INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

Reconstruction monoculaire du mouvement humain

et autres travaux, 2000–2004

Synthèse des travaux scientifiques

pour obtenir le grade de

Habilitation à Diriger des Recherches

Spécialité : Imagerie, Vision et Robotique

présentée publiquement par

William Triggs

le 7 janvier 2005, au laboratoire GRAVIR, Montbonnot, France

devant un jury composé de

M. Roger Mohr (INP Grenoble)	président
M. Philippe Cinquin (UJF Grenoble)	rapporteurs
M. Jean Ponce (Beckman Institute & U. Illinois)	
M. Luc Van Gool (KU Leuven & ETH Zürich)	(excusé)
M. Olivier Faugeras (INRIA Sophia-Antipolis)	examinateurs
M. Andrew Zisserman (Université d'Oxford)	

À mes étudiants et confrères.

Table des matières

Avant Propos

Ce document fait le bilan de mes travaux entre janvier 2000 et août 2004, dans les équipes MOVI et LEAR du laboratoire CNRS-INRIA-INPG-UJF GRAVIR à Grenoble, France. Il représente la partie synthétique de mon dossier d'Habilitation à Diriger des Recherches (HDR). Un mémoire associé qui a une structure parallèle, « Human Motion Capture from Monocular Images, and Other Works 2000–2004 », recueille les principaux articles scientifiques associés à ces travaux.

Organisation

Ce rapport est divisé en deux parties thématiques (et non-chronologiques), chacune d'elles comprend trois chapitres techniques qui font le bilan d'un ensemble de publications sur un même thème. Le chapitre correspondant du mémoire associé présente les articles eux mêmes, après un bref résumé en anglais.

La première partie considère le problème de la reconstruction du mouvement humain à partir d'images monoculaires. Elle présente trois axes de recherche complémentaires: le premier bât sur l'approche « classique » avec modèle tridimensionnel (3-D) explicite; le second considère les approches bidimensionnelles (2-D); et le dernier présente une approche 3-D novatrice, basée sur l'apprentissage direct et sans modèle 3-D explicite.

La deuxième partie du document regroupe plus brièvement mes travaux sur d'autres thèmes: (*i*) le traitement d'image et la vision bas niveau; (*ii*) la vision géométrique et la reconstruction de scène; et (*iii*) la modélisation statistique et la reconnaissance des formes.

Deux annexes détaillent mes encadrements et mes activités scientifiques non-liés aux publications, de 2000 à 2004.

Ce rapport met l'accent sur les travaux sur la perception d'humains – le thème le plus soutenu pendant cette période – ainsi que sur les collaborations fructueuses avec mes doctorants Cristian SMINCHISESCU et Ankur AGARWAL. Cependant, le mouvement humain ne représente à lui seul que la moitié de ma production scientifique pendant cette période, et quoique plus sommaire, la deuxième partie du document reste significative.

Sur le plan historique, les travaux sur la vision géométrique, et surtout les travaux algébriques, datent pour la plupart de la période qui suit directement ma thèse; ils ne sont plus le centre de mes activités. Les travaux sur l'apprentissage sont plus récents et prennent progressivement de l'ampleur. L'étude du mouvement humain, et les réflexions sur la vision de bas niveau et sur l'extraction de primitives, ont été poursuivis tout au long de cette période et doivent se maintenir à l'avenir.

Remerciements

Je voudrais remercier d'abord les collègues avec qui j'ai partagé ces petites aventures intellectuelles, et sans qui elles n'auraient jamais pu prendre leur forme actuelle. Je cite non seulement mes doctorants Ankur, Cristi, Guillaume et Marko qui voient dans ces pages leurs propres publications, mais aussi tous mes collègues de MOVI et de LEAR, qui ont contribué à créer un environnement de travail exceptionnel. Plus directement, je remercie toute particulièrement les quelques âmes patientes – parmi lesquels Roger, Cordelia et Bernard C. figurent en bonne place – qui ont toujours pris la peine de me taquiner à des moments bien choisis à fin que je fasse (enfin) le nécessaire; et également ma famille, de leur patience face à mes lubies et mes silences.

Je remercie également Roger Mohr, Jean Ponce, Michaël Sdika et Hélène Ratiney, pour avoir relu ce document afin de corriger mes fautes de français.

Les travaux ici présentés ont reçu le support généreux de plusieurs sources de financement, et notamment des projets européens CUMULI, VIBES, LAVA, ACEMEDIA et PASCAL. Un poste de Chargé de Recherche au CNRS m'a permis d'exercer le métier de chercheur pendant cette période.

Première partie

Capture du mouvement humain

Chapitre 1

Introduction à la capture de mouvement

La première partie de ce rapport développe dans trois chapitres techniques le contenu d'une série d'articles consacrés à l'estimation de la pose et du mouvement articulaire du corps humain, et ce à partir d'une seule image ou d'une séquence monoculaire d'images. Par « pose » on entend la position du corps et de ses membres majeurs (bras, jambes) – paramétrée en général par le vecteur d'angles articulaires d'un modèle du squelette d'environ 30-35 degrés de liberté. Le but ici étant le suivi fin et non pas le rendu réaliste, on se limite à une modélisation plutôt grossière du corps. Le modèle cinématique ne contient que les degrés de liberté indispensables à un suivi correct; on ne modélise ni les détails de la main et du visage, ni la forme détaillée du corps, de la peau, des vêtements. Néanmoins, même sous cette forme simplifiée, le problème de la reconstruction du mouvement humain présente un défi réel, et sa résolution pratique serait d'une utilité conséquente aux applications décrites prochainement.

1.1 Contexte

En anglais, « motion capture » ou « mocap » signifie la reconstruction du mouvement humain articulaire à partir des données sensorielles, d'où « capture du mouvement » en français. L'expression vient de l'industrie des effets spéciaux pour la production de films, de vidéos, de jeux informatiques – industrie qui est actuellement l'utilisateur le plus important et le plus exigeant de ces technologies. Les autres applications incluent: l'anthropométrie du mouvement pour la diagnostique médicale, l'entraînement sportif, et la conception de matériel sportif; et les interfaces homme-machine pour la réalité virtuelle et augmentée.

Actuellement, les systèmes de capture de mouvement les plus performants sont basés sur la vision, ou plus précisément, sur la photogrammétrie – la science de la mesure à partir d'images. La capture du mouvement 3-D multi-caméra « instrumentée » est une technologie mûre, bien commercialisée, et utilisée régulièrement pour les applications précitées. Ici, « instrumentée » veut dire que pour se servir de cette technologie dans sa forme commerciale actuelle, le sujet doit porter un harnais spécial ou des vêtements spéciaux munis des cibles géométriques réfléchissantes, et il faut travailler sous l'illumination stroboscopique dans un espace muni d'un nombre suffisant (en pratique, 4–20) de caméras spécialisées qui ont été soigneusement pre-calibrées et synchronisées. Même ainsi, on est souvent obligé de « nettoyer » à la main les traces enregistrées, avant de pou-

voir s'en servir. Une installation complète coûte dans les 50–300 kEu et est assez encombrante. De plus, en raison de l'illumination stroboscopique requise, la plupart des systèmes actuels interdisent l'acquisition d'images classiques de la scène lors de la capture du mouvement. (Et de toute façon, lorsqu'on veut des images classiques des acteurs, l'obligation de les encombrer de harnais est assez gênant. . .)

Si le système multi-caméra instrumenté a fait ses preuves, il se révèle insuffisant pour un grand nombre d'applications. Côté industrie de production graphique (film, télévision, jeux), les contraintes qu'il impose sont souvent assez gênantes, au sens où il ne s'intègre pas bien au « pipeline » de production naturel. On voudrait travailler en une seule fois, avec les acteurs réels, en costume, sur scène, sous l'illumination naturelle, et sans caméras supplémentaires [1]. Actuellement, pour monter un effet spécial, il faut travailler en deux fois, reproduisant laborieusement les éléments nécessaires de la scène dans la salle de capture de mouvement, et puis retravailler à la main les deux séquences afin de les mettre en correspondance spatiale et temporelle, corriger les occultations, ajuster l'illumination, et ainsi de suite – ce qui est assez coûteux en main d'oeuvre et, point capital, en temps de production. Il serait donc très intéressant de disposer d'un système de capture de mouvement « à travers l'objectif » (« through the lens ») – qui peut estimer le mouvement à partir de la séquence de production, prise sur scène, en costume, sous l'illumination voulue, et par une caméra de production unique. C'est sur ce problème que nous travaillons.

Signalons dès à présent que, s'il a bien d'autres avantages, un tel système « monoculaire non-instrumenté » ne peut pas prétendre à la précision métrique d'un système 3-D multi-caméra instrumentée classique. Sans la base géométrique d'un système stéréo, la triangulation devient impossible – l'estimation de la profondeur (la distance objet - caméra) doit reposer sur des indices plus subtils, et elle se révèle en pratique assez délicate. Aussi, sans cibles réfléchissantes, et sous une illumination quelconque, l'extraction d'indices de l'image et la mise en correspondance deviennent elles aussi nettement plus délicates. Du côté positif, les erreurs en profondeur sont en général peu visibles dans la caméra d'origine: elles sont par définition parfaitement alignées avec les lignes de vision de cette caméra, donc au premier ordre et jusqu'aux effets perspectifs, d'illumination, d'occultation, elles ne modifient pas l'image rendue. Cette observation suggère qu'un système fiable à travers l'objectif serait très utile pour les applications de réalité augmentée, ainsi que pour des applications 3-D de basse précision. Par voie de comparaison, pour la profondeur des membres du sujet humain relative au corps, on peut estimer qu'avec nos méthodes monoculaires non-instrumentées actuelles on arrive rarement en dessous de 5–10 cm de précision, là où un système instrumentée multi-caméra avec une géométrie correcte et une résolution image similaire aurait une erreur bien inférieure à 1 cm. Mais compte tenu de leurs autres atouts, nos travaux dans ce domaine ont suscités un vif intérêt de certains acteurs de l'industrie de production cinématographique, et ont mené à un projet de commercialisation avec la société Parisien PANDORA-Studio.

Bien d'autres applications d'un système de capture de mouvement de ce type sont envisageables. Pour les interfaces utilisateur (homme-machine, réalité virtuelle et augmentée, jeux. . .), il serait appréciable d'être en mesure de suivre en temps réel – et éventuellement de reconstruire en 3-D – les gestes et les mouvements des utilisateurs non-instrumentés dans leurs environnements quotidiens. La fiabilité et la continuité temporelle sont les enjeux critiques ici: le manque relatif de précision 3-D n'est pas forcément critique.

1. C'est-à-dire, les studios n'autorisent sur scène que les caméras qui sont indispensables pour fi lmer l'action du/des point(s) de vue voulu(s) au plan artistique. Mes discussions avec les acteurs de l'industrie de production graphique (Pandora Studio, Rhythm & Hues) ont montrées à quel point il est délicat de rajouter sur scène des caméras supplémentaires, même cachées, pour aider aux effets spéciaux. Cela s'avère trop gênant face aux contraintes de visibilité, d'illumination, de construction de la scène, de planning, de mobilité des caméras, et surtout face aux besoins artistiques (on improvise des plans plus souvent qu'on ne l'imaginait . . .)

L'analyse automatique d'actions et d'interactions humaines pour la compréhension de scène, la surveillance vidéo, l'indexation de vidéo par action, l'analyse sportive, sont d'autres domaines d'application potentiels prometteurs pour cette technologie. Cependant, au stade actuel de la technologie du suivi humain, maintenir sans décrocher un suivi 2-D approximatif lors des interactions complexes entre les sujets et leur environnement reste difficile.

1.2 Survol des contributions

1.2.1 Capture 3-D par modélisation explicite

Les travaux sur la capture de mouvement humain sont divisés en trois chapitres thématiques. Le chapitre 2 présente quatre articles d'une approche résolument « générative »: un modèle 3-D explicite du corps articulé prédit l'apparence image du sujet, et à chaque nouvelle image, un procède d'optimisation numérique ajuste les paramètres articulaires du modèle afin de « coller » au mieux aux indices image observées. En principe le processus est simple, mais il faut inverser le modèle afin de récupérer le mouvement à partir de l'apparence, et – en plus des ambiguïtés classiques de correspondance modèle-image – le problème d'ajustement des paramètres est mal conditionné, de dimension élevée, et possède un nombre inattendu de solutions inverses (en général, quelques milliers).

Ainsi, ajuster le modèle devient le coeur du problème, et nos contributions techniques principales tournent autour de ce problème d'optimisation non-convexe. Nous présentons en effet quatre méthodes de recherche non-convexe en dimension élevée: (*i*) « Covariance Scaled Sampling » – une méthode d'échantillonnage aléatoire locale adaptée aux problèmes difficiles de dimension élevée; (*ii*) « suivi de vecteur propre » et « balayage d'hyper-surfaces » – deux méthodes de recherche de points selles (« cols » à partir desquelles les minima voisins peuvent être trouvés), basées sur l'optimisation locale modifiée; (*iii*) l'échantillonnage « hyperdynamique » – une méthode de chaîne de Markov Monte Carlo consacrée également aux points selles; et les « sauts cinématiques » – un générateur structurel d'autres solutions cinématiques inverses possibles. Seule la dernière méthode est spécialisée à la structure spécifique du problème. Par contre, les méthodes point selle peuvent être d'outils intéressants pour l'optimisation non-convexe générale – surtout pour les problèmes d'estimation de modèle moindre carrés non-linéaire – et cette piste reste à développer.

En pratique, ces techniques nous ont permis d'étendre significativement la capacité de suivi de la méthode, de moins d'une seconde face aux mouvements simples pour la méthode de CONDENSATION d'origine, à une dizaine de secondes face à un mouvement de danse relativement complexe pour la méthode finale de sauts cinématiques en combinaison avec Covariance Scaled Sampling. Nous pensons que ces derniers résultats représentent aujourd'hui l'état de l'art en suivi humain monoculaire basé sur un modèle, mais il est clair qu'ils restent insatisfaisants au plan applicatif, surtout compte tenu du fait que cette méthode est assez complexe à mettre en oeuvre et lourde en temps de calcul.

1.2.2 Capture 2-D et détection de personnes

Vu ces problèmes, le chapitre 3 prend du recul sur la reconstruction 3-D, à fin de présenter deux articles consacrées aux méthodes 2-D, l'un sur la détection des personnes et l'autre sur le suivi humain. Les deux méthodes sont basées sur des modèles 2-D articulaires simplifiés de type « Scaled Prismatic Models » [CR99] – chaque membre du corps est représenté par un rectangle ou par un rhomboïde de longueur variable, et ces éléments sont articulés afin de créer un pantin 2-D. Cette représentation reste proche de l'image, et évite les problèmes de récupération de l'information 3-D.

Par contre, il faut en générale créer plusieurs modèles 2-D pour capter les différentes « aspects » 3-D possibles – vue frontale, vue de profile, ...

Le détecteur de personnes allie un jeu de détecteurs Machine à Vecteur de Support (MVS) pour les différents membres du corps à une méthode de programmation dynamique qui retrouve la meilleure configuration instantiée du modèle dans une région donnée de l'image [FE73, FH00, IF01].

En ce qui concerne la méthode de suivi, la contribution principale est d'introduire un modèle dynamique flexible (linéaire autorégressif par morceaux) qui est appris sur une base d'exemples. Ce modèle dynamique permet un suivi stable à travers les différentes phases du mouvement et les changements d'aspect du modèle 2-D (tournant de vue frontal à vue de profile, ...)

1.2.3 Capture 3-D par apprentissage

Le chapitre 4 revient au problème de la reconstruction du mouvement 3-D à partir d'images monoculaires, mais il départ de façon radicale de l'approche du chapitre 2. Le modèle 3-D explicite est remplacé par un modèle « boîte noire » implicite, construit à partir d'un ensemble d'exemples d'apprentissage sous forme de séquences de mouvement qui ont été captées par un système de capture de mouvement 3-D classique, et de leurs silhouettes dans les images correspondantes. À partir d'un descripteur robuste de la forme de la silhouette, la méthode apprend une fonction de régression qui traduit le descripteur observé directement dans une estimation de la pose 3-D correspondante. Cette méthode marche bien la plupart du temps, mais produit souvent des « hics » gênants, dus aux ambiguïtés de la reconstruction à partir de la silhouette. Afin de corriger ce problème, elle est modifiée et incluse dans une boucle de suivi qui a été conçue pour enlever les ambiguïtés. Le système final donne une reconstruction du mouvement 3-D convaincante pour un coût de calcul minime.

1.3 Travaux précédents

Il existe un nombre important de travaux sur le suivi et la reconstruction 2-D et 3-D du mouvement humain. Les articles du mémoire associé 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.

La détection 2-D de personnes: Pour la détection de personnes (corps entiers) dans les images, les approches « basées modèle » prônent les représentations basées sur les modèles 2-D articulés explicites – par exemple le « scaled prismatic model » [CR99]. Pendant la phase de détection, le modèle doit être optimisé sur l'image, ce qui est souvent fait par un processus de programmation dynamique [FH00, IF01, RST02].

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 [Pap97, MPP01, DCdB+02], d'autres [Gav00] 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.

Le suivi 2-D : En terme du suivi 2-D sans reconstruction 3-D, les approches « cardboard people » [JBY96] et « scaled prismatic model » [CR99] utilisent des modèles 2-D articulés. La corrélation image permet la mise en correspondance pendant le suivi. Une variante récente apprend automatiquement la structure et l'apparence du modèle par groupement et Expectation-Maximisation [RF03]. En comparaison, dans les approches exemplaires, le suivi se fait par l'enchaînement d'une série d'exemples types d'apparence, sans modèle explicite [TB01].

Reconstruction du mouvement 3-D : Dans ce cadre, la plupart des travaux existants se base sur les données multi-caméras. Pour deux articles de synthèse, voir [Gav99, AC99]. Les travaux sur la reconstruction monoculaire à la base d'un modèle incluent [Hog83, Roh94, BM98, DBR00, SBS02], et [RK95] pour la reconstruction de la pose de la main. Les reconstructions basées plutôt sur les exemples incluent [Bra99, MM02, SVD03, SC02, SEC02, LESC04]. Pour la plupart, ces travaux récupèrent la pose 3-D en introduisant un modèle simplifié du squelette articulaire et en adoptant la solution cinématique orthographique popularisée par Taylor *et al* [Tay00].

Autres travaux : D'autres lignes de recherche intéressantes concernent l'apprentissage des modèles pour l'extraction d'indices image [SB01], et la dynamique du mouvement humain [HLF00, SBS02].

Chapitre 2

Approche 3-D par modélisation explicite

2.1 Introduction

Ce chapitre résume le contenu de plusieurs articles qui ont été écrits avec mon doctorant Cristian SMINCHISESCU, dont la thèse [Smi02] peut être consultée pour plus de détails. Tous les articles traitent du problème de la reconstruction du mouvement humain à partir de séquences d'images monoculaires, et tous supposent qu'on dispose d'un modèle 3-D articulaire explicite du corps. Ils relèvent donc des techniques de la « vision à la base de modèle ». À partir d'un tel modèle et d'une fonction qui quantifie la précision de la correspondance entre l'image rendue du modèle et l'image observée, on peut formuler le problème du suivi et de la reconstruction du mouvement en terme de la minimisation le long de la séquence de l'erreur de correspondance totale – éventuellement avec l'adjonction de termes dynamiques qui privilégient un mouvement lisse. Ce sont les grandes lignes de la vision à la base de modèle. La particularité du suivi humain est d'avoir un modèle nettement plus complexe que ceux utilisés par ailleurs, et dont la structure induit dans la fonction de coût modèle-image de très nombreux minima locaux qui s'entremêlent souvent. Le défi principal devient donc de suivre l'évolution du modèle dans un espace paramétrique de dimension élevée, face à une ambiguïté 3-D - image considérable. Le suivi se trompe facilement de la piste et se trouve dans un minimum incorrect, et pour se rattraper lors de ces décrochages, il faut en l'essentiel retrouver le bon minimum local voisin. Nos contributions principales tournent autour de la recherche efficace de ces minima. Dans ce but, nous avons proposé et testé quatre familles d'approches, qui sont détaillées plus bas.

2.2 Les modèles

Détaillons d'abord brièvement nos modèles du corps humain, de rendu image, et de correspondance modèle-image.

Modèle du corps: Notre modèle du corps (voir la partie gauche de la figure 2.1) consiste en un squelette articulaire à 35 degrés de liberté, enrobée d'une « chair » composée de superquadriques généralisées par des paramètres de courbure et de fuselure. Il n'a pas pour but d'être réaliste au sens graphique, mais simplement de suffire au suivi du mouvement.

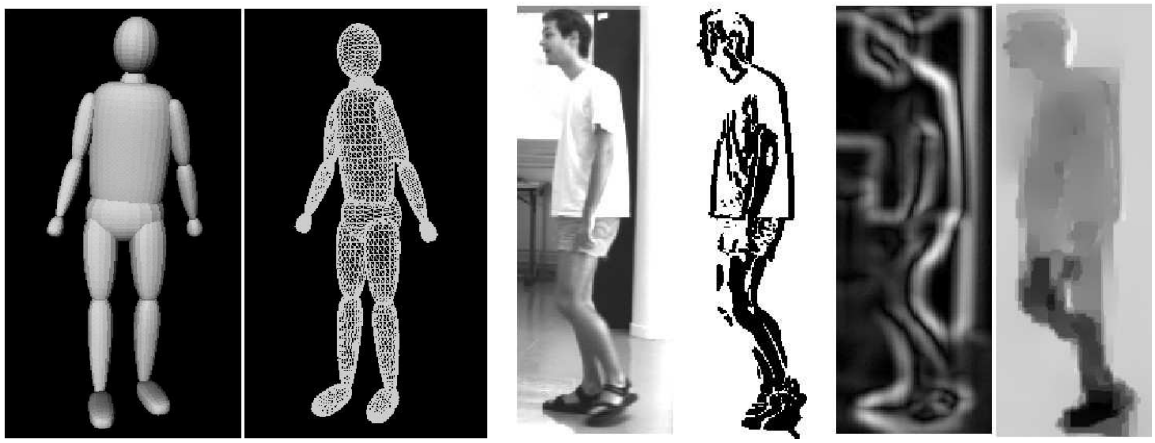


FIG. 2.1 – *À gauche: notre modèle du corps humain: une squelette articulaire d'environ 35 degrés de liberté, couvert d'une « chair » fait de superquadriques généralisées. À droite: quelques exemples de notre jeu d'indices image: (a) l'image originale; (b) frontières de mouvement; (c) l'énergie du gradient; (d) le flot optique robuste (champs horizontale).*

Les 35 paramètres libres consistent en les angles d'articulation internes du squelette plus la position 3-D du corps. On y compte une « colonne vertébrale » simplifiée de trois articulations sphériques, et un « complexe claviculaire » avec deux articulations sphériques dans chaque épaule. Ces articulations sont difficiles à observer directement, mais elles s'avèrent indispensables au suivi des mouvements d'extension de la main (sinon, l'extension effective se révélerait trop courte d'environ 5-10 cm, ce qui provoquerait facilement une décrochage du suivi).

Les variations de chaque articulation sont bornées à un intervalle angulaire. Ces limites sont mis en application par la méthode d'optimisation, en tant que contraintes d'intervalle dures. Il y a aussi des fonctions de pénalité supplémentaires, pour éviter l'interpénétration des membres du corps et pour stabiliser les articulations difficilement observables (colonne vertébrale, clavicule). Pour l'instant nous n'appliquons pas de contraintes d'équilibre et de contact au sol, qui semblent déstabiliser le suivi en raison de leur sensibilité aux imperfections du modèle.

Le modèle dispose aussi d'un jeu de paramètres qui contrôlent la forme du corps (longueur des bras, largeur des jambes, *etc*), et qui sont estimées lors de la phase d'initialisation, mais figées pendant le suivi. L'initialisation se fait manuellement en indiquant les positions image des articulations et en optimisant numériquement une fonction de coût qui exprime ces contraintes. Cette fonction est beaucoup plus rapide à évaluer que la fonction complète modèle-image, et comme elle contient déjà toutes les ambiguïtés cinématiques (voir plus bas), nous l'avons utilisée dans plusieurs de nos tests où l'aspect modèle-image n'était pas en jeu.

Rendu image: Le rendu du modèle est assez classique. Nous utilisons une chaîne de rendu OpenGL et un modèle projectif approximatif de la caméra. (La calibrage suffit à prédire la position approximative 3-D du sujet et la plupart d'effets perspectifs, mais nous avons focalisé nos efforts sur l'estimation de la pose du corps du sujet, pas de sa position 3-D absolue). Il faut aussi rendre une carte d'occultations (« item buffer »), ce qui donne en plus une carte des bords visibles de chaque membre.

Correspondance modèle-image: Afin de maximiser la fiabilité du suivi, nous utilisons une combinaison robuste d'indices d'image robustes – voir la partie droite de la figure 2.1. Une méthode de flot optique robuste [BY95] donne une première estimation du mouvement et aussi une indication

de la localisation dans l'image des bords des membres. Une reprojection de la texture de l'image précédente, sur le modèle 3-D estimé et ainsi dans l'image actuelle, donne une prédiction d'apparence de région qui peut être comparée de façon robuste avec l'image observée. Les bords des membres du modèle sont aussi mis en correspondance avec une carte d'arêtes extraite de l'image (similaire aux « snakes » classiques, mais en hypothèses multiples et prenant soin de ne pas induire dans la fonction du coût des sauts abruptes chaque fois qu'une arête sort de la fenêtre de recherche des arêtes images autour de la prédiction du modèle). Enfin, pour les expériences où le fond de scène est soit constant soit connu, une carte de distance « chanfrein » est calculée à partir des bords de la silhouette image de la personne, et utilisée en tant que fonction du coût pour les bords modèles, ce qui stabilise significativement la correspondance en échange d'une hypothèse image relativement forte.

2.3 La structure du problème

La reconstruction du mouvement humain 3-D articulaire à partir d'images monoculaires représente un cas difficile de la vision à base de modèles. On peut noter quelques sources de difficulté significatives:

Dimension élevée: Afin de bien épouser les images des personnes en mouvement, le modèle articulaire devait disposer d'au minimum 30–35 degrés de liberté (DDL). Ceci sans compter les DDL des mains, du visage, ni les mouvements complexes ou inhabituels. Par comparaison, un système de capture de mouvement classique enregistre un jeu redondant de 50–60 DDL.

Structure, contraintes, rendu et observation complexe: Les paramètres sont fortement couplés en raison de la structure cinématique chaînée, et les contraintes de limite d'articulation et de non-interpénétration des membres du corps rendent complexe l'espace paramétrique du modèle. Les occultations fréquentes entre les membres du corps diminuent l'observabilité des paramètres et brouillent la correspondance modèle-image. La forme du corps est de toute façon complexe et d'apparence très variable en raison des vêtements, qui ont souvent tendance à cacher les vraies lignes du corps et à brouiller encore plus les prédictions du modèle et la correspondance modèle-image.

Degrés de liberté en profondeur mal observés: Dans le cas monoculaire, environ un tiers des dimensions du modèle représentent des profondeurs (déplacements le long de la ligne de vision de la caméra), qui ne sont pas observables au premier ordre selon le modèle de projection orthographique. En principe, les effets perspectifs rendent observables ces paramètres, mais la matrice jacobienne correspondante reste en général assez mal conditionnée – une condition (gamme de valeurs singulières) de 10^3 à 10^4 est typique sur les 35 DDL de notre modèle, même pour une prise de vue proche où tous les membres sont bien visibles.

Minima cinématiques: Lié à cette faible observabilité en profondeur, il y a un nombre surprenant de minima locaux « cinématiques ». Supposons qu'on caractérise l'image de chaque membre par la position image des centres des articulations (supposées sphériques) à ses deux extrémités. Par exemple, un avant-bras est caractérisé par les projections des centres de son coude et de son poignet. Supposons aussi qu'on connaît déjà la position 3-D d'une de ces deux articulations, disons le coude. Il se trouve qu'il y a deux solutions possibles pour la position 3-D du poignet: les deux intersections de la ligne de vue 3-D de la position image du poignet avec la sphère 3-D centrée sur le coude et dont le rayon est la longueur de l'avant-bras. Ces solutions sont disposées de manière symétrique par rapport au plan orthogonal à la ligne de vue du coude et qui contient le coude. Ce raisonnement s'applique à chacun des 10–14 membres majeurs du corps et les ambiguïtés se

multiplient – pour chaque position d'un membre parent, il y a deux positions du membre fils: un « avant » (penchant vers la caméra) et l'autre « arrière ». Il faut donc prévoir l'existence potentielle de 2^{10} à 2^{14} solutions inverses cinématiques pour chaque configuration des centres d'articulation image donnée. En terme de la fonction de coût de correspondance modèle-image, chaque solution inverse donne lieu à un minimum local correspondant. Les contraintes de limites d'articulation et de non-interpénétration des membres éliminent un pourcentage de ces minima, mais en pratique – comme montre les méthodes de recherche de minima locaux décrit plus bas – il n'est pas rare de voir pour une configuration 2-D des articulations donnée, de 200 à 600 minima 3-D qui vérifient toutes les contraintes. Visuellement, il est souvent très difficile de lever l'ambiguïté entre ces minima. Les seules différences entre leurs images sont des effets perspectifs relativement subtils, des changements d'illumination liées à la profondeur, et éventuellement des inversions de l'ordre d'occultation.

En vue de ceci, il est légitime de se demander si on ne peut pas simplifier le problème en lui découpant dans une première étape de suivi image 2-D sans profondeur, et une deuxième étape d'estimation de profondeur qui fait face à la question des minima. Tout à fait – le chapitre prochain présente une méthode de suivi 2-D de ce type – cependant, assurer la rigidité et les contraintes 3-D est nettement plus délicat dans ce cas, parce que sans modèle 3-D de base, le modèle 2-D intermédiaire n'exprime pas la rigidité 3-D. Il est donc possible que la pose 2-D qui soit estimée ne soit pas cohérente avec aucune pose 3-D, et le suivi échoue. En pratique, on trouve que les contraintes de rigidité 3-D soient d'une aide considérable à la stabilisation du suivi 2-D – surtout pour les extrémités du corps – et ils est donc intéressant de les incorporer.

Il faut aussi souligner que ces minima restent à multiplier avec les minima produits par les ambiguïtés de correspondance modèle 2-D / image traditionnelles, qui existe indépendamment des ambiguïtés de projection 3-D à 2-D.

Minima qui s'entremêlent: Si les minima cinématiques restaient toujours bien séparés, ils ne seraient pas aussi problématiques – après initialisation, le suivi pouvait poursuivre un seul minimum sans ambiguïté. Malheureusement, les minima s'entremêlent fréquemment en pratique, ce qui augment de façon significative le probabilité de se tromper. Quand chaque membre du corps s'approche de son plan fronto-parallèle – ce qui arrive souvent en pratique – les deux minima correspondants se fusionnent dans un seul minimum qui est étendu en profondeur (la ligne de vue tangente la sphère, donc les déplacements en profondeur ne changent pas l'image au premier ordre). Quand le membre s'écarte du plan, les minima se séparent rapidement. En raison de ce comportement, l'algorithme se trompe souvent de minimum, et dans ce cas, la séparation rapide des minima implique qu'il se trouve rapidement loin de la vraie solution, d'où il est difficile de se rattraper.

2.4 Approche 1: l'échantillonnage efficace

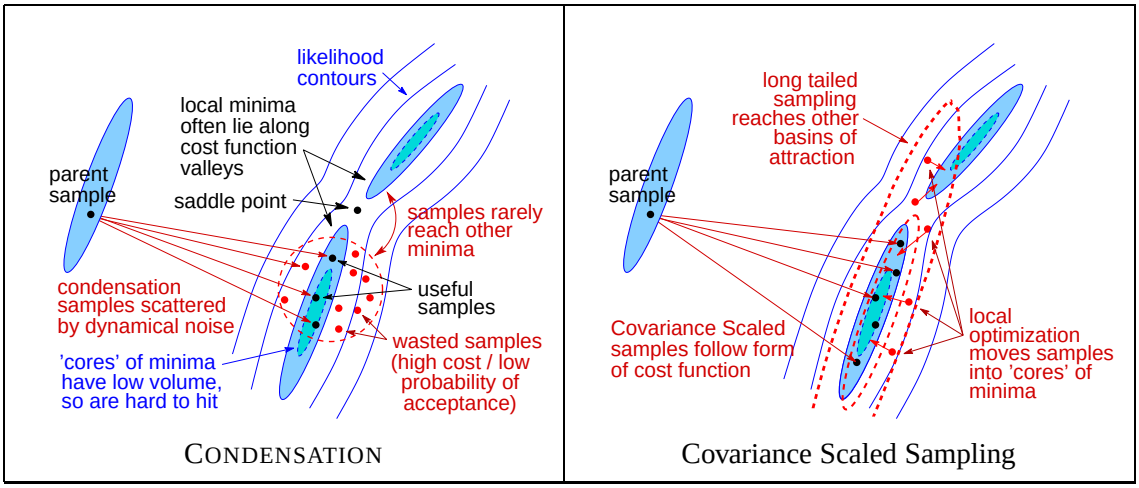
Les méthodes basées sur le filtrage particulaire bayésien telles que CONDENSATION [IB98, DdFG01, SBR+04] sont devenues la référence pour implanter le suivi visuel robuste. Pour l'essentiel, il s'agit de propager un nuage d'hypothèses possibles, chacune avec son poids probabiliste, et de re-échantillonner les hypothèses et de mettre à jour leurs poids à chaque étape, selon un modèle probabiliste d'évolution dynamique et un modèle de vraisemblance image. Le fait de disposer d'un nuage bien répandu de possibilités rend plus robuste le suivi face aux imperfections de modèle et du suivi. Cette approche se révèle assez performante en basse dimension, mais dès que la dimension augmente – et surtout quand le problème est mal conditionné – elle devient rapidement moins fiable. Et ceci malgré quelques résultats théoriques qui tendent à indiquer que la performance de la méthode reste correcte en haute dimension [DdFG01], au moins dans les problèmes uni-modaux.

Nous avons déjà évoqué la source principale de la difficulté – le fait que la géométrie du suivi monoculaire du mouvement 3-D humain implique l'existence d'un nombre important de minima locaux qui s'entremêlent fréquemment et puis se séparent abruptement. Pour pallier aux problèmes créés par les minima locaux, les implantations de suivi particulaire en vision adoptent souvent l'heuristique d'étendre artificiellement leur zone d'échantillonnage dans l'espace des paramètres, afin de se donner plus de chances de rattraper lors des décrochages. En basse dimension, une simple augmentation du bruit dynamique suffit. Malheureusement, cette heuristique ne suffit plus en haute dimension: elle a tendance à perdre la plupart de ses échantillons dans les régions de coût élevé, sans arriver à échantillonner là où il le faut pour rattraper – voir la figure 2.2. Nous avons mené une analyse de cette situation qui conclut que trois éléments sont indispensables pour assurer un échantillonnage adéquat dans les problèmes mal conditionnés en dimension élevée:

1. **Un échantillonnage qui épouse la forme de la fonction de coût:** Quand le problème est mal conditionné, la fonction de coût a des bassins très allongés et il ne suffit pas d'échantillonner de façon isotrope. Un échantillonnage isotrope qui reste dans le bassin à forcément un petit diamètre, ce qui ralentit significativement l'exploration aléatoire du bassin, et si le diamètre est choisi pour couvrir la dimension la plus étendue du bassin, la plupart des échantillons sont perdus dans les zones de haut coût sur les côtés du bassin. En dimension élevée il n'y a pas de compromis possible entre ces deux possibilités parce que le volume d'un ellipsoïde allongé est minuscule face au volume de sa sphère englobante. Cet effet peut être très significatif. Pour notre cas d'étude on voit souvent des facteurs de volume de l'ordre de $10^{50} : 1$. La solution est donc de redéfinir l'échantillonnage afin de mieux épouser la forme locale de la fonction du coût, ce qui peut être fait en pratique en estimant le Hessien (la courbure locale) de la fonction près du minimum, et en appliquant un échantillonnage dont la covariance est égale au Hessien. Autrement dit, on fait une approximation gaussienne de type « point selle » de la fonction au voisinage du minimum, et on échantillonne cette gaussienne. .
2. **Un échantillonnage avec des queues longues:** En pratique, les minima voisins sont souvent séparés de dizaines, voire de centaines, d'écarts standards. Afin d'arriver jusqu'aux bassins d'attraction d'autres minima avec un probabilité significative, il faut augmenter systématiquement la taille de la distribution d'échantillonnage. Augmenter l'écart standard de cette distribution sans changer sa forme aide un peu, mais il est plus judicieux de remplacer la forme non-robuste (par exemple gaussienne) de cette distribution avec une forme plus robuste avec des queues longues. Ceci permet de garder un bon échantillonnage au centre de la distribution (et donc à la zone probable de suivi), tout en éparpillant bien un pourcentage des échantillons afin de donner à la recherche un caractère un peu plus global.
3. **L'optimisation locale:** L'échantillonnage en soi même ne suffit pas. En haute dimension, le volume d'une sphère croît très rapidement avec son rayon, donc le volume du « noyau » (zone de coût bas) d'un minimum est toujours minuscule en comparaison avec celui de son bassin d'attraction. Ceci veut dire que même quand un échantillon tombe dans le bassin, il ne tombera presque jamais dans la zone de coût bas, donc il serait presque certainement éliminé par l'étape de re-échantillonage chaîne de Markov en raison de son coût élevé / vraisemblance minime. On peut contourner ce problème en lançant quelques pas d'optimisation locale à partir de chaque échantillon, afin de voir à quel point le coût est susceptible d'être réduit avant de procéder à l'étape de re-échantillonnage.

À partir de ces trois observations, il est facile de créer une fonction d'échantillonnage avec les propriétés nécessaires. Chacune de ces trois étapes modifie le modèle du bruit dynamique original; ceci peut nuire à l'interprétation probabiliste stricte du modèle, mais en revanche, la robustesse pratique du suivi s'améliore nettement.

Nous appelons cette méthode « Covariance Scaled Sampling (CSS) ». Appliquée au problème



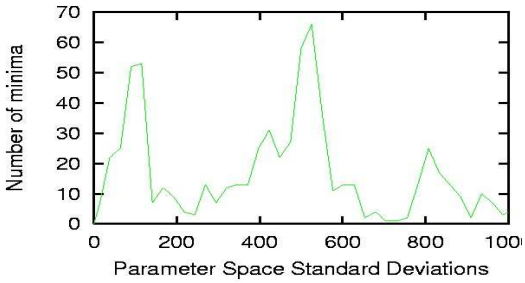


FIG. 2.2 – En haut: *pourquoi l'échantillonnage* CONDENSATION *classique devient insuffisant pour les problèmes mal conditionnés en haute dimension. (i) Si la distribution d'échantillonnage ne suit pas la forme de la vraisemblance (fonction de coût), la plupart des échantillons sont perdus dans les régions de coût élevé aux bords du minimum local (petit rapport de volume entre une ellipsoïde et sa sphère englobante en dimension élevée). (ii) Si la distribution d'échantillonnage n'a pas de queues étendues, les échantillons arrivent trop rarement jusqu'aux voisinages des minima voisins, qui sont typiquement distant d'un grand nombre d'écarts types. (iii) En dimension élevée, il est rare de tomber droit dans le noyau central d'un minimum en raison du petit volume du centre par rapport au bassin d'attraction, donc si après l'échantillonnage on n'optimise pas localement les échantillons avant de décider de les retenir ou pas, la plupart d'entre eux sont perdus en raison de leur coût élevé.* En bas: *histogramme typique de la séparation d'un minimum donné des minima locaux voisins, en écarts types. Un échantillonnage Gaussien classique de 35 DDL n'arrivera presque jamais à pénétrer jusqu'à ces minima.*

de suivi humain monoculaire, elle fournit d'assez bons résultats, ce qui nous a permit d'étendre le temps moyen entre décrochements d'une demi-seconde avec CONDENSATION jusqu'à environ 2 secondes / 50 trames avec la CSS. La figure 2.3 montre un exemple des résultats. Cependant, même si cette performance reste honorable [1] en comparaison avec l'état de l'art du suivi monoculaire à l'époque où elle était introduite, il est clair qu'elle ne suffit pas pour un suivi pratique. En effet, avec un nombre d'échantillons réalistes, l'échantillonnage n'arrive pas à pénétrer suffisamment loin dans l'espace de solutions alternatives, et tôt ou tard, le suivi échoue.

Ce travail a été publié dans un numéro spécial du journal *International Journal of Robotics*

1. Rappelons ici que ce problème est nettement plus délicat que ses analogues 2-D et multi-caméra 3-D, et ne doit pas être confondue avec eux.

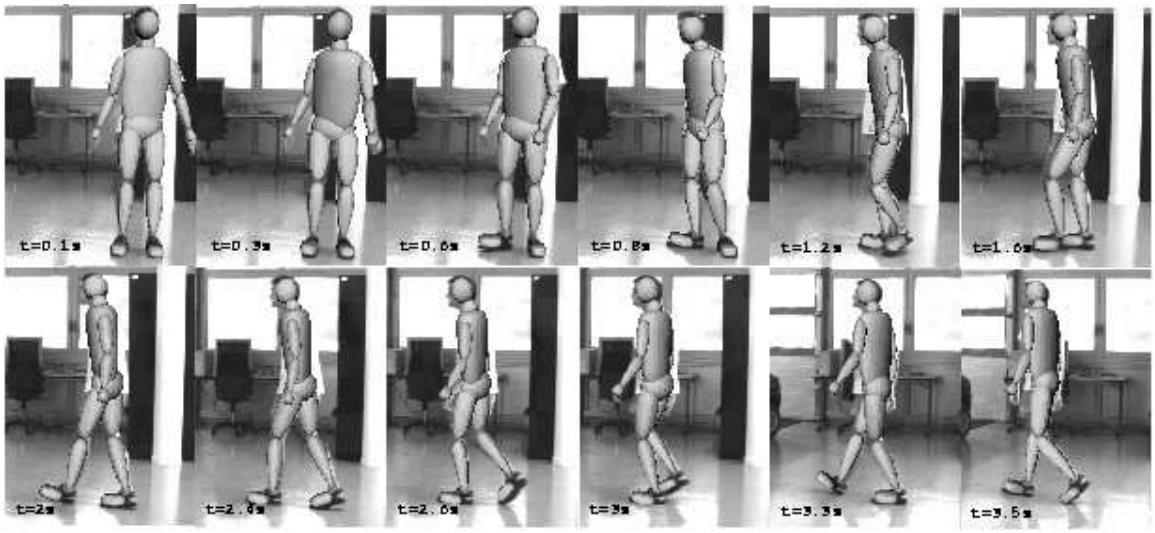


FIG. 2.3 – *Suivi d'un mouvement humain sur fond encombré d'objets, avec la méthode Covariance Scaled Sampling. Voir la référence [ST03a] (article #1 du mémoire associé) pour les détails.*

Research (IJRR), consacré à la reconstruction du mouvement humain, [ST03a] (article #1 dans le mémoire des publications associé). Deux versions préliminaires ont été publiées, dans un rapport technique INRIA [ST01b] et au *2001 International Conference on Computer Vision and Pattern Recognition (CVPR)* [ST01a].

2.5 Approche 2 : rechercher des « états de transition »

Afin d'étendre le suivi, il faut essentiellement développer des algorithmes plus efficaces de recherche de minima locaux voisins. Une piste consiste à observer que chaque sentier entre deux minima voisins doit traverser un « col » ou un sommet à son point le plus haut dans le paysage du coût, et si le sentier a un coût maximale qui est localement minimale, ce point doit être un col et pas un sommet. Si on pouvait retrouver de tels cols voisins, il suffirait simplement de descendre le gradient de coût à l'autre côté par minimisation locale, afin de retrouver les minima correspondants. Plus précisément, « col » veut dire un point selle de la fonction du coût dont précisément une des courbures principales est négative (la direction de passage du col) et les autres sont positives. (Sinon, le coût de passage – la valeur maximum de la fonction – peut être minoré localement).

Il semble que ce genre d'approche n'apparaisse pas (ou bien, a été oublié...) dans la bibliographie d'optimisation globale[2], mais on peut trouver quelques exemples de méthodes de ce type dans la littérature de la chimie computationelle, où l'étude des transitions d'énergie mini-

2. Voir par exemple l'état de l'art du projet EU COCONUT *http://www.mat.univie.ac.at/ñeum/glopt/coconut/StArt.html*, le tour d'horizon plus récent de A. Neumaier [Neu04], et plus généralement les sites web *http://solon.cma.univie.ac.at/ñeum/glopt.html* et *http://www-unix.mcs.anl.gov/otc/Guide/faq/nonlinear-programming-faq.html*. Les approches qu'on voit citées régulièrement les plus similaires à les notres semblent être les méthodes « tunnelling » [LM85] – où on « bouche » systématiquement les minima qui ont déjà été trouvés en ajoutant des pôles (fonctions de forme $f(\mathbf{x})/\|\mathbf{x}-\mathbf{x}^*\|$ ou similaire, qui sont infinies au minimum $\mathbf{x}^*$) – et « filled function » [Ge87] – qui appliquent une fonction du type $1/(f(\mathbf{x})+\text{const.})$ ou pareil, afin de transformer le minimum local dans un maximum avant de poursuivre l'optimisation. Ni l'une ni l'autre est bien adaptée au traitement d'un grand nombre de minima parce que la série correspondante de transformations composées déforme de façon progressive la fonction du coût. La méthode de « balayage de hyper-surface » qui est décrite prochainement peut être vue comme une version plus sophistiquée de l'approche tunnelling.

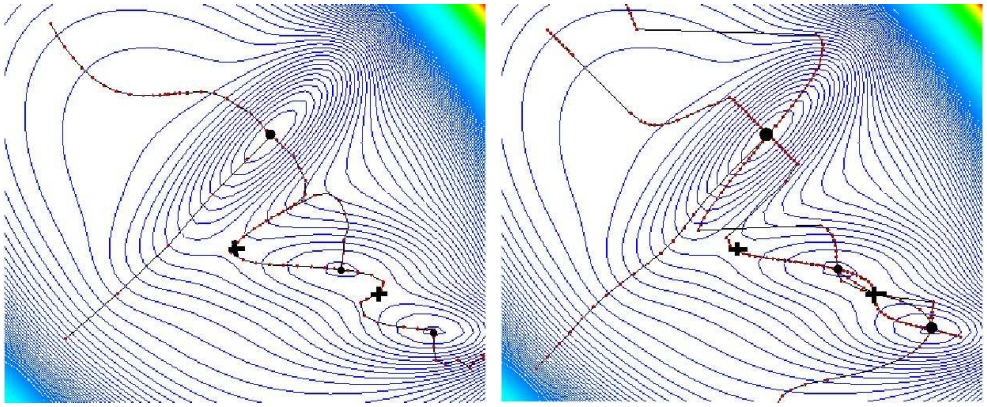


FIG. 2.4 – *Les trajectoires de suivi de vecteur propre (à gauche), et de balayage d'hyper-surface (à droite), dans un problème synthétique 2-D simple. Les points marquent les trois minima de la fonction de coût, et les croix marquent les deux points de selle. Les trajectoires qui sont lancées à partir de chaque minimum trouvent les autres minima.*

mum entre deux états (d'un molécule ou d'une réaction, par exemple) est un des principaux enjeux [CS71, Hil77, CM81, Wal89, WW96, MW99, HJ99, SJTO83, NTSS90, Hel91, CDNG92, Bof94, JJH88, SR93, Jen95, AR94, ART94, Bar96, MB98]. Suivant cette littérature, on dira « état de transition » à la place de col.

Nous avons développé et étudié deux classes de méthodes de ce type dans le contexte du suivi humain: deux méthodes déterministes à la base de l'optimisation locale dont on parlera maintenant, et une méthode aléatoire de type chaîne de Markov dont on parlera plus bas.

Les approches basées sur l'optimisation locale modifie les algorithmes de minimisation locale afin qu'ils convergent vers un point de selle du type voulu, et non vers un minimum. L'itération Newton de base converge déjà indifféremment vers n'importe quel type de point de selle. L'essentiel est donc de modifier la méthode stabilisante de contrôle de boucle qui garantit normalement la descente vers un vrai minimum, afin de monter du minimum, à la recherche d'un point de selle de la signature voulue.

Suivi de vecteur propre: L'approche la plus élémentaire est de garder la stabilisation Newton amorti / Levenberg-Marquardt telle qu'elle est, mais de la « tromper » afin qu'elle monte à la place de descendre. On travaille dans le repère des vecteurs propres du Hessien local de la fonction de coût, on choisit un de ces vecteurs (celui qu'on veut pousser vers une courbure négative), et on inverse le signe du gradient et de la courbure de ce vecteur avant de calculer le pas Newton amorti. L'effet est de monter vers une courbure négative en cette direction, lorsqu'on continue de descendre vers une courbure positive en toutes les directions orthogonales. Cette méthode s'appelle « suivi de vecteur propre » (eigenvector tracking). Elle est simple à mettre en oeuvre et relativement performante quand elle réussit, mais elle est souvent erratique: afin d'assurer le progrès il faut modifier « le même » vecteur propre à chaque pas, et cette notion d'identité des vecteurs propres n'a pas de sens absolue – elle est nécessairement heuristique et parfois instable [3].

3. Quand deux valeurs propres s'approchent afi n de se croiser, les vecteurs propres correspondants ne peuvent plus être distingués – ils devient en effet instables et tournent rapidement sur 90°, ce qui a pour effet d'échanger leurs identités et d'éviter le croisement des valeurs propres correspondants. Vue de grande échelle, les valeurs propres ont l'apparence de se croiser et de garder à peu près leurs vecteurs propres originaux. Vue de petite échelle, il y a une déviation rapide des vecteurs, sans croisement des valeurs. Les deux comportements sont diffi ciles a distinguer, et pour l'application actuelle, le comportement inexact à grande échelle semble être la chose voulue.

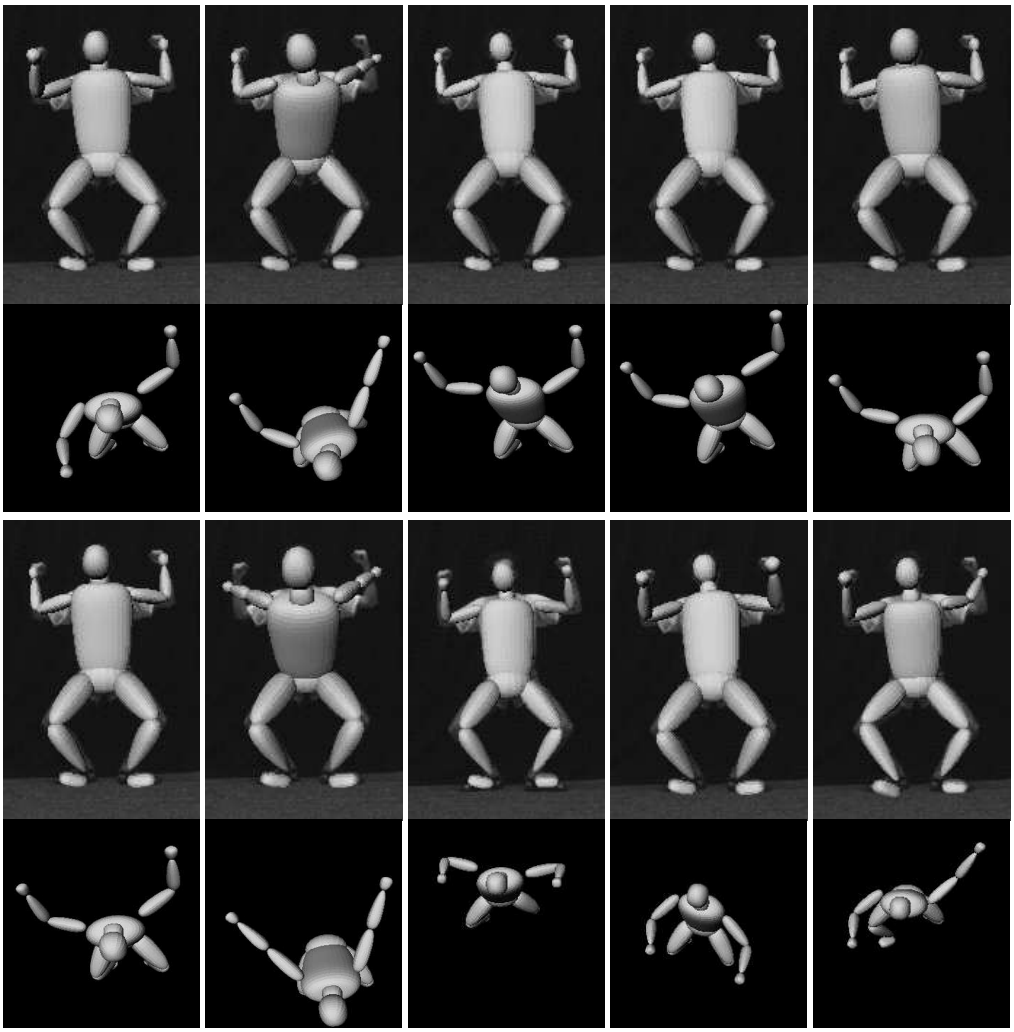


FIG. 2.5 – *Quelques exemples des centaines de minima « cinématiques » qui sont trouvés par une recherche par suivi de vecteur propre à partir d'une seule image et pose 3-D d'origine. Ici la fonction de coût fixe les centres des articulations dans l'image, donc la plupart des minima sont « cinématiques » – il n'y a ni d'ambiguïté de correspondance modèle-image, ni d'ambiguïté d'étiquetage des membres du corps dans l'image. On voit que l'image du corps change peu (seulement par les effets d'illumination, de perspective), comme il se doit. Seulement, vu d'en haut, les poses 3-D sont très différentes, et un suivi lancé à partir d'une pose incorrecte s'échouera rapidement.*

Balayage d'hyper-surface : Une approche un peu plus sophistiquée consiste à définir une famille d'hyper-surfaces qui balayent l'espace – une famille de sphères ou d'ellipsoïdes croissants, ou d'hyper-plans balayant – et de suivre un ou plusieurs des minima locaux de la fonction de coût dans ces hyper-surfaces lors du mouvement, jusqu'à ce que le coût du minimum passe par un maximum local. L'idée est que le minimum est balayé vers un col (la minimisation dans l'hyper-surface l'assure), où son coût atteint un maximum local. Les maxima indiquent donc des états de transition. Cette méthode de « balayage d'hyper-surface » (« hypersurface sweeping ») implique une série de minimisations contraintes aux hyper-surfaces, mais elle peut être implantée de façon efficace avec les techniques numériques standards. Elle est plus fiable que le suivi des vecteurs propres au sens où elle ne peut pas boucler sans avancer, mais elle a aussi deux défauts significatifs : (*i*) elle

est systématiquement « aveugle » aux points de selle que l'hyper-surface ne traverse pas de façon montante (c-à-d. où la direction du mouvement local de l'hyper-surface n'est pas dans la cône des directions de courbure négative du point de selle – à de tels points, l'hyper-surface ne coupe pas la fonction de coût dans un minimum); (*ii*) parmi les maxima qui sont trouvés, certains représentent les événements topologiques (annihilation d'un minimum hyper-surfacique avec un point de selle) et non pas des états de transition – quoique la recherche des *minima* voisins peut toujours poursuivre à partir de ces points. En pratique, il est conseillé d'essayer les deux méthodes.

La figure 2.4 illustre l'action de ces deux méthodes sur un exemple 2-D synthétique simpliste. En ce cas chaque méthode trouve tous les trois minima locaux et tous les deux points de selle de la fonction d'objectif.

Appliqué au problème de suivi monoculaire humain, chaque méthode donne des résultats intéressants. Les minima qui sont trouvés ont un caractère légèrement différent dans les deux cas, mais il n'est pas facile de dire laquelle des méthodes est la plus performante. En chaque cas on trouve des centaines de minima locaux voisins de bas coût – voir la figure 2.5 – et ceci pour une charge de calcul relativement modeste. Implantée dans notre système de suivi du mouvement humain, chaque méthode permet d'étendre la période moyen de suivi avant décrochement, de quelques secondes maximum avec Covariance Scaled Sampling, à environ 4–6 secondes.

Ces travaux seront publiés dans *International Journal of Computer Vision* en janvier 2005 [ST05a] (article #2 du mémoire associé). Une version préliminaire fut publiée au *2002 European Conference on Computer Vision* [ST02b].

2.6 Approche 3 : la chaîne de Markov « hyperdynamique »

Une deuxième approche issue de la chimie est l'échantillonnage « hyperdynamique ». Il s'agit d'une méthode chaîne de Markov Monte-Carlo (MCMC) développé pour accélérer la simulation des réactions chimiques. L'astuce consiste à modifier la fonction d'énergie selon laquelle le système évolue, afin de biaiser les échantillons plutôt vers les zones de transition que vers les bassins des minima. En alternant les phases d'hyperdynamique et de dynamique MCMC normale, on peut augmenter énormément la vitesse à la quelle le calcul MCMC approche à l'équilibre thermique, et ainsi, trouver les minima voisins de bas coût. La modification de la fonction du coût consiste à l'adjonction d'un terme conçu pour augmenter le coût près des minima, sans le changer auprès des zones de transition (les zones où le gradient est relativement petit et la hessienne a exactement une valeur propre négative). Cette modification dépend du Hessien de la fonction de coût originale – voir les articles pour les formules précises. La figure 2.6 montre l'effet des deux paramètres qui déterminent la contribution hyperdynamique, dans un exemple synthétique en 2-D.

En pratique, appliquée au suivi humain, cette approche fonctionne, mais – comme d'autres méthodes de chaîne de Markov pour ce problème – elle est beaucoup plus coûteuse que les méthodes à la base de l'optimisation locale citées précédemment. Elle serait probablement plus intéressante dans les applications où les méthodes de chaîne de Markov de base sont déjà les méthodes de choix.

Ce travail est accepté à un numéro spécial du journal *Image and Vision Computing* consacré aux meilleures articles d'ECCV'02, qui apparaîtra en début 2005 [ST05b] (article # 3 du mémoire associé). Une version préliminaire fut publiée au *2002 European Conference on Computer Vision* [ST02a].

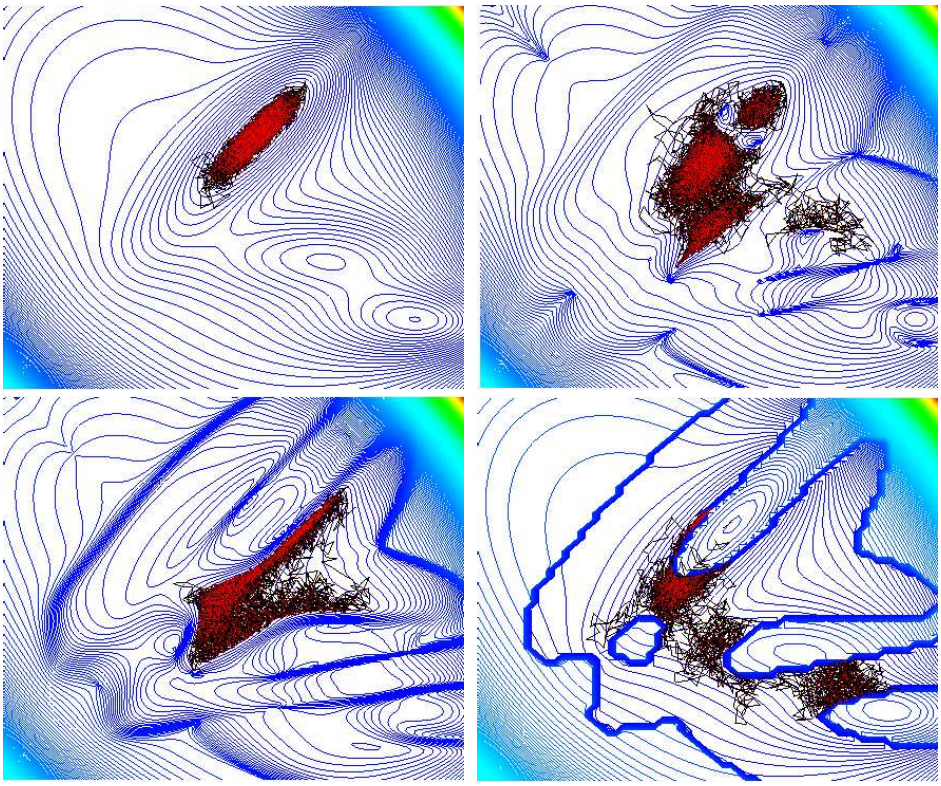


FIG. 2.6 – *L'effet des deux paramètres qui déterminent la forme de la fonction de coût hyperdynamique. Première ligne: à gauche, sans modification hyperdynamique, l'échantillonnage chaîne de Markov est coincé – il n'arrive pas à sortir du minimum initial. À droite, une petit contribution hyperdynamique commence à faire sortir la recherche du minimum. Deuxième ligne: une contribution hyperdynamique plus forte focalise l'échantillonnage sur les zones des deux points de selle. À gauche: avec une échelle de transition spatiale grande, l'échantillonnage se focalise sur le premier point de selle. À droite: une échelle spatiale plus petite rend la fonction de coût plus raide et plus convulsée, et disperse les échantillons sur les deux points de selle.*

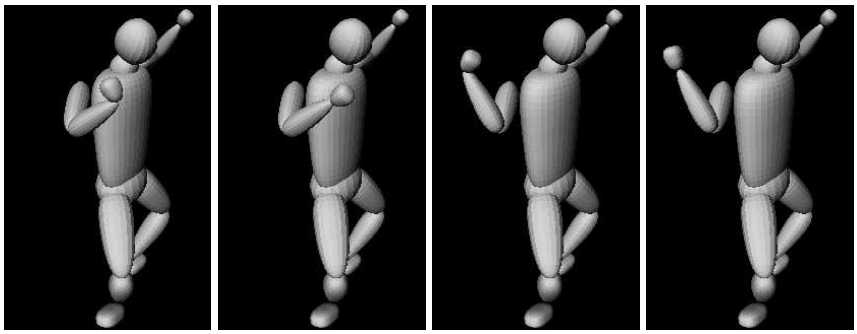


FIG. 2.7 – *Quelques configurations possibles engendrées pour un avant-bras et une main (sans déplacer le coude) lors d'un processus « sauts cinématiques » à partir d'une vue d'en face de la personne. La première configuration ne se conforme pas à la contrainte d'angle d'articulation du poignet, et sera éliminée.*

2.7 Approche 4: les sauts cinématiques

Les méthodes citées précédemment sont génériques – elles s'appliquent à n'importe quelle fonction de coût lisse qui possède des minima locaux multiples. Une approche alternative consiste

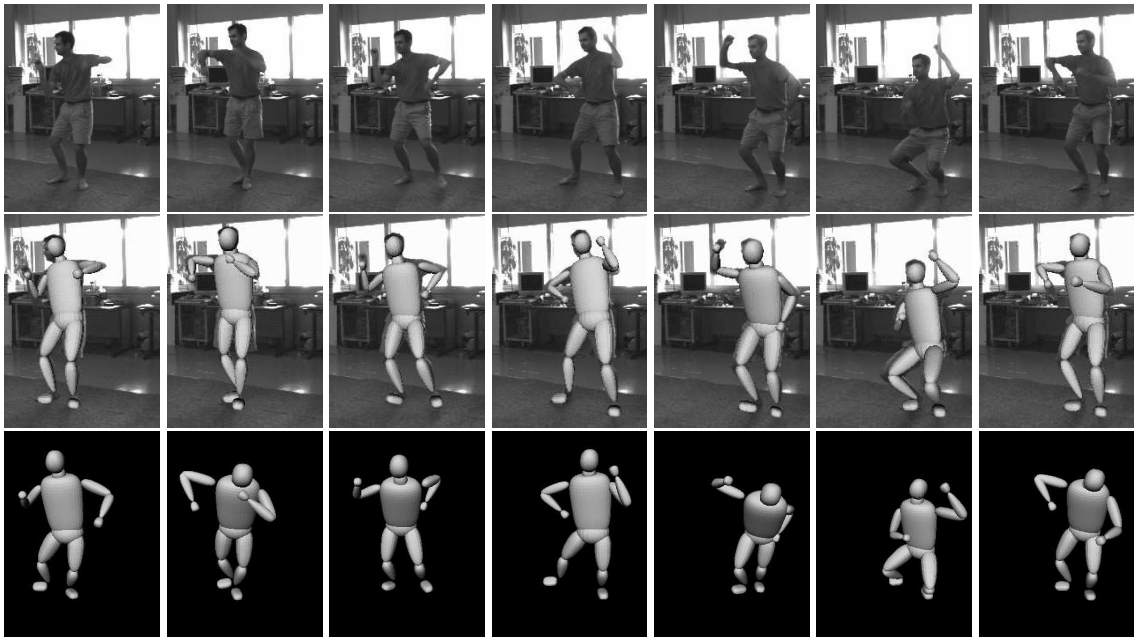


FIG. 2.8 – *Le suivi « sauts cinématiques » sur une séquence de danse. Première ligne : images originales. Deuxième ligne : reprojection du suivi dans l'image d'origine (le suivi est multi-hypothèses, mais seulement la meilleure configuration est montrée à chaque instant). Dernière ligne : une vue différente du mouvement reconstruit.*

à exploiter la structure spécifique du problème de suivi humain monoculaire. En particulier, on sait (au moins en rétrospectif) que cette structure induit un grand nombre de minima locaux « cinématiques », qui représentent une source majeur de décrochement.

L'astuce consiste en parcourir explicitement l'ensemble des minima cinématiques qui correspondent à la configuration 2-D observée, afin de retrouver celui qui est le plus apte à être la continuation du mouvement observé. À partir d'une configuration 3-D donnée et de sa configuration image correspondante (positions 2-D des centres d'articulation), il est facile de générer une série de « sauts cinématiques » qui font le tour des autres configurations 3-D possibles. Il suffit simplement d'énumérer (ou, selon le cas, d'échantillonner de façon aléatoire) les alternatives avant/arrière possibles, et de produire de haut en bas les solutions cinématiques correspondantes en utilisant la géométrie d'intersection rayon-sphère. La figure 2.7 illustre quelques mouvements de ce type d'un bras.

Cette méthode se révèle assez performante en pratique. À partir d'un minimum 3-D donné, les sauts cinématiques prédisent très bien les positions des autres minima locaux possibles, et la exploration explicite des ces minima est beaucoup plus rapide et plus sûr que les méthodes génériques de recherche de minima. Cependant, cette méthode ne sait parcourir que les minima cinématiques : il faut utiliser en parallèle une méthode générale comme Covariance Scaled Sampling, qui peut examiner les minima dûs aux ambiguïtés de correspondance modèle-image. La combinaison des ces deux méthodes nous a permis de suivre et de reconstruire des mouvements de danse relativement complexes pendant une dizaine de secondes – voir la figure 2.8.

Ce travail fut présenté au *2003 International Conference on Computer Vision and Pattern Recognition* [ST03b] (article #4 du mémoire associé). Une version journal est en cours.

2.8 Conclusions et perspectives

Un défi majeur de la reconstruction monoculaire du mouvement humain est éviter les décrochements provoqués par l'entremêlement des nombreux minima locaux de la fonction de coût de correspondance modèle-image. Ce chapitre a présenté quatre familles de méthodes de recherche de minima locaux voisins consacrées à la résolution de ce problème. Trois de ces approches sont en effet génériques et susceptibles d'être appliquées à d'autres problèmes de recherche de minima locaux voisins: *Covariance Scaled Sampling* perfectionne la recherche locale aléatoire pratiquée par les méthodes de suivi particulaires – il exploite la forme locale de la fonction du coût, l'échantillonnage à queues longes, et l'optimisation locale, afin d'atteindre plus souvent aux minima voisins sans gaspiller trop d'échantillons dans les régions de coût élevé; *suivi de vecteur propre* et *balayage d'hyper-surface* sont deux méthodes basées sur l'optimisation locale modifiée qui retrouvent les « cols de montagne » (points selle) voisins qui mènent aux minima voisins; et *l'échantillonnage hyperdynamique* est une méthode aléatoire chaîne de Markov du même type. La quatrième approche, les *sautes cinématiques*, exploite la structure spécifique du problème de suivi monoculaire afin d'examiner systématiquement tous les minima d'une famille de solutions cinématiques (qui ont toutes la même projection image).

En pratique, une combinaison de la méthode de sautes cinématiques (afin de choisir de façon efficace, pour une projection image donnée, la bonne solution cinématique) avec soit Covariance Scaled Sampling soit suivi de vecteur propre / balayage d'hyper-surface (afin d'adresser les ambiguïtés de correspondance 2-D modèle-image) semble donner les meilleurs résultats [4]. Cette combinaison nous a permis de suivre et de reconstruire les gestes et les mouvements relativement complexes pendant une dizaine de secondes, ce qui est, à l'heure actuelle, à l'état de l'art pour le problème de capture de mouvement 3-D monoculaire à base de modèle.

Quoique cette performance permet d'envisager d'applications semi-automatiques telles que la capture de mouvement « manuelle assistée » pour la production de films, elle ne suffit pas pour les applications plus automatisées où l'intervention manuelle n'est pas permis. Aussi: la méthode actuelle demande l'initialisation manuelle; il y a toujours de passages où le suivi est délicat; l'étape d'optimisation est plutôt lourde (jusqu'à quelques minutes par trame – mais ceci reste à optimiser); il y a un modèle complexe articulé à construire à la main; et implanter l'algorithme est complexe lui aussi. Pour ces raisons – et malgré le fait que disposer du mouvement 3-D explicite d'un modèle 3-D explicite permet de viser le nombre maximum d'applications – les deux prochains chapitres abordent deux approches alternatives: la modélisation 2-D sans 3-D explicite; et une approche apprentissage « boîte noire » à la reconstruction du mouvement 3-D, sans modèle explicite.

J'ai déjà signalé le fait que trois de nos méthodes de recherche de minima locaux voisins soient génériques et susceptibles d'application dans d'autres problèmes d'optimisation non-convexe. Les approches du type suivi de vecteur propre et balayage d'hyper-surface semblent être particulièrement prometteuses dans ce sens, et sont à valider dans d'autres problèmes. En particulier, elles semblent être bien adaptées aux problèmes d'estimation moindre carré non-linéaire (MCN) qu'on voit si souvent en pratique, par exemple en vision géométrique et en régression statistique. Ces problèmes sont lisses, non-convexes et souvent mal conditionnés, et ils ont des nombres de variables et de minima locaux qui peuvent être considérables sans être démesurés (disons de deux à quelques milliers dans chaque cas). Deux des particularités des problèmes MCN pratiques sont les faits que la fonction du coût est souvent assez convolutée et difficile de borner, et que les minima sont produits

4. À noter que ces deux sources d'ambiguïté sont presque indépendantes. Il est aussi possible de s'adresser aux ambiguïtés image avec un modèle 2-D du type présenté au chapitre prochain, avant même de s'attaquer au problème 3-D. Cependant, dans ce cas il faut deux modèles, et assurer les contraintes et la dynamique 3-D est délicat parce que le modèle 2-D intermédiaire ne les exprime pas.

en l'essentiel par biais d'annulation locale entre plusieurs termes, ce qui implique une forte corrélation entre les variables correspondantes, et assez souvent le mauvais conditionnement. Dans ce cas – et malgré beaucoup de progrès – les méthodes d'optimisation globale « certifiées » telles que les approches « branch and bound » marchent mal, parce qu'elles n'arrivent pas à bien borner la valeur de la fonction (complexité, annulation) et la décomposition spatiale qu'elles pratiquent est à la fois exponentielle en la dimension du problème et mal alignée avec les corrélations qui déterminent le comportement local de la fonction [Neu04, section 23]. Aussi, les approches non-convexes plus heuristiques (recuite simulée, *etc*) n'exploitent pas le fait que la fonction soit lisse avec un nombre de minima modéré, et elles ont forte tendance à être piégé dans un minimum sous-optimal. Par contre, les approches suivi de vecteur propre / balayage d'hyper-surface héritent le comportement exemplaire des méthodes de Newton face aux corrélations et au mauvais conditionnement, le nombre de variables modéré permet l'évaluation explicite du Jacobien et ainsi l'évaluation de l'approximation Gauss-Newton du Hessien sans avoir à calculer les dérivés de second ordre, et le nombre modéré de minima locaux doit permettre de les rechercher un par un avec des bonnes chances de succès.

Chapitre 3

Approche 2-D et détection de personnes

Quoique l'information 3-D soit en générale le but final, en pratique l'approche 3-D est assez lourde à mettre en oeuvre et à initialiser, et elle souffre du fléau des minima locaux cinématiques. Il est donc intéressant d'étudier le cas 2-D où la correspondance modèle-image est beaucoup plus directe et où les minima cinématiques n'interviennent pas. Ce chapitre présente deux contributions en cette direction: un détecteur articulaire de personnes; et une méthode de suivi de modèle 2-D dont l'apport principal est une approche de la modélisation dynamique basée sur l'apprentissage.

L'approche 2-D coupe les liens avec la reconstruction en profondeur et rend difficile l'application des contraintes de forme et de texture 2-D qui sont engendrées par la rigidité 3-D. La rigidité 2-D ne s'applique pas, en raison des mouvements 3-D hors du plan de la caméra qui engendrent le raccourcissement visuel de l'image du membre. Les modèles 2-D prennent donc la forme d'un ensemble de régions 2-D qui représentent les membres du corps, connectées par des articulations comme dans le cas 3-D, et sujet soit aux déformations 2-D affines (modèle « cardboard people » [JBY96]), soit aux déformations de raccourcissement relative le long de leur axe principale (par exemple, le « scaled prismatic model » de [CR99]). Ce deuxième modèle est moins flexible que le premier mais considérablement plus stable, et nous l'avons adopté ici.

3.1 Détecteur articulaire d'humains

La détection des personnes dans les images reste un des défis majeurs de la reconnaissance visuelle. La difficulté principale est la grande variabilité d'apparence du corps humain articulé. Les approches qui adoptent les gabarits plus ou moins rigides – on y compte les gabarits moindre carrées classiques, les approches « exemplaires » [Gav00, TB01] et aussi les « détecteurs de piétons » qui sont basées sur d'indices image extraites d'une fenêtre fixe [Pap97, MPP01, DCdB+02, VJS03] – s'adaptent difficilement aux variations de forme engendrées par les articulations, au moins avec les bases de données d'apprentissage de taille abordable à présent.

Une autre approche consiste à modéliser les articulations et en rechercher les configurations articulées qui collent au mieux à l'image – un problème de recherche de dimension élevée, mais qu'on peut simplifier en exploitant la structure articulaire arborescent du modèle. En particulier, on peut retrouver la configuration la plus probable du modèle 2-D dans une région donnée de l'image par une approche programmation dynamique récursive [FE73, FH00, IF01]. Cette méthode fonctionne

FIG. 3.1 – *À gauche: Une image d'apprentissage pour notre détecteur articulaire de personnes, étiquetée à la main par la pose correspondante du modèle articulaire. À droite: les sous-images normalisées des 15 segments du corps (y compris une image à base résolution du corps entier). Un détecteur d'objet de type machine à vecteur de support est appris pour chaque classe de segment. Ces détecteurs fournissent la vraisemblance image pour l'étape de programmation dynamique qui assemble dans chaque image de test la configuration du corps la plus probable.*

de la façon suivante. Avec chaque membre (représenté par une région 2-D avec texture, transitions de bord, *etc*), on parcourt l'image à toutes les positions, orientations et échelles possibles (une discrétisation est nécessaire), en calculant une carte dont les entrées sont les vraisemblances totales maximales qui sont atteignables « à cette position » au membre et à son arbre de descendants cinématiques. C'est-à-dire, chaque position de la carte contient la vraisemblance de la configuration optimale de l'arbre, si elle est enracine à cette position et si les descendants prennent leurs positions relatives optimales correspondantes. Pour un membre terminal (une main, un pied), la carte contient simplement la vraisemblance image du membre à la position donnée. Pour un membre interne, sa vraisemblance image est combiné avec les maxima des vraisemblances arborescentes qui sont possibles à chacun de ses fils directs, sur l'ensemble de leurs positions relatives possibles. (L'utilisation de cartes de vraisemblance pre-calculées évite une recherche dont le coût est exponentielle lors de ce calcul). Arrivé à la racine de l'arbre cinématique, on peut évaluer directement la meilleure position de la racine, et redescendre l'arbre afin de tracer la configuration correspondante.

Nous avons étudié une méthode de ce type, dont l'originalitée est de remplacer les mesures de vraisemblance image relativement simplistes qui ont été utilisées jusqu'à présent – par exemple, [FH00] supposent que les membres sont vêtus d'une couleur claire connue – par des détecteurs de membre issus des techniques de l'apprentissage. Voir la figure 3.1. Nous avons développé deux types de détecteurs basés sur les descripteurs image de type grille normalisée d'énergie de gradient et de laplacien, l'un appris par une Machine à Vecteur de Support (MVS) et l'autre par notre algorithme de machine à vecteur de pertinence (« Relevance Vector Machine » – une méthode d'apprentissage bayésienne creuse analogue à la MVS). Ce genre d'approche donne des résultats intéressants pour des problèmes de détection d'objets rigides ou presque rigides tels que les gabarits

FIG. 3.2 – *Détections de personne triées par ordre de probabilité décroissant, et leurs log-vraisemblances correspondantes, selon la méthode programmation dynamique sur les détecteurs machine à vecteur de support. Trois détections chacune pour trois personnes sont montrées, sur la base de piétons de MIT.*

« piétons » rigides [Pap97, MPP01]. En pratique, l'approche articulée – voir la figure 3.2 – permet de étendre ces résultats sur une gamme plus étendue de poses, mais en revanche elle est significativement plus relativement lourde à mettre en oeuvre parce qu'il faut lancer les détecteurs de tous les membres à toutes les orientations possibles (en plus du parcours position - échelle traditionnel). Dans cette application, les détecteurs basés sur la MVS semblent marcher mieux que ceux de la machine à vecteur de pertinence.

Ce travail fut publié au *2002 European Conference on Computer Vision* [RST02] (article #5 du mémoire associé).

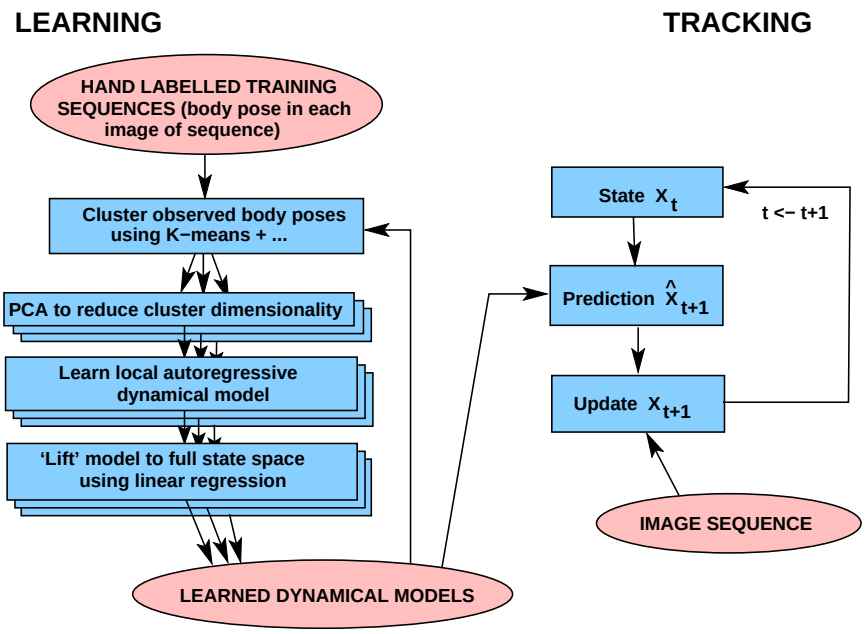


FIG. 3.3 – *Une vue d'ensemble du processus d'apprentissage de notre modèle dynamique pour le suivi 2-D.*

3.2 Modélisation dynamique pour le suivi 2-D du mouvement humain

Les mouvements humains peuvent être rapides et très variés, et il est souvent difficile de les suivre en raison de leur complexité, des occultations fréquents qui limitent l'observabilité, et de leur vitesse qui provoque non seulement de grandes déplacements entre images, mais aussi un flou de mouvement significatif. Un bon suivi demande un modèle dynamique qui est à la fois prédictif et bien adaptée aux variations d'apparence. On ne peut pas utiliser le même modèle d'apparence 2-D pour les vues de face et de côté, et le suivi doit aussi s'adapter aux changements d'aspect du modèle. Par exemple, quand le sujet avance, tourne, et retrace ses pas, le modèle doit suivre non seulement le mouvement complexe du tournant, mais aussi deux changements d'aspect – de la vue du côté gauche, à la vue d'en face, à la vue du côté droit.

Pour faire face à ce problème, nous avons développé une approche adaptative qui permet l'apprentissage d'un modèle dynamique performant qui incorpore les transitions entre aspects. Un modèle 2-D de type « scaled prismatic model » qui suffit pour représenter les différents poses et aspects est créé et ajusté (pour l'instant à la main) sur une base de séquences d'apprentissage. À partir de cette information, un modèle auto-régressif linéaire par morceaux est appris, dont les différentes régions linéaires encodent les différentes zones dynamiques et/ou aspects du modèle. L'apprentissage se fait de la façon suivante (voir la figure 3.3). (*i*) L'espace des paramètres est partitionné par un groupement « K-means » initial. Cette partition segmente les trajectoires en morceaux, et associe les segments similaires des différentes trajectoires. (*ii*) Dans chaque partition, la dimension de l'espace des paramètres est réduite par une analyse en composantes principales (ACP) linéaire; un modèle auto-régressive linéaire stabilisé de deuxième ordre est appris; et le modèle qui résulte est « haussé » à l'espace original en inversant l'ACP. (Afin de ne pas reprojeter à chaque étape le système dans le sous-espace ACP, le modèle haussé prédit en effet le changement d'état et pas l'état lui même). Ce processus de réduction-rehaussement permet une estimation plus stable du modèle auto-régressif. (*iii*) Un processus itératif analogue à l'Expectation-Maximisation met à jour la par-

tition, en regroupant les exemples d'apprentissage selon le modèle linéaire qui les prédit le mieux, et en re-estimant les modèles selon les exemples qui y sont attribués.

Le modèle final est capable de suivre la marche, la course, et quelques transitions simples entre les aspects du modèle comme un tournant. La figure 3.4 montre quelques exemples. Cependant, la méthode reste expérimentale et une implantation plus performante reste à faire.

Ce travail avec mon doctorant Ankur AGARWAL fut publié au *2004 European Conference on Computer Vision* [AT04d] (article #6 du mémoire associé).

3.3 Conclusions et perspectives

Ce chapitre a présenté deux méthodes de traitement d'images d'humains fondées sur la modélisation 2-D articulaire: un détecteur où la programmation dynamique sélectionne et articule des hypothèses de pose des membres du corps issues des détecteurs de membre de type machine à vecteur de support; et une méthode de suivi articulaire 2-D dont l'originalité est d'apprendre un modèle dynamique non-linéaire qui épouse les détails du mouvement humain.

Le détecteur actuel a plusieurs limitations. Même avec une discrétisation grossière de l'espace de recherche, il est lourd en temps de calcul parce qu'il lance tous les détecteurs de base à tous les angles possibles (ainsi qu'à toutes les positions et à toutes les échelles image, comme ailleurs). Aussi, il n'intègre ni modèle d'occultation ni modèle d'interaction d'apparence entre les différents membres, ce qui limite ses performances. Afin d'alléger le calcul, il serait intéressant de voir si on ne pouvait pas remplacer les détecteurs individuels par un classificateur unifié (par exemple une méthode de type arbre de décision) qui traite à la fois toutes les poses de tous les membres. Aussi, une discrétisation fine de l'espace de recherche (position-échelle-angle) étant hors de question en raison de sa dimension élevée, on peut probablement améliorer la qualité des résultats en incorporant dans la recherche de sous-arbres cinématiques une étape d'optimisation locale qui raffine les solutions partielles prometteuses.

En ce qui concerne la méthode de suivi 2-D, il est prévu de réimplanter les fonctions traitement d'image, d'y ajouter une (re)initialisation automatique basée sur la détection de personnes, et ainsi d'augmenter significativement l'ensemble d'exemples d'apprentissage afin de rendre plus robuste le suivi et de traiter une gamme plus étendue de mouvements. Nous voudrons aussi étendre l'approche au suivi de geste, où le corps entier n'est pas forcément visible et il devient plus important de suivre les détails des mains et du visage.

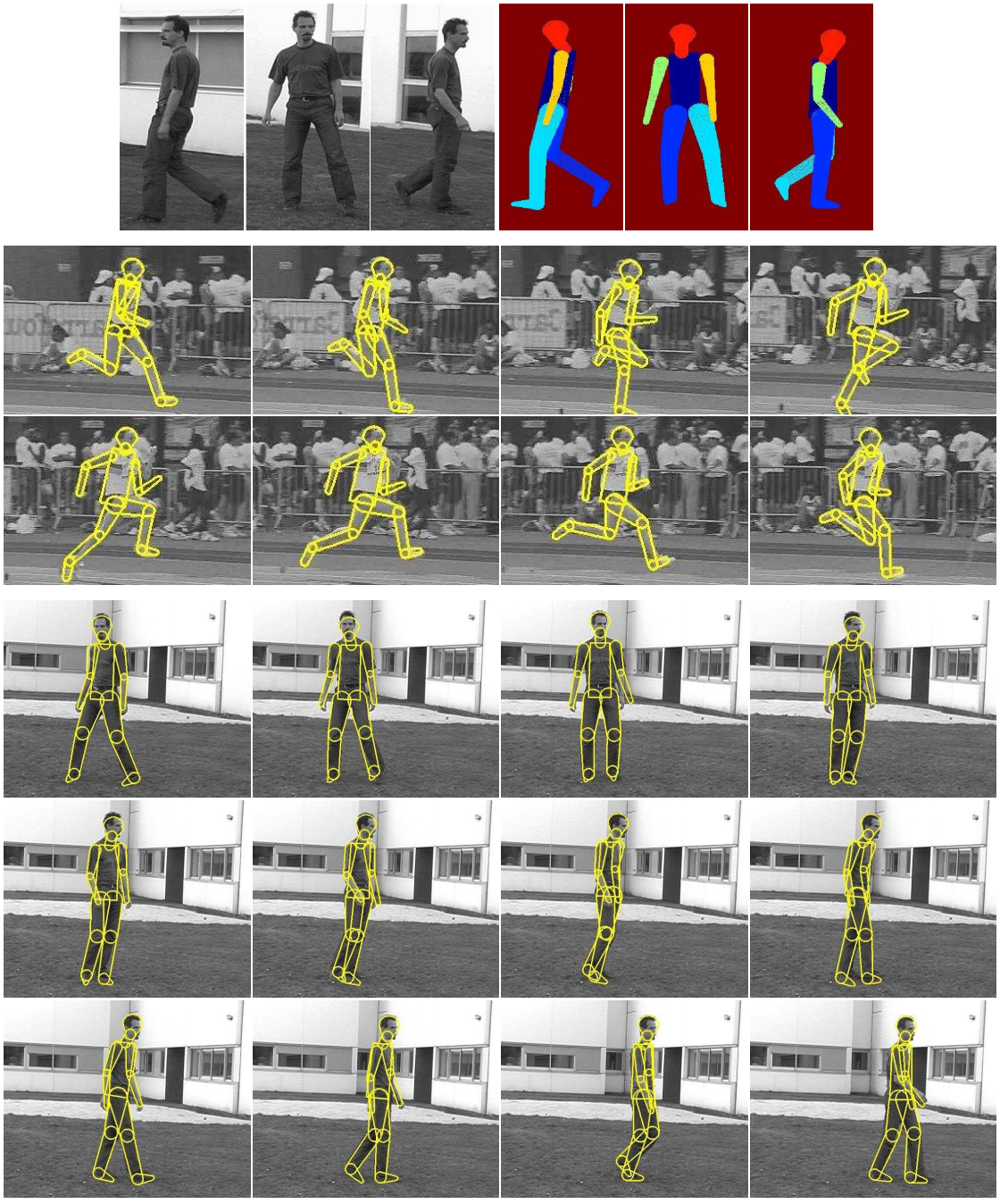


FIG. 3.4 – *Quelques exemples du suivi 2-D avec notre modèle dynamique. Première ligne: Avant la phase d'apprentissage, la configuration du modèle est marqué à la main dans chaque image d'apprentissage. Ici on montre trois images et les configurations correspondantes du modèle, avec leurs cartes de visibilité. Lignes 2–3: Une athlète qui court. Le modèle a été appris sur une autre athlète, mais suit bien le mouvement de celle ci, sauf le bras gauche qui était invisible lors de l'étape d'initialisation. Dernière trois lignes: le suivi d'un tournant pendant la marche. Le modèle 2-D du corps change d'aspect – de la vue d'en face à la marche à droite – mais le modèle dynamique appris arrive à suivre le mouvement.*

Chapitre 4

Approche 3-D par apprentissage

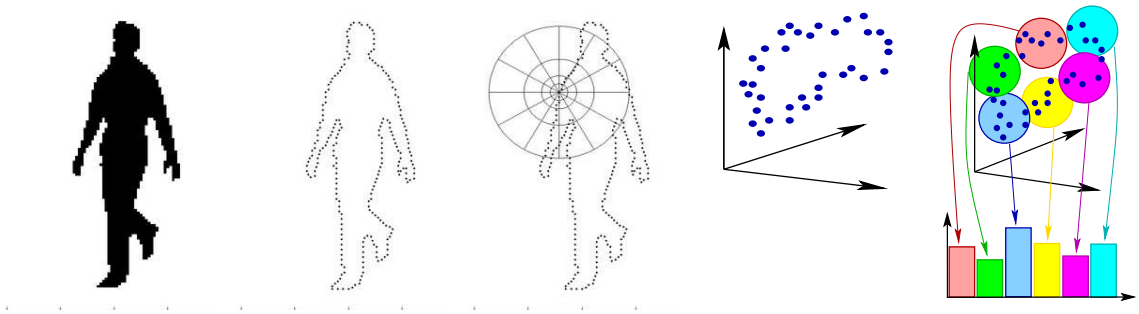
4.1 Introduction

Revenons sur la question de la capture du mouvement 3-D, monoculaire et non-instrumentée déjà abordée au chapitre 2, où nous avons adopté une approche basée sur la modélisation explicite du corps humain articulé, l'enjeu principal étant l'optimisation des paramètres du modèle face aux problèmes de dimension, de minima locaux, *etc*. Cette optimisation s'est montrée possible grâce au développement d'algorithmes performantes, mais elle reste complexe a mettre en oeuvre et lourde en temps de calcul. La robustesse pratique du suivi à été améliorée considérablement, mais ne saurait pas encore satisfaire à la plupart des applications réelles. On peut légitimement se demander si l'approche basée modèle est la bonne.

Heureusement on sait que les humains savent suivre et reconstruire le mouvement humain sans difficulté, au moins de façon qualitative. Aussi, on pouvait constater que les méthodes citées précédemment passent beaucoup de leur temps à examiner des configurations qui – quoiqu'en principe *possibles* pour une personne au plan cinématique – n'ont rien d'habituel ni de confortable. Leur modèle est donc trop général: il faut trouver une façon de représenter ce qui est « typique » où « caractéristique » d'une personne, et ne pas essayer de représenter en premier lieu tout ce qui est possible en principe.

Ceci remet en question la nature du modèle. Pour décrire les poses qui sont possibles au plan cinématique, il suffit de se limiter à la « géométrie » (biomécanique, biométrie), mais pour la question plus subtile de ce qui est typique, il faut plutôt étudier le comportement humain *in vivo*. La géométrie seule ne suffit plus, et il devient très difficile de construire à la main un modèle adéquat. La solution est de faire appel aux techniques de l'apprentissage et de la modélisation flexible statistique, afin d'apprendre une représentation effective à partir d'une base d'exemples.

À ce point on peut aussi remettre en question la démarche « générative », où le modèle est surtout utilisé pour la synthèse d'images hypothétiques, qui sont ainsi comparées avec les images réelles afin d'inverser le processus et d'en déduire les paramètres cachés du modèle qui correspondent aux images observées. Cette démarche a certes ses avantages – elle est explicite et intuitive, et elle permet une utilisation très flexible du modèle – mais elle est plutôt indirecte, et en particulier elle ne met pas en évidence quels aspects du modèle sont critiques pour une inversion réussie, et lesquels sont superflus. Ne serait il pas possible d'apprendre un modèle (une fonction) « inverse » qui estime directement les paramètres voulus à partir de l'image d'entrée donnée? C'est-à-dire, ne serait il pas possible de créer une méthode de reconstruction du mouvement de forme « diagnos-



(a) extraire la silhouette

(b) échantillonner les points le long les bords

(c) évaluer leurs shape contexts

(d) distribution dans l'espace des shape contexts

(e) quantifi cation vectorielle de la distribution

FIG. 4.1 – *Le processus d'extraction de descripteur de silhouette. Les descripteurs « shape context » sont calculés à des points régulièrement espacés le long de la silhouette, et la distribution des réponses est encode dans une histogramme 100-D par quantification vectorielle.*

tique » plutôt que « générative ».

Le travail présenté dans ce chapitre représente un premier pas dans cette direction, le but du jeu étant de voir jusqu'où peut mener l'apprentissage pur adoptant une approche d'estimation de fonction entrée-sortie « boîte noire » minimaliste, sans modèle 3-D explicite. La formulation est entièrement diagnostique: on renonce ainsi à l'approche générative et à tous ses accessoires (le rendu d'image, l'optimisation de pose et de correspondance, *etc*). En revanche, on suppose l'existence d'un ensemble d'exemples à partir de laquelle on peut apprendre un modèle effectif qui est adéquat à la tâche à exécuter – dans ce cas, l'estimation de la pose et du mouvement 3-D d'une personne à partir d'images 2-D monoculaires.

4.2 Descripteurs de silhouette

Renoncer aux modèles explicites veut aussi dire renoncer à l'ajustement du modèle sur l'image. Ainsi, il faut extraire directement de l'image, sans l'aide d'un modèle, un jeu de descripteurs qui caractérisent bien la configuration image du corps du sujet. À présent notre approche se base sur la silhouette du sujet, parce que celle ci contient déjà une bonne partie de l'information géométrique disponible, et – au moins quand le fond est statique ou connu à l'avance – peut être extraite de façon relativement fiable.

Pour la méthode décrite ici, il nous a fallu un codage numérique robuste de la forme de la silhouette. Nous avons choisi la méthode suivante (voir la figure 4.1): à chaque point de la silhouette, un descripteur « shape context » [MM02] – un histogramme des pixels arêtes observés et dont les cases sont disposées de façon log-polaire autour du point central – est calculé; ceci engendre un ensemble de vecteurs dans un espace histogramme à 60 dimensions, où chaque vecteur code l'apparence d'une partie de la silhouette; la distribution de ces vecteurs est à son tour codée dans un histogramme à 100 dimensions [1] par la quantification vectorielle (selon 100 centres qui sont appris sur toutes les silhouettes de l'ensemble d'apprentissage par le groupement « K-means ») – en faisant le totale du nombre de vecteurs qui tombent dans la case de chaque centre et en normalisant, on produit un histogramme 100-D, ce qui est notre descripteur de silhouette.

En pratique, on trouve que cette représentation capture bien l'information géométrique qui est

1. Le nombre de dimensions / de cases ici n'est pas critique, pourvu que la forme de la distribution 60-D soit codée.

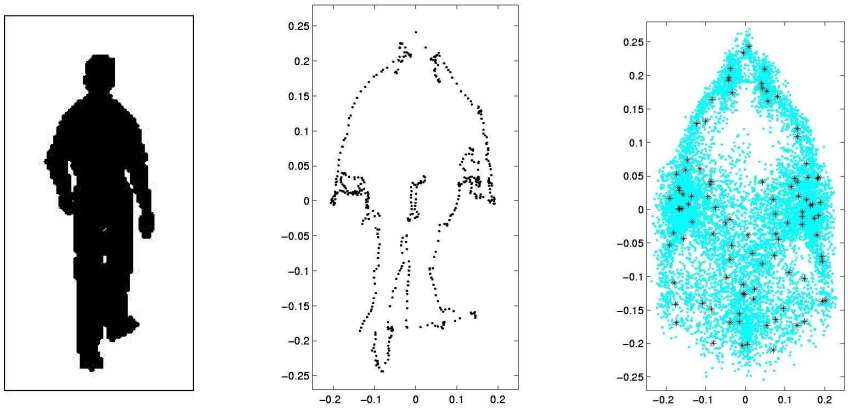


FIG. 4.2 – *Le descripteur réussit à encoder la forme de la silhouette. Par exemple, la projection de la réponse d'une silhouette sur les deux premiers composantes principales de l'espace des shape contexts redonne une image de la silhouette « perceptuellement déformée », ce qui montre que la position le long de la silhouette a été bien captée. À gauche, on montre une silhouette d'entrée, au centre, sa reprojection déformée 2-D, et à droite, la distribution des réponses de toutes les silhouettes de la base d'apprentissage selon cette reprojection, avec les centres des bennes de quantification vectorielle indiqués par des étoiles.*

disponible à partir de la silhouette. On choisit les shape contexts à environ la taille d'une jambe, pour qu'ils présentent un bon degré de localité et donc de robustesse contre les défauts de segmentation de la silhouette. En effet, on trouve que les shape contexts encodent non seulement la forme locale de la silhouette, mais aussi la position de leurs centres le long d'elle : une projection sur les deux premiers composantes principales récupère une version « perceptuellement déformée » de la silhouette – voir la figure 4.2.

4.3 Approche statique

Nous décrivons d'abord la version statique de notre méthode, qui reconstruit la pose 3-D du sujet humain à partir d'une seule image statique. La pose 3-D est encodée par un vecteur d'angles d'articulation 54-D – une représentation redondante à trois angles par articulation, utilisé par les systèmes de capture de mouvement classiques [2]. On cherche à estimer ce vecteur à partir du vecteur 100-D de descripteurs de silhouette robustes qui a été décrit précédemment. L'estimation se fait par une fonction de régression multi-dimensionnelle non-linéaire régularisée, apprise sur une base d'exemples d'apprentissage. La base comporte une variété de séquences de capture de mouvement réelles, avec leurs vecteurs de pose 54-D et les silhouettes correspondantes [3]. Pour chaque exemple,

2. La méthode peut utiliser n'importe quelle représentation et la redondance ne lui pose pas de problème particulier. Ici on adopte la représentation capture de mouvement d'origine pour être simple, et parce qu'elle peut être importée directement dans les logiciels de modélisation graphique comme POSER et MAYA.

3. Malheureusement, rares sont les systèmes de capture de mouvement actuels qui peuvent sortir à la fois les images et les vecteurs de pose. Quand nous avons débuté ces travaux, nous ne disposions pas de séquences de ce type, et nous avons dû synthétiser les silhouettes d'apprentissage à partir des poses réelles captées, en utilisant un logiciel de modélisation humaine (POSER de Curious Labs). Quoique cette approche n'est pas idéale, elle permet l'apprentissage à partir d'une gamme de points de vue et de gabarits de corps variés – ce qui aide à la généralisation – et elle fournit aussi des séquences d'essais dont les poses sont précisément connues. Nous soulignons que ceci concerne seulement la génération des données d'apprentissage : le système final ne fait aucun appel aux modèles de POSER.

(a) (b) (c) (d) (e) (f)

FIG. 4.3 – *Quelques reconstructions de pose fait par la méthode silhouettes statique. Les reconstructions (a–c) sont correctes, tandis que (d–f) montrent quelques erreurs caractéristiques de la méthode statique, liées aux ambiguïtés de la représentation silhouette. Dans (d), les jambes ont été inversées. Dans (e–f), deux solutions sont possibles et une solution intermédiaire entre les deux a été retournée. En pratique, environ 15% des résultats sont sujet à de telles ambiguïtés, les autre 85% étant corrects.*

le descripteur 100-D qui correspond à la silhouette est calculé, et la régression est apprise à partir des paires pose-silhouette (54-D, 100-D).

Nous avons évalué plusieurs différentes méthodes de régression. Notre descripteur de silhouette est déjà très non-linéaire et il se trouve qu'une régression linéaire aux moindres carrés régularisée donne déjà des résultats tout à fait intéressants. Cependant, la régression noyau – c'est-à-dire, à partir d'une base de fonctions de régression gaussiennes centrées sur les exemples d'apprentissage – améliore légèrement les résultats. Nous avons aussi étudié les régressions linéaire et noyau apprises par la machine à vecteur de pertinence (Relevance Vector Machine – RVM [Tip01]) et par la machine à vecteur de support régressive (Support Vector Regression – SVR [Vap98]). Ces méthodes sophistiquées donnent des résultats très similaires à l'apprentissage linéaire aux moindre carrés régularisé, mais – surtout dans le cas de la machine à vecteur de pertinence – la fonction de régression est significativement plus creuse, donc plus rapide à évaluer, ce qui devrait permettre une reconstruction en ligne de la pose humaine en temps réel [4]. Nous avons sélectionné comme méthode de référence la version noyau de la machine à vecteur de pertinence, mais les différences sont minimes au plan de la précision de reconstruction.

En pratique, bien que cette méthode donne des erreurs de reconstruction qui semblent être assez intéressantes, en terme de rendu graphique elle n'est pas très satisfaisante. En effet, le long d'une séquence, environ 85% des poses reconstruites sont correctes, mais les autres 15% sont aberrantes,

4. En effet, la régression tourne déjà en quelques millisecondes sous MATLAB. Seule l'extraction de descripteur de silhouette reste à optimiser.

FIG. 4.4 – *Quelques exemples des poses reconstruites par la méthode silhouette dynamique. En comparaison avec la méthode statique (voir la figure 4.3), toutes les reconstructions sont correctes. L'erreur moyen dans cette séquence de teste synthétique où le sujet marche dans une spirale décroissante est environ* 4.1° *par degré de liberté.*

ce qui provoque des « hics » visuels fréquentes qui sont très perturbants. La raison est assez évidente. Le problème de reconstruction de pose à partir d'une silhouette est intrinsèquement ambigu, non seulement en terme des ambiguïtés cinématiques citées précédemment, mais aussi en raison des ambiguïtés d'étiquetage des jambes et des bras, et des vues d'avant et d'arrière. Par exemple, pour une silhouette vue de côté, il n'est pas toujours possible de décider si c'est la jambe gauche ou la jambe droite qui est devant l'autre – voir la figure 4.6(d,e). La régression doit faire un choix forcé entre les solutions possibles, et elle choisit parfois mal, ce qui provoque les hics. Cette ambiguïté est intrinsèque à la représentation par des silhouettes et elle ne peut pas être enlevée à partir d'une seule image. La prochaine section explique comment une approche basée sur le suivi dynamique la corrige, et la section suivante présente une approche « mélange de régressions » qui sait au moins retourner les différentes solutions possibles.

4.4 Approche dynamique

Même dans le cas où plusieurs solutions pour la pose sont probables à priori, on peut espérer que la solution soit *localement* unique, c'est-à-dire que dans une région de taille suffisamment limitée de l'espace des paramètres, il n'y aura qu'une seule branche de l'espace de solutions, et l'ambiguïté disparaîtra. Dans ce cas, et si on peut apprendre une fonction de régression locale pour chacune de ces régions, une première estimation, même très grossière, de la pose peut être exploitée à fin de sélectionner la régression à utiliser, et la reconstruction devient unique. Dans le contexte du suivi du mouvement, la prédiction dynamique à partir des images précédentes fournit une estimation initiale appropriée.

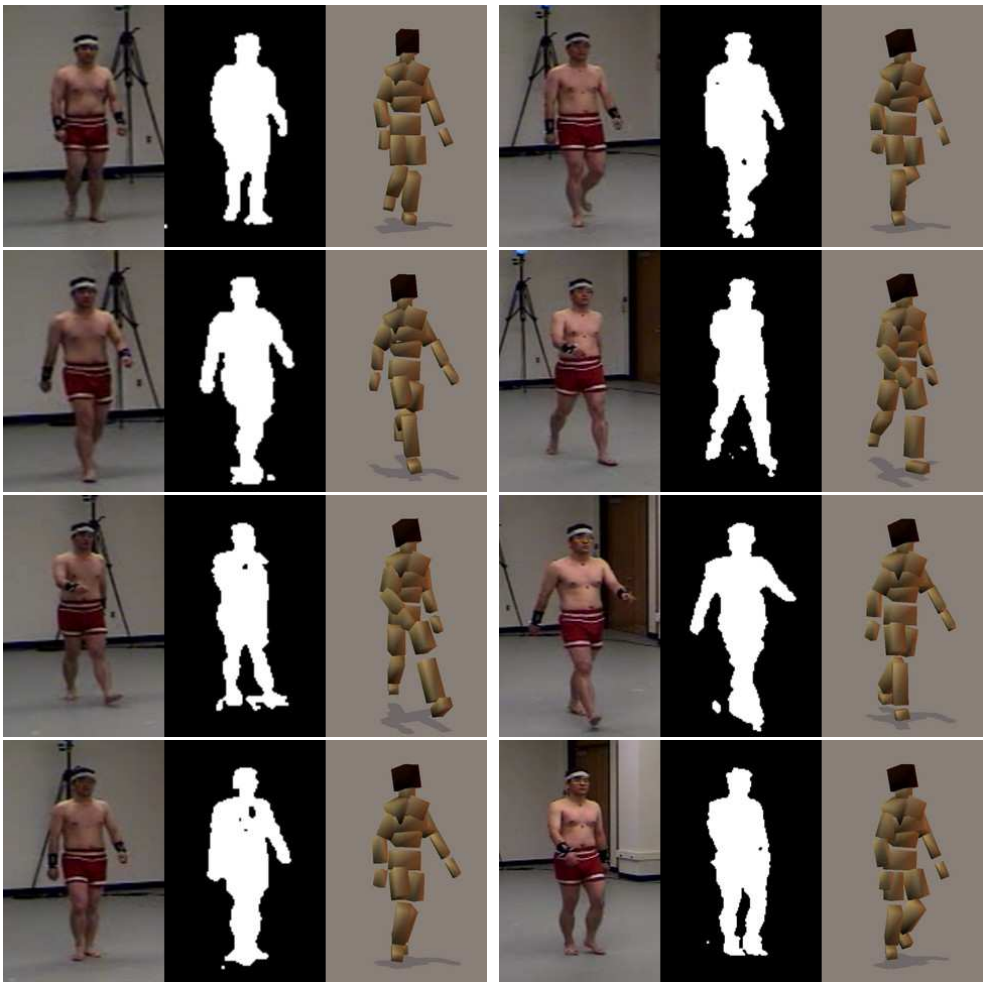


FIG. 4.5 – *Quelques résultats de la méthode silhouette dynamique sur une séquence réelle (obtenue de http://mocap.cs.cmu.edu/).*

Quoiqu'on puisse implanter la décomposition en régions et l'apprentissage des régressions correspondantes de façon explicite, la méthode des noyaux offre la possibilité de le faire de façon implicite sans avoir à décider combien de régions et de régressions sont nécessaires. Supposons que $\mathbf{x}$ est (le vecteur de descripteurs de) la silhouette observée, $\mathbf{y}$ est (le vecteur d'angles d'articulation de) la pose voulue, et $\mathbf{y}_0$ est une estimation initiale grossière de la pose (obtenue par la prédiction dynamique, par exemple). Une régression non-linéaire directe, $\mathbf{y} = \mathbf{f}(\mathbf{x})$, est sujette à l'ambiguïté, mais une régression de la forme $\mathbf{y} = \mathbf{f}(\mathbf{x}, \mathbf{y}_0)$ et basée sur les fonctions bien localisées en $(\mathbf{x}, \mathbf{y}_0)$ – par exemple les gaussiennes conjointes en $(\mathbf{x}, \mathbf{y}_0)$ – enlève l'ambiguïté. La localisation en $\mathbf{y}_0$ « sélectionne » la bonne région de $\mathbf{y}$ – c'est-à-dire, les $\mathbf{y}_0$ de la base d'apprentissage qui sont pertinents – et la localisation en $\mathbf{x}$ permet l'apprentissage d'une régression $\mathbf{x} \rightarrow \mathbf{y}$ correspondante. Il suffit de trouver un noyau en $\mathbf{y}_0$ qui est suffisamment étendu pour couvrir la région voulue en $\mathbf{y}$ (la zone d'une régression $\mathbf{x} \rightarrow \mathbf{y}$ locale), sans être si grande que les différentes solutions possibles en $\mathbf{y}$ y soient confondues.

Nous avons développé une méthode de ce type où le modèle dynamique est une auto-régression linéaire de deuxième ordre sur le vecteur d'angles d'articulation 3-D, et dont les coefficients ont été acquis par apprentissage. Dans ce cas, la prédiction dynamique $\mathbf{y}_0$ apporte aussi de l'information

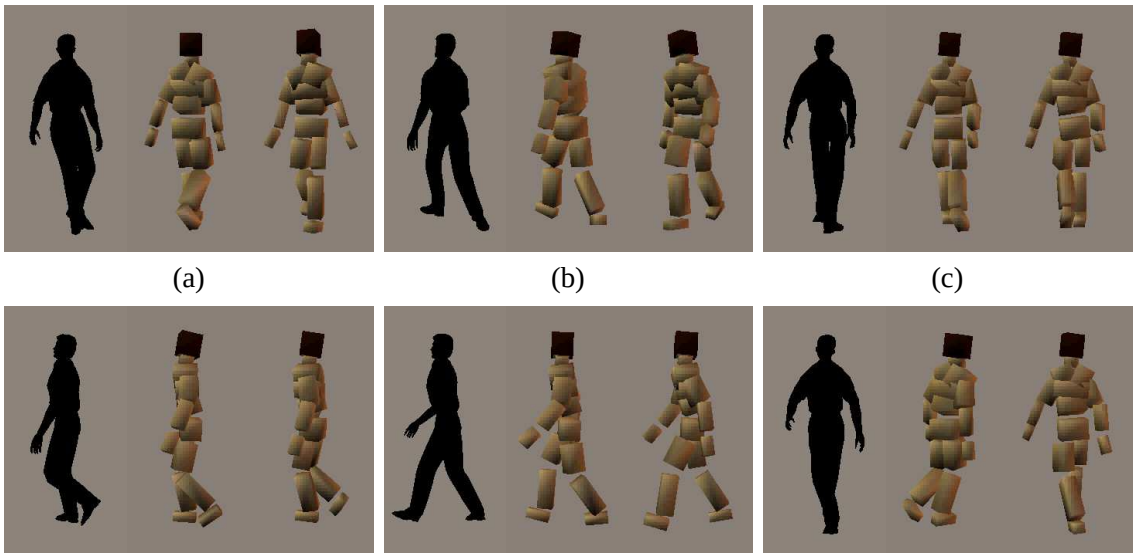


FIG. 4.6 – *Quelques exemples de la reconstruction silhouette multi-hypothèse. Chaque panneau montre une silhouette d'origine et les deux reconstructions qui sont jugées les plus probables par la méthode. (a,b) montrent des ambiguïtés face-dos; (c) montre une saute cinématique (échange avant-arrière) des deux jambes; (d,e) montrent des ambiguïtés d'étiquetage entre les deux jambes; et (f) montre un cas où la première solution est erronée suite à une mis-estimation de probabilité (la bonne solution est aussi trouvée, mais elle reçoit une probabilité plus faible).*

quantitative utile sur la pose $\mathbf{y}$, donc notre régression finale pour $\mathbf{y}$ est une combinaison linéaire de $\mathbf{y}_0$ et de $\mathbf{f}(\mathbf{x}, \mathbf{y}_0)$, la prédiction issue de la silhouette. Les coefficients sont acquis par apprentissage.

En pratique, cette méthode fonctionne très bien. Elle supprime presque entièrement les « hics » de la méthode statique, et elle donne une impression de reconstruction de pose 3-D continue et fluide pour un temps de calcul raisonnable. Les figures 4.4 et 4.5 montrent quelques exemples. En contrepartie, pour amorcer la chaîne de suivi de la méthode dynamique, il faut une valeur $\mathbf{y}_0$ initiale approximative. Nos expériences initiales indiquent que la méthode statique serait en mesure de fournir cette valeur, et une implantation automatique en ce sens est en cours. Aussi, l'incorporation de termes dynamiques semble avoir introduit une ralentissement dans certains sections du mouvement reconstruit, ce qu'on aimerait corriger.

4.5 Approche hypothèses multiples par mélange de régressions

Une autre façon de gérer l'ambiguïté de la silhouette statique est de proposer à l'utilisateur une choix de solutions possibles – choix qui s'intègrera naturellement, par exemple, au suivi multi-hypothèse robuste de type CONDENSATION. Nous avons implanté une méthode de ce type basée sur la régression multi-valeurs. La méthode est actuellement en cours d'amélioration, mais la version initiale utilise une mélange de plusieurs régressions linéaires globales (huit en l'occurrence), initialisée par une partitionnement K-means de l'espace des poses 3-D, et ensuite optimisée par Expectation-Maximisation (le modèle est en l'essentiel une mélange de gaussiennes). Pour les détails, voir [AT04c].

Cette méthode fonctionne relativement bien dans le sens où parmi les huit hypothèses qui sont

retournées, on trouve en général une bonne approximation de la vraie solution. La figure 4.6 montre quelques exemples. Cependant, les probabilités postérieurs qu'attribue la méthode aux solutions proposées ne sont pas dans tous les cas un guide fiable à leurs vraisemblances relatives, et le fait de proposer un nombre fixe de solutions à chaque étape n'est pas très satisfaisant au plan intellectuel. Nous travaillons actuellement sur une méthode de régression multi-valeurs plus raffinée, qui semble résoudre ces problèmes.

4.6 Conclusions et perspectives

Ce chapitre a présenté trois méthodes de reconstruction de pose 3-D humaine à partir de silhouettes image monoculaires. Leur spécificité est de renoncer à la modélisation 3-D explicite et de se baser sur un ensemble d'exemples d'apprentissage (une pose 3-D réelle et la silhouette image correspondante). Ceci les permet de caractériser les comportements humains « typiques » – là où les approches à base de modèle n'arrivent qu'à limiter leurs recherches à la gamme beaucoup plus étendue de comportements qui sont « possibles en principe ». Nos méthodes réduisent la reconstruction de pose à une régression « boîte noire » apprise, basé sur un descripteur robuste de la forme de la silhouette. La première méthode implante la régression mono-hypothèse à partir d'images individuelles. Parfois elle marche bien, mais elle produite un taux significatif de valeurs aberrantes en raison d'ambiguïtés de la représentation silhouette. La deuxième méthode incorpore la première dans une boucle dynamique afin de corriger ce phénomène. Elle permet une bonne reconstruction du mouvement humain à partir d'une séquence d'images. La troisième méthode reprend l'approche statique mais implante une régression multi-hypothèse qui propose plusieurs solutions possibles au utilisateur, annotées avec leurs probabilités d'occurrence.

Ces travaux ont été fait en collaboration avec mon doctorant Ankur AGARWAL. Les deux premières méthodes ont été publiés dans deux papiers de congrès, au *2004 International Conference on Computer Vision and Pattern Recognition (CVPR)* [AT04a] (version statique), et au *2004 International Conference on Machine Learning* [AT04b] (version dynamique), et les trois méthodes sont le sujet d'un article soumis au journal *Pattern Analysis and Machine Intelligence (PAMI)* [AT04c] (article #7 du mémoire associé).

Les méthodes actuelles exigent l'extraction préalable de silhouettes plus ou moins nettes, et elles ne exploitent pas d'autres indices image – telles que la visibilité du visage et l'ordre apparent d'occultation des membres – qui peuvent lever l'ambiguïté de la représentation silhouette. Il serait intéressant de combiner l'approche régression – surtout la version multi-hypothèses qui fournit non seulement une prédiction de la pose, mais aussi une probabilité associée – avec d'autres jeux de descripteurs image (notamment les descripteurs de histogramme de gradient issus de certains de nos travaux récents), afin de créer un détecteur de personnes qui retourne à la fois la probabilité d'avoir vu une personne dans la fenêtre donnée, et une estimation de la pose 3-D de cette personne, si elle a été détectée. En effet, pourvu qu'il n'y ait pas trop de fond, la méthode shape contexts actuelle donne déjà des résultats intéressants si on remplace la silhouette d'origine par une image naturelle du sujet…

Il faut aussi souligner que par leur nature – et en contraste de l'approche modélisation explicite – les méthodes à base d'exemples ne sont pas capables de reconstruire les mouvements qui s'écartent trop de ceux de la base d'apprentissage. Les prédictions sont satisfaisantes à l'intérieur de l'enveloppe des poses d'apprentissage, mais (au moins en ce qui concerne les méthodes à noyau) leur capacité d'extrapolation reste assez limitée. Par exemple, dans notre base d'exemples d'apprentissage principale, le sujet déplace relativement peu ses bras quand il marche, et si on essaie de reconstruire un sujet qui tourne beaucoup plus ses bras, on trouve que la reconstruction suit le

mouvement de façon naturelle, mais avec un déplacement trop limité des bras. Il reste à voir si une base d'exemples adéquate au suivi d'une gamme de mouvements complexes (la danse, les sports) serait de taille abordable. Déjà on peut dire que les méthodes à noyau creuses telles que la machine à vecteur de pertinence réduisent de façon très significative le nombre d'exemples à stocker, et que la taille du modèle actuel (qui est limité pour l'essentiel à plusieurs types de marche et de tournant) reste assez modeste.

En effet, les approches à base de modèles et à base d'exemples ont chacune leurs propres atouts, et on peut envisager une combinaison des deux, où une méthode à base d'exemples fournit l'initialisation robuste d'un modèle explicite, qui est ensuite ajusté afin d'optimiser le rapport modèle-image.

Chapitre 5

Perspectives et problèmes ouverts

Les trois chapitres précédents ont présentés trois façons de s'approcher de la reconstruction du mouvement humain à partir d'images monoculaires: respectivement, l'approche 3-D à base de modèle explicite, l'approche 2-D à base de modèle image, et l'approche 3-D par apprentissage, sans modèle explicite. Pris dans l'ensemble, ces travaux et leurs analogues dans d'autres équipes [SBS02, SBR+04, DBR00, IF03, FH00, CR99, HLF00, IF01, Bra99, MM02, SVD03, LC01, SC02, SEC02, LESC04] ont avancé considérablement l'état de l'art en reconstruction monoculaire du mouvement humain, mais le problème reste ouvert. Les étapes principales – l'initialisation de pose et d'apparence, le suivi 2-D, et la reconstruction 3-D – restent toutes les trois délicates, et on est toujours assez loin d'un traitement automatique et fiable des séquences typiques de la vie quotidienne, des films, des évènements sportifs.

Il va sans dire que nos algorithmes de traitement d'image, de mise en correspondance, de modélisation, seraient toujours perfectibles, mais pour moi, la question la plus pressante à l'heure actuelle est comprendre comment sortir le problème du laboratoire, où un modèle unique, monolithique et solitaire suit des mouvements exécutés en isolation dans un environnement simplifié. Il suffit de regarder un centre-ville, un bureau, un film pour comprendre à quel point il est rare de voir une personne bien isolée, entièrement visible sans occultations, à la bonne résolution image, exécuter une série de gestes nettes. Pour s'approcher plus de la réalité, il va falloir comprendre comment gérer les niveaux de détail, les vues partielles et occultées, les interactions entre sujets. On ne peut pas s'attendre à tout voir, donc il faut trouver comment travailler avec des modèles imprécis et partiels ou partiellement initialisés. Si on accepte que ce qui n'a pas été vu ne peut pas être reconstruit, au moins il ne devrait pas gêner la reconstruction de ce qui est bien visible, comme c'est le cas avec nos modèles actuels.

Ceci implique la disparation du modèle monolithique du corps, ce qui se verra remplacé – au moins dans un premier temps – par un troupeau de modèles partiels spécialisés: dans une classe de gestes, un aspect visuel, un niveau de détail; d'une zone du corps reconnaissable au préalable (un bras, les deux pieds, l'espace gestuel des mains et du visage, une vue d'ensemble à basse résolution...). Dans le cas idéal, les modèles seraient relativement simples – comme c'est le cas pour le modèle de reconstruction régressif actuel, par exemple – mais ils seraient capables de reconnaître quand et où ils s'appliquent, et en partie de s'adapter aux apparences. Ils peuvent avoir un caractère hybride – à la fois diagnostique afin de suggérer, et génératif afin de vérifier – avec une composante apprise importante. Un réseau de liens spatiaux faibles entre les modèles assurera une cohérence globale approximative, et un processus consacré à l'articulation au sens fort « serrera les vis » quand l'évidence image le permet.

On peut voir ici une convergence avec la reconnaissance visuelle de classes d'objets où les mêmes motifs – un ensemble relâché de modèles locaux appris, chacun spécialisé dans un élément simple et reconnaissable de la scène – émergent. Seulement, dans le cas de la reconstruction du mouvement humain, l'articulation des contraintes et la cohérence globale sont les éléments centraux.

Deuxième partie

Autres travaux, 2000–2004

Chapitre 6

Introduction

La deuxième partie de ce document résume plus brièvement en trois chapitres thématiques les travaux sur divers sujets *autres* que le mouvement humain, que j'ai effectué entre janvier 2000 et septembre 2004. Quoique ce document mette l'accent sur la perception d'humains, les travaux présentés ici représentent plus que la moitié de ma production scientifique pendant cette période et je n'ai pas voulu les passer sous silence.

Le premier chapitre parle du traitement d'image et de la vision au bas niveau, le deuxième de la vision géométrique et de la reconstruction de scène, et le troisième de la modélisation statistique et de la reconnaissance des formes.

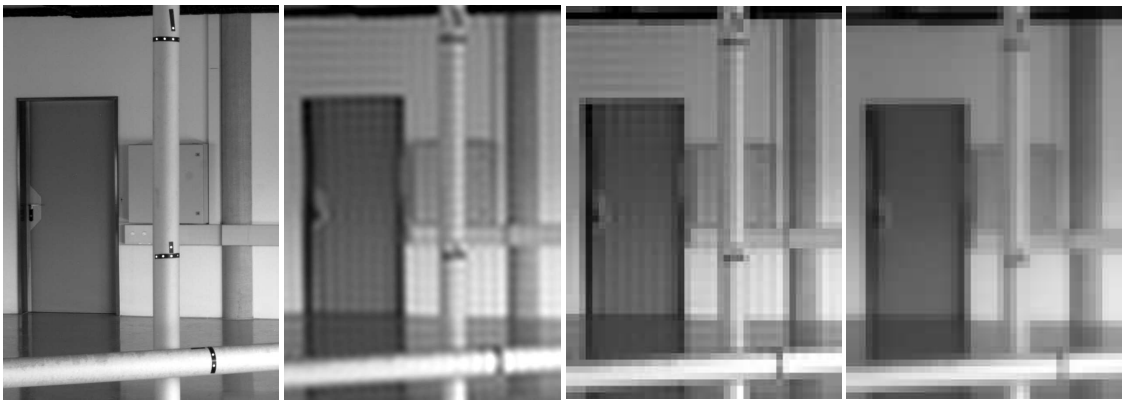
Chapitre 7

Vision de bas niveau et traitement d'image

Ce chapitre résume trois articles sur la vision de bas niveau. Le premier donne une approche apprentissage / optimisation empirique de l'interpolation et du re-échantillonnage d'images naturelles. Le deuxième présente un détecteur de points saillants multi-échelles, qui généralise l'approche populaire de Förstner et de Harris. Le dernier remonte le niveau un peu pour parler d'un modèle probabiliste de mise en correspondance géométrique.

7.1 L'interpolation d'image et le re-échantillonage sous-pixélique

En vision de bas niveau et en traitement d'image, il faut souvent soit interpoler, soit re-échantillonner, soit mettre en correspondance avec une précision sous-pixélique les images naturelles. Compte



(a) image d'origine *(b)* bande passante limitée *(c)* b.p. limitée + décimée *(d)* lissée + décimée

FIG. 7.1 – *L'effet de troncation abrupte de la bande passante. L'image d'origine (a) est limitée à* $1/15$ *de la bande passante (b), et puis décimée (c). Dans (b,c) on voit clairement les oscillations caractéristiques d'une image de bande passante abruptement tronquée. Ces oscillations gênent les algorithmes de mise en correspondance et de traitement d'image. Un lissage gaussien préalable de* $\sigma = 7$ *pixels avant la décimation donne une image de sortie légèrement plus floue, mais enlève les oscillations (d).*

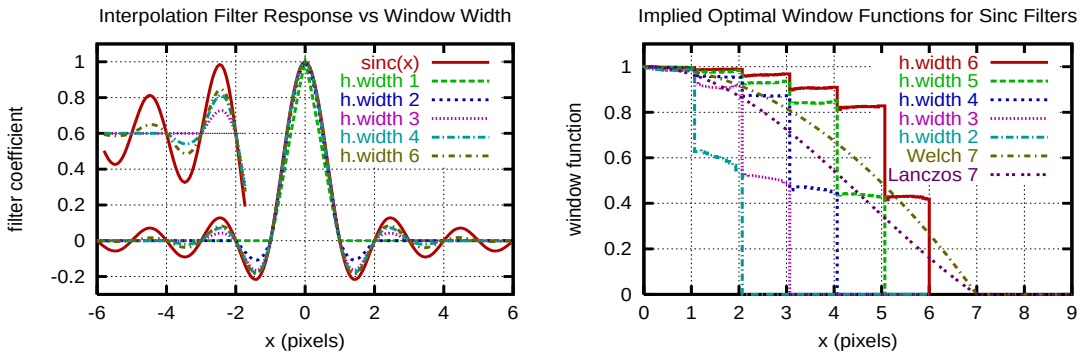


FIG. 7.2 – *À gauche: les filtres d'interpolation linéaires qui minimisent l'erreur d'interpolation empirique ont une forme oscillante similaire aux filtres classiques « sinc », mais avec des queues nettement moins étendues – 2–3 oscillations suffisent en pratique – et ils ont aussi de petits discontinuités de gradient à chaque traversée de zéro. À droite: si les filtres empiriques sont exprimés en terme des filtres* sinc *fenêtrés équivalents, les fonctions de fenêtrage descendent en escalier, et non pas selon l'enveloppe lisse prônée par les textes classiques de traitement du signal.*

tenu du nombre considérable des méthodes d'interpolation et de re-échantillonnage qui ont déjà été proposé, on peut légitimement se demander laquelle est la meilleure en pratique. Nous avons voulu trouver une réponse pragmatique à cette question, indépendante du traitement orthodoxe de l'échantillonnage, où la bande passante du signal est censée avoir été tronquée abruptement à la fréquence de Nyquist avant l'étape d'échantillonage, ce dont on déduit que les méthodes d'interpolation et de re-échantillonnage basées sur la fonction « sinc » ($\frac{\sin x}{x}$) sont les références par rapport à lesquelles toute autre méthode doit être jugée. Notre point de départ est le fait que les images naturelles ne se conforment pas à cet idéal « bande tronquée » – elles contiennent de nombreuses transitions d'intensité abruptes, ce qui engendre dans leurs spectres une décroissance $\mathcal{O}(1/f)$ caractéristique, non tronquée aux hautes fréquences. Réciproquement, implanter l'idéal Nyquist ne laisse pas aux images une allure très naturelle: tronquer abruptement la bande passante a tendance à créer des oscillations (dépassements) significatives, qui brouillent le signal et qui peuvent nuire aux performances des algorithmes visuels (voir la figure 7.1).

S'il faut tronquer plus doucement, comment et combien? Trop de lissage rend l'image floue et réduit la précision géométrique, et trop peu la rend oscillatoire et nuit aussi à la précision. En l'absence d'une théorie satisfaisante, nous avons décidé d'estimer empiriquement la forme de la fonction de re-échantillonnage qui minimise l'erreur d'interpolation sur une base d'exemples d'apprentissage. L'estimation se fait par optimisation numérique explicite.

Les résultats correspondent plus ou moins à une fonction séparable de forme analogue au « sinc » classique, mais limité à quelques oscillations seulement de chaque coté, et dont les dérivées ne sont pas continues lors des traversées de zéro. Ceci implique que les fonctions de fenêtrage équivalentes qui transforment le sinc idéal dans les filtres empiriques, ont une forme non-classique « en escalier ». Voir la figure 7.2. En effet, chaque intervalle des coefficients du filtre sert à estimer le pixel dans un intervalle différent à partir d'un ensemble différent de pixels d'entrée, donc on ne peut pas s'attendre à ce que les gradients soient continus et les fonctions de fenêtrage lisses.

Une autre conclusion de l'étude est le fait (connu des photogrammètres) qu'un flou optique léger – c'est-à-dire, appliqué *avant* l'échantillonnage pixélique de l'image – augment significativement non seulement la précision d'interpolation, mais aussi celle de la mise en correspondance spatiale par corrélation, en raison de la suppression d'aliasing le long des transitions abruptes d'intensité.

(Un flou d'environ $\sigma = 0.5$ pixels est l'idéal – juste suffisant pour supprimer l'aliasing visible le long des transitions fortes).

Ce travail fut publié aux actes du congrès « 2001 International Conference on Computer Vision » [Tri01a] (article #8 du mémoire associé).

7.2 Détection de points clés qui sont stables par rapport aux transformations géométriques

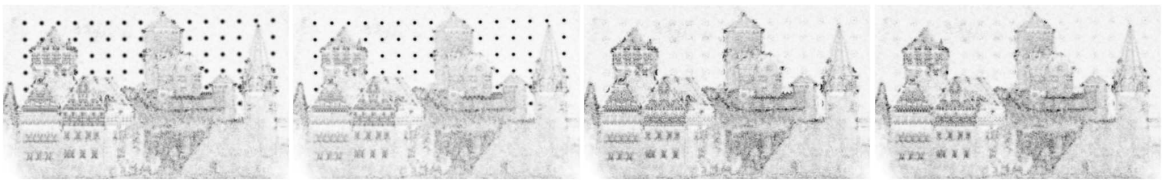
L'approche des tâches visuelles « indices locaux » a fait ses preuves pendant une vingtaine d'années, notamment pour la mise en correspondance d'images de scènes naturelles face à des variations importantes d'illumination et de point de vue [För86, TZ00, HZ00, SMB00, MS04, MTS+04]. L'idée est d'extraire de l'image un ensemble de taille modeste de « points clés » (c'est-à-dire, de régions locales saillants et isolés – on dit aussi « points d'intérêt »), et de baser la mise en correspondance, la reconnaissance, *etc*, sur ces points. La localité assure une bonne résistance aux changements d'illumination, de point de vue et d'occultation, et la saillance limite l'ambiguïté de la mise en correspondance. Se concentrer sur un ensemble limité de points allège le calcul, et facilite la relation avec la géométrie 3-D et la reconstruction de scène.

Un bon nombre d'opérateurs ont été développés pour la tâche d'extraction de points clés [Mor77, För86, HS88, Low99, Bau00, MS01, SZ02, MS02, KB01]. Chaque opérateur a ses propres atouts que nous n'allons pas détailler ici, mais un motif récurrent est la stabilité géométrique des points extraits. Afin de faciliter la recherche des points similaires, l'étape de détection est souvent suivi par une extraction de descripteurs locaux d'image autour des points clés. Plus ces descripteurs sont caractéristiques de l'apparence locale de l'image et invariants par rapport aux transformations et aux perturbations indésirables, plus la tâche de mise en correspondance est facile, et pour cela, la bonne stabilité géométrique des points de base est indispensable. Il est important de développer des détecteurs et des descripteurs qui assurent ensemble le niveau de stabilité nécessaire. Nous nous concentrons ici sur la conception de détecteurs.

Des détecteurs de points clés qui ont été proposés jusqu'à présent, ceux qui souscrivent à l'approche de Förstner [För86] sont parmi les plus réussis. Ces détecteurs *définissent* comme points clés, les points dont la stabilité géométrique locale est maximale. Par exemple, les détecteurs de Moravec [Mor77], de Förstner [För86, FG87, För94] et de Harris & Stevens [HS88] prennent tous comme métrique de saillance des critères qui caractérisent la *stabilité géométrique en translation* de la mise en correspondance locale – ils sélectionnent les points qui semblent offrir la meilleure « prise » géométrique en translation pour la mise en correspondance entre images. Sur le plan conceptuel, chaque région de l'image subit une recherche de mis en correspondance locale contre elle même – en l'occurrence, la somme des différences carrées non-normalisée est adoptée comme mesure d'erreur – et les régions dont l'erreur augmente le plus rapidement sous des déplacements infinitésimaux sont sélectionnées comme points clé. Sur le plan pratique, cette évaluation peut être réduite aux formules différentielles qui tournent autour du « tenseur de structure » – on dit aussi « matrice de moments quadratiques » – de la région image: $\int_{\text{région}} (\nabla I)\,(\nabla I)^\top \mathbf{dx}$, où $I(\mathbf{x})$ est le niveau de gris / vecteur de couleur de l'image au pixel $\mathbf{x}$.

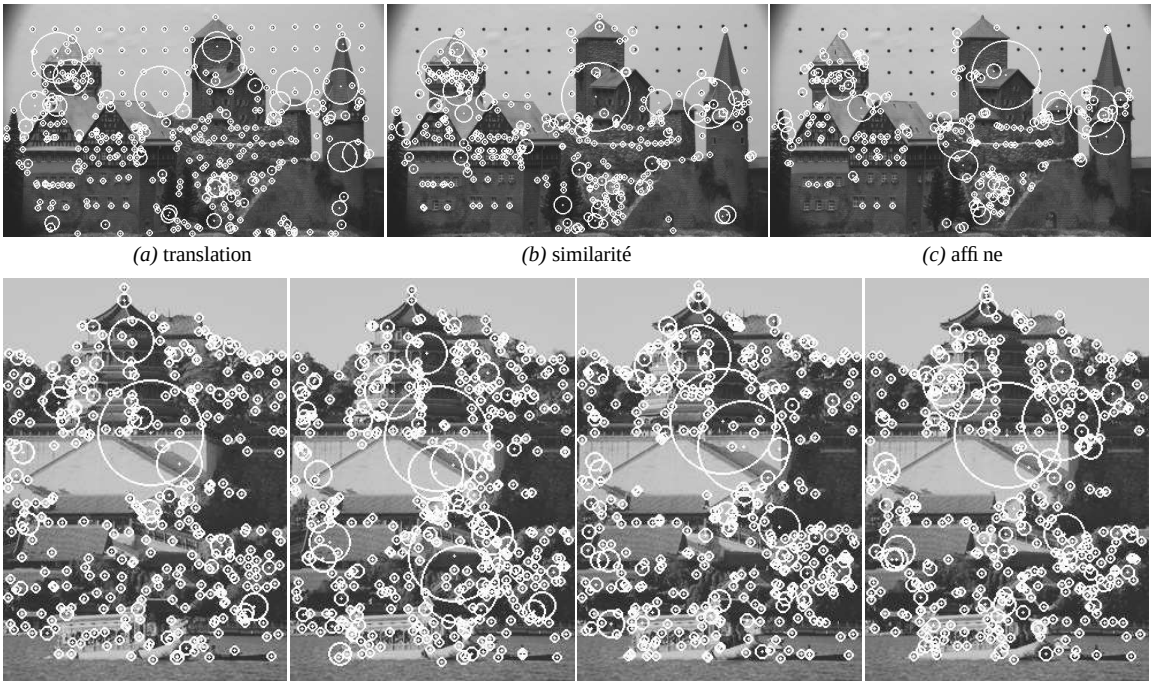
Ces approches donnent des résultats satisfaisants, mais elles n'assurent que la stabilité en translation. Or, on cherche aujourd'hui typiquement à avoir des descripteurs qui sont localisés non seulement en position, mais aussi en échelle, et ceci souvent sous des transformations de normalisation rotationelles et affines et les changements d'illumination.

Afin de produire des détecteurs qui garantissent la stabilité en présence des rotations, des changements d'échelle, *etc*, nous avons étendu l'approche de Förstner et Harris aux transformations



(a) translation *(b)* translation + échelle *(c)* translation + rotation *(d)* similarité

FIG. 7.3 – *Un exemple des cartes de stabilité de mise en correspondance locale, sous les transformations indiquées. Par exemple, si on demande que la mise en correspondance soit stable sous les rotations image, la réponse aux points circulaires baisse nettement – comparez les images (a) et (c). Chaque carte présente la valeur propre minimale du tenseur de structure généralisé correspondant. Nos détecteurs déclarent un point clé à chaque maximum suffisamment grand en position et en échelle de la carte de stabilité correspondante.*



(a) translation *(b)* similarité *(c)* affi ne

(d) translation *(e)* translation + rotation *(f)* translation + échelle *(g)* similarité

FIG. 7.4 – *Quelques exemples de points clé trouvés par nos détecteurs, sélectionnés pour être stables sous les transformations infinitésimales indiquées.*

géométriques locales générales, et en plus augmenté son invariance par rapport aux changements d'illumination. Le détecteur final permet de sélectionner – par exemple – les points qui ont un degré de stabilité désiré en présence des transformations de similarité, et qui résistent aussi aux changements d'illumination affines. La méthode incorpore au tenseur de structure des composant qui correspondent aux différentes degrés de liberté des transformées géométriques et de l'illumination que l'on considère. Suite à l'évaluation du tenseur généralisé à chaque point, un processus « complément de Schur » annule les effets de changement d'illumination, et une décomposition par valeurs propres du tenseur réduit permet d'estimer la stabilité géométrique généralisée de chaque point. Comme dans le détecteur de Harris multi-échelle [MS01], on parcourt la pyramide position-échelle des valeurs de stabilité, afin de trouver les points dont la stabilité locale est à la fois suffisante et maximale localement. Ces points sont les points clés retournés.

Le détecteur final donne des résultats intéressants, en particulier en ce qui concerne sa capacité à éliminer les points qui sont instables par rapport aux rotations 2-D de l'image. Voir la figure 7.3 pour un exemple des cartes de stabilité / saillance, et la figure 7.4 pour quelques exemples de détections. Pour l'instant ces résultats sont préliminaires. Une implantation perfectionnée est en cours, avant de lancer une évaluation plus conséquente.

Ce travail fut publié au congrès « 2004 European Conference on Computer Vision » [Tri04] (article #10 du mémoire associé).

7.3 Une approche probabiliste de la mise en correspondance géométrique

La mise en correspondance d'éléments homologues entre les différentes images d'un même objet ou d'une même scène est l'une des problèmes de base de la vision par ordinateur. Dans le cas rigide, les « contraintes d'appariement » tensorielles classiques – les contraintes géométriques qui lient entre elles les positions dans différentes images d'un même élément de la scène – sont l'outil de base. Cependant, ces contraintes s'avèrent parfois trop rigides pour décrire la scène – par exemple, si elle contient des mouvements indépendants; et parfois, elles ne peuvent pas être estimées de façon stable. En particulier, quand la parallaxe devient petite – soit pour une scène presque plane, soit pour une translation insuffisante de la caméra – un modèle plan (une homographie 2-D) peut être plus adaptée qu'un modèle 3-D (la géométrie épipolaire), mais il faut à ce point faire un choix de modèle avec toutes les risques et les complexités que cela implique.

Nous avons développé une approche novatrice au problème de mise en correspondance multi-images, qui s'applique indifféremment aux scènes rigides et non-rigides, profondes et peu profondes. Au lieu de se baser sur les contraintes géométriques idéales, elle adopte un modèle empirique et probabiliste. Pour une classe de correspondances donnée – par exemple les points image entre trois images données – elle cherche à caractériser leur loi de correspondance empirique par l'estimation de leur « Distribution Conjointe de Features (DCF) » – la distribution de probabilité conjointe de leurs positions image. La notion de la DCF capte de façon implicite mais efficace la trace image de la géométrie rigide de la scène, mais aussi son incertitude. Tous les aspects de la géométrie tensorielle ont leur analogue en terme de la DCF. La probabilité pour un ensemble quelconque de features d'être en correspondance est donnée directement par la valeur de la DCF sur l'ensemble. La recherche des correspondances potentielles et le transfert de features entre images sont exprimés très naturellement en terme de la probabilité conditionnelle – il suffit de conditionner la DCF sur les positions des features déjà connus afin de donner une DCF conditionnelle réduite pour les features non encore connus. La figure 7.5 illustre le processus. La reconstruction 3-D peut également être exprimée de façon naturelle en terme des distributions conjointes.

Quoique le modèle de base ne fasse aucune hypothèse de rigidité, on peut toujours proposer une forme paramétrique pour la DCF qui exprime une préférence pour la quasi-rigidité, la rigidité par morceaux, ou d'autres propriétés géométriques voulues. Dans l'article décrit ici, nous montrons comment convertir chacune des classes de contraintes d'appariement classiques tensorielles rigides dans un modèle DCF quasi-rigide équivalent, et aussi comment estimer ce modèle par une méthode analogue à l'estimation « linéaire » classique du tenseur d'appariement correspondant. Par exemple, pour la DCF des paires de points correspondants entre deux images, on peut proposer une forme « épipolaire » quasi-rigide, estimée à partir d'au moins 8 paires de points en correspondance (l'analogue de la méthode classique quasi-linéaire « 8 points » pour la matrice fondamentale).

Le modèle DCF « épipolaire » est considérablement plus général que le modèle épipolaire classique, mais il garde tout son intérêt dans une scène rigide où le modèle classique s'applique.

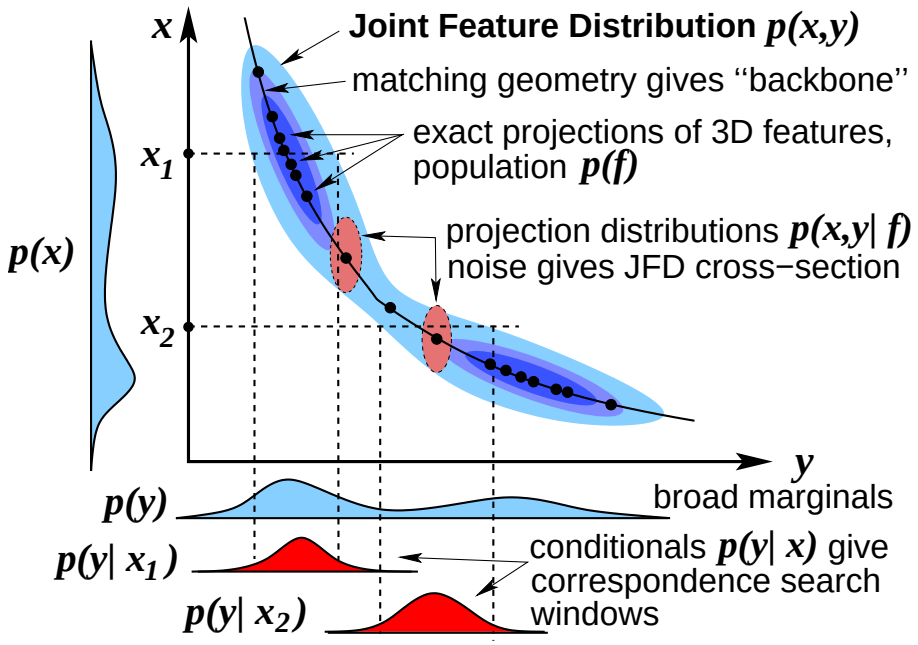


FIG. 7.5 – *La mise en correspondance « Distribution Conjointe de Features (DCF) » est basée sur l'apprentissage d'un modèle de la distribution conjointe des positions image des paires, triples,* etc, *de primitifs correspondants entre deux, trois,* etc, *images. En raison des contraintes (géométriques ou autres) qui lient les primitifs correspondants entre images, la DCF est en générale assez étroite dans l'espace conjoint, quoique ses marginales (projections) dans les images individuelles soient floues. (Dans le cas des contraintes géométriques rigides, l'épaisseur de la DCF est donnée par le bruit d'extraction de primitif). À partir du modèle, conditionner sur la position d'un primitif image donné prédit la distribution des positions de son/ses correspondant(s) dans la/les autre(s) image(s). Cette information peut être utilisée pour guider la recherche des primitifs correspondants.*

Nous avons déjà fait allusion aux faits que l'estimation de la géométrie épipolaire devient mal conditionnée dès que la parallaxe observée devient insuffisante, et que l'estimation d'un modèle homographique peut être intéressant en ce cas, mais que le choix entre ces deux modèles peut être délicat. Typiquement, il y a une zone de paramètres où aucun des deux modèles n'est satisfaisant. Le modèle DCF épipolaire évite ces problèmes parce que son estimation linéaire reste stable même dans le cas des scènes presque planes. Au lieu postuler une ligne épipolaire entière correspondante à chaque point dans l'autre image, la DCF estime plutôt une ellipse (une gaussienne) correspondante. Si la scène est profonde, ces ellipses sont très étendues et se rapprochent aux lignes épipolaires entières, et par contre, quand la profondeur se réduite, les ellipses se raccourcissent progressivement le long des lignes épipolaires. Si la scène devient plane, les ellipses deviennent des cercles centrés sur le point homographique correspondant, et dont la taille est le bruit image. Le modèle DCF épipolaire paramètre donc de façon stable non seulement le cas épipolaire classique, mais aussi le cas homographique et tous les cas intermédiaires. En plus, il prédit pour chaque point dans une image, une région de recherche du point correspondant dans l'autre image qui reflet la géométrie et le bruit observé. La figure 7.6 illustre les résultats sur des scènes rigides profondes et peu profondes, pour deux classes de mouvement de la caméra.

Ce travail a été publié dans les actes du congrès « 2001 International Conference on Computer Vision » [Tri01b] (article #11 du mémoire associé). Un appendice résume la géométrie algébrique qui sous-tend la construction.

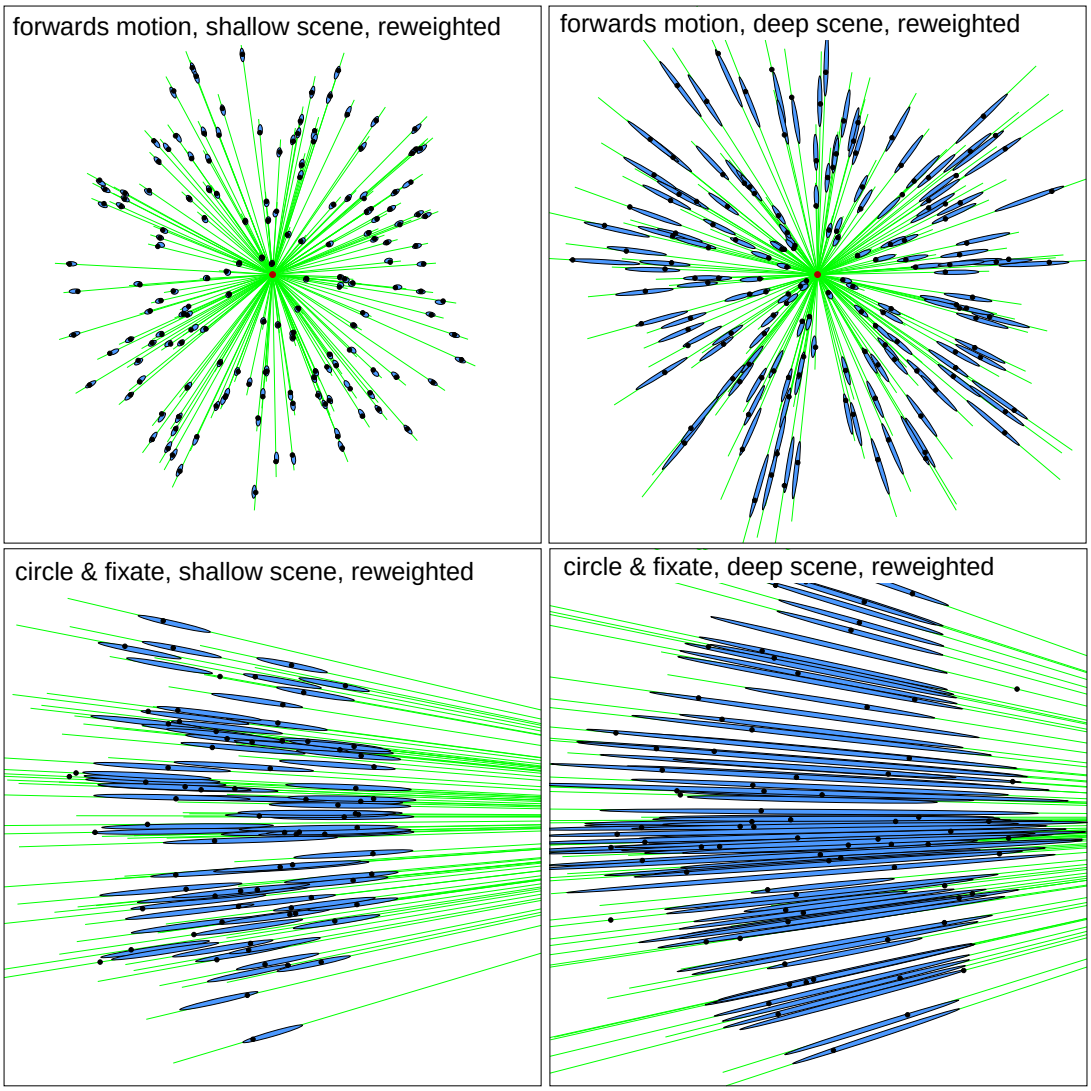


FIG. 7.6 – *Quelques exemples des prédictions de la position probable du point correspondant, produites par le modèle Distribution Conjointe de Features projectif deux images (l'analogue DCF d'une matrice fondamentale). La caméra avance (première ligne) et va de gauche à droite (deuxième ligne) dans une scène presque plane (à gauche) et profonde (à droite). Dans le cas de scènes peu profondes, le modèle s'approche d'une homographie image à laquelle s'ajoutent les ellipses d'incertitude de transfert de points. Pour les scènes plus profondes, les ellipses s'étendent progressivement pour s'approcher d'une géométrie épipolaire classique (à laquelle s'ajoutent encore une fois les incertitudes). Contrairement aux modèles homographiques et épipolaires classiques, l'estimation de la DCF reste stable dans tous ces cas, sans sélection de modèle, et la recherche des points correspondants le long des droites épipolaires s'adapte naturellement à la profondeur de la scène observée.*

7.4 Conclusions et perspectives

Ce chapitre a présenté trois articles sur le traitement d'image et la mise en correspondance: la première propose une approche empirique à la conception de filtres d'interpolation sous-pixelique d'image; la deuxième un détecteur de points clé qui généralise l'approche « stabilité translatio-

nelle » de Förstner et de Harris aux transformées géométriques et photométriques plus générales; et la troisième propose un modèle de mise en correspondance probabiliste qui généralise et stabilise les contraintes de mise en correspondance géométriques classiques.

La méthode d'interpolation d'image reste à valider sur un ensemble plus varié d'images, notamment sur des images dont les translations sous-pixeliques sont reelles et connues (et non pas synthétisées sélon la méthode de l'article). De nombreuses extensions sont possibles: aux transformées géométriques non-translationelles (changement d'échelle, rotation); aux images couleur (en fonction de la matrice d'échantillonnage des couleurs des pixels – grille Bayer RV/VB par exemple); et à l'évaluation du gradient et d'autres dérivés (dans ce cas il faut définir precisément ce qu'on entend par le gradient, c'est-à-dire ce qu'on entend reproduire avec la méthode). Plus généralement, la méthode actuelle reste linéaire, et il serait intéressant de voir si un modèl non-linéaire ne pouvait pas « reconnaître » le contenu local de l'image (une arête à telle position et à tel angle, par exemple) afin de mieux reproduire l'image voulue. Dans ce contexte, les travaux (beaucoup plus sophistiqués) de super-résolution baysienne de Freeman et de ses collaborateurs sont à citer [FPC00]. Aussi, il serait très intéressant de confronter ces méthodes essentiellement empiriques avec les approches théoriques « modernes » issues de la théorie d'ondlettes et de splines – notamment les travaux d'Unser et de Blu à Lausanne [TBU00b, TBU00a, Uns00].

En ce qui concerne le détecteur de points clé, une reimplantation raffinée est en cours avant de faire une validation plus complète. Comme pour tous les autres détecteurs que je connais (et malgré la dérivation rélativement sophistiquée de celui ci), la stabilité de détection sous les changements d'échelle de l'image ne semble pas encore être satisfaisante. Je soupçonne l'influence d'effets d'échantillonnage lors du calcul des dérivés. De toute façon, je suis convaincu que si on se limite à « l'échantillonnage cru », on n'arrivera jamais à l'extraction précise et fiable d'indices image géométriques – afin de s'addresser aux questions sous-pixeliques, il faut passer par un modèle d'interpolation d'image qui remplit de façon cohérente les lacunes entre les échantillons, et ainsi par les outils (dérivés, intégrals, *etc*) basés sur l'image continue résultante.

La mise en correspondance Distribution Conjointe de Features est assez générale et l'article ne traite que les formes paramétriques et les méthodes d'estimation « linéaires » les plus simples. En pratique il faut adopter une version robuste de ce modèle. Par exemple, afin de réduire la sensibilité aux correspondances aberrantes, on peut ajouter à la distribution actuelle les queues longues (voir constantes), et adopter une initialisation robuste de type recherche aléatoire RANSAC [FB81]. Aussi, quand la scène combine plusieurs modèles – par exemple une scène rigide dont les corréspondances se trouvent sur plusieurs plans, ou une scène 3-D qui contient plusieurs mouvements indépendants – le modèle DCF correspondant est un mélange et pour l'estimer il faut mettre en oeuvre l'Expectation-Maximisation robuste ou pareil.

Chapitre 8

Vision géométrique et reconstruction de scène

Ce chapitre résume quatre articles sur la vision géométrique et la théorie de reconstruction des scènes. Au plan thématique ils datent de la période directement après ma thèse.

8.1 Configurations critiques pour l'auto-calibrage

Si on étudie les méthodes de reconstruction de scène et de calibrage de caméra, on trouve qu'elles sont souvent assez sensibles aux détails obscurs de la géométrie, et qu'elles ont tendance à échouer pour certains géométries insuffisamment générales. Ces « configurations critiques » – configurations de la position et/ou des paramètres internes des caméras qui sont intrinsèquement singulières ou ambiguës – arrivent en pratique plus souvent qu'on ne l'imagine, et il est important de les comprendre, ne serait ce que pour les éviter. En particulier, nous avons été impliqués dans le développement de l'« auto-calibrage » – l'estimation de paramètres de calibrage de la caméra et/ou de la géométrie métrique de la scène, à partir des contraintes faibles telles que le fait que la calibrage interne de la caméra ne change pas pendant ses déplacements. Il s'agit d'un processus délicat où les configurations critiques sont particulièrement gênantes, et afin de mieux comprendre la situation, nous avons entrepris une étude algébrique de ses configurations critiques. Une combinaison de géométrie projective et d'algèbre géométrique effective (approche bases de Gröbner) nous a permis de caractériser les ensembles critiques de plusieurs problèmes d'auto-calibrage. Le problème est difficile et la caractérisation reste implicite dans la plupart des cas étudiés, mais dans quelques cas relativement simples – notamment les problèmes à deux images où seulement les deux focales restent inconnues – nous avons pu donner une caractérisation détaillée et des algorithmes de reconstruction explicites. Par exemple, la figure 8.1 montre (une partie de) l'ensemble critique des configurations de deux caméras dont seules les positions et les focales sont inconnues.

Ce travail fut publié au « Journal of Mathematical Imaging and Vision » [KTÅ00] (article #12 du mémoire associé), et représente la culmination d'une collaboration débutée quelques années plus tôt [KT99] avec Fredrik Kahl et Kalle Åström de l'Université de Lund.

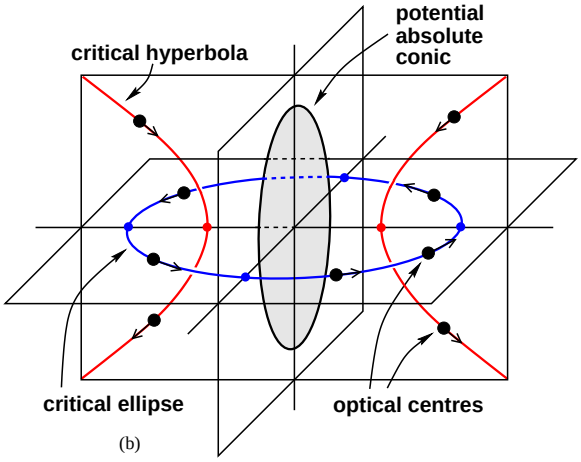


FIG. 8.1 – *Un exemple d'une configuration critique. À partir de deux images, et quand les positions et les focales des deux caméras sont inconnues, si les deux caméras tombent dans une configuration relative où elles sont toutes les deux sur une courbe composée d'une hyperbole et d'une ellipse dont la géométrie relative est connue, et avec leurs axes optiques tangents à cette courbe, la configuration est critique et les positions, les focales et la géométrie euclidienne de la scène ne peuvent pas être récupérées de façon stable.*

8.2 Méthodes algébriques d'estimation de pose de caméra

La résection – l'estimation de la position 3-D d'une caméra à partir de points observés dont les positions 3-D sont connues – est l'un des éléments de base de la reconstruction visuelle photogrammétrique. Quoique il existe déjà de nombreuses méthodes algébriques pour les cas les plus simples de ce problème, il fournit toujours un bon champs d'essai pour de nouvelles méthodes algébriques qui sont susceptibles d'être appliquées plus tard à d'autres problèmes plus difficiles.

Dans cette étude, nous avons appliqué quelques techniques récentes de l'algèbre géométrique effective, et notamment une méthode basée sur les résultants creux, au problème de l'estimation de la pose d'une caméra calibrée à partir de trois points 3-D connus (4 solutions en général) et de quatre points ou plus connus (solution unique en général - le défi étant de formuler une méthode algébrique numérique stable et efficace pour un problème redondant mais bruité).

Ce travail avec mon doctorant Marc-André AMELLER et avec Long QUAN fut publié aux actes du congrès « Reconnaissance des Formes et Intelligence Artificielle (RFIA) 2002 » [AQT02] (article #13 du mémoire associé).

8.3 Les liens entre les approches projective-tensorielle et plan + parallaxe

L'approche tensorielle fournit un formalisme complet pour exprimer la géométrie de la mise en correspondance et de la reconstruction projective multi-images, mais sa généralité a tendance à cacher la structure géométrique essentielle qui sous-tend l'algèbre. Une approche plus élémentaire est la réduction « plan + parallaxe », qui cherche à aligner toutes les images sur un plan de référence 3-D réel ou virtuel par l'application d'une homographie (transformation projective 2-D) appropriée, avant d'aborder les questions de mise en correspondance et de reconstruction 3-D. L'alignement rend considérablement plus élémentaire les calculs, au prix d'introduire une représentation parti-

tionnée où les éléments intimement liés sont séparés – un point 3-D est représenté par son point de référence sur le plan et son élévation (« parallaxe ») au dessus du plan – ce qui cache encore une fois la structure géométrique essentielle qui sous-tend l'algèbre.

Cependant, appliquer les homographies image ne change en rien la géométrie 3-D essentielle du problème. On peut donc voir chaque équation tensorielle comme une équation plan + parallaxe de base, modifié par des homographies additionnelles, et inversement. Aussi, si on choisit un système de coordonnées 3-D projectif où le plan d'alignement est placé à l'infini, la géométrie devient (une déformation projective de) celle des caméras calibrées en translation pure – un cas particulièrement simple où, moyennant l'estimation des facteurs d'échelle inconnus, l'algèbre devient celle des vecteurs 3-D standards. Avec cette représentation, les paramétrisations des objets géométriques principaux – points, droits, plans, caméras – et ainsi les contraintes d'appariement géométriques bi-, tri-, et quadri-focales, sont faciles à dériver.

Ces observations permettent de traduire entre les langages tensorielle et plan + parallaxe, et de mieux comprendre la structure géométrique essentielle des relations tensorielles. Elles suggèrent aussi la possibilité de développer de nouveaux algorithmes. Par exemple, l'article décrit une nouvelle méthode de reconstruction projective multi-images, par « factorisation des profondeurs projectives » plan + parallaxe.

Ce travail fut publié au « 2000 European Conference on Computer Vision (ECCV) » [Tri00] (article #14 du mémoire associé).

8.4 L'ajustement des faisceaux pour la reconstruction de scène

Cet article étendu (presque une petite monographie) présente un état de l'art de l'algorithmique du problème photogrammétrique d'« ajustement des faisceaux ». En reconstruction visuelle de scène, la plupart des algorithmes actuels qui prétendent à une précision élevée complètent leur solution initiale par une phase de raffinement des paramètres estimées – en l'occurrence, la structure 3-D de la scène et les paramètres internes et externes des caméras. Cette phase d'optimisation locale s'appelle « ajustement des faisceaux » (bundle adjustment), soit parce que tous les paramètres sont ajustés ensembles (donc « en faisceau »), soit parce que les rayons optiques point-caméra sont ajustés pour coïncider exactement « en faisceaux » aux caméras et aux points.

L'ajustement des faisceaux occupe une position centrale dans la reconstruction visuelle photogrammétrique. Quoique simple en principe, en pratique elle exige un déploiement considérable de la technologie de l'optimisation numérique, en raison de son caractère, sa structure, et la taille importante des problèmes à résoudre. Il n'est pas rare de avoir à estimer 10^5 paramètres de scène (coordonnées de points 3-D) et 10^3 paramètres de caméra.

Afin de réduire le coût de l'optimisation à une taille abordable, il est indispensable d'exploiter la structure caractéristique bipartite et creuse du problème, où – au moins dans le cas le plus simple – les points 3-D ne sont pas liés entre eux, sauf indirectement par le biais des caméras, et inversement les caméras ne sont liées entre elles que par le biais des points. Le problème a souvent d'autres niveaux de structure creuse. Par exemple, en photogrammétrie aérienne, le recouvrement du terrain local et par morceaux limite les liens indirectes entre caméras aux caméras voisines, et la connectivité au deuxième ordre (après l'élimination linéarisée des points 3-D) devient celle d'une grille irrégulière de caméras.

La division entre points et caméras suggère immédiatement l'algorithme naïf de base du domaine – une alternance entre les étapes de résection (reconstruction des caméras à partir des points 3-D connus) et intersection (reconstruction des points 3-D à partir des caméras connues). Cette alternance de base – elle a de nombreuses variantes – suffit pour les problèmes les plus simples, mais

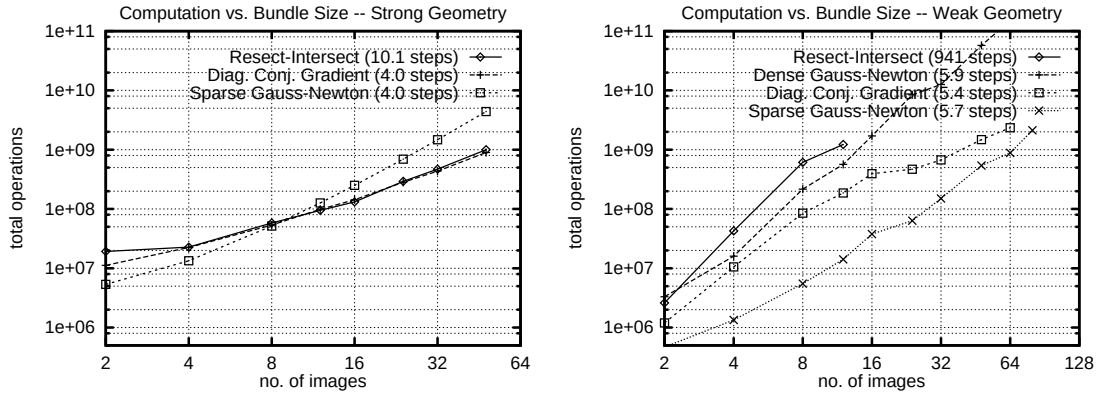


FIG. 8.2 – *L'efficacité relative des algorithmes d'ajustement de faisceaux a une forte dépendance sur la géométrie du problème. À gauche: les algorithmes simplistes comme l'alternance classique résection - intersection marchent bien quand la géométrie est forte – ici, une nuage sphérique de points 3-D, contournée par un quart de cercle des caméras qui voient tous les points. À droite: quand la géométrie devient plus faible – ici une ligne de caméras qui voient chacune une zone locale d'une bande étendue de points, pour simuler un « strip » de photogrammétrie aérien – les méthodes simplistes deviennent trop inefficaces, et il faut adopter un méthode de type Newton qui exploite la structure du Hessien. Voir la référence [TMHF00] pour plus de détails.*

dans le cas de problèmes plus difficiles elle montre très vite ses limitations. En effet, la plupart de problèmes de reconstruction visuelle réels et de taille significative sont assez mal conditionnés en raison des limites de recouvrement entre les caméras, et les approches par alternance gèrent mal ce mauvais conditionnement. À la place, il faut utiliser une méthode d'optimisation plus sophistiquée, et les algorithmes de type Newton Cholesky creux dominent dans les applications pratiques. L'article détaille les méthodes principales qui s'applique au problème, et donne quelques conseils pour choisir entre elles. La figure 8.2 illustre les comportements de plusieurs différentes méthodes dans les problèmes faciles et difficiles.

L'article considère également: la paramétrisation à adopter; l'intégration des contraintes; le traitement des dégénérescences « de jauge » (liées au choix du système de coordonnées, 3-D ou autre); la robustesse; l'estimation de niveau d'incertitude; et le contrôle de la fiabilité de la solution. En appendice, il fait un bilan historique du domaine, et donne quelques détails sur les méthodes numériques, un glossaire et une bibliographie étendue.

Cet article fut publié dans la collection [TMHF00] (article #15 du mémoire associé). Il a été conçu lors d'un panel sur le sujet au workshop ICCV'99 « Vision Algorithms: Theory and Practice ». Son but principal est de mieux répandre dans la communauté vision les connaissances photogrammétriques – entre autres, sur les limitations de l'approche résection-intersection et sur la choix d'algorithmes numériques – développés pendant une cinquantaine d'années, et éparpillés dans une littérature spécialiste parfois difficile d'accès.

8.5 Conclusions et perspectives

Ce chapitre a présenté quatre articles sur la vision géométrique et la reconstruction de scène: la première et la deuxième présentent respectivement deux études algébriques, sur les configurations critiques de l'auto-calibrage et sur l'estimation initiale de la pose d'une caméra calibrée; la troisième combine les approches tensorielle et « plan + parallaxe » de la reconstruction projective de

scène; et la quatrième présente un état de l'art de « l'ajustement des faisceaux » – l'optimisation numérique de reconstruction de scène visuelle.

Pour ma part, les études algébriques sont quelque peu décevants. Malgré une ingéniosité considérable et l'application des algorithmes état-de-l'art de l'algèbre géométrique, nous n'avons pu résoudre de façon satisfaisante que quelques problèmes les plus simples de l'auto-calibrage et de la pose de caméra. Nous avons en effet investi un effort considérable dans d'autres problèmes de ce type, sans aboutir à des résultats publiables – notamment: l'analogue trois images des contraintes d'auto-calibrage « de Kruppa » (c'est-à-dire, la partie *essentiellement* trois images de ces contraintes – indépendante des contraintes de Kruppa entres les 3 paires d'images); le problème « 4P3 » de la pose relative de 3 caméras qui voient 4 points 3-D quelconques; et le problème générale de l'incorporation de coefficients bruités et de données redondantes dans les méthodes d'élimination algébrique (résultants, bases de Gröbner...) Il semble qu'en raison du caractère combinatoire de l'élimination, on passe très vite du trivial au inabordable avec les outils de calcul algébriques actuels. Ce qui est dommage, parce qu'initialiser les algorithmes de reconstruction de scène de façon efficace et comprendre leurs configurations singulières a une utilité pratique réelle.

Inversement, l'approche plan + parallaxe me plaît. Je pense qu'elle représente une des façons les plus aptes à aborder la théorie des structures et des contraintes multi-images. En particulière, elle permet une séparation nette des structures 2-D et 3-D, ce qui facilite considérablement la compréhension du domaine.

L'étude d'ajustement des faisceaux montre les défauts d'une composition trop rapide – en particulier elle n'est pas aussi abordable aux débutants que j'aurais voulu – mais elle représente néanmoins un travail de synthèse conséquent dont certains résultats sont difficiles de trouver ailleurs. Il serait bien de lui refaire dans une forme plus abordable.

Chapitre 9

Modélisation statistique et reconnaissance des formes

Ce chapitre présent deux articles plus statistiques: l'un sur la théorie de la discrimination et l'autre sur la modélisation statistique pour la reconnaissance des formes en vision.

9.1 Moyennage entre les approches génératives et diagnostiques

Un débat statistique actuel porte sur les relations entre la modélisation générative / conjointe et la modélisation diagnostique / conditionnelle. Supposons que nous avons une variable observée $\mathbf{x}$ et une variable non-observée $\mathbf{y}$ à prédire (une classe, une dépendance régressive). L'approche générative modélise la distribution conjointe $p(\mathbf{y}, \mathbf{x}) = p(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})$, et assigne donc une probabilité à chaque observation et à chaque évènement, réelle ou virtuelle. Par comparaison, l'approche diagnostique ne modélise que la partie $p(\mathbf{y} \mid \mathbf{x})$. Elle ne cherche pas à modéliser les observations $\mathbf{x}$ elles mêmes, seulement les évènements $\mathbf{y}$ conditionnés sur elles.

L'approche générative fait plus d'hypothèses que l'approche diagnostique, et elle a donc tendance à être plus stable et plus informative, mais aussi plus biaisée, que celle ci. En particulier, elle modélise la distribution d'observations $p(\mathbf{x})$ – ce qui est sans rapport avec le diagnostic direct, mais souvent utile, par exemple pour la détection des valeurs aberrantes. Dans les applications où une $\mathbf{y}$ cachée engendre l'$\mathbf{x}$ observée, le modèle $p(\mathbf{y}, \mathbf{x}) = p(\mathbf{x} \mid \mathbf{y})\, p(\mathbf{y})$ est plus simple et plus naturel que le modèle « inverse » $p(\mathbf{y} \mid \mathbf{x}) = p(\mathbf{x}, \mathbf{y})/p(\mathbf{x}) = p(\mathbf{x}, \mathbf{y}) / \left(\sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y})\right)$, et souvent nettement plus facile à estimer. Cependant, dans les problèmes réels de caractère diagnostique, les biais de l'approche générative deviennent problématique. La modélisation imparfaite de $p(\mathbf{x})$ détériore l'estimation des probabilités $p(\mathbf{y} \mid \mathbf{x})$ voulues, et l'approche diagnostique est en général à préférer, au moins à la limite d'un grand nombre d'observations. Inversement, avec très peu d'observations, la stabilité de l'approche générative a tendance à gagner.

Quelque part entre les deux il doit y avoir une zone de transition où la stabilité et le biais sont tous les deux en jeu, et nous nous sommes demandés s'il n'y avait pas moyen de faire mieux que les deux approches en « moyennant » entre elles. L'idée est de voir le terme génératif supplémentaire – le $\log p(\mathbf{x})$ dans la log-vraisemblance – comme une espèce de « prior » ou de « régularisation » qui stabilise mais qui biaise – et de régler la force de cette régularisation afin d'optimiser les résultats. Nous proposons une famille de modèles de log-vraisemblance pénalisée de la forme $\log p(\mathbf{y} \mid \mathbf{x}) + \lambda \log(\epsilon + p(\mathbf{x}))$, où $\lambda \in [0, 1]$, $\epsilon \geq 0$ sont deux paramètres réglables « de régularisation ». (Dans

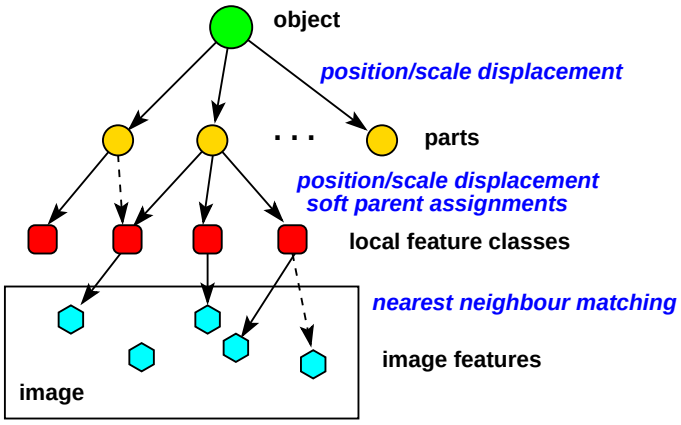


FIG. 9.1 – *La structure de notre modèle de reconnaissance d'objets par parties. Le modèle est hiérarchique avec l'objet à la racine, les parties en fils, éventuellement les sous-parties, et au fond un ensemble de classes type – classées selon l'apparence et la position spatiale – de features image. Entre chaque sous-partie et son parent, il y a une transformation incertaine en position et en échelle qui caractérise le déplacement relative de la sous-partie, et la souplesse / le rigidité de ses liens avec son parent. Les attributions des parents des sous-parties sont aussi incertaines (apprises pendant la phase d'apprentissage). Au fond de l'arbre, chaque classe de features est mis en correspondance avec le feature image observé le plus proche.*

les expériences que nous avons menées, nous utilisons soit $\lambda = 1$ soit $\epsilon = 0$). Le paramètre λ a pour effet d'« aplatir » la distribution $p(\mathbf{x})$, et donc de réduire à la fois l'influence du biais et de la stabilisation générative. Par contre, ϵ « robustifie » le modèle effectif de $p(\mathbf{x})$, amoindrissant l'influence des zones de petit probabilité où le biais est souvent le plus nocif.

Comme pour le cas diagnostique, les modèles sont appris par l'optimisation directe de la log-vraisemblance pénalisée, sur l'espace paramétrique du modèle. Le cas génératif donne souvent une bonne initialisation. Les paramètres λ, ϵ sont choisis par validation croisée.

En pratique, cette approche améliore parfois les approches générative et diagnostique qu'elle généralise, mais l'amélioration semble être relativement modeste pour la plupart, au moins dans le petit nombre d'expériences que nous avons menées jusqu'à la présente.

Ce travail avec mon thésard Guillaume BOUCHARD fut publié aux actes du congrès « 2004 IASC International Symposium on Computational Statistics (COMPSTAT) » [BT04b] (article #16 du mémoire associé).

9.2 Reconnaissance des classes visuelles – approche part - sub-part - features locaux

Cet article décrit une approche statistique à la reconnaissance d'objets en vision, basée sur les primitives locales organisés selon un modèle hiérarchique spatiale. Ici, les indices image sont des petites régions d'image distinctives, sélectionnées au préalable par un détecteur générique et sans connaissance de l'objet à reconnaître, mais qui sont susceptibles à caractériser l'apparence locale de cet objet. Plusieurs types de détecteur ont été conçues pour les problèmes de ce type. Ici, nous utilisons les descripteurs SIFT [Low99] sur les points clés multi-échelles de Harris [MS04].

Afin d'utiliser ces « features locaux » pour la reconnaissance, il y a deux approches populaires:

– Les modèles « sac de features » intègrent les features comme si ils étaient des indices indé-

pendants. Il peut y avoir un grand nombre de features en jeu (des centaines), mais il n'y a aucun modèle de cohérence spatiale qui les lient entre eux.

– Les modèles de type « constellation » prennent l'approche inverse. Ils ont un modèle fort de la cohérence spatiale – par exemple une distribution de probabilité pour la « constellation » (la configuration spatiale conjointe) de tous les features – mais ceci limite significativement le nombre de features qui peuvent entrer en jeu. (Environ 5–7 au maximum, pour le modèle spatial complet).

Notre article présente une approche qui est intermédiaire entre ces deux extrêmes, basée sur un modèle spatial hiérarchique de type partie / sous-partie – voir la figure 9.1. Au plan spatial, cette approche n'a pas la même précision qu'un modèle conjoint de type constellation, mais elle assure toujours un positionnement spatiale relative approximative des éléments, sans limiter le nombre de features qui peuvent être incorporés dans une partie. (En pratique, on peut en incorporer plusieurs centaines). Le modèle final est un mélange gaussien sur les apparences (types) et les positions des features, mais dont la covariance spatiale a une forme structurée. Le modèle est initialisée par une méthode de Hough multi-échelle hiérarchique, et ensuite optimisée par Expectation-Maximisation. L'approche à été validée sur plusieurs classes de la base d'images de CalTech. La figure 9.2 montre quelques exemples.

Ce travail avec mon thésard Guillaume BOUCHARD [BT04a] (article #16 du mémoire associé) est soumis au congrès « Computer Vision and Pattern Recognition ». Le travail a déjà été présenté au workshop invité « 2004 International Workshop on Object Recognition ».

9.3 Conclusions et perspectives

Ce chapitre a présenté deux articles sur la modélisation statistique: l'une sur la relation entre l'apprentissage diagnostic et génératif; l'autre sur une approche statistique générative de la reconnaissance de formes visuelles.

En ce qui concerne la question diagnostique / générative, l'apport de la méthode présentée semble être réel mais modeste, mais la question générale – comment combiner la commodité et la souplesse de l'approche générative avec la précision l'approche diagnostique – reste ouverte. Nous étudions à présent d'autres méthodes pour combiner ces deux approches.

L'article sur la reconnaissance s'inscrit dans le débat actuel sur comment combiner les aspects géométriques et apparences locales afin de mieux reconnaître les catégories visuelles. La question se pose en particulier pour les catégories souples et/ou articulées telles que les personnes et les formes naturelles (animaux, arbres,...). Le modèle hiérarchique représente un premier jet dans cette direction. Il peut en principe s'adapter à une grande variété de formes – en particulier, les modèles articulaires humains 2-D du chapitre 3 représentent un cas spécial de ce modèle, construit à la main. Je voudrais poursuivre le développement de modèles arborescents de ce type qui incorporent un choix d'indices image plus riches et plus aptes à décrire les formes naturelles.

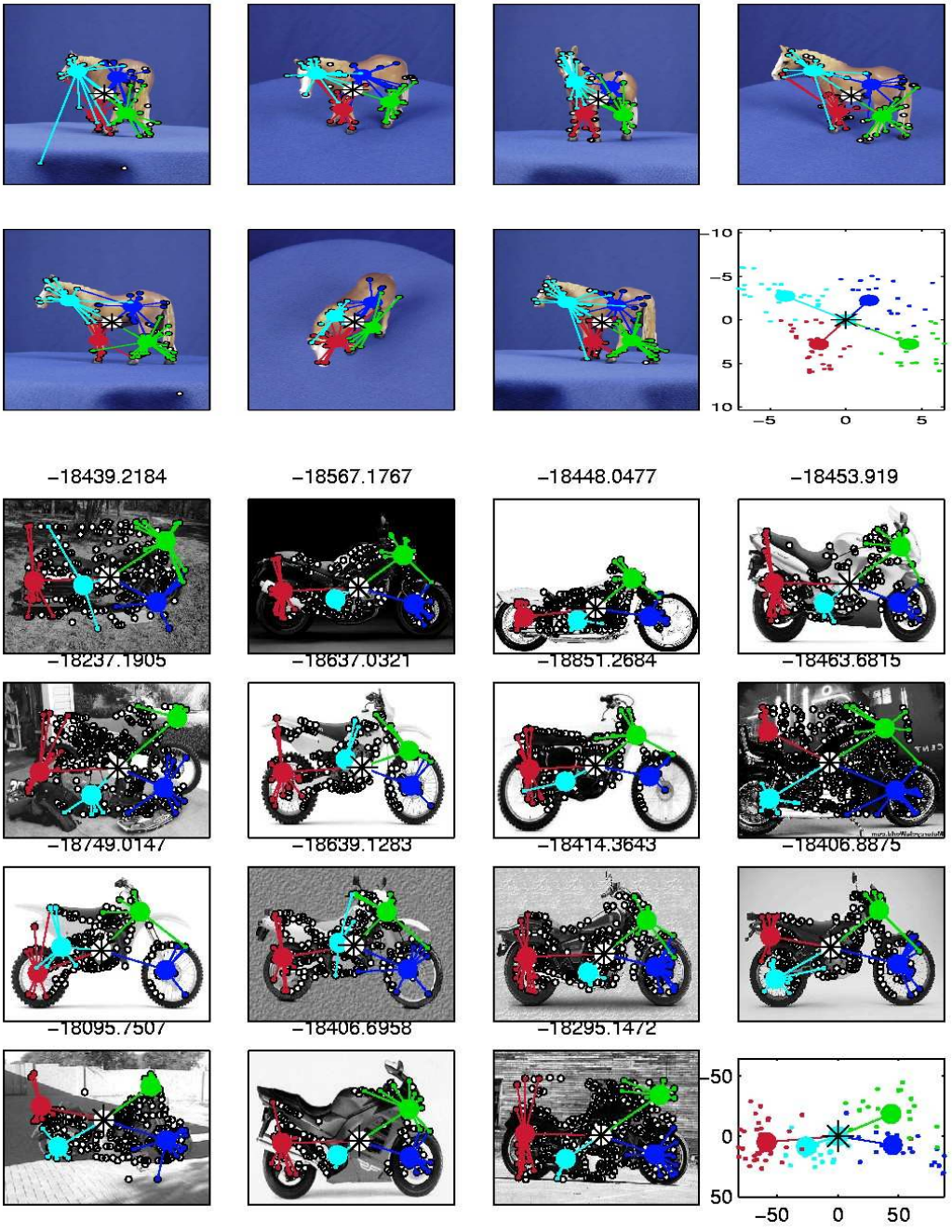


FIG. 9.2 – *Quelques exemples de notre modèle de reconnaissance hiérarchique en action. En haut, on voit que la géométrie hiérarchique flexible du modèle permet une bonne résistance aux changements de point de vue. En bas, le modèle s'adapte aux différentes types de motocycle.*

Troisième partie

Annexes

Annexe A

Encadrements de thèse et de stage

Cette annexe résume mes encadrements de thèse et de stage. Depuis janvier 2000 j'ai encadré ou co-encadré cinq doctorants – deux qui ont soutenu en 2002, un qui soutiendra en janvier 2005, et deux qui sont toujours en cours – plus trois DEA et plusieurs autres stages. Les sujets encadrés touchent tous à la vision, mais mettent en jeu des techniques assez diverses – la vision classique, l'algèbre géométrique, l'optimisation numérique locale et globale, la modélisation du corps humain, l'apprentissage et l'estimation statistique – diversité qui se voit aussi dans les formations des doctorants, pour la plupart informaticiens, mais aussi un mathématicien pur et un statisticien. À signaler en plus que deux de ces étudiants, G. Bouchard et A. Agarwal, sont devenus membres doctorants du Réseau d'Excellence européen PASCAL, sur l'apprentissage machine et la modélisation statistique.

A.1 Doctorants

Cristian SMINCHISESCU, « Estimation algorithms for ambiguous visual models — 3-D human modelling and motion reconstruction in monocular video sequences ». Informaticien de formation, C. Sminchisescu à été doctorant à l'INPG sur une bourse Eiffel de décembre 1999 à juillet 2002, puis post-doctorant INRIA sur le contrat européen VIBES de septembre à novembre 2002. J'ai été son co-directeur de thèse avec Long QUAN à partir de décembre 1999, et – par dérogation de l'école doctorale de l'INPG – son directeur unique à partir de juillet 2001 (et en pratique, dès l'été de l'année 2000). Sa thèse [Smi02] traite principalement du suivi et de la reconstruction du mouvement humain [ST05b, ST05a, ST03a, ST03b, ST02b, ST02a, ST01a, ST01b] – voir le chapitre 2 – mais elle contient aussi des appendices sur d'autres problèmes de suivi visuel. Soutenue le 16 juillet 2002 devant un jury de caractère international, elle a reçu la mention très honorable avec félicitations. À présent Dr Sminchisescu est « Research Fellow » à l'Université de Toronto.

Marc-André AMELLER, « Applications de la géométrie algébrique effective à la vision ». De formation mathématiques pures, M-A. Ameller était doctorant à l'INPG sur une bourse de l'École Normale de Rennes entre octobre 1999 et juillet 2002. J'ai co-encadré sa thèse avec Long QUAN du CNRS à Grenoble et Bernard MOURRAIN de l'INRIA Sophia Antipolis. La thèse traite des méthodes algébréo-géométriques d'estimation de pose des caméras [AQT02, ABQ02b, ABQ02a]. Elle a été soutenue le 16 juillet 2002 [Ame02]. À présent Dr Ameller est professeur de mathématiques dans un lycée Parisien.

Guillaume BOUCHARD. « Generative models in supervised statistical learning with applications

to digital image categorization and structural reliability ». G. Bouchard est doctorant à l'UJF. Il a été financé d'abord sur un contrat EDF (octobre 2001 – août 2002, équipe IS2) puis sur notre contrat européen LAVA (septembre 2002 – août 2004, équipe LEAR). Statisticien de formation, il fait une thèse interdisciplinaire entre les équipes IS2/SELECT (statistique) et MOVI/LEAR (vision et apprentissage). Sa thèse est encadré par Gilles CELEUX, et, depuis septembre 2002, co-encadré par moi même. Son sujet de thèse est l'estimation de modèles statistiques, et en particulier les mélanges de forme restreinte et les liens entre les approches discriminantes et génératives [BT04b]. La soutenance de sa thèse est prévu en janvier ou en février 2005, et il a déjà commencé un postdoc sur les méthodes statistiques pour le diagnostic des pannes à Xerox Research Centre Europe à Grenoble.

Navneet DALAL. Doctorant à l'INPG de formation informaticien, financé sur les projets européens VIBES et ACEMEDIA depuis septembre 2003. N. Dalal est co-encadré par Cordelia SCHMID et par moi même. Son sujet de thèse est la détection et la localisation des personnes dans les images.

Ankur AGARWAL. Informaticien de formation, son sujet est l'application des méthodes de l'apprentissage au suivi et à la reconstruction du mouvement humain. Officiellement, il vient de commencer sa thèse, financé par une bourse du ministère et encadré (par dérogation de l'école doctorale) par moi même. En pratique, j'ai aussi encadré son DEA et la thèse est déjà bien entamée, avec trois publications apparues dans les congrès internationaux [AT04d, AT04a, AT04b] et un article journal soumis [AT04c] – voir le chapitre 4.

A.2 Stages de DEA / Masters

Marc-André AMELLER, « Méthodes de résolution de systèmes de polynômes et application à un problème de vision ». DEA d'algèbre de l'Université de Rennes 1, effectué à l'équipe MOVI entre décembre 1999 et septembre 2000. J'ai co-encadré ce stage de DEA dont Long QUAN a été le directeur principal. Il traite de l'application des méthodes de l'algèbre géométrique à l'estimation de pose d'une caméra calibrée à partir des points 3-D connus.

Ankur AGARWAL, « Learning Dynamical Models for Tracking Complex Motion ». Projet de DEA fait sous ma direction dans l'équipe LEAR entre novembre 2002 et juillet 2003 (avec, vu que A. Agarwal n'est pas francophone d'origine, une prolongation à juillet 2004 pour finir les cours de DEA et pour commencer le travail de thèse). Ce stage traite du suivi du mouvement humain articulaire dans les séquences d'images. La contribution principale est un algorithme pour apprendre un modèle dynamique auto-régressive et linéaire par morceaux, qui permet une modélisation plus flexible du comportement humain, et donc un suivi plus fiable et plus précis du mouvement de la personne. Les résultats ont été publiés au congrès ECCV 2004 [AT04d] – voir le chapitre 3.

Aurélie BUGEAU, « Attention visuelle multi-échelle ». Stage de DEA/Masters fait sous ma direction dans l'équipe LEAR entre décembre 2003 et juillet 2004. Le but principal était de réimplanter le modèle d'attention visuelle de Itti et Koch [IKN98, Itt00, IC01], mais en multi-échelle. C'est-à-dire, la méthode produit une carte de saillance visuelle qui est une pyramide et non seulement une image à l'échelle fixe, afin de coder l'échelle spatiale à laquelle la saillance a été mesurée. Le modèle a été implanté et testé. Le temps disponible n'a pas permis une étude poussée, mais il semble être acquis que l'analyse multi-échelle améliore nettement les résultats en comparaison avec la méthode d'origine. A. Bugeau à commencée une thèse à Rennes en septembre 2004.

A.3 Autres Stages

Jean GOFFINET, « Machines à vecteur de support pour la détection et le suivi de personnes sur des séquences vidéo ». Stage d'option scientifique de l'École Polytechnique, fait dans l'équipe MOVI entre avril et juillet 2001, et co-encadré par Cordelia SCHMID et par moi même. Ce stage est en l'essentiel une réimplantation et une étude de la performance du « détecteur des piétons SVM » fait au MIT par Papageorgiou et Poggio.

Olivier GALIZZI et Laure HEIGEAS, « Reconstruction temps réel d'une scène dynamique ». Stage de magistère informatique de l'UJF, fait sous ma direction dans l'équipe MOVI entre décembre 2000 et septembre 2001, et co-encadré par Edmond BOYER. Ce stage est à cheval entre la vision et la graphique temps-réel. Le but était de synthétiser, à partir de plusieurs caméras vidéo qui regardent une scène dynamique, et en temps réel, une vidéo virtuelle de la scène, par exemple prise d'un point d'observation différent des caméras réelles. Deux techniques approximatives à la base d'« enveloppes visuelles » ont été implantées: une reconstruction 3D voxélique classique, avec un re-rendu lissé et texturé à la base de « marching cubes »; et la méthode épipolaire de Matusik, Buehler, McMillan, Raskar et Gortler, qui ne fait pas de reconstruction 3D explicite, mais qui fournit directement l'image voulue, rendue à partir d'un calcul basé sur la géométrie épipolaire. L'implantation comporte ces deux méthodes, et les étapes préalables de calibrage des caméras (en l'occurrence, environ 4 webcams USB), d'apprentissage et de soustraction de fond. Les deux méthodes tournent en temps réel et donnent des résultats relativement satisfaisants. Ce projet a été classé premier ex aequo entre les projets du magistère 2001. Les deux etudiants travaillent toujours en infographie à l'INRIA Grenoble: Laure Heigeas est ingénieur expert et Olivier Galizzi doctorant dans l'équipe ÉVASION.

Nippun KWATRA, « Where was this taken? – semi-supervised kernel based learning for location recognition ». Stage INRIA-IIT encadré par moi-même, mai – juillet 2003. Ce stage étudie l'application des méthodes d'apprentissage semi-supervisée dit « transductives » à la classification d'images en vision par ordinateur. Le problème de base était de classer un ensemble d'images selon le type de la scène qu'elles figurent (plage, forêt, ville...). Le défi était d'apprendre un critère de décision efficace à partir de très peu de données d'apprentissage pre-étiquetées avec leur classe type, en exploitant au maximum l'information implicite structurelle sur les classes qui est fournie par une deuxième base d'images *non*-étiquetées. L'étude a combinée une approche « sac de features » bâti sur les descripteurs d'image local affines, avec l'apprentissage machine transductive. La conclusion principale est que quoique la transduction offre une amélioration modeste, elle ne permette pas à elle seule, de réduire le nombre d'images étiquetées de façon très significative. N. Kwatra à commencé une thèse aux États Unis au Georgia Institute of Technology.

Diane LARLUS-LARRONDO, « La segmentation d'image ». Stage TER de l'UFR IMA encadré par moi-même, janvier – juin 2004. Le but de ce court stage était de faire une étude bibliographique des méthodes algébriques de segmentation d'image, et d'en implanter quelques unes, y compris la méthode « normalized cut » de Shi et Malik.

A.4 Postdocs et Ingénieurs

J'ai encadré le postdoc de Cristian SMINCHISESCU (reconstruction du mouvement humain, sur le projet EU VIBES, septembre – novembre 2002), et co-encadré avec Cordelia SCHMID le postdoc de Ragini CHOUDHURY (détection et suivi des visages, VIBES, avril 2001 – janvier 2002). J'ai aussi encadré l'ingénieur industriel Marius MALCIU de la société PANDORA-Studio entre juin et

juillet 2003. Actuellement, je co-encadre (toujours avec Cordelia SCHMID) l'ingénieur de recherche Michaël SDIKA (détection d'objets en temps réel, projet EU LAVA).

Annexe B

Autres Activités Scientifiques

B.1 Transferts et commercialisations

PANDORA-Studio: Le travail présenté au chapitre 2 sur le suivi et la reconstruction du mouvement humain par modélisation explicite est le sujet d'un accord de développement industriel entre l'INRIA et la société Parisien PANDORA-Studio depuis juin 2003. Le but est de développer cette technologie de capture de mouvement au contexte de la production des vidéos et des jeux. PANDORA a investi un an de travail d'ingénieur dans le projet, mais à présent l'avenir du projet reste incertain suite au départ de l'ingénieur chargé du travail.

B.2 Projets de recherche

• Co-responsable scientifique (avec Cordelia Schmid) des travaux dans notre équipe LEAR sur les projets européens suivants: VIBES – Video Indexing, Browsing and Exploration (12/2001 – 04/2004); LAVA – Learning for Adaptable Visual Assistants (05/2002 – 04/2005); ACEMEDIA – Integrating Knowledge, Semantics and Content for User-centred Intelligent Media Services (01/2004 – 12/2007).

• Un des instigateurs principaux, avec le coordinateur John Shawe-Taylor de Southampton, du réseau d'excellence européen « Pattern Analysis, Statistical Modelling, and Computational Learning (PASCAL) » (12/2003 – 11/2007). Dans ce réseau je suis: membre du comité central; responsable du Core Site INRIA Grenoble; responsable du programme « Balance & Integration » (équilibrage globale de l'activité de réseau); responsable des groupes d'activité en vision par ordinateur et en théorie et algorithmes; et co-responsable de l'activité thématique sur les interfaces multi-modales.

• Expert invité du consultation EU FP6 sur la priorité thématique « MultiModal Interfaces », avril 2004.

• Participant à l'Action Concertée Incitative vision/statistique « MoviStar », qui est conjointe entre les équipes LEAR et Mistis de l'INRIA Rhône-Alpes, l'équipe SMS du LMC-IMAG Grenoble, et Heudiasyc, UTC Compiègne.

B.3 Jurys de thèse

• « Examinateur externe » du PhD de Frederick SCHAFFALITZSKY, Université d'Oxford, février 2002. (Ce rôle est équivalent à celui d'un rapporteur externe unique. La défense se fait à huis clos par deux examinateurs seulement, l'un interne et l'autre externe. Elle dure de 3 à 4 heures environ, avec des questions sur le domaine de travail et la thèse.)

B.4 Comités d'organisation

• General Chair et animateur principal du congrès « Ninth IEEE International Conference on Computer Vision », Nice, France, octobre 2003.

• Workshops Chair du congrès « Tenth IEEE International Conference on Computer Vision », Beijing, China, octobre 2005.

• Organisateur du workshop « Pattern Recognition and Machine Learning in Computer Vision » du Réseau d'Excellence EU PASCAL, Grenoble, France, mai 2004.

• Organisateur de la section PASCAL du workshop « International Workshop on Object Recognition », Taormina, Sicile, octobre 2004.

B.5 Comités de relecture

• Area Chair des congrès internationaux: European Conference on Computer Vision (2000, 2002); Computer Vision and Pattern Recognition (2003); International Conference on Computer Vision (2005).

• Relecture régulière pour les journaux: Pattern Analysis and Machine Intelligence (PAMI); International Journal of Computer Vision (IJCV); Journal of Mathematical Imaging and Vision (JMIV); Pattern Recognition Letters (PRL); Image & Vision Computing (IVC).

• Relecture pour les congrès: International Conference on Computer Vision (ICCV, 1998–2005); Computer Vision and Pattern Recognition (CVPR, 1997–2005); European Conference on Computer Vision (ECCV, 1998–2004); Neural Information Processing Systems (NIPS, 2003–2004); International Conference on Machine Learning (ICML, 2004).

Note sur les journaux: Je fais campagne pour une littérature scientifique en ligne et ouverte à tous, et en particulier pour l'archivage ouvert publique et systématique de toutes les publications scientifiques. Ceci étant incompatible avec la politique de limitation d'accès des éditeurs commerciaux, j'ai malheureusement été obligé de refuser en 2003–2004 l'invitation de devenir l'un des Associate Editors du journal « International Journal of Computer Vision » (Kluwer-Springer), et de repousser depuis 2003 les suggestions de devenir Associate Editor du journal « IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) » (IEEE Press).

Annexe C

Bibliographie

[ABQ02a] M.-A. Ameller, A. Bartoli, and L. Quan. Minimal metric structure and motion from three affine images. In *Asian Conf. Computer Vision*, pages I 356–361, January 2002.

[ABQ02b] M.-A. Ameller, A. Bartoli, and L. Quan. Reconstruction métrique minimale à partir de trois caméras affines. In *Reconnaissance des Formes et Intelligence Artificielle*, pages II 471–477, Angers, January 2002.

[AC99] J.K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440, March 1999.

[Ame02] M.-A. Ameller. *Applications de la géométrie algébraique effective à la vision*. PhD thesis, INPG, July 2002.

[AQT02] M.-A. Ameller, L. Quan, and B. Triggs. Le calcul de pose : de nouvelles méthodes matricielles. In *Reconnaissance des Formes et Intelligence Artificielle*, Angers, January 2002.

[AR94] Y. Abashkin and N. Russo. Transition state structures and reaction profiles from constrained optimization procedure. implementation in the framework of density functional theory. *J. Chem. Phys.*, 1994.

[ART94] Y. Abashkin, N. Russo, and M. Toscano. Transition states and energy barriers from density functional studies: Representative isomerization reactions. *Int. J. Quantum Chemistry*, 1994.

[AT04a] A. Agarwal and B. Triggs. 3d human pose from silhouettes by relevance vector regression. In *Int. Conf. Computer Vision & Pattern Recognition*, pages II 882–888, Washington, June 2004.

[AT04b] A. Agarwal and B. Triggs. Learning to track 3d human motion from silhouettes. In *Int. Conf. Machine Learning*, pages 9–16, Banff, July 2004.

[AT04c] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. Under review for PAMI, September 2004.

[AT04d] A. Agarwal and B. Triggs. Tracking articulated motion with piecewise learned dynamical models. In *European Conf. Computer Vision*, pages III 54–65, Prague, May 2004.

[Bar96] G.T. Barkema. Event-based relaxation of continuous disordered systems. *Phys. Rev. Lett.*, 77(21), 1996.

[Bau00] A. Baumberg. Reliable feature matching across widely separated views. In *Int. Conf. Computer Vision & Pattern Recognition*, pages 774–781, 2000.

[BM98] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Int. Conf. Computer Vision & Pattern Recognition*, 1998.

[Bof94] J.M. Bofill. Updated hessian matrix and the restricted step method for locating transition structures. *J. Computational Chemistry*, 15(1):1–11, 1994.

[Bra99] M. Brand. Shadow puppetry. In *Int. Conf. Computer Vision*, pages 1237–1244, 1999.

[BT04a] G. Bouchard and B. Triggs. Hierarchical part-based visual object categorization. Submitted to CVPR'05, November 2004.

[BT04b] G. Bouchard and B. Triggs. The tradeoff between generative and discriminative classifiers. In *COMPSTAT — IASC Int. Symp. Computational Statistics*, pages 721–728, Prague, August 2004.

[BY95] M.J. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *Int. Conf. Computer Vision*, pages 374–381, 1995.

[CDNG92] P. Culot, G. Dive, V.H. Nguyen, and J.M. Ghuysen. A quasi-newton algorithm for first-order saddle point location. *Theoretica Chimica Acta*, 82:189–205, 1992.

[CM81] C.J. Cerjan and W.H. Miller. On finding transition states. *J. Chem. Phys.*, 75(6), 1981.

[CR99] T. Cham and J. Rehg. A multiple hypothesis approach to figure tracking. In *Int. Conf. Computer Vision & Pattern Recognition*, pages II 239–245, 1999.

[CS71] G.M. Crippen and H.A. Scheraga. Minimization of polypeptide energy. xi. the method of gentlest ascent. *Archives of Biochemistry and Biophysics*, 144:462–466, 1971.

[DBR00] J. Deutscher, A. Blake, and I. Reid. Motion capture by annealed particle filtering. In *Int. Conf. Computer Vision & Pattern Recognition*, 2000.

[DCdB+02] V. Depoortere, J. Cant, B. Van den Bosch, J. De Prins, F. Fransens, and L. Van Gool. Efficient pedestrian detection: a test case for svm based categorization. In *ECVision Workshop on Cognitive Vision Systems*, Zürich, September 2002.

[DdFG01] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo in Practice*. Springer-Verlag, 2001.

[FB81] M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Computer Graphics and Image Processing*, 24(6):381–395, 1981.

[FE73] M.A. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. Computing*, C-22:67–92, 1973.

[FG87] W. Förstner and E. Gülch. A fast operator for detection and precise location of distinct points, corners and centres of circular features. In *ISPRS Intercommission Workshop*, Interlaken, June 1987.

[FH00] P. Felzenszwalb and D. Huttenlocher. Efficient matching of pictorial structures. In *Int. Conf. Computer Vision & Pattern Recognition*, pages 66–75, 2000.

[För86] W. Förstner. A feature-based correspondence algorithm for image matching. *Int. Arch. Photogrammetry & Remote Sensing*, 26 (3/3):150–166, 1986.

[För94] W. Förstner. A framework for low-level feature extraction. In *European Conf. Computer Vision*, pages II 383–394, Stockholm, 1994.

[FPC00] W. Freeman, E. Pasztor, and O. Carmichael. Learning low-level vision. *Int. J. Computer Vision*, 40(1):25–48, 2000.

[Gav99] D.M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, January 1999.

[Gav00] D. M. Gavrila. Pedestrian detection from a moving vehicle. In *European Conf. Computer Vision*, 2000.

[Ge87] R. Ge. The theory of the filled function method for finding a global minimizer of a nonlinearly constrained minimization problem. *J. Comp. Math.*, 1987.

[Hel91] T. Helgaker. Transition-state optimizations by trust-region image minimization. *Chemical Physics Letters*, 182(5), 1991.

[Hil77] R.L. Hilderbrandt. Application of newton-raphson optimization techniques in molecular mechanics calculations. *Computers & Chemistry*, 1:179–186, 1977.

[HJ99] G. Henkelman and H. Jonsson. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *J. Chem. Phys.*, 111(15):7011–7022, 1999.

[HLF00] N. R. Howe, M. E. Leventon, and W. T. Freeman. Bayesian reconstruction of 3d human motion from single-camera video. In *Neural Information Processing Systems*, 2000.

[Hog83] D. Hogg. Model-based vision: a program to see a walking person. *Image & Vision Computing*, 1(1):5–20, 1983.

[HS88] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, 1988.

[HZ00] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[IB98] M. Isard and A. Blake. Condensation — conditional density propagation for visual tracking. *Int. J. Computer Vision*, 29(1):5–28, 1998.

[IC01] L. Itti and C.Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3), 2001.

[IF01] S. Ioffe and D. Forsyth. Probabilistic methods for finding people. *Int. J. Computer Vision*, 43(1):45–68, 2001.

[IF03] S. Ilic and P. Fua. Implicit Mesh Models for Modeling and Tracking. In *Int. Conf. Computer Vision & Pattern Recognition*, Madison, WI, June 2003. Submitted for publication.

[IKN98] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Analysis & Machine Intelligence*, November 1998.

[Itt00] L. Itti. *Models of Bottom-Up and Top-Down Visual Attention*. PhD thesis, California Institute of Technology, January 2000.

[JBY96] S. Ju, M. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated motion. In *Int. Conf. Automatic Face & Gesture Recognition*, pages 38–44, 1996.

[Jen95] F. Jensen. Locating transition structures by mode following: A comparison of six methods on the ar_8 lennard-jones potential. *J. Chem. Phys.*, 102(17):6706–6718, 1995.

[JJH88] P. Jorgensen, H.J.A. Jensen, and T. Helgaker. A gradient extremal walking algorithm. *Theoretica Chimica Acta*, 73:55–65, 1988.

[KB01] T. Kadir and M. Brady. Scale, saliency and image description. *Int. J. Computer Vision*, 45(2):83–105, 2001.

[KT99] F. Kahl and B. Triggs. Critical motions in euclidean structure from motion. In *Int. Conf. Computer Vision & Pattern Recognition*, Fort Collins, Colorado, 1999.

[KTÅ00] F. Kahl, B. Triggs, and K. Åström. Critical motions for auto-calibration when some intrinsic parameters can vary. *J. Mathematical Imaging & Vision*, 13(2):131–146, October 2000.

[LC01] D. Liebowitz and S. Carlsson. Uncalibrated motion capture exploiting articulated structure constraints. In *Int. Conf. Computer Vision*, July 2001.

[LESC04] G. Loy, M. Eriksson, J. Sullivan, and S. Carlsson. Monocular 3d reconstruction of human motion in long action sequences. In *European Conf. Computer Vision*, 2004.

[LM85] A. Levy and A. Montalvo. The tunneling algorithm for the global minimization of functions. *SIAM J. Stat. Comp.*, 1985.

[Low99] D. Lowe. Object recognition from local scale-invariant features. In *Int. Conf. Computer Vision*, pages 1150–1157, 1999.

[MB98] N. Mousseau and G.T. Berkema. Traveling through potential energy lanscapes of disordered materials: The activation-relaxation technique. *Phys. Rev. E*, 57(2), 1998.

[MM02] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *European Conf. Computer Vision*, volume 3, pages 666–680, 2002.

[Mor77] H.P. Moravec. Towards automatic visual obstacle avoidance. In *IJCAI*, page 584, 1977.

[MPP01] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 23(4), 2001.

[MS01] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Int. Conf. Computer Vision*, pages 525–531, 2001.

[MS02] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *European Conf. Computer Vision*, pages I.128–142, 2002.

[MS04] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *Int. J. Computer Vision*, 60(1):63–86, 2004.

[MTS+04] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. Under review for the International Journal of Computer Vision, 2004.

[MW99] L.J. Munro and D.J. Wales. Defect migration in crystalline silicon. *Physical Review B*, 59(6):3969–3980, 1999.

[Neu04] A. Neumaier. Complete search in continuous global optimization and constraints satisfaction. In A. Iserles, editor, *Acta Numerica*, pages 271–369. Cambridge University Press, 2004.

[NTSS90] J. Nichols, H. Taylor, P. Schmidt, and J. Simons. Walking on potential energy surfaces. *J. Chem. Phys.*, 92(1), 1990.

[Pap97] C. Papageorgiou. Object and pattern detection in video sequences. Master's thesis, Massachusetts Institute of Technology, 1997.

[RF03] D. Ramanan and D. A. Forsyth. Finding and tracking people from the bottom up. In *Int. Conf. Computer Vision & Pattern Recognition*, 2003.

[RK95] J. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *Int. Conf. Computer Vision*, 1995.

[Roh94] K. Rohr. Towards model-based recognition of human movements in image sequences. *Computer Vision, Graphics & Image Processing*, 59(1):94–115, 1994.

[RST02] R. Ronfard, C. Schmid, and B. Triggs. Learning to parse pictures of people. In *European Conf. Computer Vision*, Copenhagen, 2002.

[SB01] H. Sidenbladh and M. Black. Learning image statistics for bayesian tracking. In *Int. Conf. Computer Vision*, pages II 709–716, 2001.

[SBR+04] L. Sigal, S. Bhatia, S. Roth, M.J. Black, and M. Isard. Tracking loose-limbed people. In *Int. Conf. Computer Vision & Pattern Recognition*, pages I 421–428, June 2004.

[SBS02] H. Sidenbladh, M. Black, and L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In *European Conf. Computer Vision*, pages I–784–800, 2002.

[SC02] J. Sullivan and S. Carlsson. Recognizing and tracking human action. In *European Conf. Computer Vision*, 2002.

[SEC02] J. Sullivan, M. Eriksson, and S. Carlsson. Automating multi-view tracking and reconstruction of human motion. In *ECCV Workshop on Vision and Modelling of Dynamic Scenes*, 2002.

[SJTO83] J. Simons, P. Jorgensen, H. Taylor, and J. Ozmen. Walking on potential energy surfaces. *J. Phys. Chem.*, 87:2745–2753, 1983.

[SMB00] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *Int. J. Computer Vision*, 37(2):151–172, 2000.

[Smi02] C. Sminchisescu. *Estimation Algorithms for Ambiguous Visual Models — Three Dimensional Human Modeling and Motion Reconstruction in Monocular Video Sequences*. PhD thesis, INPG, July 2002.

[SR93] J.Q. Sun and K. Ruedenberg. Gradient extremals and stepest descend lines on potential energy surfaces. *J. Chem. Phys.*, 98(12), 1993.

[ST01a] C. Sminchisescu and B. Triggs. Covariance scaled sampling for monocular 3D body tracking. In *Int. Conf. Computer Vision & Pattern Recognition*, pages I 447–454, Hawaii, 2001.

[ST01b] C. Sminchisescu and B. Triggs. A robust multiple hypothesis approach to monocular human motion tracking. Research Report 4208, INRIA, June 2001.

[ST02a] C. Sminchisescu and B. Triggs. Hyperdynamic sampling for articulated estimation. In *European Conf. Computer Vision*, Copenhagen, 2002.

[ST02b] C. Sminchisescu and B. Triggs. Mapping minima and transitions of visual models. In *European Conf. Computer Vision*, Copenhagen, 2002.

[ST03a] C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *Int. J. Robotics Research*, 22(6):371–391, June 2003. Special issue on Visual Analysis of Human Movement.

[ST03b] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3d human tracking. In *Int. Conf. Computer Vision & Pattern Recognition*, pages I 69–76, June 2003.

[ST05a] C. Sminchisescu and B. Triggs. Building roadmaps of minima and transitions in visual models. *Int. J. Computer Vision*, 61(1):81–101, January 2005.

[ST05b] C. Sminchisescu and B. Triggs. Hyperdynamic sampling. *J. Image & Vision Computing, special issue on ECCV 2002 papers*, 2005. Accepted, to appear in early 2005.

[SVD03] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter sensitive hashing. In *Int. Conf. Computer Vision*, pages 750–757, 2003.

[SZ02] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?". In *European Conf. Computer Vision*, pages I 414–431, Copenhagen, 2002.

[Tay00] C. J. Taylor. Reconstruction of articulated objects from point correspondences in a single image. *Computer Vision and Image Understanding*, 80(3):349–363, 2000.

[TB01] K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *Int. Conf. Computer Vision*, July 2001.

[TBU00a] P. Thévenaz, T. Blu, and M. Unser. Image interpolation and resampling. In I.N. Bankman, editor, *Handbook of Medical Imaging, Processing and Analysis*, chapter 25, pages 393–420. Academic Press, San Diego CA, USA, 2000.

[TBU00b] P. Thévenaz, T. Blu, and M. Unser. Interpolation revisited. *IEEE Transactions on Medical Imaging*, 19(7):739–758, July 2000.

[Tip01] M. Tipping. Sparse bayesian learning and the relevance vector machine. *J. Machine Learning Research*, 1:211–244, 2001.

[TMHF00] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment — a modern synthesis. In B. Triggs, A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practice*, pages 298–372. Springer-Verlag, 2000.

[Tri00] B. Triggs. Plane + parallax, tensors and factorization. In *European Conf. Computer Vision*, pages 522–538, Dublin, 2000.

[Tri01a] B. Triggs. Empirical filter estimation for subpixel interpolation and matching. In *Int. Conf. Computer Vision*, pages II 550–557, Vancouver, 2001.

[Tri01b] B. Triggs. Joint feature distributions for image correspondence. In *Int. Conf. Computer Vision*, pages II 201–208, Vancouver, 2001.

[Tri04] B. Triggs. Detecting keypoints with stable position, orientation and scale under illumination changes. In *European Conf. Computer Vision*, pages IV 100–113, May 2004.

[TZ00] P. H. S. Torr and A. Zisserman. Feature based methods for structure and motion estimation. In B. Triggs, A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practice*, pages 278–294, Corfu, Greece, 2000. Springer-Verlag LNCS.

[Uns00] M. Unser. Sampling — 50 years after Shannon. *Proceedings of the IEEE*, 88(4):569–587, April 2000.

[Vap98] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[VJS03] Paul Viola, Michael J. Jones, and Daniel Snow. Detecting pedestrians using patterns of motion and appearance. In *Int. Conf. Computer Vision*, pages 734–741, Nice, France, 2003.

[Wal89] D.J. Wales. Finding saddle points for clusters. *J. Chem. Phys.*, 91(11), 1989.

[WW96] D.J. Wales and T.R. Walsh. Theoretical study of the water pentamer. *J. Chem. Phys.*, 105(16), 1996.