probabilistic simultaneous pose and non-rigid shape recovery," in *CVPR*. IEEE, pp. 1289–1296.2011

[60] R. Urtasun, D. Fleet, A. Hertzmann, and P. Fua, "Priors for people tracking from small training sets," in *ICCV*, vol. 1.IEEE, pp. 403–410.2005

[61] A. Fossati, M. Salzmann, and P. Fua, "Observable subspaces for 3d human motion recovery," in *CVPR*.Ieee, pp. 1137–1144.2009.

[62] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade, "Non rigid structure from motion in trajectory space," in *NIPS*, pp. 41–48.2008.

[63] H. Park, T. Shiratori, I. Matthews, and Y. Sheikh, "3d reconstruction of a moving point from a series of 2d projections," *Computer Vision–ECCV 2010*, pp. 158–171, 2010.

[64] L. Sigal and M. Black, "Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion," Brown Univertsity, Tech. Rep., brown Univertsity TR.2006.

[65] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *CVPR*. IEEE, pp. 17–24.2010.

[66] M. Andriluka and L. Sigal, "Human context: Modeling human-human interactions for monocular 3d pose estimation," *AMDO*, pp. 260–272, 2012.

[67] P. Noriega, Thèse de doctorat " Modèle du corps humain pour le suivi de gestes en monoculaire ", l'université Pierre et Marie Curie – Paris 6.2007.

[68]G.Mori , J.Malik. Estimating human body con_gurations using shape context matching. In ECCV (3), pp 666_680, 2002.

[69] O.Bernier, P.Cheung-Mon-Chang. Real-time 3d articulated pose tracking using particle _ltering and belief propagation on factor graphs. In British Machine Vision Conference, volume 01, pp 005_008, 2006.

[70] J. Roberts, J. McKenna, and W. Ricketts. Human pose estimation using learnt probabilistic region similarities and partial con_gurations. In ECCV (4), pp 291_303, 2004.

[71] R.Plänkers , P.Fua. Articulated soft objects for multiview shape and motion capture.IEEE Trans. Pattern Anal. Mach. Intell., 25(9) :pp1182_1187, 2003.

[72] C.Bregler ,J.Malik. Tracking people with twists and exponential maps.In CVPR, pp 8_15, 1998.

[73] D.Demirdjian, T. Ko, T.Darrell. Constraining human body tracking.In ICCV '03 : Proceedings of the Ninth IEEE International Conference on Computer Vision,pp 1071, Washington, DC, USA, IEEE Computer Society.2003.

[74] R.Urtasun ,P.Fua. 3d Human Body Tracking using Deterministic Temporal Motion Models. Technical report, 2004.

[75] R.Plänkers , P.Fua. Articulated soft objects for multiview shape and motion capture.IEEE Trans. Pattern Anal. Mach. Intell., 25(9) :pp1182_1187, 2003.

[76] D. Demirdjian, L.Taycher, G.Shakhnarovich, K.Grauman,T.Darrell. Avoiding the "streetlight e_ect" : Tracking by exploring likelihood modes. In ICCV, pp 357_364, 2005.

[77] M.W.Lee and I.Cohen. Proposal maps driven mcmc for estimating human body pose in static images. In CVPR (2), pp 334_341, 2004.

[78] T. Gevers ,A. Smeulders. A comparative study of several color models for color image invariant retreival. In Proc. 1st Int. Workshop on Image Databases & Multimedia Search, Amsterdam, Netherlands., pp 17_24, 1996.

[79] J.Gao ,J.Shi. Multiple frame motion inference using belief propagation.In FGR, pp 875_882, 2004.

[80] M. Dimitrijevic, V. Lepetit, P. Fua.Human body pose recognition using spatiotemporaltemplates. In ICCV, 2005.

[81] X.Lan , P. Huttenlocher. A unified spatio-temporal articulated model for tracking.cvpr, 01:pp 722_729, 2004.

[82] R.Navaratnam, A.Thayananthan, H. S. Torr, R.Cipolla. Hierarchical part-based human body pose estimation. In British Machine Vision Conference, volume 00, 2005.

[83] J. La_erty, A. McCallum, F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. of 18th Int. Conf. on Machine Learning, pp 282_289, 2001.

[84] S. Yedidia, T. Freeman, Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. IEEE Trans. on Information Theory, 51(7) :pp2282_2312, July 2005.

[85] Y.Weiss. Correctness of local probability propagation in graphical models with loops. Neural Computation, 12(1) :pp1_41, 2000.

[86]] O.Bernier , P. Cheung-Mon-Chang. Real-time 3d articulated pose tracking using particle filtering and belief propagation on factor graphs. In British Machine Vision Conference, volume 01, pp 005_008, 2006.

[87] L.Sigal, S. Bhatia, S. Roth, J. Black, M.Isard.Tracking loose-limbed people. In CVPR (1), pp 421_428, 2004.

[88] M. Isard. Pampas : Real-valued graphical models for computer vision. In CVPR (1), pp 613_620, 2003.

[89] Y. Wu, G. Hua, T. Yu. Tracking articulated body by dynamic markov network.In ICCV, pp 1094_1101, 2003.

[90] X. Ren C. Berg, J.Malik. Recovering human body configurations using pairwise constraints between parts.In Proc. 10th Int'l. Conf.Computer Vision, volume 1, pp 824_831, 2005.

[91] G.Mori, X.Ren, A. Efros, J.Malik. Recovering human body configurations : Combining segmentation and recognition. In CVPR (2), pp 326_333, 2004.

[92]C.Leignel , E.Viallet. A blackboard architecture for the detection and tracking of a person. In RFIA, 2004.

[93] Timothy J. Roberts, Stephen J. McKenna, and Ian W. Ricketts.Human pose estimation using learnt probabilistic region similarities and partial configurations.In ECCV (4), pp 291_303, 2004.

[94]http://fr.wikipedia.org/wiki/Histogramme_de_gradient_orient%C3%A9

[95] H.SIBT UL. Machine learning methods for visual object detection.Thèse docteur de l'université de Grenoble . 2011

[96] J.Guyomard, Spécifications des images de Cassini pour la reconnaissance automatique de symboles et premiers résultats , rapport, Géopeuple, mars 2012.

[97] http://fr.wikipedia.org/wiki/Bootstrap_(statistiques)

[98]http://www.gimp.org/

[99]svmlight.joachims.org

[100]http://pascal.inrialpes.fr/data/human/

[101] http://sanchom.wordpress.com/2011/09/01/precision-recall/

[102]J.Lafferty, A.McCallum, F.C. N Pereira, , A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In: Proceedings of the IEEE 77 : No. 2, pp. 257-286.1989.

[103] D.John; C. N.Fernando: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In:Proceedingsof the Eighteenth International Conference on Machine Learning (ICML 2001), Morgan Kaufmann Publishers, pp. 282-289.2001

[104]K.Tomanek, J.Wermter, U.Hahn,: A Reappraisal of Sentence and Token Splitting for Life Science Documents. In:.Proceedings of the 12th World Congress on Medical Informatics (MEDINFO 2007).Brisbane, Australia, pp. 524-528.2007.

[105]S. Fei; F.Pereira: Shallow parsing with Conditional Random Fields. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL '03). Morristown, NJ, USA: Association for Computational Linguistics, pp. 134-141.2003.

[106]E.Buyko, K.Tomanek, U.Hahn: Resolution of Coordination Ellipses in Named Entities in Scientific Texts. In: Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING 2007). Melbourne, Australia, pp. 163-171.2007.

[107]B.Settles: Biomedical Named Entity Recognition using Conditional Random Fields and Rich Feature Sets. In: Collier, Nigel; Ruch, Patrick; Nazarenko, Adeline (Editors.): COLING 2004 International Joint workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP) 2004. Geneva, Switzerland: COLING, pp. 107-110.2004.

[108] R.McDonald, F.Pereira: Identifying gene and protein mentions in text using conditional random fields. In: BMC Bioinformatics 6 (Suppl 1) , May, No. S6.2005.

[109]R.Klinger, C.Friedrich,J.Fluck, M.Hofmann-Apitius: Named Entity Recognition with Combinations of Conditional Random Fields. In: Proceedings of the Second BioCreative Challenge Evaluation Workshop. Madrid, Spain, pp. 89-91.2007.

[110] R.McDonald,R. S.Winters, M.Mandel, Jin, Yang; White, Peter S.; Pereira, Fernando: An entity tagger for recognizing acquired genomic variations in cancer literature. In: Bioinformatics 20 ,pp. 3249-3251.2004.

[111]D.DeCaprio, J.Vinson, M.Pearson, P.Montgomery,M.Doherty, J.Galagan,: Gene Prediction using Conditional Random Fields. In: GENOME RESEARCH 17 , No. 9, pp. 1389-1396.2007.

[112]X.He,S. Zemel, M.Carreira-Perpinan,: Multiscale conditional random fields for image labeling. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Vol. 2, pp. 695-702.2004.

[113]A.Quattoni, M.Collins, T.Darrell: Conditional Random Fields for Object Recognition. In: Saul, Lawrence K.; Weiss, Yair; Bottou, Leon (Editors.): Advances in Neural Information Processing Systems 17. Cambridge, MA: MIT Press, pp. 1097-1104.2005.

[114]K.Gupta,B. Nath, K.Ramamohanarao: Conditional Random Fiels for Intrusion Detection. In: AINAW '07: Proceedings of the 21st International Conference on Advanced Information Networking and Applications Workshops. Washington, DC, USA: IEEE Computer Society, pp. 203-208.2007.

[115]X.Zhang, D.Aberdeen,S.Vishwanathan: Conditional random fields for multi-agent reinforcement learning. In: ICML '07: Proceedings of the 24th international conference on Machine learning. New York, NY, USA: ACM, pp. 1143-1150.2007.

[116]L.Rabiner: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In: Proceedings of the IEEE 77 ,No. 2, pp. 257-286. 1989.

[117]T.Jaynes: Information Theory and Statistical Mechanics. In: Physical Review 106 May, No. 4, pp. 620-630. 1957.

[118]A.Korn: Mathematical Handbook for Scientists andEngineers: Definitions, Theorems, and Formulas for Reference and Review. 2 Revised. New York: Dover Publications Inc., 2000.

[119]https://computing.ece.vt.edu/~santol/projects/zsl_via_visual_abstraction/parse/index.html