Estimation algorithms for ambiguous visual models : Three Dimensional Human Modeling and Motion Reconstruction in Monocular Video Sequences

Cristian Sminchisescu

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

THÈSE

pour obtenir le grade de

DOCTEUR DE L'INPG

Spécialité : **Imagerie Vision et Robotique**

préparée au laboratoire GRAVIR - IMAG - INRIA
dans le cadre de **l'École Doctorale**
«Mathématiques, Sciences et Technologies de l'Information, Informatique»

présentée et soutenue publiquement

par

Cristian SMINCHISESCU

le 16 julliet 2002

ESTIMATION ALGORITHMS FOR AMBIGUOUS VISUAL MODELS

Three-Dimensional Human Modeling and Motion Reconstruction in Monocular Video Sequences

---

JURY

M.	Roger MOHR	Président
M.	Michael BLACK	Rapporteur
M.	Andrew BLAKE	Rapporteur
M.	William TRIGGS	Directeur de thèse
M.	Long QUAN	Co-encadrant
M.	Richard HARTLEY	Examinateur

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

THÈSE

pour obtenir le grade de

DOCTEUR DE L'INPG

Spécialité : **Imagerie Vision et Robotique**

préparée au laboratoire GRAVIR - IMAG - INRIA
dans le cadre de **l'École Doctorale**
«Mathématiques, Sciences et Technologies de l'Information, Informatique»

présentée et soutenue publiquement

par

Cristian SMINCHISESCU

le 16 julliet 2002

ALGORITHMES D'ESTIMATION POUR DES MODÈLES VISUELS AMBIGUS

Modélisation Humaine Tridimensionnelle et Reconstruction du Mouvement dans des Séquences Vidéo Monoculaires

---

JURY

M.	Roger MOHR	Président
M.	Michael BLACK	Rapporteur
M.	Andrew BLAKE	Rapporteur
M.	William TRIGGS	Directeur de thèse
M.	Long QUAN	Co-encadrant
M.	Richard HARTLEY	Examinateur

Abstract

This thesis studies the problem of tracking and reconstructing three-dimensional articulated human motion in monocular video sequences. This is an important problem with applications in areas like markerless motion capture for animation and virtual reality, video indexing, human-computer interaction or intelligent surveillance.

A system that aims to reconstruct 3D human motion using single camera sequences faces difficulties caused by the lossy nature of monocular projection and the high-dimensionality required for 3D human modeling. The complexities of human articular structure, shape and their physical constraints, and the large variability in image observations involving humans, render the solution non-trivial.

We focus on the *general* problem of 3D human motion estimation using *monocular* video streams. Hence, we can not exploit the simplifications brought by using multiple cameras or strong dynamical models such as walking, and we minimize assumptions about clothing and background structure. In this unrestricted setting, the posterior likelihoods over human pose space are inevitably highly multi-modal, and efficiently locating and tracking the most prominent peaks is a major computational challenge.

To address these problems, we propose a model that incorporates realistic kinematics and several important human body constraints, and a principled, robust and probabilistically motivated integration of different visual cues like contours, intensity or silhouettes. We then derive three novel continuous multiple-hypothesis search techniques that allow either deterministic or stochastic localization of nearby peaks in the high-dimensional human pose likelihood surface: **Covariance Scaled Sampling**, **Eigenvector Tracking** and **Hypersurface Sweeping** and **Hyperdynamic Importance Sampling**. The search methods give general, principled approaches to the deterministic exploration of the non-convex error surfaces so often encountered in computational vision problems. The combined system allows monocular tracking of unconstrained human motions in clutter.

Résumé

Cette thèse s'intéresse au problème du suivi et de la reconstruction tridimensionnelle de mouvements articulés humains dans des séquences vidéo monoculaires. Cette problématique est importante et comporte un champ d'applications assez large qui touche des domaines tels que la capture du mouvement sans cibles pour l'animation et la réalité virtuelle, l'indexation vidéo, les interactions homme-machines ou la télésurveillance.

La reconstruction de mouvement 3D humain à partir de séquences monoculaires est un problème complexe en raison de la perte d'informations due à la projection monoculaire, et en raison de la dimensionnalité importante nécessaire à la modélisation du corps humain. En effet, la complexité de la structure articulaire et volumétrique du corps humain, ses contraintes physiques, ainsi que la grande variabilité dans les observations images, rendent la recherche d'une solution à ce problème difficile.

L'objectif principal de cette thèse est donc d'étudier dans quelle mesure l'estimation de *mouvement générique humain* est réalisable à partir d' *une seule caméra*. Par conséquent, nous ne faisons pas d'hypothèses sur le mouvement ou l'habillement du sujet suivi, sur la structure du fond, ou sur la présence de plusieurs caméras. Cette formulation non-restrictive résulte en une distribution de probabilité dynamique et fortement multimodale dans l'espace des configurations (les poses). Le défi majeur réside ici dans la localisation temporelle effective des modes les plus importants de cette distribution.

Pour aborder cette étude, nous proposons un cadre de modélisation qui tient compte des contraintes physiques du corps humain et qui permet une intégration cohérente, robuste et statistiquement justifiée des différentes informations visuelles comme les contours, les intensités ou les silhouettes. Dans ce cadre, nous décrivons trois nouvelles méthodes de recherche continues reposant sur des hypothèses multiples qui nous permettent à la fois une localisation déterministe et un échantillonage stochastique des modes multiples sur les surfaces de probabilité associées aux poses du corps humain: **Covariance Scaled Sampling**, **Eigenvector Tracking** et **Hypersurface Sweeping** et **Hyperdynamic Importance Sampling**. Ces méthodes nous permettent a la fois un suivi et une reconstruction efficace du mouvement humain dans des contextes naturels, ainsi qu'une étude systématique et déterministe des surfaces d'erreurs multimodales souvent rencontrées dans des problèmes de vision par ordinateur.

Contents

List of Figures

List of Tables

Tuturor celor pe care îi iubesc.

All things are lawful for me, but not all things are profitable.

1 CORINTHIANS 6,12

Acknowledgments

First, I would like to thank my advisor and friend Bill Triggs for excellent scientific supervision during this thesis. Bill's devotion to my scientific development has spanned the range of what one can realistically expect from an advisor: explain difficult concepts and show how to recognize and pursue promising research directions, help analyze results and identify failures or explain how to write clear scientific arguments that preserve what is essential and discard what is not. I am therefore grateful to Bill for giving me the freedom to choose and explore my directions and style of research, but for continuously keeping me aware that one has the moral duty to strive for perfection when obtaining or communicating research results. This only says little of what I've learned during many long, late, discussions together.

I am also grateful to my co-advisor Long Quan, who was a friendly and supportive host during my initial thesis years in Grenoble and to whom I owe scientific advice as well as much open and enthusiastic encouragement. My thesis in MOVI probably wouldn't have been possible without the generous help of Roger Mohr, who encouraged me to come to INRIA and provided invaluable support during the application for my EIFFEL fellowship, even though he was preparing for a difficult surgery during that period. After my arrival here, Roger continued to be interested and follow my work through regular and enthusiastic discussions, and today he is the president of my jury, for which I thank him once more.

I would also like to express my gratitude to my thesis rapporteurs, Andrew Blake and Michael Black for kindly accepting to read and review the thesis, for their patient and detailed comments that helped to improve the manuscript, and for their previous encouragement and criticism of this research while in its earlier stages of development. I also want to thank my examiner, Richard Hartley, for accepting to participate in my jury and for his interest in my work.

Prior to being in France, I studied for a MSc. in the U.S., where I have also learned about vision, object modeling and optimization. I would like to thank my advisor Sven Dickinson, for great scientific help both while in U.S. and in France. Sven has been the first to introduce me to the complexities of research in vision, and his devotion, patience and understanding of me as a person (and not only as a scientist) is something I cannot forget. I am also grateful to my co-advisor in U.S., Dimitris Metaxas, to whom I owe many friendly and open discussions, and emphasis of the importance of modeling in vision. I would also like to thank Doug DeCarlo who kindly provided many detailed explanations on practical aspects of deformable models that have been very useful in

my later developments.

I am grateful to all the members of the MOVI team for being very supportive hosts and for creating a very motivating scientific environment during my stay here. Special thanks go to my office friend Fred, who helped in a variety of eclectic contexts including video capture, posing and dancing as a tracked subject, and rendering my French writing readable. I also cannot forget the friendly and open discussions and lunches with Peter, Edmond, Cordelia, Radu, Rémy, Markus, Navneet and Adrien. Neither can I overlook the constant help and joyful presence of Veronique and her patience and backup of my clumsy handling of administration.

I would also like to thank my good friend Alex Telea, who has contributed to the software design and implementation I have done, and has given me many limpid technical explanations in long e-mails over time. I am indebted to my other friends that have not only been a warm and stimulating company, but have always been helpful when I needed them: Andi, Oana and Bogdan, Vlad, Florin, Tudor, Cristi and Andrea, Gabi and Dragos, Edi, David, Catalin.

Last but not least, my thanks go to those few presences that, from far away or from nearby, have surrounded me with their love, encouragement and trust that made all this long endeavor possible.

ALGORITHMES D'ESTIMATION POUR DES MODÈLES VISUELS AMBIGUS

Modélisation Humaine Tridimensionnelle et Reconstruction du Mouvement dans des Séquences Vidéo Monoculaires

Notre but est de développer des modèles et des algorithmes pour résoudre des tâches visuelles à partir de séquences spatiales ou temporelles d'images. Plusieurs applications concrètes ont été envisagées pendant ce these: la modélisation, reconstruction ou suivi du mouvement de structures déformables ou articulés (des humaines dans des scènes naturelles) et la modélisation incrementale et le suivi du mouvement d'objets paramétriques et déformables complexes.

Commun à tous ces domaines, il existe un modèle fonctionnel direct, qui prédit certain aspects d'images ou signaux, que l'on peut mesurer, et souvent on veut estimer la structure cachée et les paramètres de ce modèle. La modélisation conduit parfois à résoudre un grand problème d'optimisation inverse, et la faisabilité de la solution dépend d'une manière critique de notre capacité d'intégrer de manière consistante, des hypothèses incertaines multiples sur l'espace des configurations, des méthodes de recherche efficaces en hautes dimensions, et des connaissances a-priori sur la structure et les régularités du domaine. Notre travail, gravite, par conséquent, autour de ce cadre commun de modélisation et étudie ses comportements et compromis calculatoires et algorithmiques dans des domaines et applications spécifiques.

Introduction et Approche

Notre domaine de recherche est la modélisation et l'apprentissage visuel et nous nous intéressons particulièrement dans la création des algorithmes pour la construction de modèles paramétriques complexes à partir de séquences d'images spatielles ou temporelles.

On voit **l'approche unitaire** sur ces problèmes à partir d'une structure tripartie: I) un modèle génératif du domaine qui prédit des signaux, images ou des descripteurs - il peut être com-

plexe, 'physique', avec plusieurs paramètres continues et discrèts, géométrique, photométrique, avec occultation, etc.; II) une mesure de ressemblance ou fonction de coût d'appariement modèle-image/observations qui associe implicitement ou explicitement des prédictions génératives avec des primitives images - cette fonction devrait être robuste et parfois motivée probabilistiquement; III) un processus de recherche qui découvre des configurations (les paramètres et parfois la structure cachée du modèle) de haut probabilité sur la surface de coût résultante. Ce cadre est suffisamment général pour accommoder plusieurs paradigmes courants en vision, comme la modélisation Bayesiene, la théorie des formes ('pattern or information theory') ou les formulations basées sur des énergies. Une telle modélisation conduit parfois à résoudre un grand problème d'optimisation inverse, et la faisabilité de la solution dépend d'une manière critique de notre capacité a intégrer de manière consistante, des hypothèses incertaines multiples sur l'espace des configurations, des méthodes de recherche efficaces en hautes dimensions, et des connaissances a-priori sur la structure et les régularités du domaine. Notre travail, gravite, par conséquent, autour de ce cadre commun de modélisation et d'estimation et étudie ses comportements et compromis calculatoires et algorithmiques dans des domaines et applications spécifiques.

Contexte de la These

L'estimation et la modélisation de mouvement humain à partir d'images a déjà fait l'objet de nombreuses études. D'un côté, il existe depuis longtemps des systèmes de « motion capture » commerciaux, qui sont utilisés dans la réalité virtuelle ou augmenté, dans les applications médicales, et plus récemment pour l'animation des personnages virtuels dans les films et les jeux vidéo. Ces systèmes sont à présent très performants, mais ils ont besoin de plusieurs caméras calibrées et synchronisées, d'une illumination contrôlée, et surtout de vêtements et de maquillages spéciaux, muni de cibles actives ou passives qui facilitent les étapes de suivi et de mise en correspondance. De l'autre côté – et c'est de ce côté que cette thèse se positionne – il y a les études qui tentent de travailler avec des séquences vidéo plus « naturelles » : non-calibrées, non-synchronisées, pris dans un milieu naturel et non-instrumenté, et sans cibles ou autres aides à l'appariement. En particulier, nous nous intéressons à la reconstruction (a) des mouvements et (b) des modèles cinématiques / articulaires, des personnages dans des sequence video (par example, les films dramatiques grand public, où aucune information préalable n'est disponible).

Travaux Effectuées Pendant la These

Pendant la these, j'ai travaillé sur l'estimation incrémentale de structure et de mouvement pour des modèles déformables et sur la reconstruction tridimensionnelle du mouvement articulé humain dans des séquences vidéo monoculaires. ce qui a impliqué:

1. Dériver des représentations génératifs articulés et déformables pour des objets complexes rigides, non-rigides et articulés et pour la fusion incrémentale, dans la représentation, des primitifs image non-modélisées au début. Nous avons introduit un cadre theorique pour lier les techniques basées sur des modèles et ceux basées sur des contraintes géométriques entre les images (jusqu'a présent séparées) pour la création incrémentale et robuste des modèles complexes. L'effort de modélisation a-priori d'un objet est relaxé car des parties inconnus (non-modelisees initialement) sont découvertes, reconstruites et intègrées pendant le processus de suivi du mouvement à l'aide de contraintes géométriques de consistance modèle-image.

2. Construire des fonctions consistantes et probabilistiquement motivées d'appariement, par l'intégration des informations basées sur des données contour, texture et silhouette, pour la robustification du processus d'estimation. J'ai aussi initié des travaux sur l'utilisation des couplages de haut niveau ('Gestalt') des données, comme la symétrie ou la colinéarité sur les configurations appariées, pour produire une surface de coût avec plus de cohérence globale.

3. Développer une méthode d'estimation robuste, mixte (continue/discrète) basée sur des hypothèses multiples, pour la recherche dans les espaces de haut dimension comme celui associé avec la modélisation humain monoculaire. Le problème est très difficile à cause du mauvaise conditionnement de la surface de coût, et de la structure complexe, contrainte, induite par des connaissances a-priori sur l'espace de modélisation humain, concernant des limites d'articulations et des contraintes de non-pénétration. Pour ces problèmes, il reste très important de trouver des distributions d'échantillonnage efficaces et creuse pour la génération des hypothèses. Notre algorithme d'échantillonnage basé sur la structure de covariance, (*Covariance Scaled Sampling*), utilise l'incertitude dans la structure locale de la surface de coût pour la génération des hypothèses dans des régions de haute probabilité. On évite par conséquence l'inefficacites des méthodes d'échantillonnage aléatoires classiques comme le filtrage de particules ou CONDENSATION en haute dimension. Ces methodes utilisent le niveau dynamique du bruit comme un paramètre empirique pour concentrer l'effort de recherche sur les paramètres.

4. Proposer une méthode d'échantillonage par importance basée sur la hyperdynamique (*Hyperdynamic Importance Sampling*), qui accélère l'exploration multi-modale des méthodes d'échantillonnage séquentielles Monte Carlo basées sur les chaînes de Markov (Markov Chain Monte Carlo). La méthode utilise le gradient et la courbure locale de la distribution originelle pour construire un échantillonnage par importance, concentrée sur les états de transitions – des points de selle de codimension-1, réprésentant des 'passages en montagne', liant des bassins de coût avoisinés – ou, plus précisément – sur les régions de courbure négative et de coût très bas qui ont plus de chance de contenir des états de transition. Cette méthode

est complémentaire par rapport aux méthodes basée sur le recuit ('annealing'), qui échantillonnent sans discrimination dans un certain niveau énergétique, sans prendre en compte si les points échantillonnés sont susceptibles de conduire vers une autre bassin de potentiel (donc, un autre minimum), ou, toute simplement vers un mur de potentiel de plus en plus haut. La méthode est complètement nouvelle en vision, ou les méthodes basées sur le recuit ou le filtrage de particles ou CONDENSATION etaient, jusqu'à présent utilisées. La méthode a de nombreuses applications dans le domaine d'optimisation globale et calcul Bayesien.

5. Dériver des algorithmes déterministes d'optimisation pour 'cartographier' les minima et les états de transitions dans les modèles visuels ('Mapping Minima and Transitions of Visual Models'). Ces algorithmes tracent des cartes locales des minima voisins et des routes de transition qui les lient entre eux. Les routes de transition sont des routes conduisant d'un minimum, sur un passage bas (ou col) dans la surface de coût et descendant en bas vers un autre minimum voisin. Nous avons introduit deux familles de méthodes numériques pour les calculer: le première est une formulation modifiée de minimisation de Newton, basée sur le suivi des vecteurs propres (*Eigenvector Tracking*), l'autre balayant l'espace avec une hypersurface en mouvement et suivant les minima sur l'hypersurface (*Hypersurface Sweeping*). Les méthodes sont entièrement nouvelles en vision et en optimisation numérique, ou à notre avis, jusqu'à maintenant le problème de trouver les minima multiples d'une maniere deterministe était considéré comme insoluble. Les algorithmes ont donc de nombreuses applications dans des problèmes non-linéaires avec des minima multiples.

Futures Applications Envisagées

1. **La modélisation, reconstruction ou suivi du mouvement des structures complexes déformables ou articulés**. Nous nous intéressons à l'acquisition de modèles complexes de scènes naturelles dynamiques, texturées, contenant des organismes biologiques articulés, déformables ou avec une structure articulaire cachée par des vêtements fortement déformable, etc. De nombreuses applications existent à la fois dans le domaine de la reconstruction photorealiste mais aussi le domaine de la reconstruction robuste (factoriser les effets induits d'une si grande variabilité) pour l'interprétation vidéo de haut niveau. Du point de vue technique, il reste encore des directions ouvertes à explorer sur les représentations géométriques et photométriques nécessaires pour reconstruire la structure et le mouvement de telles scènes complexes. Certaines directions semblent intéressantes, notamment les méthodes qui se situent entre la modélisation physique et l'apprentissage des aspects, comme les travaux récents sur la synthèse de textures a partir d'exemples, la reconstruction de textures 3D, les représentations stratifiées, etc.

2. **La reconnaissance des activités ou la compréhension de scène**. Nous visons des applications comme un bureau ou kiosk intelligent, pour la modélisation et la reconnaissance automatique des activités, des gestes, pour la création du contexte annoté dans la réalité augmentée, ou la fusion de données en utilisant d'autres modalités sensorielles (comme la parole) pour les interfaces de haut niveau. On s'intéresse aussi à la segmentation automatique et l'apprentissage des modèles statistiques en-ligne (e.g. des modèles du mouvement pour la biometrie, la segmentation d'une vidéo basée sur le contenu, et son découpage en plans et scènes pour l'indexation, etc.).

3. **La modélisation incrementale et active des objets ou des sites complexes**. La modélisation géométrique rigide à partir d'images est déja un domaine bien étudié, mais il reste encore des problèmes à la fois pratique comme l'initialisation des algorithmes de reconstruction géométrique, ou théorique comme l'acquisition de modèles plus symboliques des objets (qui sont finalement essentielles pour l'interprétation de haut niveau de la scène). Nous voulons étudier la modélisation et la reconstruction incrémentale des objets à partir des modèles minimaux a-priori, et le découpage automatique de nouvelles données reconstruites en primitives de haut niveau (nous relaxons donc, certains contraintes sur les deux cotes modèle/image: a la fois on ne suppose pas un modèle a-priori complètement connu, mais un modèle partiellement reconstruit en-ligne à base de primitives images découvertes comme consistantes avec son mouvement et/ou sa géométrie).

Problèmes de Recherche et Directions Ouverts

1. **Dériver des algorithmes d'approximation et d'optimisation hybrides quasi-globales en hautes dimensions**, qui utilisent la structure de la surface de coût pour accélérer le processus de recherche sur les paramètres des modèles. Plusieurs problèmes de vision (e.g. le suivi du mouvement) ont une forte cohérence spatiale et temporelle. La multi-modalité est un grand problème, et il faut trouver efficacement des minima voisins avec un minimum donné (courant). On s'intéresse à la fois à des algorithmes probabilistiques et déterministes, basées sur des informations de hauts degrés sur la surface de coût (extrêmes de gradient, etc.), pour concentrer les efforts de recherche sur les paramètres.

2. **Estimation et apprentissage en-ligne dans des modèles graphiques**. Notre but est de pouvoir apprendre en-ligne à la fois la structure cachée des modèles et leur parametres, à partir des données non-étiquetées, dans un seul processus automatique et flexible d'estimation (e.g. suivre les activités dans une séquence vidéo, appendre leurs modèles individuelles,

et segmenter la séquence en morceaux correspondant a chaque activité). Traditionelment, les modèles sont construits hors-ligne, à partir des intuitions du domaine ou heuristiques et leur structure est fixée a-priori. Cela limite toujours leur adéquation aux données réelles en tant que leur robustesse pour une application spécifique. Par contre, l'approche 'flexible' augmente énormement la difficulté du problème d'estimation (nombres de paramètres et de minima locaux, plusieurs modèle donc un caractère mixte, continu/discrèt, etc.) et il n'est pas toujours claire de dire quels sont les algorithmes de recherche et les contraintes images que l'on devrait rajouter pour rendre faisable la solution. Certaines approches variationnelles récentes sont basée sur le découplage des dépendances entre les variables du modèle graphique, et l'exploitation de sa structure locale, dans de calcules pour trouver des minima locaux (e.g. les approximations du champ moyen, etc.). Néanmoins, les erreurs introduit par les approximations et leur impact sur le résultats sont pour l'instant obscure. En plus, dans la plupart de problèmes, trouver un minimum local n'est pas du tout suffisant pour la roubustesse.

3. **L'apprentissage de descripteurs visuelles, pour la construction des mesures de ressemblance statistiques en localisation et classification des objets**. Nous nous intéressons à l'apprentissage efficace des descripteurs pour certaines classes d'objets, à partir des données étiquetées ou non-étiquetées, et dans la discrimination efficace avec le changement de point de vue, occultation, illumination ou des variations entre les classes. Nous voulons aussi examiner l'apprentissage de couplage de haut niveau sur les configurations impliquées dans l'appariement ou les associations des données (e.g. en utilisant des réseaux de neurones qui accomplissent l'achevement des formes de haut niveau a partir de primitives image de bas niveau). Le but est une surface de coût avec relativement peu des minima locaux interpretables, qui peut être recherchée efficacement. La plupart des surfaces de coût utilisées aujourd'hui peuvent avoir un nombre exponentiel de minima locaux (générées par le nombre des choix pour la mise en correspondance modèle-image), et ils manquent une structure continue immédiate.

4. **L'apprentissage de distributions a-priori sur les paramètres de modèles pour stabiliser la solution des problèmes inverses et pour l'initialisation des algorithmes de recherche non-linéaires**. Plusieurs problèmes inverses ont des paramètres qui sont mal contrôlés par des observations, et une distribution a-priori ou un modele plus simple peut être utile pour stabiliser la solution. Néanmoins, obtenir une distribution a-priori représentative n'est pas facile, et son usage peut parfois introduire des biais et augmenter la connectivité du problème, et donc, sa difficulté. Par contre, des techniques pour sélectionner des modeles ou imposer des contraintes existent, mais décider automatiquement quand utiliser ou comparer des résultats obtenus avec des modeles différentes n'est pas facile et des développements sont

encore nécessaires. Les techniques d'apprentissage peuvent être aussi utiles pour initialiser les algorithmes de recherche non-linéaires, e.g. à partir des distributions de probabilité appris sur des fonctions inverses image-modèle.

5. **Combiner différentes modalités visuelles et dériver des mécanismes d'attention active**. Plusieurs formulations, ici compris le Bayesien, permettent la fusion des mesures visuelles différentes ('cues', e.g. des contours, ombres, stéréo ou mouvement), mais décider quelles sont les modalités fiables, les utiliser adaptivement, ou combiner leur résultats (parfois conflictuelles) reste encore un problème ouvert. Comme dans le cas des distributions a-priori, leur usage simultané peut introduire des ambiguités, en incrémentant la cardinalité des minima ou l'ambiguité dans la fonction du coût.

6. **Évaluation des acteurs Bayesiens**. Plusieurs cadres de modélisation existent en vision (Bayesiene, energetique, la théorie des formes) et, comme nous avons déja expliqué dans l'introduction, ils peuvent être réduits à définir un modèle en terme de mesures de ressemblance avec les observations (images) et connaissances a-priori sur son espace paramétrique, ou on recherche des configurations bien supportées. Néanmoins, il reste une problème ouvert comment ces trois components (mesure de ressemblance, connaisances a-priori, complexité de l'algorithme de recherche sur les parametres) interagissent dans des contextes génériques ou spécifiques. Leur importance relative n'est connue que d'une manière qualitative. Par exemple si l'on veut localiser un objet on sait pas quelles connaisances a-priori rajouter sur sa structure, quelles sont les descripteurs images à utiliser pour la mesure de ressemblance et quelle est la relation de dependence entre eux, et finalement quelles sont les limites sur l'effort de recherche dans l'espace paramétrique, et dans quelle conditions l'objet est détectable ou pas. Des éclaircissements sur ces problèmes sont bienvenues non seulement sur le plan théorique mais aussi pratique pour pouvoir quantifier les performances des différentes algorithmes et leur régimes de fonctionnement optimals.

Conclusions et Perspectives

Dans cette thèse nous avons etudié le problème du suivi et de la reconstruction tridimensionnelle de mouvements articulés humains dans des séquences vidéo monoculaires. Dans ce cadre, nous avons proposé trois nouvelles méthodes de recherche continues reposant sur des hypothèses multiples qui nous permettent à la fois une localisation déterministe et un échantillonage stochastique des modes multiples sur les surfaces de probabilité associées aux poses du corps humain: **Covariance Scaled Sampling**, **Eigenvector Tracking** et **Hypersurface Sweeping** et **Hyperdynamic Importance Sampling**. Ces méthodes permettent a la fois un suivi et une reconstruction du mouvement

humain dans des contextes naturels, ainsi qu'une étude systématique et déterministe des surfaces d'erreurs multimodales souvent rencontrées dans des problèmes de vision par ordinateur.

Notre but de recherche est de finalement construire des systèmes qui peuvent, d'une manière plus au moins automatique extraire l'information utile à partir des images. Nous nous intéressons à l'estimation des structures 3D déformables et articulés, la raisonnement vidéo de haut niveau, et la reconnaissance des gestes et actions. Ces applications presentent encore des difficultés au présent, à cause d'une modélisation complexe, de fortes non-linéarites, d'un grand nombre de paramètres, tant que une grande variabilité dans les conditions de capture d'images et observations, qui rend difficile la construction des mesures de ressemblance et des modèles bien adaptées pour l'estimation efficace. A long terme, on s'intéresse aux relations entre les aspects biologiques, calculatoires et algorithmiques de la vision, notamment dans une théorie calculatoire sur la perception visuelle et ses liaisons avec les autres processus neuro-cognitifs.

Chapter 1

Introduction

The major goal of **computational vision** is to provide a quantitative understanding of visual perception that will eventually lead to the construction of artificial systems able to perform visual tasks. Such tasks may involve navigating or observing an environment and extracting representations about the structure, motion or action of the classes of objects and people or animals therein, based on images acquired with video cameras. Many such environments are occupied by humans and extracting meaningful information about their structure, motion or actions is of interest for applications like intelligent human-computer interfaces, biometrics, virtual reality or video surveillance.

In this chapter we give a broad overview of the problem of tracking and reconstructing human motion using sequences of images acquired with a single video camera. We explain the difficulties involved, motivate our approach and our perspective on the problem and conclude with a brief presentation of the major contributions of the work, and the structure of this thesis.

1.1 The problem

The problem we address in this thesis is the incremental tracking and reconstruction of full-body 3D human motion in monocular video sequences. By *incremental*, we mean that the images are available one at a time and we want to update our estimate about the human state after each such new observation[1].

It is legitimate to ask why one should restrict attention to only one camera, in order to attack an already difficult 3D inference problem ? The answers are both practical and philosophical. On the practical side, it is often the case that only an image sequence shot by a single camera is available, for instance when processing and reconstructing movie footage, or when cheap devices are used as interface tools devoted to gesture or activity recognition. From a philosophical viewpoint,

[1]This is in contrast with *batch* approaches that estimate the human state at any given time, using an entire sequence of images, prior and posterior to that time step. This may constrain the tracking, but under such assumptions a real-time implementation would not be possible, even in principle.

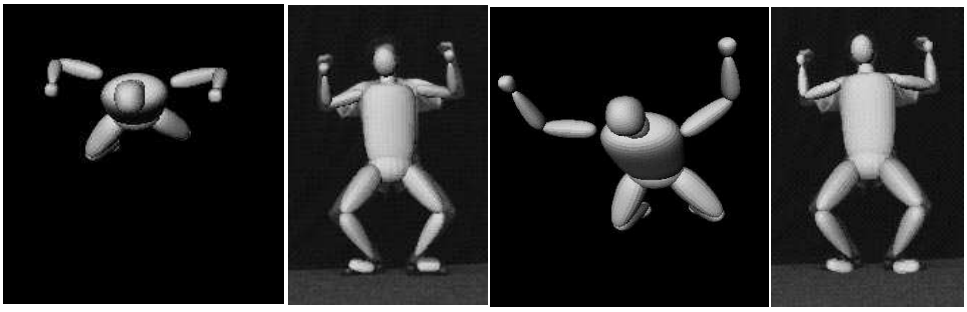


Figure 1.1: Reflective Ambiguities (a,b,c,d). Two very different configurations of a 3D model (a and c) generate good image fits (b and d).

reconstructing 3D motion by using only one eye is something that we, as humans, can do. We don't yet know how much is direct computation on 'objective' image information, and how much is prior knowledge in such skills, or how are these combined. Our intuition says that there are probably those skills that make biological vision systems flexible and robust, despite being based on one eye or many. It was, thus, our hope that by attacking the 'general' problem instead of focusing on problem simplifications, we would make progress towards identifying components of these robust and efficient visual processing mechanisms.

1.2 Difficulties

Extracting monocular 3D human motion poses a number of difficulties that we shall briefly review below. Some are due to the use of a single camera, others are generic computational vision difficulties that would probably arise in any sufficiently complex image understanding or model-based vision problem.

Depth 3D-2D Projection Ambiguities: Projecting the world into images suppresses depth information. This is a fundamental difficulty in computational vision. The fact that we want to make inferences about the world based on this type of information from *only one camera*, firmly places our research in the class of science dealing with inverse and ill-posed problems[2]. The non-uniqueness of the inverse mapping for monocular human pose is apparent in the "forward-backward ambiguities" caused by reflectively symmetric positioning of the human limbs, sloping forwards of backwards, with respect to the camera 'rays of sight' (see fig. 1.1). Reflecting the limb angles in the frontoparallel plane leaves the image unchanged to first order. Such ambiguities can generate likelihood surfaces that have a symmetric, multiple global peak structure.

High-Dimensional Representation: Reconstructing human motion inevitably raises the problem

[2]For vision, an ill-posed problem, is, in the sense of Hadamard, a problem for which inverting the direct 3D-2D mapping operator leads to situations where a solution doesn't exist, is not unique or doesn't depend continuously on the 2D image data (Bertero et al., 1988).

of what one wants to recover and how to represent this. A-priori, a representation in which the 3D human is discretized as densely as possible, with a set of 3D point coordinates, with independent structures and motions is as natural as any other, and could be the most realistic one. Nevertheless, in practice, such a representation would be computationally intractable, since it has too many degrees of freedom that the monocular images cannot account for (or can account for in almost any way)[3]. Representing the entire human as a blob with a couple of centroid coordinates is the opposite extreme, that might give more efficient and constrained estimates at the price of not being particularly informative for 3D reasoning[4]. Consequently, a middle-ground has to be found. At present, this selection is based mostly on intuition and on physical facts from human structural anatomy. For 3D human tracking the currently preferred choice remains a kinematic representation with a skeletal structure covered with 'flesh' of more or less complex type (cones, cylinders, globally deformable surfaces)[5]. For motion estimation, the model needs to have in the order of 30 joint angle parameters, that can explain or reproduce a reasonable class of human motions with some accuracy. However, estimation over a 30-dimensional parameter space is computationally demanding and exhaustive or entirely random search is practically infeasible. The key problem in such a space is therefore the efficient localization of good cost minima (likelihood peaks) by tracing routes, or jumping, between them.

Appearance Modeling, Clothing: We have already emphasized the issue of representation and dimensionality and pointed out that certain compromises must be made in order to balance modeling power and computational tractability. This means that the body modeling will not be entirely accurate. Another difficulty is the presence of deformable clothing which produces very strong variability in shape and appearance and it is not clear how such variability can be accounted for. In principle, cloth simulation and reconstruction techniques do exist, but they are expensive computationally, and difficult to constrain from an estimation viewpoint. Even ignoring the geometry, state of the art texture reconstruction methods require carefully controlled multi-camera and lighting environments (Carceroni and Kutulakos, 2001).

Physical Constraints: Physically inspired representations like the kinematic/volumetric ones just described also need to model the physical constraints of real human bodies. In particular, they have to ensure that the body parts do not penetrate each other and that the joints only have limited intervals of variation (see fig. 1.2). From the viewpoint of estimation, the presence of constraints

[3]The issue here is not the size of the parameter space *per se*, but the lack of 'supporting' image cues that can induce good attractors therein. Even if exhaustive search were available, the space would probably look flat, like an ocean with shallow minima and ridges everywhere.

[4]Apart from tractability constraints, the choice of a representation is also application dependent. Our goal here is to reconstruct the human motion in 3D to a relatively high-level of kinematic or volumetric detail, but not to reconstruct a visually realistic 3D shape model.

[5]From a statistical perspective, more rigorous would be to follow a learned data-driven approach *i.e*. a minimal representation with intrinsic dimension based on its capacity to synthesize the variability of human shapes and poses present in the tracking domain. Given the intrinsic high-dimensionality and non-linearity of the problem, such an approach is at present computationally un-feasible and requires future investigations.

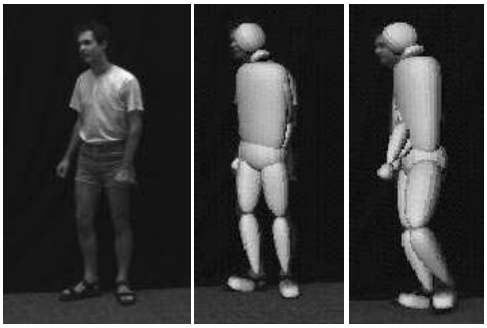


Figure 1.2: Physical Constraint Violations.

is both good and bad news. The good news is that the admissible volume of the parameter space is smaller than initially designed, because certain regions are not physically reachable, and many otherwise possible false solutions may be pruned. The bad news is that the handling the constraints automatically is far from trivial, especially for continuous optimization-based methods, which are among the most efficient methods available for finding peaks on high-dimensional cost surfaces. Even for discrete (sampling) methods, handling constraints in a principled manner is not straightforward given that, besides the need to check which constraints are active, one also needs to decide where to look next during the search (or the sample proposal process). Simply rejecting samples falling outside the admissible parameter space, does not help much and often leads to cascades of rejected moves that slow down the sampler.

Self-Occlusion: Given the highly flexible structure of an articulated human body, self-occlusion between different body parts occurs rather frequently in human tracking and has to be accounted for. Two aspects of this are important. The first is accurate occlusion prediction, so as to avoid the misattribution of image measurements in occluded model regions that have not generated any contribution to image appearance. The second and potentially deeper issue is the construction of a likelihood surface that realistically reflects the probability of different configurations under partial occlusion and viewpoint change[6]. Independence assumptions are usually used to fuse likelihood terms from different measurements, but this conflicts with occlusion, which is a relatively coherent phenomenon. For a realistic likelihood model, the probabilities of both occlusion and measurement have to be incorporated. Building and sampling such a model in the general case is close to being in-tractable without further simplifications, although good ideas can be found in (MacCormick and Blake, 1998).

General Unconstrained Motions: Another important difficulty in reconstructing human motion is that humans move in a complex, rich, but also highly structured ways. Some motions have a repetitive structure (*e.g.* running or walking) or others represent 'cognitive routines' of various levels of complexity (*e.g.* gestures during a discussion, or crossing the street by checking for cars

[6]This issue is particularly important for hypothesis selection in sampling methods or for higher level applications like object recognition.

to the left and to the right, or entering one's office in the morning, sitting down and checking e-mail, *etc*.). It is reasonable to think that if such routines could be identified in the image, they would provide strong constraints for tracking and reconstruction with image measurements serving merely to adjust and fine tune the estimate. Nevertheless, there are several problems with this strategy. Human activities are not simply preprogrammed – they are parameterized by many cognitive and external un-expected variables (goals, locations of objects or obstacles) that are nearly impossible to recover from image data. Also, several activities or motions are often combined. Interesting solutions do exist for some of the problems considered in isolation (dynamic time warping, learning individual motion models, switching models), but their combination is at present computationally intractable, even under the assumption of multiple known motion models.

Kinematic Singularities: These arise when the kinematic Jacobian looses rank and are particularly disturbing for continuous estimation methods since numerical instability is expected and this may lead to tracking failure.[7]

A notable example is the non-linear rotation representation used for kinematic chains, for which no singularity-free minimal representation exists [8]. The set of all such rotation matrices in $\mathbf{R}^3$ is called SO_3. The problem is that there is no global diffeomorphism from (an open sub-set) of $\mathbf{R}^3$ to SO_3. The presence of such a mapping would imply that small changes in the parameters representing a rotation will result in small changes in the attitude of the body (and vice-versa). Such a diffeomorphism does not exist, owing to the different topologies of $\mathbf{R}^3$ and SO_3. In particular, SO_3 is closed and bounded, but $\mathbf{R}^3$ is not. The practical consequence is that every differentiable map from a subset of $\mathbf{R}^3$ into SO_3 has singular points. At these configurations, the differential of the map is not invertible, so in such neighborhoods, small changes in the rotation result in large changes in the parameters of the rotation.

Observation Ambiguities: These ambiguities arise when certain model parameters cannot be inferred from the current image observations. They include but are by no means limited to kinematic ambiguities. Observability depends on the design of the cost surface and image features observed. For instance when a limb is straight and an edge-based cost function is used, rotations around the limb's own axis cannot be observed with a coarse, symmetric body part model. The occluding contour changes little when the limb rotates around its own axis. It is only when the elbow starts moving that the value of the uncertain axial parameter can be observed. This would *not* be an ambiguity under an intensity-based cost function, where the texture on the arms allows the rotation to be observed.

In ambiguous situations, when parameter estimation is performed using non-linear continuous

[7] Different types of behavior may be expected depending on whether both the cost gradient is zero and the cost Hessian is singular, or only the Hessian is singular. In the first case, zero parameter updates are predicted or convergence may occur anywhere near the singularity. For the second case, very large parameter updates are predicted (the gradient being weighted by the local curvatures). The steps may be shortened if the second-order update is inaccurate, but one expects numerical instability there.

[8] Non-singular over-parameterizations are known, but they are not unique.

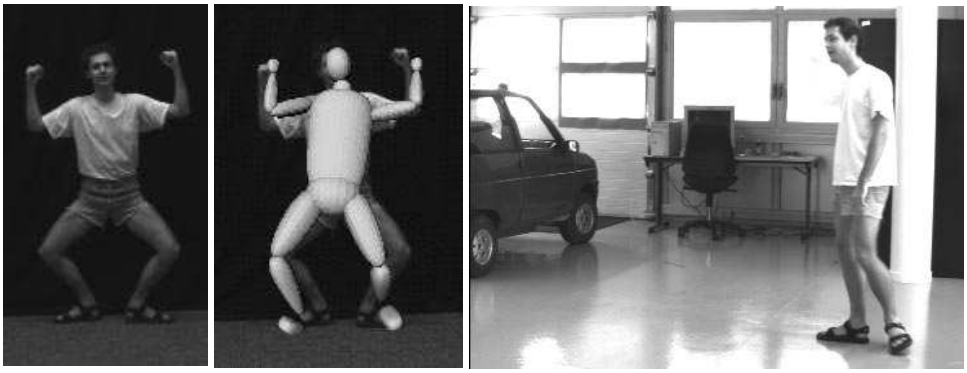


Figure 1.3: Edge matching errors (a,b) and cluttered image containing many distracting, parallel limb-like structures (c).

techniques, there is large uncertainty along difficult to observe parameter combinations, *i.e*. the cost function is almost flat (modulo noise effects) along these directions. For single hypothesis methods (local optimization, extended Kalman filtering), this situation can lead to later tracking failure, due to uncertain initialization at the beginning of the state update iteration in each time frame. The model state estimate may be far away from the true configuration (which could be anywhere in the uncertain/flat cost region), when the singularity is exited. When the ambiguity diminishes as a result of an observable motion, the nonlinear estimator may converge to an undesirable local minimum due to this poor initialization that is out of the basin of attraction of the true state.

Image Clutter, Lighting and Matching Ambiguities and Motion Blur: The problem of matching a generative human model to noisy image observation, belongs to a more general class of vision difficulties. Part of the problem is that the images often contain many distracting clutter features that partly resemble human body parts (for instance various types of edges, ridges or pillars, trees, *etc*., encountered in man-made and natural environments). This is exacerbated by our rather limited current ability to model and extract coherent higher-level structures from image. Most of the likelihood surfaces built to date are assembled from low-level image descriptors such as points or edges, treated naively as if they were independent observations. Using a model to drive the detection process offers a certain degree of robustness and *local* coherence, but still doesn't guarantee the coherence of the *global* result (see also fig. 1.3).

Lighting changes form another source of variability for likelihoods based on edges or intensities. Artificial edges can be created by cast shadows and the inter-frame lighting variations could lead to complicated, difficult to model changes in image texture. The extent to which lighting variation can be accounted for as unstructured noise is limited, although more systematic methods designed to model lighting variations do, in principle, exist (Hager and Belhumeur, 1998).

Finally, for systems with a long shutter time, or during rapid motion, image objects appear blurred or blended with the background at motion boundaries. This has impact on the quality of both static feature extraction methods, and of frame to frame algorithms, such as the ones that

compute the optical flow.

1.3 Approach to Visual Modeling

For the human body reconstruction and tracking problem under consideration, our goal is to recover a complex parametric human model from temporal image sequences. Summarizing the previous section, by complex, we mean that the problem has many parameters, an internal structure with complex chains of functional dependencies, a hybrid composition with several aspects to model (geometry, photometry, prior knowledge, statistics, etc.), and owing to its size and multi-modality efficient search methods are needed to render the solution feasible.

We approach the problem in terms of a three part "model-based vision" structure: I) a generative human model that predicts images or descriptors – this is complex, partly 'physical', with many continuous parameters, geometry, photometry, occlusion, *etc.*; II) a model-image matching cost function that implicitly or explicitly associates generative model predictions with extracted image features – this is robust and where possible probabilistically motivated; III) a search or optimization process that discovers well-supported configurations on the resulting cost surface. This framework is general enough to cover several current vision paradigms, including Bayesian modeling (Geman and Geman, 1984; Isard and Blake, 1998), pattern theory (Mumford, 1994), and energy-based formulations (Mumford, 1996). The formulation typically leads to solving a large non-convex inverse problem whose robust solution requires the consistent integration of prior knowledge and the use of efficient multiple hypothesis search methods to locate possible solution groups. When tracking image sequences, the problem has a strong temporal dimension, the results of the previous time frame providing a dynamical initialization for the search at the current time step [9].

1.4 Perspective on the Search Problem

From a temporal perspective, the human tracking problem can be viewed as the process of following multiple peaks on a high-dimensional surface that evolves dynamically, but in discrete time steps. From both a distribution sampling and global optimization viewpoint, the crucial problems are: (*i*) Locating the attraction regions of important peaks[10] and, once such basins of attraction have been identified, (*ii*) Efficient sampling the corresponding peaks (when using sampling methods), or their precise location (when using optimization).

Given the problem being studied, our emphasis is on tracking and reconstructing general human motion with minimal assumptions about the clothing, the background and especially the motion of

[9] The fact that such priors are useful computationally is mainly motivated by the assumption of temporal (quasi)coherence in the evolution of the process we are modeling (human motion).

[10] Theoretically, both processes are shadowed by weak convergence guarantees. Practically, in high-dimensions, the search process is limited to prior neighborhoods or stopped long before any expectations about exploring all peaks could be made.

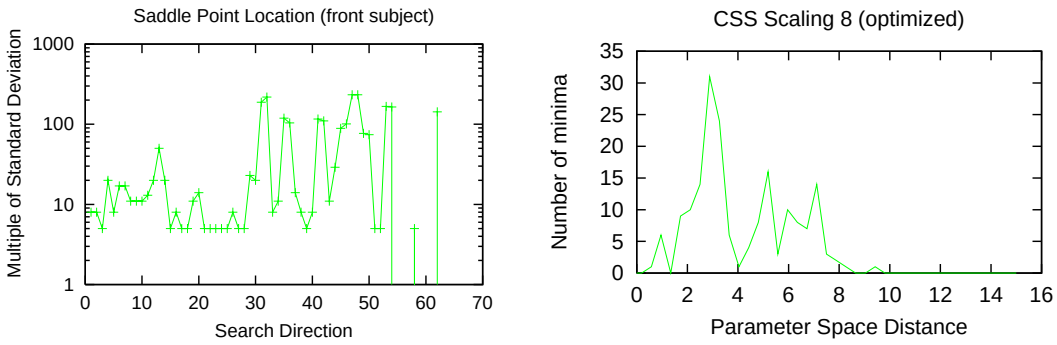


Figure 1.4: Inter-minimum transition regions along different directions away from a minimum (left) and minima distribution (right) for human poses. Note that the saddle points leading to other minima are far in parameter space. The minima themselves are even further, the distances in the plot are for a parameter space distances in radians and meters for 3D human position and joint angles.

the human subject. The goal can thus be formulated more clearly:

How can we derive multi-modal search methods that are admissible (that respect the internal constraints of the representation) and that efficiently exploit the structure of the high-dimensional likelihood function without making restrictive assumptions about the dynamics of its peaks ?

Until now, the process of jumping between peaks at any given time step or tracking them between two successive time steps[11] has relied on either known, fairly precise dynamical models or random sampling using dynamical noise, when no information about the motion of the subject was available. Note that, in practice, likelihood peaks are usually separated by many standard deviations (see also fig. 1.4). When falling onto a spurious peak[12], *e.g.* due to incorrect assignments of a model edge to an incorrect limb side of a target subject, many model-image assignments will be wrong. Distant sampling is usually necessary to decouple the wrongly assigned model features and lead the model towards a different set of assignments that will attract it into the correct state. Nevertheless, under unknown dynamics, and for high dimensional problems with strong multi-modal structure, an escape strategy based on a wide but uniform dynamical noise is un-effective. This is because a small noise level produces insufficient sample scatter to leave the current peak, while a sufficiently wide level wastes an unaffordably large number of samples in fruitless regions.

The reasons why designing efficient escape strategies is non-trivial can now be formulated more precisely: (*i*) The extent of the basins of attraction is not known a-priori. (*ii*) Distant sampling is necessary to escape the basins and this leads to large volumes that need to be sampled efficiently[13].

[11] Precisely, this means jumping from the central neighborhood of a prior peak to it's basin of attraction after it has moved in-between two observation time frames.

[12] Such peaks are artifacts of global modeling incoherences or matching ambiguities.

[13] Given the high-dimensionality of the problem, a dense coverage will be usually not feasible, so sparse sample

(*iii*) For ill-conditioned problems, the minimum shape is non-spherical and usually not aligned with the axes of the parameter space – instead it is highly elongated along directions corresponding to difficult to observe parameter combinations.[14]. (*iv*) Even if a sample escapes out of the current basin of attraction, in high-dimensions, the probability of hitting the high-cost surroundings of another minimum is much larger than that of hitting its low cost core – this is due to the large increase of volume with radius, which makes the volume of the core of the minima tiny in comparison with their full basins of attraction. (*v*) For discrete sampling methods, high-cost samples are un-likely to be selected for resampling, unless the entire distribution is dominated by them – which will usually happen only if the track has already been lost.

During this thesis, we will argue that efficient search methods must use the local cost surface in order to optimally drive the search trajectories or shape the sampling distributions. This is necessary in order to sample sufficiently distantly to escape minima, but, at the same time, sufficiently approximately so as not to waste too many samples. We consider that some form of continuous optimization, or local descent is almost indispensable to detect true low-cost regions, given the low probability of finding them by purely random sampling[15]. We show then that optimization-based deterministic trajectories can be constructed that precisely locate transition neighborhoods linking adjacent cost basins. This leads to a method for systematically identifying nearby minima. Finally, we explain how a random search can be implicitly and adaptively focused near transition regions by using gradient and cost curvature information to build a modified cost function. Sampling such modified function, gives substantially increased minimum exploration behavior.

A more detailed formulation of the contributions is given in the next section.

1.5 Contributions

Our contributions are motivated by what we believe to be the crucial problems discussed above, namely robust and physically consistent likelihood modeling and efficient, concentrated, high-dimensional search.

1.5.1 Modeling

In chapter 3, we describe our human body model and its consistent physical priors, and our image likelihood design. Together, these allow efficient continuous search in the space of admissible human parameter combinations, with good resistance to mismatches and outliers.

---

proposals are needed.

[14]For the human monocular pose case, such combinations of parameters will generate difficult to observe motions towards the camera. Multiple reflective minima would therefore be aligned with the camera rays of sight.

[15]In fact, complex high-dimensional distributions cannot be efficiently sampled without this. Some influential sampling strategies like Hybrid Monte Carlo or Langevin method, do actually use gradient information to generate moves.

Constrained Consistent Anthropometric Body Model

We introduce a continuously parameterized human body model that incorporates prior knowledge on anatomical joint angle limits and physical body-part non-interpenetration constraints. We also include anthropometric priors that stabilize the estimation of specific human body parameters by biasing them towards default human dimensions. The formulation supports both efficient second-order continuous optimization and sampling, in a way that is coherent with the priors and physical constraints. Optimization and sampling methods based on these properties, are proposed in chapters 4, 5 and 6.

Robust Coherent Likelihood Design

We propose robust, coherent and probabilistically motivated matching functions that integrate contours, intensity and (if available) silhouette data, to robustify the estimation process. We have also initiated studies on using higher-level 'Gestalt' feature couplings like symmetry or co-linearity over matched configurations to produce a globally coherent cost surface. Our likelihood measure combines robustness and consistency properties, and integrates multiple visual cues and model-image assignment uncertainty in order to limit the number of spurious responses in parameter space.

1.5.2 Covariance Scaled Sampling

In chapter 4, we derive a robust multiple hypothesis estimation method that efficiently searches the high-dimensional space associated with monocular human modeling. The method addresses the problem of the ill-conditioned nature of the cost surface, and the complicated, constrained nature of the human pose search space. In this context, it is essential to find efficient, sparse proposal distributions for hypothesis generation. Our *Covariance Scaled Sampling* strategy uses the local uncertainty structure of the cost surface to generate samples in probable regions. This is in contrast with classical sampling methods (particle filtering, CONDENSATION), that use the dynamical noise level as an empirical search focusing parameter.

1.5.3 Building Deterministic Trajectories for Finding Nearby Minima

In chapter 5 we derive deterministic optimization algorithms for finding nearby minima in visual models. These algorithms construct local 'roadmaps' of the nearby minima and the 'transition pathways' linking them. These pathways are routes leading from a minimum, over a low 'pass' in the cost surface, and down into a neighboring minimum. As far as we know, such methods are unknown in computer vision and numerical optimization, but some do exist in computational chemistry, where transitions are important for predicting reaction parameters. We present two families of methods, originally inspired from computational chemistry, but here generalized, clarified and adapted to the requirements of model based vision: one is a modified form of Newton minimization based on eigenvector following (*Eigenvector Tracking Algorithm*), the other propagates a moving

hypersurface through space and tracks minima on it (*Hypersurface Sweeping Algorithm*). The algorithms are potentially useful in many nonlinear vision problems, including structure from motion or shape from shading. They are also applicable in global optimization or statistical calculations.

1.5.4 Hyperdynamic Importance Sampling

In chapter 6, we propose an importance sampling method that accelerates the inter-mode exploration behavior of sequential Markov Chain Monte Carlo samplers using 'hyperdynamics'. This method uses the local gradient and curvature of the input distribution to construct an adaptive importance sampler focused on 'transition states' – codimension-1 saddle points representing 'mountain passes' linking adjacent cost basins – or more precisely on low cost negative curvature regions that are likely to contain transition states (*Hyperdynamic Importance Sampling*). This is complementary to annealing methods that sample indiscriminately within a certain energy band, regardless of whether the points sampled are likely to lead out of the basin to another minimum, or whether they lead further up in an ever-increasing potential wall. The method was originally developed in computational chemistry, but as far as we know this is its first application to vision or Bayesian inference. It also has applications in global optimization or probabilistic reasoning computations.

1.6 Thesis Outline

This chapter has given an overview of the problem of study, as well as its difficulties. We discussed our modeling approach for human motion reconstruction and tracking, as well as our perspective on taming the difficulties involved.

We continue with **chapter 2**, where we briefly review related approaches to human body tracking. We present both a broad review arranged along several major modeling and estimation axes, and a more in-depth review of relevant systems that track and reconstruct 3D body models. Also, in each of the below chapters, we compare our methods to selected existing approaches.

In **chapter 3**, we explain the problem of modeling, including our representation of volumetric and kinematic structures, the constraints and priors used, and the probabilistic contour, intensity and silhouette likelihoods that we use in different experimental contexts in subsequent chapters. We also show several experiments revealing the motivation behind the proposed likelihood design and review other approaches to human likelihood construction in human modeling.

Chapter 4 describes a Bayesian tracking framework of a representation based on Gaussian mixture distributions. The method combines robust constraint-consistent local optimization for efficient mode seeking with covariance scaled heavy tailed sampling for broad hypotheses generation in probable, low-cost regions. We perform experiments for an arm and full body tracking involving general motions against both a simple and highly cluttered backgrounds.Relationships to other high-dimensional search and sampling strategies are also discussed.

Chapter 5 presents two deterministic methods for finding nearby local minima. These methods are based on constructing trajectories that locate the first-order saddle points that link the basins of adjacent minima, followed by local descent in neighboring minima themselves. Experiments showing the efficiency of the methods on different likelihood surfaces for both inter-frame tracking and human pose estimation are presented. Reviews and relationships with relevant methods in computational chemistry and global optimization are discussed.

Chapter 6, describes an 'hyperdynamic' importance sampling method that modifies the likelihood surface adaptively using gradient and curvature information. This results in automatic focusing of the search on transition regions linking adjacent minima. We discuss relationships to alternative sampling methodologies and present pose estimation experiments that show significant increase in the inter-minimum exploration behavior for MCMC methods.

Chapter 7 presents several comparative sampling experiments and analyzes the complexity of the search problem. We discuss the connexions between the structure and properties of the likelihood, and classify the strategies and solution techniques that can be used for different classes of search problems. We characterize problem difficulty and give suggestions on which methods among the ones we have proposed (as well as other known ones) may be effective in each problem context.

In **chapter 8**, we present a technique for tracking deformable objects, that incrementally detects and fuses initially un-modeled low level image features into the representation. Such methods can be combined with human tracking methods in environments where the human is interacting with objects, or where a full model representation is not apriori available. This work is complementary to the mainstream human tracking research in the thesis.

We conclude in **chapter 9**, with a summary of open problems and perspectives for future research.

1.7 List of Relevant Publications

The research described in this thesis is based on material from the following publications:

1. C.Sminchisescu, B.Triggs: Building Roadmaps of Local Minima of Visual Models, *European Conference on Computer Vision*, Copenhagen, June 2002.

2. C.Sminchisescu, B.Triggs: Hyperdynamics Importance Sampling, *European Conference on Computer Vision*, Copenhagen, June 2002.

3. C.Sminchisescu, B.Triggs: Estimating Articulated Human Motion with Covariance Scaled Sampling, *International Journal of Robotics Research*, 2002, submitted.

4. C.Sminchisescu: Consistency and Coupling in Human Model Likelihoods, *IEEE International Conference on Automatic Face and Gesture Recognition*, Washington D.C.,May 2002.

5. C.Sminchisescu, A.Telea: Human Pose Estimation from Silhouettes. A Consistent Approach Using Distance Level Sets, *International Conference on Computer Graphics, Visualization and Computer Vision*, WSCG, Czech Republic, February 2002.

6. C.Sminchisescu, B.Triggs: Covariance-Scaled Sampling for Monocular 3D Body Tracking, *IEEE International Conference on Computer Vision and Pattern Recognition*, CVPR, Hawaii, December 2001.

7. C.Sminchisescu, D.Metaxas, S.Dickinson: Improving the Scope of Deformable Model Shape and Motion Estimation, *IEEE International Conference on Computer Vision and Pattern Recognition*, CVPR, Hawaii, December 2001.

8. C.Sminchisescu, A.Telea: A Framework for Generic State Estimation in Computer Vision Applications, *International Conference for Computer Vision, International Workshop on Computer Vision Systems*, ICVS, Vancouver, July 2001.

9. C.Sminchisescu, B.Triggs: A Robust Multiple Hypothesis Approach to Monocular Human Motion Estimation, *INRIA Research Report No. 4208*, June 2001.

10. C.Sminchisescu, D.Metaxas, S.Dickinson: Incremental Model-Based Estimation Using Geometric Consistency Constraints, *INRIA-Research Report No.4209*, June 2001.

Chapter 2

State of the Art

Estimating human motion from image sequences is a challenging and active research topic (see (Gavrila, 1999; Moeslund and Granum, 2001; Aggarwal and Cai, 1999) for partial reviews) with a large spectrum of potential applications including human-computer interfaces, animation, special effects and virtual reality, automated biometrics, model-based image coding and intelligent surveillance, *etc*. Different applications require different levels of representation detail. For instance, surveillance applications often require low-level 2D image blobs or centroid coordinates, motion capture systems extract precise 3D angles of the human articulations, while scene interpretation applications infer high-level information about human activity type or behavior (walking, turning, running, object interaction, anger, impatience, *etc*.). This chapter gives a brief overview of the state of the art with particular attention to articulated 3D body tracking methods which are the focus of this thesis.

2.1 Modeling and Estimation

We view the formulation of the human motion estimation problem as consisting of two major components: **modeling** and **estimation**. By modeling, we ultimately understand, the construction of the likelihood function. This involves the human body model (volumetric and kinematic representation), the camera type, the types of model and image descriptors extracted, and the way they are matched and fused. Additionally, it has to account for the structure of the parameter space and its internal or physical constraints. Once a likelihood surface is available, estimation is a matter of searching it for low cost configurations (model instantiations that 'explain' image well). The likelihood surfaces in monocular human modeling are usually multi-modal, so robust, multiple hypothesis methods are needed if one aims for solution quality. From the viewpoint of estimation the human tracking problem can be also divided in the two sub-problems of **initialization** and **tracking**. However, there is not really any qualitative difference between these two processes and the distinction between them should be less blurred in a robust system owing to the need to recapture lost tracks. In practice, initialization involves a more global search over the image and/or in model

parameter space, while tracking assumes temporal coherence and small inter-frame motions. Under these assumptions, the search for parameters can be localized to the neighborhood of optimal states (*i.e.* posterior peaks or minima basins) of the previous tracking time-step.

2.1.1 Classification Axes

Several criteria can be used to classify human body tracking methods: the representation used to model body parts (cylinders, quadrics, cones, deformable), and kinematics (Euler angles, twists, independent parts); the image features used (edges, intensities, silhouettes); or the motion models (noisy Gaussian dynamics, linear auto-regressive processes, dedicated walking or running models); the parameter estimation method (single or multiple hypothesis methods). In the following sections we give a broad overview of some relevant approaches, according to these modeling and estimation ideas.

2.1.2 Body Models

A large variety of 2D and in 3D human models have been proposed. For surveillance purposes, 2D image blob models (Papageorgiu and Poggio, 1999; Haritaoglu et al., 1998; Comaniciu et al., 2000; Isard and MacCormick, 2001) have proved effective for approximate 2D human tracking. Other 2D1/2 models include planar articulated representations where the model is built in terms of limbs represented as intensity patches and joints modeled as planar articulations (Ju et al., 1996; Morris and Rehg, 1998; Cham and Rehg, 1999). Foreshortening effects due to depth variation during tracking, are modeled to a certain extent by allowing limb sizes to vary. Such models have been employed both for approximate human localization, body part labeling (Ioffe and Forsyth, 2001; Ronfard et al., 2002) and tracking (Ioffe and Forsyth, 2001; Cham and Rehg, 1999; Ju et al., 1996), but despite their effectiveness, they are fundamentally represent only 2D information and can not represent 3D constraints as well as a true 3D model. However, the 2D output in terms of the human image position or body labeling information could be a valuable component for higher level modules (*e.g.* it could form the basis for constructing 3D more coherent likelihood surfaces).

Instead, we will follow the line of research that directly estimates the human motion in 3-space and using 3D volumetric and kinematic models (Rehg and Kanade, 1995; Kakadiaris and Metaxas, 1996; Gavrila and Davis, 1996; Bregler and Malik, 1998; Deutscher et al., 2000; Wachter and Nagel, 1999; Sidenbladh et al., 2000; Plankers and Fua, 2001; Drummond and Cipolla, 2001; Sminchisescu and Triggs, 2001b). Volumetric representations include cylinders (Bregler and Malik, 1998; Sidenbladh et al., 2000), tapered cones (Deutscher et al., 2000; Wachter and Nagel, 1999) and more complex deformable shapes (Kakadiaris and Metaxas, 1996; Gavrila and Davis, 1996; Plankers and Fua, 2001; Sminchisescu and Triggs, 2001b). Most of the authors have used minimal representations with implicit articulation constraints, modeling the degrees of freedom associated to each joint (up to 3), although others (Kakadiaris and Metaxas, 1996; Drummond and Cipolla, 2001) allow independent rigid displacements for each body part and model coincidence of articulation

centers separately. Such non-minimal, internally constrained parameterizations are usually more difficult to handle during parameter estimation but they avoid the parameterization singularities (Morris and Rehg, 1998; Deutscher et al., 1999) encountered with minimal representations.

In principle, very complex and realistic 3D body models (muscle and articulation) are possible and these might be effective in controlled scenarios, especially where the subject wears tight fitting clothing. In practice, the presence of loose clothing and the significant body variability between tracked subjects, mean that simpler models have to be used, at least in the current state of the art.

2.1.3 Image Descriptors

Many different image descriptors have been used to build likelihood surfaces for human tracking systems. Some authors use sparse edge information extracted in the neighborhood of the individual human model contour predictions (Kakadiaris and Metaxas, 1996; Gavrila and Davis, 1996; Deutscher et al., 2000; Wachter and Nagel, 1999; Drummond and Cipolla, 2001; Sminchisescu and Triggs, 2001b). Edges offer advantages in terms of partial invariance to viewpoint and lighting and good detectability properties in humans. Nevertheless, for human tracking, edges have certain limitations. In particular, when limbs rotate around their 3D symmetry axes there is little edge change in the image and such motions tend to ultimately pass undetected (*i.e*. due to partial invariance of a limb projected occluding contour with respect to its 3D axial motion).

Other systems use dense intensity cost surfaces based on either the image texture under the model projection at initialization stage (Rehg and Kanade, 1995) or the texture observed at the optimal estimated model configuration in the previous tracking time step (Bregler and Malik, 1998; Wachter and Nagel, 1999; Sidenbladh et al., 2000; Sminchisescu and Triggs, 2001b). Implicitly the reference texture is mapped onto the model surface and then matched against the image texture in the neighborhood of the model prediction for the current time frame. Although dense methods exploit rich sources of information that can help to disambiguate in the axial limb motion case previously described, the presence of intensity variations, deformable clothing or the lack of texture may significantly decrease their performance and lead to tracking drift.

Silhouette information, has also proved useful for human tracking in case it is available from a background subtraction, motion segmentation or contour tracking process (Deutscher et al., 2000; Delamarre and Faugeras, 1999; Sminchisescu, 2002; Brand, 1999; Haritaoglu et al., 1998). In general, silhouettes provide strong global lock information, although they can also be un-informative for certain camera viewpoints, when limbs are projected inside the body contour.

A number of robust and effective tracking systems integrate image cues providing complementary information: (Wachter and Nagel, 1999) use edge and intensity cues; (Deutscher et al., 2000) use silhouette and edge information; (Sidenbladh and Black, 2001) use edge, ridge and intensity likelihoods; (Sminchisescu and Triggs, 2001b) used edge, model-based intensity and if available silhouette information (Sminchisescu, 2002) from motion segmentation (Black and Anandan, 1996) or adaptive background subtraction processes (Haritaoglu et al., 1998), *etc*.

2.1.4 Number of Cameras

Most of the 3D tracking systems proposed to date use multiple cameras to alleviate the effects of occlusion and depth ambiguities (whereas for 2D systems, several cameras are often useful but not mandatory). Full body tracking systems based on monocular cameras include (Deutscher et al., 2000; Sidenbladh et al., 2000) for periodic motion and (Sminchisescu and Triggs, 2001b) for unconstrained motion. Monocular systems that track unconstrained lower-dimensional arm or upper body motions are (Gonglaves et al., 1995; Wachter and Nagel, 1999; Sidenbladh and Black, 2001).

Another relevant aspect concerns knowledge of camera external pose and internal parameters (*i.e.* focal length, aspect ratio). Most methods use a camera with fixed external orientation and internal parameters that are fixed and estimated off-line. Reliable methods do currently exist for off-line camera calibration (Hartley and Zisserman, 2000). In fact, external calibration information is not strictly necessary in the monocular case, as the motion of the human subject can be always estimated with respect to the camera considered fixed. Consequently, the motion of the camera can be 'absorbed' in the global rigid motion of the human[1].

2.1.5 Prior Knowledge

The use of prior information about human poses, constraints and motions can add an important degree of robustness to a human tracking system, especially in the 3D case. Such knowledge can come from several sources. Statically, the parameter space defined by the joint angles is highly constrained and for full consistency one has to enforce the anatomical joint angle limits (Sminchisescu and Triggs, 2001b; Wachter and Nagel, 1999) and also the physical non inter-penetration constraints (Sminchisescu and Triggs, 2001b) between the body parts. For discrete estimation, enforcing anatomical joint angle limits is relatively straightforward (although care has to be taken with dynamics on the admissible domain boundary) and this has been addressed by (Sminchisescu and Triggs, 2002b; Deutscher et al., 2000; Sidenbladh et al., 2000). For the continuous case, the situation is more delicate and involves constrained optimization (Sminchisescu and Triggs, 2001a,b). Care needs to be taken in the design of the error surface for first and second order continuity conditions if the constraints are soft. Alternatively, constraint projection methods could to be used in the optimizer when implementing inequality hard constraints[2]. Robust and efficient ways for doing this are presented in chapter 3 and are among the contributions of this thesis. Note also that such knowledge is difficult to enforce in 2D systems (Ju et al., 1996; Morris and Rehg, 1998; Cham and Rehg, 1999), as their representations do not directly support 3D consistency reasoning.

Additional priors on the parameter space come from statistical information about the human proportions (Sminchisescu and Triggs, 2001a,b) or the motion performed (Sidenbladh and Black,

[1] This is only true when using one camera. For a scene observed by multiple moving cameras, more complex calibration procedures would be necessary.

[2] Other methods are available for implementing more complex types of non-linear constraints, but constraint projection is a particularly effective one in this problem context.

2001; Deutscher et al., 2000; Rohr, 1994). Motion modeling represents a good source of constraints that can be added to stabilize tracking, especially in situations where information about the motion of the human subject is available. Models studied include example-based motion models (Leventon and Freeman, 1998; Sidenbladh et al., 2002), specific walking models (Rohr, 1994; Sidenbladh et al., 2000) and higher-oder autoregressive processes (Deutscher et al., 2000). It is possible to track using multiple motion models (North and Blake, 1998; Blake et al., 1999; Pavlovic et al., 2001), and to estimate the transitions between them, although the approach becomes significantly more expensive. One challenging problem in motion modeling remains the construction of priors that are sufficiently generic to be useful in realistic scenarios. A recent step in this direction is (Sidenbladh et al., 2002).

2.1.6 Estimation

Estimation methods must search of the likelihood surface for good configurations. Strictly local search does not usually suffice when dealing with the multi-modal distributions encountered in human modeling. In practice, however, for high-dimensional parameter spaces, the search must necessary remain fairly local owing to the complexity of exhaustive search in many dimensions[3]. Estimation methods proposed for tracking include single hypothesis continuous Kalman filtering or local-optimization methods (Rehg and Kanade, 1995; Kakadiaris and Metaxas, 1996; Bregler and Malik, 1998; Wachter and Nagel, 1999; Plankers and Fua, 2001; Drummond and Cipolla, 2001), discrete single hypothesis methods (Gavrila and Davis, 1996), discrete CONDENSATION-based multiple hypotheses methods (Deutscher et al., 2000; Sidenbladh and Black, 2001), continuous multiple hypothesis methods (Choo and Fleet, 2001; Sminchisescu and Triggs, 2002a,b) and mixed continuous/discrete multiple hypothesis methods (Heap and Hogg, 1998; Cham and Rehg, 1999; Sminchisescu and Triggs, 2001a,b). The continuous search strategies become more valuable as the dimension increases, owing to the increasingly low probability of reaching probable regions by purely random sampling.

[3]Note that, in fact, the relevant quantity here is not the parameter space volume *per se*, but the ratio between the volume of support of the prior and likelihood, respectively. Nevertheless, the fact remains that even this volume is in most cases prohibitively large and cannot be searched exhaustively.

2.2 Articulated Full-Body Tracking Methods

This section presents a more detailed review of research directions that study on full-body 3D tracking. We pay particular attention to methods of extracting 3D information about the motion of the body parts. Note, that only few authors have addressed the *monocular 3D full-body tracking* problem (Deutscher et al., 2000; Sidenbladh et al., 2000; Sminchisescu and Triggs, 2001b).

In the presentation, we group methods based on the estimation strategy used. We have selected this criterion as, ultimately, as a result of modeling, one is left with a high-dimensional, multi-modal error surface that has to be searched robustly and efficiently. Furthermore, deriving such robust and efficient search methods that recover likely configurations of the target has been one of the major goals of this thesis line of research.

2.2.1 Single Hypothesis Methods

Gavrila and Davis use an articulated three-dimensional model built from tapered superquadrics and perform tracking hierarchically, based on a ring of four inwards looking calibrated cameras and edge distance transforms in the form of Chamfer maps (Gavrila and Davis, 1996). They also use color labeling, the tracked subject wearing tightly fitting clothing with sleeves of contrasting colors. The authors propose a top-down discrete search-space decomposition technique that first finds solutions for the body parts higher-up in the hierarchy of the kinematic chain (like the torso for instance), followed by recursive search for the configurations of the body parts lower down in the hierarchy (e.g. the individual limbs). The method encounters difficulties with occlusions (in general, only partial body configurations are available at any given time, so visibility order is not available) and biases owing to premature fixing of subsets of parameters on the articulated chains during fitting. Experiments involving general upper-limb and body tracking were reported.

Kakadiaris and Metaxas represent individual body parts as deformable superquadrics, under general rigid displacements, parameterized independently (Kakadiaris and Metaxas, 1996). Kinematic constraints are enforced with soft point to point springs between individual 3D body parts for spherical joints and hinge-axis to hinge-axis ones for hinge joints. Motion estimation is performed using Lagrangian dynamics, embedded in a Kalman filtering predictor, where ‘forces’ derived from image contours are used to align the model occluding contours with the data. From an optimization viewpoint, the Lagrangian approach is equivalent to a non-linear least squares estimator (either gradient descent or Gauss-Newton depending on the defined mass, damping and stiffness matrices), with soft penalty functions corresponding to the body part adjacency constraints. They use 3 orthogonally calibrated cameras and a stationary background to track the planar motion of an arm. A heuristic active viewpoint method is used for camera selection, in order to avoid potential image degeneracies during tracking based on the percentage of self-occlusion of the body parts, as well as the amount of first order model predicted occluding contour change in different views. A shape initialization algorithm is also proposed in (Kakadiaris and Metaxas, 1995) based on a motion pro-

Authors	Estimation			Cam	Shape Model	Physical Constraints		Occl Pred	Image Cues	Motion Model
	Hyp	Type	Method			Joint Limits	Body Collision			
This work	M	C	Covariance Scaled-Sampling	1	deformable superquadrics	Y	Y	Y	probabilistic edge + intensity	-
Deutscher et al.	M	D	Annealed Condensation	3	cones	Y	-	Y	edge + silhouette	multiple ARP
Sidenbladh et al.	M	D	Condensation	1	cylinders	Y	-	Y	intensity	walk
Plankers & Fua	1	C	least-squares	3	deformable meta-balls	-	-	-	stereo + silhouette	-
Drummond & Cipolla	1	C	weighted least-squares	3	quadrics	-	-	-	edges	-
Delamarre & Faugeras	1	C	energy minimization	3	cones	-	-	Y	silhouette	-
Wachter & Nagel	1	C	Kalman filter	1	cones	Y	-	-	edge+intensity	linear
Wren & Pentland	1	C	Kalman filter	2	ellipsoid blobs	-	-	Y	color blobs	linear
Bregler & Malik	1	C	least-squares	3	cylinders	-	-	-	intensity	-
Kakadiaris & Metaxas	1	C	least-squares	3	deformable superquadrics	-	-	Y	edge	linear
Gavrila & Davis	1	D	best-first search	4	tapered superquadrics	Y	-	Y	edge + color label	linear
Cham & Rehg	M	C	MHT	1	2D template	-	-	-	intensity template	linear
Ju et al.	1	C	gradient descent	1	2D template	-	-	-	flow	linear
Brand	1	C	entropy minimization	seq	-	-	-	-	silhouette	HMM-based
Rosales & Sclaroff	1	C	EM	1	-	-	-	-	silhouette	-
Leventon & Freeman	1	C	MAP	seq	-	-	-	-	hand-marked joints	PCA-based
Liebowitz & Carlsson	1	C	linear SVD	3/seq	-	-	-	-	hand-marked joints	-
DiFranco et al.	1	C	least-squares	seq	-	Y	-	-	2D&3D hand-marked joints	smooth

Table 2.1: Comparison of different human tracking systems. 'M'=multiple hypothesis based, 'C'= uses continuous non-linear optimization, 'D' = uses discrete sampling.

tocol for human subject motion and extracted silhouette in the cameras (experiments for arm and leg acquisition are reported).

Bregler and Malik represent the kinematic chain using twists, and use cylinders to model the shape of individual body parts (Bregler and Malik, 1998). The relationship between the joint evolution and image motion is, as in many other cases (joint angles, rigid displacements, Denavit-Hartenberg), based on linearizing the kinematic and projection model around the current configuration, using Taylor expansion to pass from twists to rigid displacements. Consequently, the approximation is *always* linear and iteration is needed to reach a minimum in configuration space. The authors linearize a Gaussian brightness error model, and employ a model-based optical flow iteration based on a Gauss-Newton pseudo-inverse approximation (which is undamped and hence may fail to converge to a local minimum, see also the chapter 5 in this thesis for a related discussion). The authors use 3 cameras to track a subject walking parallel to the image plane, using only a half skeleton (up to 19 d.o.f.), modeling the visible side of the body.

Wachter and Nagel use articulated kinematics and a shape model built from of truncated cones, and estimate motion in a monocular sequence, using edge and intensity (optical flow) information using an extended Kalman filter (Wachter and Nagel, 1999). Anatomical joint limits are enforced at the level of the filter prediction, but not during the update step, where they could be violated. They show experiments in an unconstrained environment for a subject wearing clothing and track motion parallel with the image plane using articulated models having 10-15 d.o.f.

Wren and Pentland perform a coarse modeling of the hands and head of the human body from ellipsoid surfaces, reconstructed by tracking color blobs in the image (Wren and Pentland, 1998). They use a Lagrangian dynamics framework to fuse information in the form of hard kinematic constraints and soft constraints, derived from 2D color blobs measurements, and embedded this in a Kalman filtering framework. They report experiments involving limb and head motion, and use 2 cameras to disambiguate occlusion and drive the detection of color blobs in the image.

Plankers and Fua use a kinematic model with flesh built from deformable meta-balls defining an implicit surface attached to each limb (and consisting of around 4-5 ellipsoids). Pose estimation and tracking is performed using a least-squares framework combining points reconstructed with a trinocular stereo system and additional human contour information from silhouettes (Plankers and Fua, 2001). Joint angle limits and similar physical constraints are not enforced (Plankers, 2001). The authors report experiments with general upper-body limb motions involving self-occlusion against a stationary background and with a short motion tracking sequence involving a full subject, moving parallel to the image plane but against a cluttered background.

Drummond and Cipolla employ a kinematic skeleton and shape model built from quadric (ellipsoidal) surfaces and fuse edge information from 3 cameras (Drummond and Cipolla, 2001). The model limbs on the kinematic chain are over-parameterized in terms of independent general rigid displacements and a weighted least-squares scheme is employed to estimate the parameters. The authors emulate the behavior of a robust distribution by switching between a Gaussian and a Lapla-

cian error model to limit the influence of distant measurements (outliers). Technically speaking, a single robust distribution, automatically providing the desired behavior, could be used (Black and Anandan, 1996; Triggs et al., 2000; Sminchisescu and Triggs, 2001a,b), see also chapter 4 in this thesis. Note that the parameterization has internal constraints and the coercion of kinematic constraints has to be performed after each iteration making the process delicate with respect to optimal re-projection onto the constraint surface. The authors report experiments involving the motion of an upper part of a human body tracked using an 18 d.o.f model against a cluttered background.

Delamarre and Faugeras build a volumetric model in terms of parallelepipeds, spheres and truncated cones and estimate dynamics using energy minimization, based on forces inferred from silhouette contours (Delamarre and Faugeras, 1999). They use 3 calibrated cameras and derive 2D spring-like forces between the model predicted and extracted silhouette using 'Maxwell demons', a matching formalism based on the Iterative Closest Point algorithm (ICP) and knowledge of normal contour directions. The 2D forces resulting from each camera are combined to obtain 3D forces to apply on the 3D model and align it with the data. The dynamical simulation is embedded in a Kalman filter predictive framework. Experimental results of a running subject wearing unconstrained clothing and against an interior background are reported. Overall, the method appears to give good global results, although it seems to be sensitive to the accurate detection of the silhouettes and in certain cases these do not seem to be sufficient to precisely constrain the limb estimates.

Batch Methods

Liebowitz and Carlson reconstruct articulated motion given an image sequence acquired with at least 2 un-calibrated cameras, and manually provided joint correspondences across all video frames (Liebowitz and Carlson, 2001). They perform linear geometric reconstruction in a stratified approach: affine from factorization followed by metric rectification, exploiting the temporal constancy of limb lengths. With the given input assumptions, they show very good 3D reconstructions of athletes performing complex motion during sporting events.

In a related approach DiFranco et al. employ additional "3D given key-frames" (reconstructed body positions) and 2D joint correspondences data over an entire monocular sequence to estimate 3D motion (DiFranco et al., 2001). They use a non-linear least squares framework and smooth motion and joint angle constraints, implemented as discontinuous barrier functions. They show reconstructions from motion capture data and sports events. Note that if joint correspondences in 2 cameras were available, the method of (Liebowitz and Carlson, 2001) could provide a linear initialization for the non-linear refinement of (DiFranco et al., 2001).

2.2.2 Multiple Hypothesis Methods

Deutscher et al. employs an articulated 3D body model based on truncated cones and a cost function based on edge and silhouette information (Deutscher et al., 2000). In their full experiments, the images are acquired using 3 calibrated cameras against a black background. They perform tracking

using the annealed particle filter, a probabilistic density propagation method related to CONDENSATION, but involving, additional processing for each frame, designed to reduce trapping in poor minima, specifically: the error surface is progressively annealed, followed by re-sampling (noisy dynamics and sample re-weighting) at each annealing level. The annealing is used as an importance sampling distribution. Correction weighting could be done at each layer in order to improve mixing (as in (Neal, 1998, 2001)), but (Deutscher et al., 2000) perform it in the final annealing layer when sampling the original density, so that the entire annealing protocol is used as a sophisticated artificial dynamics, for sample mixing. The authors report very good tracking results for general, unconstrained motions, of a 30 d.o.f. human model. It would be also interesting to study how much multi-modality persists in the error surface given the number of cameras and the background settings used.

Another relevant sampling method, known as partitioned sampling (MacCormick and Isard, 2000), represents a statistical formulation of the search space decomposition considered by (Gavrila and Davis, 1996), and is actually a form of Gibbs sampling (Neal, 1993). The variables of the density are alternately (or recursively) updated in blocks, using a heuristic decomposition of the human body into torso, upper limbs, lower limbs for the human body, *etc.* Nevertheless, only relatively low-dimensional models are considered and the extension of the method for models with strong coupling between variables, bias and occlusion is still to be proved. Recent results of (Deutscher et al., 2001) for human tracking in conjunction with annealing and a cross-over operator inspired by genetic optimization appear promising.

Sidenbladh et al. perform tracking in monocular cluttered sequences, using an articulated 3D model with shape represented in terms of cylinders and intensity-based image cues (Sidenbladh et al., 2000). They use CONDENSATION for density propagation, sometimes using importance sampling from a weakened original density (Sidenbladh and Black, 2001) – essentially one layer of annealing in (Deutscher et al., 2000). The authors use prior cyclic (walking) models for tracking full body walking motion. Later work (Sidenbladh and Black, 2001) integrates flow, edge and ridge cues using Laplace like error distributions learned from training data, and shows improved upper body tracking for a subject performing planar motion in a cluttered scene acquired with a moving camera. Their more recent work (Sidenbladh et al., 2002) extends (Leventon and Freeman, 1998; Howe et al., 1999) snippets example-based motion models, learned using PCA from small pieces of motion, in the spirit of example-based texture synthesis methods (Efros and Leung, 1999; Wei and Levoy, 2000). The current estimated joint trajectory is matched against a database of examples and extrapolated to obtain a predicted configuration. Matching the estimated trajectory against example trajectories is performed using probabilistic nearest-neighbor search, in the PCA model data-base. The authors showed improved tracking behavior for a planar motion of an arm and a walking subject. The method has to restrict to relatively low dimensional motion representations, given the complexity of nearest-neighbor queries in spaces of dimension higher than 15.

Choo and Fleet (Choo and Fleet, 2001) use a stick model (without any shape model) for which

joint to image correspondences from motion capture data are available and propose a sampling approach more efficient than CONDENSATION, which is based on Hybrid Monte Carlo. Hybrid Monte Carlo uses the gradient of the density to place samples in parameter space. However, despite its superior focusing of samples near the modes, the method is still prone to trapping in sub-optimal local minima, as in (Choo and Fleet, 2001) (see chapter 6 in this thesis for a discussion). The authors report reliable convergence to the correct pose, presumably due to a favorable initialization, but tracking is difficult since the pose space is highly multi-modal under this type of image data [4].

A number of multiple hypothesis methods have also been proposed for 2D tracking (Toyama and Blake, 2001; Heap and Hogg, 1998; Cham and Rehg, 1999). Although these methods are beyond the scope of our research that is focused on 3D, we will review these briefly. More detailed specific comparisons will appear in chapter 4.

Heap and Hogg learn a statistical model for the modes of variation of a hand shape and track this using CONDENSATION and local sample refinement (Heap and Hogg, 1998). Cham and Rehg use a 2D "scaled prismatic model" (SPM) of a human body parameterized by limb lengths (to account for foreshortening), planar rotation angles, and planar intensity templates attached to each limb (Cham and Rehg, 1999). They assume a multimodal parametric distribution representation based on piece-wise Gaussians. Image plane tracking is performed using CONDENSATION style sampling from the parametric density and least-squares refinement as in (Heap and Hogg, 1998). Both (Heap and Hogg, 1998) and (Cham and Rehg, 1999) describe mixed methods combining continuous optimization and sampling-based state-space search.

Toyama and Blake track 2D humans by clustering in the the space of silhouette appearances (Toyama and Blake, 2001). They use a representation that combines learned prior knowledge for human silhouette shape (the space of silhouette appearances is clustered using a Gaussian mixture) with analytic rigid displacement representations to insure viewpoint invariant behavior. Tracking is based on CONDENSATION and uses Chamfer edge image cues. Good 2D silhouette tracking results are reported under viewpoint changes.

2.3 Learning-Based Methods

This section summarizes several interesting and promising learning-based approaches to human tracking, pose estimation and joint detection.

Brand attempts to capture the natural dynamical motion constraints embedded in the 3D articulated pose space, as do (Howe et al., 1999; Leventon and Freeman, 1998), by learning a Hidden Markov Model (HMM) with Gaussian states, using 3D motion sequences, obtained with a motion capture system (Brand, 1999). The observables are silhouette moments over the entire sequence. The standard HMM Viterbi algorithm is used to find the sequence of states that optimally explains

[4] In fact, many global minima exist, and hundreds of other local minima can be found, as we show in chapter 5, chapter 6 and chapter 7.

the observables (this computation can be done in closed-form). Learning the HMM structure is done using the iterative EM algorithm. The M-step performs an iterative MAP estimate, using an entropic prior encouraging low-dimensional, concise models. The HMM learning solution is not closed-form since iteration and initialization are required in both the EM and the MAP estimation layers. The reconstruction results approximately capture the qualitative motion in silhouette sequences and the method is robust to fast motions. However, the reconstructions are extremely prior driven as the silhouette cues alone are likely to produce ambiguous results, despite temporal constraints.

Howe et al. also attempt to capture natural dynamics learning a mixture of Gaussian representation, based on 'snippets', small pieces of motion of body joint angles, acquired using a motion capture system (Howe et al., 1999; Leventon and Freeman, 1998). They assume that 2D joint positions are available or tracked throughout in an entire image sequence and recover the most probable corresponding 3D trajectory by performing a MAP estimation, with the learned mixture prior, using the EM algorithm. By using more precise image joint information the authors achieve better reconstruction results that (Brand, 1999). The critical part of the system remains the robustness of the 2D joint tracker over longer sequences.

Pavlovic et al. employ the same 2D scaled prismatic model as (Cham and Rehg, 1999) to classify motions using switching linear dynamical systems (Pavlovic et al., 2001). Each dynamical system is learned off-line and models a particular type of motion. Motion segmentation can in principle be performed (after tracking the entire sequence) using the Viterbi algorithm or variational inference. A slightly different approach proposed by (North and Blake, 1998; Blake et al., 1999) is on-line learning. The assumption of entirely known motion models is relaxed. Only the form of model need be known, for instance an auto-regressive process of order k, not its parameters. As before, states representing individual models are connected by transition probabilities and both individual model parameters and the inter-model transition matrix are estimated using CONDENSATION tracking, sequence smoothing and the EM algorithm, but the method is expensive computationally. Results for simple models were shown: ballistic, catch and carry for the motion of a ball.

Rosales and Sclaroff learn mappings between image silhouettes and 2D body joint configurations (which are called 'body poses', but note that no 3D information is extracted by the algorithm) (Rosales and Sclaroff, 2000). They use 3D motion capture data to generate training image silhouettes and corresponding 2D joint configurations. Joint configurations are clustered in 2D by fitting a Gaussian mixture using the EM algorithm. An inverse mapping is subsequently learned between image silhouette moments and 2D joint configurations, for each joint cluster. New image silhouettes are mapped to joint configurations and the most probable configuration is selected (presumably using another direct mapping between the joint configurations and silhouette moments to select the most probable cluster). The authors report good results. The method does not support occlusion and is not directly applicable to 3D reconstruction (mapping joints from 2D to 3D is highly ambiguous – see chapter 5) but might provide input for building more coherent cost surfaces for

3D reconstruction. Note that approaches for 3D pose recovery using joint 3D-2D correspondences do exist (Lee and Chen, 1985) but they require the construction and pruning of an interpretation tree of possible (reflective) joint configurations, using physical constraints, other prior information or heuristics. More recent approaches make certain simplifications by using scaled orthographic camera projection (Taylor, 2000) or restricted human poses in the image (Barron and Kakadiaris, 2000).

2.4 Additional Specific State of the Art Reviews

Further specific details and relationships to relevant work will be given later in the thesis in chapters 3-6, when our modeling choices are made and our estimation methods are presented.

Chapter 3

Modeling

In this chapter we discuss the modeling aspects of the problem of 3D human pose from monocular video. In particular, we describe how we construct the likelihood function we use for human motion estimation and detail the types of constraints and priors we use to delineate the admissible parameter space search regions. Given the high-dimensionality of the problem, it's chronic ill-conditioning, and the presence of clutter, we pay particular attention to the robust and probabilistically consistent integration of contour, intensity and (if available) silhouette information and to prior knowledge that either constrains the search space (joint angle limits, body part inter-penetration avoidance), or that provides valuable anthropometric information about the relative proportions of generic humans hence stabilizing the estimation of difficult to observe but useful modeling parameters.

3.1 Human Body Model

Our human body model (fig. 3.1a,b) consists of a kinematic 'skeleton' of articulated joints controlled by angular **joint parameters** $\mathbf{x}_a$, covered by 'flesh' built from superquadric ellipsoids with

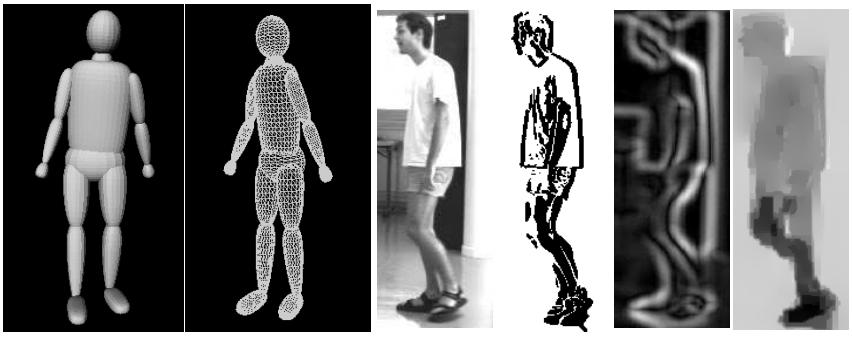


Figure 3.1: Two views of our human body model, and examples of our robust low-level feature extraction: original image (c), motion boundaries (d), intensity-edge energy (e), and robust horizontal flow field (f).

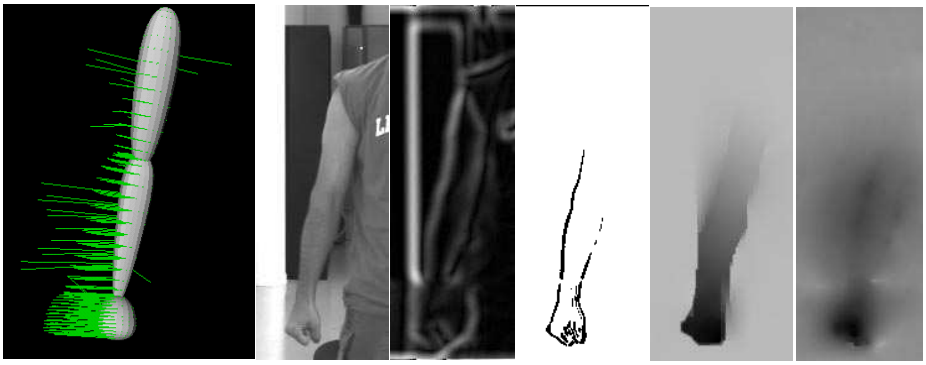


Figure 3.2: A view of an arm model (a), and examples of our robust low-level feature extraction: original image (b), intensity edge-energy (c), motion boundaries (d), robust flow field in horizontal (e) and vertical direction (f).

additional tapering and bending parameters (Barr, 1984). A typical model has around 30 joint parameters, plus 8 **internal proportion** parameters $\mathbf{x}_i$ encoding the positions of the hip, clavicle and skull tip joints, plus 9 **deformable shape** parameters for each body part, gathered into a vector $\mathbf{x}_d$. A complete model can be encoded as a single large parameter vector $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_d, \mathbf{x}_i)$. During tracking we usually estimate only joint parameters, but during initialization the most important internal proportions and shape parameters are also optimized, subject to a soft prior based on standard humanoid dimensions obtained from (Group, 2002) and updated using collected image evidence. Although this model is far from photo-realistic, it suffices for high-level interpretation and realistic occlusion prediction, and it offers a good trade-off between computational complexity and coverage.

The model is used as follows. Superquadric surfaces are discretized as meshes parameterized by angular coordinates in a 2D topological domain. Mesh nodes $\mathbf{u}_i$ are transformed into 3D points $\mathbf{p}_i = \mathbf{p}_i(\mathbf{x})$ and then into predicted image points $\mathbf{r}_i = \mathbf{r}_i(\mathbf{x})$ using composite nonlinear transformations:

$$\mathbf{r}_i(\mathbf{x}) = P(\mathbf{p}_i(\mathbf{x})) = P(A(\mathbf{x}_a, \mathbf{x}_i, D(\mathbf{x}_d, \mathbf{u}_i))) \tag{3.1}$$

where D represents a sequence of parametric deformations that construct the corresponding part in its own reference frame, A represents a chain of rigid transformations that map it through the kinematic chain to its 3D position, and P represents perspective image projection. During model estimation, robust prediction-to-image matching cost metrics are evaluated for each predicted image feature $\mathbf{r}_i$, and the results are summed over all features to produce the image contribution to the overall parameter space cost function. We use both direct image-based cost metrics such as robustified normalized edge energy, and extracted feature based ones. The latter associate the predictions $\mathbf{r}_i$ with one or more nearby image features $\bar{\mathbf{r}}_i$ (with additional subscripts if there are several matches). The cost is then a robust function of the prediction errors $\Delta\mathbf{r}_i(\mathbf{x}) = \bar{\mathbf{r}}_i - \mathbf{r}_i(\mathbf{x})$.

3.2 Problem Formulation

We aim towards a probabilistic interpretation and optimal estimates of the model parameters by maximizing the total probability according to Bayes rule:

$$p(\mathbf{x}|\bar{\mathbf{r}}) \propto p(\bar{\mathbf{r}}|\mathbf{x})\, p(\mathbf{x}) = \exp(-\sum_i e(\bar{\mathbf{r}}_i|\mathbf{x}))\, p(\mathbf{x}) \tag{3.2}$$

where $e(\bar{\mathbf{r}}_i|\mathbf{x})$ is the cost density associated with the observation of node i and $p(\mathbf{x})$ is a prior on model parameters. In our MAP approach, we discretize the continuous problem and attempt to minimize the negative log-likelihood for the total posterior probability, expressed as the following cost function:

$$f(\mathbf{x}) = -\log(p(\bar{\mathbf{r}}|\mathbf{x})\, p(\mathbf{x})) = -\log p(\bar{\mathbf{r}}|\mathbf{x}) - \log p(\mathbf{x}) = f_o(\mathbf{x}) + f_p(\mathbf{x}) \tag{3.3}$$

3.3 Model Priors

The complete prior penalty over model parameters is a sum of negative log likelihoods $f_p = f_{an} + f_s + f_{pa}$ corresponding to the following prior densities p_{an}, p_s, p_{pa}:

Anthropometric data p_{an}: The internal proportions for a standard humanoid (based on statistical measurements) are collected from (Group, 2002) and used effectively as a Gaussian prior, $p_{an} = \mathcal{N}(\boldsymbol{\mu}_{an}, \boldsymbol{\Sigma}_{an})$, to estimate a concrete model for the subject to be tracked. Left-right symmetry of the body is assumed: only "one side" of the internal proportions parameters are estimated while collecting image measurements from the entire body.

Parameter stabilizers p_s: Certain details are far more important than intuition would suggest. For example, it is impossible to track common turning and reaching motions unless the clavicle joints in the shoulder are modeled accurately. However, such parameters have fairly well defined equilibrium positions and leaving them unconstrained would often lead to ambiguities corresponding to nearly singular (flat) cost surfaces. We control the curvature of these hard-to-estimate parameters with robust (sticky) prior stabilizers based on scales of the Gaussian equilibria, $p_s = \mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$. This effectively ensures that in the absence of strong measurements, the parameter is determined by the prior. As more constraints from measurements are available, they may move the parameter value away from its equilibrium and the prior contribution will then effectively turn-off.

Anatomical joint angle limits C_{bl}: 3D consistency requires that the values of joint angles evolve within anatomically consistent intervals. Similarly, when estimating internal body proportions during initialization, we ensure they remain in a certain range of variation around the standard humanoid (typically 10%). We model this with a set of inequalities of the form $C_{bl} \cdot \mathbf{x} < 0$, where C_{bl} is a 'box-limit' constraint matrix.

Body part interpenetration avoidance p_{pa}: Physical consistency requires that different body parts do not interpenetrate during estimation. We avoid this by introducing repulsive potentials that decay

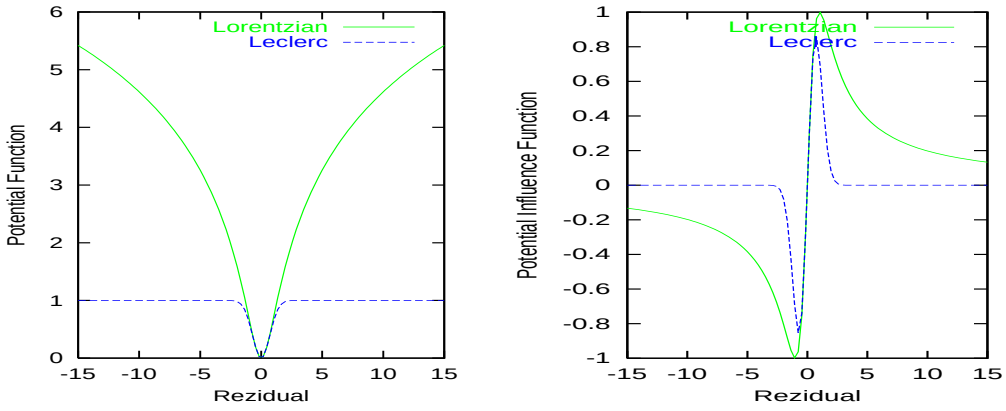


Figure 3.3: Robust Leclerc and Lorentzian error potentials and corresponding influence functions

rapidly outside the surface of each body part, $f_{pa} = \exp(-f(\mathbf{x})\,|f(\mathbf{x})|^{p-1})$ where $f(\mathbf{x})$ defines the implicit surface of the body part and p controls the decay rate.

3.4 Observation Likelihood

Whether continuous or discrete, the search process depends critically on the observation likelihood component of the parameter space cost function. Besides smoothness properties, the likelihood should be designed to limit the number of spurious local minima in parameter space. Our method employs a combination of robust edge and intensity information within a multiple assignment strategy based on a weighting scheme that focuses attention towards motion boundaries. The likelihood terms are based on robust (heavy-tailed) error distributions. Note that both robustly extracted image cues and robust parameter space estimation are used: the former provides "good features to track", while the latter directly addresses the model-image association problem.

3.4.1 Robust Error Distributions

MAP parameter estimation is naturally robust so long as it is based on realistic 'total likelihoods' for the combined inlier and outlier distributions of the observations (see fig. 3.3). We model these as robust penalty functions $\rho_i(s_i)$ of the normalized squared errors $s_i = \|\mathbf{\Delta r}_i\|^2/\sigma_i^2$. Each $\rho_i(s)$ is an increasing sublinear function with $\rho_i(0) = 0$ and $\frac{d}{ds}\rho_i(0) = 1$, corresponding to a radially symmetric error distribution with a central peak of width σ. Here we used the 'Lorentzian' $\rho(s) = \nu \log(1 + s/\nu)$ and 'Leclerc' $\rho(s) = \nu(1 - \exp(-s/\nu))$ potentials, where ν is a strength parameter related to the frequency of outliers.

Normalizing by the number of nodes N in each mesh, the cost adopted for the i^{th} observation is $e(\bar{\mathbf{r}}_i|\mathbf{x}) = \frac{1}{N}e_i(\mathbf{x})$, where $\mathbf{W}_i$ is a positive definite weighting matrix and:

$$e_i(\mathbf{x}) = \begin{cases} \frac{1}{2}\rho_i(\mathbf{\Delta r}_i(\mathbf{x})\,\mathbf{W}_i\,\mathbf{\Delta r}_i(\mathbf{x})^\top) & \text{if } i \text{ is assigned} \\ \nu_{bf} = \nu & \text{if back-facing} \\ \nu_{occ} = k\nu, \;\; k > 1 & \text{if occluded} \end{cases} \tag{3.4}$$

The total robust observation likelihood is thus:

$$f_o(\mathbf{x}) \;=\; -\log p(\bar{\mathbf{r}}|\mathbf{x}) \;=\; f_a(\mathbf{x}) + N_{bf}\,\nu_{bf} + N_{occ}\,\nu_{occ} \tag{3.5}$$

where $f_a(\mathbf{x})$ represents the term associated with the image assigned model nodes, while N_{occ} and N_{bf} are the numbers of occluded and back-facing (self-occluded) model nodes.

Notice that occluded model predictions are not simply ignored. They contribute a constant penalty to the overall observation likelihood. This is necessary in order to build likelihoods that preserve their response properties under occlusion and viewpoint change. For instance, good fits from both frontal and side views should ideally have similar peak responses, but it is clear that the number of occluded model points is in general larger in a side view, than in a frontal one. This can lead to down-weighting of peaks for side views if only the visible nodes are taken into account. An additional difficulty arises in cases when the legs pass each-other (in a side-view) and the model 'locks' both of its legs onto the same image leg. To avoid such situations, we include all of the model nodes when fusing the likelihood, but we slightly penalize occlusions in order to make them less attractive. A way to choose ν is to fit the model to the data and compute an approximate error per node. By using a slightly higher value for occluded nodes, we make them more attractive than a bad fit but less attractive than other non-occluded states that can exist in the neighborhood of the parameter space. We find this heuristic gives good results in practice[1], although a more rigorous treatment of occlusion would be desirable in the general case. At present, this is computationally too expensive, but interesting approximations can be found in (MacCormick and Blake, 1998).

3.5 Contour Image Cues

Related Work: Generating, detecting and matching image contours for a particular object class is difficult. Methods can be generally classified as bottom-up (based on *correspondence space search*) or top-down (*model space search*). A correspondence space is just a Cartesian product of all image features of a particular kind (edges, corners, *etc*.) together with the set of predicted model-features against which they have to be matched. Note that both model and correspondence spaces are exponential with respect to a (usually large) scale parameter: the model-space is exponential in its dimension while the correspondence space is exponential in the numbers of model and image

[1]This is particularly effective when combined with the Covariance Scaled Sampling (CSS) algorithm presented in chapter 4. Loss of visibility of certain body parts leads to increased uncertainty in related parameters, and CSS automatically ensures broader sampling in those parameter space regions.

features. Hence, for high-dimensional spaces or in the presence of clutter the search process has to be pruned in some way, and good heuristics are necessary in order to drive the search process towards well supported, target configurations.

Bottom-up MRF methods (Geiger et al., 1995; Liu and Geiger, 1998; Perez et al., 2001; Zhu, 1999) define an energy functional that encodes the basic properties of the contour (smoothness, edge responses) and solve using dynamic programming, or Monte-Carlo methods for more general types of couplings. These methods are elegant, but lead to expensive high-dimensional feature space representations for which finding good proposal distributions is non-trivial. Their generalizations to 3D and the related invariance issues also seem problematic. *Top-down model based approaches* (Terzopoulos et al., 1988; Deutscher et al., 2000; Sminchisescu and Triggs, 2001b), use a valid predicted model configuration (smoothness properties are built into the model parameterization) to generate independent local search processes in the neighborhood of individual model predictions[2]. Matching configurations is fast, but the matching process is essentially local, so arbitrarily wrong global results due to gradual loss of model regularization effects are obtained as the search window increases. Additionally, probabilistic approaches (MacCormick and Blake, 1998; Sullivan et al., 1999; Sidenbladh and Black, 2001) exploit prior knowledge about the foreground and background distributions or model the inherent ambiguities involved in the matching process. Our work builds on probabilistic top-down contour techniques using model-based consistency constraints to drive the image contour detection process. Model-image edge assignment strategies and the consistency and weighting issues involved in contour likelihood construction are the subject of the next 2 sections.

Simple image preprocessing operations are used for feature extraction in contour likelihood construction, as follows: the images are smoothed with a Gaussian kernel, contrast normalized, and a Sobel edge detector is applied.

3.5.1 Multiple Assignment Probabilistic Contour Likelihood

For visible nodes on model occluding contours ($\mathcal{O}$), we perform line search along the model predicted normal direction and retain all possible edge responses within the search window S. Also, motion boundaries are independently extracted from the outlier map of a robust multi-scale optical flow computation, based on Black & Anandan's implementation using affine motion models, (Black and Anandan, 1996). The motion boundaries are processed similarly to edges, to obtain a smooth 2D potential field S_p. This outlier map conveys useful information about the motion boundaries and is used to weight the significance of the intensity edges (see fig. 3.1d). We typically use diagonal weighting matrices associated with the predicted feature $\mathbf{r}_i$ and corresponding matched observation

[2]Note that, in a strict sense, this is not a valid Bayesian strategy. Formally the set of measurements should not depend on the hidden state, but be defined in advance, *e.g.* by using a fixed image grid and computing intersections between the contour and the grid. Nevertheless, in practice, the approximation appears to be sufficiently good as shown in the experiments of (MacCormick and Blake, 1998).

$\bar{\mathbf{r}}_i$, of the form: $\mathbf{W}_i(\mathbf{r}_i) = 1 - kS_p(\bar{\mathbf{r}}_i)$, where k is a constant that controls the emphasis on and confidence in the motion boundary estimation. For instance, $k = 0$ weights all the edges uniformly, while $k = 1$ entirely excludes edge responses that are not on motion boundaries (this is because the smoothed motion boundary image is a real image with values between 0 and 1 as in fig. 3.1d). In practice, we find that values of k in the range $k = 0.2 - 0.4$ work well.

Suppose that the probability of detecting n clutter edge responses on a measurement line of length S is modeled by a distribution $\mathcal{C}(n)$ that can be learned from training data (for instance by generating different configurations of the model and building histograms corresponding to the number of contour responses, then fitting an analytic distribution to the empirical histograms). Suppose also that the n detected features are distributed uniformly on the measurement line (so the probability of a clutter detection at any position on the line is $1/S$). One can then easily verify that the probability of detecting n clutter features at arbitrary positions on the measurement line is $\mathcal{C}(n)/S^n$ (MacCormick and Blake, 1998). Suppose that we are in a situation in which n edge responses have been detected and that the probability of non-detection is zero, that is, we are sure that the true assignment is in the detected edgel set. There are then $n!$ possible ways in which the $n-1$ clutter and 1 true edge response can be combined. Furthermore, for each possible position of the true edge response (out of n) there are $(n-1)!$ clutter combinations. Consequently, the probability of selecting a random permutation corresponding to a target detection event for the model prediction "i" at configuration "$\mathbf{x}$" can be derived as the probability of detecting $n-1$ clutter edges and the probability of detecting 1 true one (integrated over all such possibilities) and normalized (by the total number of possibilities $n!$):

$$p_i(\mathbf{x}) = \frac{\mathcal{C}(n-1)}{S^{n-1}} \cdot \frac{(n-1)! \sum_{k=1}^{n} e^{-f_{ik}(\mathbf{x})}}{n!} = \frac{\mathcal{C}(n-1)}{nS^{n-1}} \sum_{k=1}^{n} e^{-f_{ik}(\mathbf{x})} \tag{3.6}$$

where:

$$f_{ik}(\mathbf{x}) = \frac{1}{2}\rho_{ik}(\Delta\mathbf{r}_{ik}(\mathbf{x})\,\mathbf{W}_{ik}\,\Delta\mathbf{r}_{ik}(\mathbf{x})^\top) \tag{3.7}$$

where the subscripts "ik" denote the edgel k assigned to model prediction i.

The edge assigned data term over the model predicted occluding contour (3.5) thus becomes:

$$f_c(\mathbf{x}) = -\log p_{\mathcal{E}}(\mathbf{x}) = -\log \prod_{i=1}^{O} p_i(\mathbf{x}) = -\log \prod_{i=1}^{O} \frac{\mathcal{C}(n_i-1)}{n_i S^{n_i-1}} \sum_{k=1}^{n_i} e^{-f_{ik}(\mathbf{x})} \tag{3.8}$$

The robustified gradient and Hessian $\mathbf{g}_{\mathcal{E}}, \mathbf{H}_{\mathcal{E}}$ corresponding to edge contributions can be derived from (3.8), using the chain rule, in terms of individual gradient and Hessian operators corresponding to model prediction i and edge feature k, using the model-image Jacobian matrix[3], $\mathbf{J}_i = \frac{\mathbf{dr}_i}{\mathbf{dx}}$:

$$\mathbf{g}_{ik} = \mathbf{J}_i^\top \rho'_{ik} \mathbf{W}_{ik} \Delta\mathbf{r}_{ik} \tag{3.9}$$

$$\mathbf{H}_{ik} \approx \mathbf{J}_i^\top (\rho'_{ik}\mathbf{W}_{ik} + 2\rho''_{ik}\,(\mathbf{W}_{ik}\Delta\mathbf{r}_{ik})(\mathbf{W}_{ik}\Delta\mathbf{r}_{ik})^\top)\,\mathbf{J}_i \tag{3.10}$$

[3]We compute the Jacobian analytically, using the chain rule, by backpropagation on the kinematic chain.

3.5.2 Symmetry Consistent Contour Likelihood

Compared to the previous section, for visible nodes on model occluding contours ($\mathcal{O}$), we perform line search along the projected contour normal in order to uniquely match the corresponding prediction with the closest image edgel[4]. As the search and matching processing for each model prediction is performed independently, it is not guaranteed that the matched image features form admissible configurations. Robust estimation is one way this problem can be alleviated. There are however, two problems with it. The first is caused by nearby outliers whose influence can't be entirely switched off by the robust cost. This is particularly apparent in contexts where the inter-frame motion is sufficiently important that a larger noise process has to be used. Secondly, the independence assumptions during search define an incorrect (or actually weaker) model to robustify and robustness can't tackle this.

The solution we propose is to use robust estimation but to provide it with the correct model of the target to be localized. This should reflect the non-independent relations-couplings between model elements. We particularize this in the case of human motion analysis for the models of limbs, but the idea applies far more generally in the sense that the semantics between any model predicted points has to be preserved between individual corresponding image matched features.

Matching Consistency and Data Coupling

The consistency term exploits the property that a configuration of matched points has to have similar appearance with the model prediction of the corresponding points. While this property might appear trivially fulfilled for model-driven search, independence assumptions and wider noise models might generate wrong matched configurations with good cost. Properties like symmetry (of a limb, for instance) can be violated during the search process. The uniqueness assignment principle is also, no longer guaranteed.

We propose building likelihood terms that encode not only the probabilities of edge responses but also the model symmetries or other non-independent model properties. For instance for any predicted model node $\mathbf{r}_i(x)$, lying on the model occluding contour, there exist a symmetric node (with respect to a projected limb axis) $\mathbf{r}_{i_c}(x)$ in the model. One can derive a symmetry corrected negative log-likelihood term, based on the deviation between the model predicted symmetries and the matched ones: $\Delta\mathbf{r}_{i_c} = (\mathbf{r}_i(\mathbf{x}) - \mathbf{r}_{i_c}(\mathbf{x})) - (\bar{\mathbf{r}}_i - \bar{\mathbf{r}}_{i_c})$. The term is essentially exploiting the semantics present in the model which has to directly transfer to the results of correspondence process (see also fig. 3.4):

$$f_c(\mathbf{x}) = \frac{1}{2}\Big(\sum_{i\in\mathcal{O}} \rho_i(\Delta\mathbf{r}_i(\mathbf{x})\mathbf{W}_i\Delta\mathbf{r}_i(\mathbf{x})^\top) + \sum_{i\in\mathcal{O}} \varphi_{i_c}(\Delta\mathbf{r}_{i_c}(\mathbf{x})\mathbf{W}_{i_c}\Delta\mathbf{r}_{i_c}(\mathbf{x})^\top)\Big) \tag{3.11}$$

[4]We restrict our treatment to unique assignments for the sake of clarity in the mathematical treatment. In practice, the probabilistic assignment strategy presented in §3.5.1 and the symmetry consistency constraints of the likelihood in §3.5.2 can be combined.

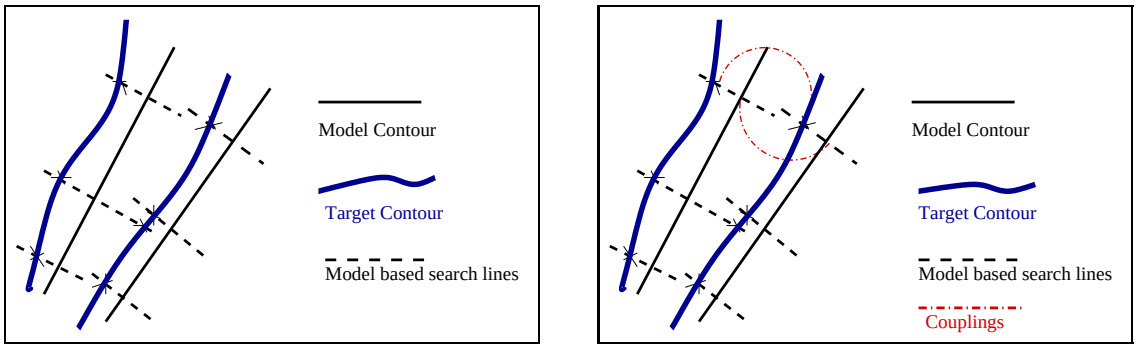


Figure 3.4: (a) Model-based contour search process and (b) Contour detection using model-based symmetry constraints. The use of symmetric couplings removes some of the incorrect assignments (see text).

where φ is a negative log-likelihood of a robust error potential with a wide attraction zone. The gradient and Hessian corresponding to the predicted assignments and couplings, can be derived using the columns of the model-image Jacobian matrix, $\mathbf{J}_i = \frac{\mathbf{dr}_i}{\mathbf{dx}}$:

$$\mathbf{g}_c = \sum_{i \in \mathcal{O}} \mathbf{J}_i^\top \rho_i' \mathbf{W}_i \Delta \mathbf{r}_i + \sum_{i \in \mathcal{O}} \mathbf{J}_i^\top \varphi_{ic}' \mathbf{W}_{ic} \Delta \mathbf{r}_{ic} \tag{3.12}$$

$$\mathbf{H}_c \approx \sum_{i \in \mathcal{O}} \mathbf{J}_i^\top (\rho_i{}' \mathbf{W}_i + 2\rho_i'' (\mathbf{W}_i \Delta \mathbf{r}_i)(\mathbf{W}_i \Delta \mathbf{r}_i)^\top) \mathbf{J}_i \tag{3.13}$$

$$+ \sum_{i \in \mathcal{O}} \mathbf{J}_i^\top (\varphi_{ic}{}' \mathbf{W}_{ic} + 2\varphi_{ic}'' (\mathbf{W}_{ic} \Delta \mathbf{r}_{ic})(\mathbf{W}_{ic} \Delta \mathbf{r}_{ic})^\top) \mathbf{J}_i \tag{3.14}$$

For a human body model, the above "symmetrized" edge term will add over pairs of symmetric visible nodes lying on occluding contours for all limbs (the increase in cost is not significant, for n edge based-terms an extra $\mathcal{O}(n/2)$ terms are added). In fact other properties present in the model that could be potentially lost during correspondence search, could be built in order to drive the matching process towards valid configurations (like for instance further "Gestalt" properties like collinearity, co-circularity, *etc.*). The coupling of model and observation nodes can be of arbitrary order, although with the expected increase in computational cost. The problem can be formulated, as above, such that the continuous properties of the likelihood model are preserved. For instance collinearity properties for nodes on limbs can be built into a determinant and require $\mathcal{O}(n^2)$ terms. Conic properties involve deriving $2nd$ order parametric curves, *etc.*

Experiments : We built symmetry constraints into the likelihood model. The first experiment we show involves passing a lower arm over an image arm. As one can see in fig.3.6, the cost has multiple minima, owing to incorrect assignments of both model edges to the same image edge. The problem is indeed alleviated by the introduction of modeling constraints on the matched configurations. A similar experiment is performed for the motion of a foot (fig.3.7), and the pure independent correspondence search gets stuck into a local minima owing to incorrect assignments due to the break in model symmetries.

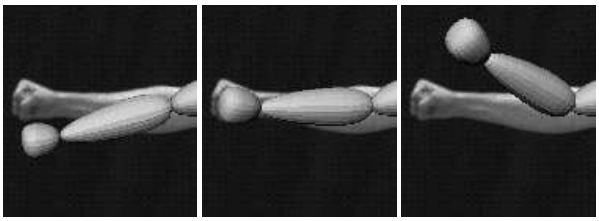


Figure 3.5: Cost Function Experiment. The elbow joint is varied and the corresponding edge cost is monitored. Multi-modal behavior arises due to incorrect matches during independent edgels search. Matching under model consistency constraints alleviates the problem see Fig. 3.6.

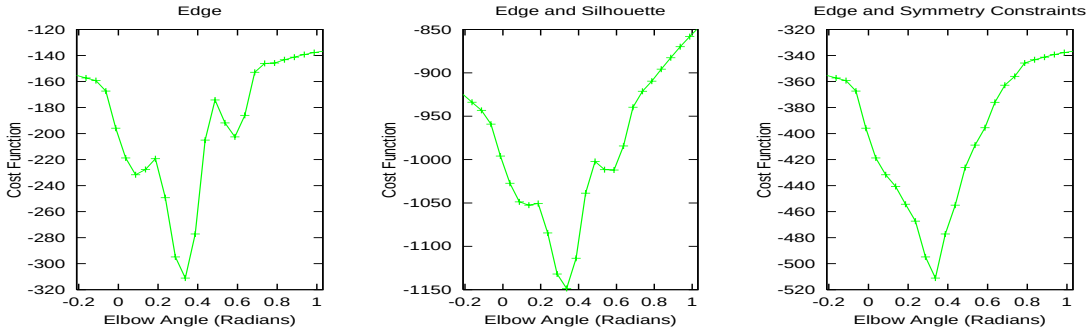


Figure 3.6: Multi-modality in the cost surface (plot left and middle) due to inconsistent edge assignments is removed when using an edge likelihood with coupled/symmetry constraints (right).

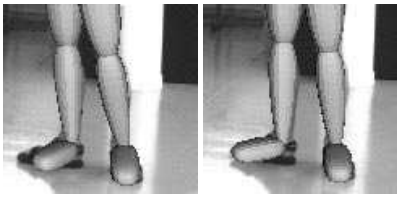


Figure 3.7: Contour symmetry alleviates data association.

3.6 Intensity Image Cues

Related Work: Intensity image cues have been used by many authors (Rehg and Kanade, 1995; Ju et al., 1996; Hager and Belhumeur, 1998; Bregler and Malik, 1998; Sidenbladh et al., 2000; Sminchisescu and Triggs, 2001a,b) for human model-based tracking problems. Most of the systems use the model-based optical flow approach §3.6.2, in which an implicit texture is assigned to the 3D model and subsequently used for registration. In the next section §3.6.1, we show that low-level extracted flow could also be used as a pure correspondence field. Our experience with both

methods shows nevertheless that the errors induced by different geometric models are less important compared to the usually restrictive assumptions on lighting variations (typically not modeled – but a notable exception is (Hager and Belhumeur, 1998)) as well as inter-frame shape deformation due to clothing in realistic scenarios.

3.6.1 Image Flow Correspondence Field

The first intensity-based likelihood is used as a dense correspondence field in the image. More specifically, a robust multi-scale method based on Black & Anandan's implementation (Black and Anandan, 1996) is used to compute an optical flow field based on translational with regularization image motion models. The model is projected into the image and, for all the visible nodes lying inside the object ($\mathcal{I}$) (and not on the occluding contour) we directly assign the already computed displacement vector at that image coordinate. Although the weaker constraints of such a displacement field do not necessarily guarantee consistency under the 3D articulated and volumetric constraints (as the likelihood presented in the next section), we find that they give good results in practice, avoiding texture visibility issues and preventing an undesirable accumulation of modeling errors [5].

The assigned intensity data term for visible nodes lying inside the object ($\mathcal{I}$) thus becomes:

$$f_I(\mathbf{x}) = \sum_{j \in \mathcal{I}} \rho_{j_f}(\Delta \mathbf{r}_{j_f}(\mathbf{x})\, \mathbf{W}_{j_f}\, \Delta \mathbf{r}_{j_f}(\mathbf{x})^\top) \tag{3.15}$$

where the subscripts "j_f" denote the flow term assigned to model prediction j.

3.6.2 Model-Based Intensity Matching

The above models apply not only to geometric image features like edges or already computed feature correspondence fields, but also to intensity-based matching of image patches. In this case, the observables are image gray-scales or colors $\mathbf{I}$ rather than feature coordinates $\mathbf{r}$, and the error model is based on intensity residuals. To get from a point projection model $\mathbf{r} = \mathbf{r}(\mathbf{x})$ to an intensity based one, we simply compose with the assumed local intensity model $\mathbf{I} = \mathbf{I}(\mathbf{r})$ (*e.g.* obtained from an image template corresponding to the model projection, or another image that we are matching against), premultiply point Jacobians by point-to-intensity Jacobians $\frac{\mathbf{dI}}{\mathbf{dr}}$, *etc*. Note that intensity matching is the real basis of feature based methods. Feature detectors are optimized for detection not localization. To localize a detected feature accurately we need to match (some function of) the image intensities in its region[6] against either an idealized template or another image of the feature (in our case the image patch corresponding to the 3D model prediction during initialization or to

[5] Despite being relatively complex, our modeling of the human skeleton and volumetric structure is still crude in a physical perspective. In practice, we don't model clothing, and anatomically, the articulations have more complex structure, inter-displacements, and degrees of freedom that the ones we are representing.

[6] As explained before, our surface model is already discretized, so we work on individual mesh nodes that are visible and not on an occluding contour. In the case of detected features, precise localization would involve working on the surface nodes in their neighborhood.

the previous image for tracking applications), using an appropriate geometric deformation model, in our case the volumetric model with part articulation constraints, *etc*. For example, suppose that the intensity matching model is:

$$f_i = \frac{1}{2}\rho(\|\Delta\mathbf{I}(\mathbf{r}_i)\|^2) \tag{3.16}$$

where $\Delta\mathbf{I}$ is the current intensity prediction error, $\mathbf{r}_i$ parameterizes and constrains the local geometry (translation & warping), and $\rho(\cdot)$ is some intensity error robustifier. Then the cost gradient in terms of $\mathbf{r}_i$ is:

$$\mathbf{g}_i = \frac{\mathbf{d}f_i}{\mathbf{d}\mathbf{r}_i} = \rho' \, \Delta\mathbf{I}^\top \frac{\mathbf{dI}}{\mathbf{d}\mathbf{r}_i} \tag{3.17}$$

Similarly, the cost Hessian in $\mathbf{r}_i$ in a Gauss-Newton approximation is:

$$\mathbf{H}_i = \frac{\mathbf{d}^2 f_i}{\mathbf{d}\mathbf{r}_i^2} \approx \rho'' \, (\frac{\mathbf{dI}}{\mathbf{d}\mathbf{r}_i})^\top \frac{\mathbf{dI}}{\mathbf{d}\mathbf{r}_i} \tag{3.18}$$

We express $\mathbf{r}_i = \mathbf{r}_i(\mathbf{x})$ as a function of the model parameters, so, as before, if $\mathbf{J}_i = \frac{\mathbf{d}\mathbf{r}_i}{\mathbf{dx}}$ we have a corresponding intensity cost gradient $\mathbf{g}_I$ and Hessian $\mathbf{H}_I$ contribution assembled over all observations:

$$\mathbf{g}_I = \sum_i \mathbf{J}_i^\top \mathbf{g}_i \tag{3.19}$$

and

$$\mathbf{H}_I = \sum_i \mathbf{J}_i^\top \, \mathbf{H}_i \, \mathbf{J}_i \tag{3.20}$$

In other words, the intensity matching model is locally equivalent to a quadratic feature matching one on the 'features' $\mathbf{r}(\mathbf{x})$, with effective weight (inverse covariance) matrix $\mathbf{W}_i = \mathbf{H}_i$. As feature covariances are a function of intensity gradients $\rho'' \, (\frac{\mathbf{dI}}{\mathbf{d}\mathbf{r}_i})^\top \frac{\mathbf{dI}}{\mathbf{d}\mathbf{r}_i}$, they can be both highly variable between features (depending on how much local gradient there is), and highly anisotropic (depending on how directional the gradients are). *E.g.*, for points along a 1D intensity edge, the uncertainty is large along the edge direction and small across it.

3.7 Silhouette Image Cues

Related Work: Deutscher *et al* uses a discrete silhouette based term for his cost function design in a multi-camera setting (Deutscher et al., 2000). The term peaks if the model is inside the silhouette without demanding that the silhouette area is fully explained (see Sec. 3.7.1). Consequently, an entire family of un-informative configurations situated inside the silhouette will generate good costs under this likelihood model. The situation is alleviated by the use of additional edge cues and multiple cameras. Delamarre & Faugeras use silhouette contours in a multi-camera setting and computes

assignments using an Iterative Closest Point algorithm and knowledge of normal contour directions (Delamarre and Faugeras, 1999). The method is local and doesn't enforce globally consistent assignments, but relies on fusing information from many cameras to ensure consistency. Brand uses silhouette sequences to infer temporal human poses (Brand, 1999) while Rosales & Sclaroff use them as inputs to a system which directly learns 2D silhouette to 2D human joint mappings (Rosales and Sclaroff, 2000).

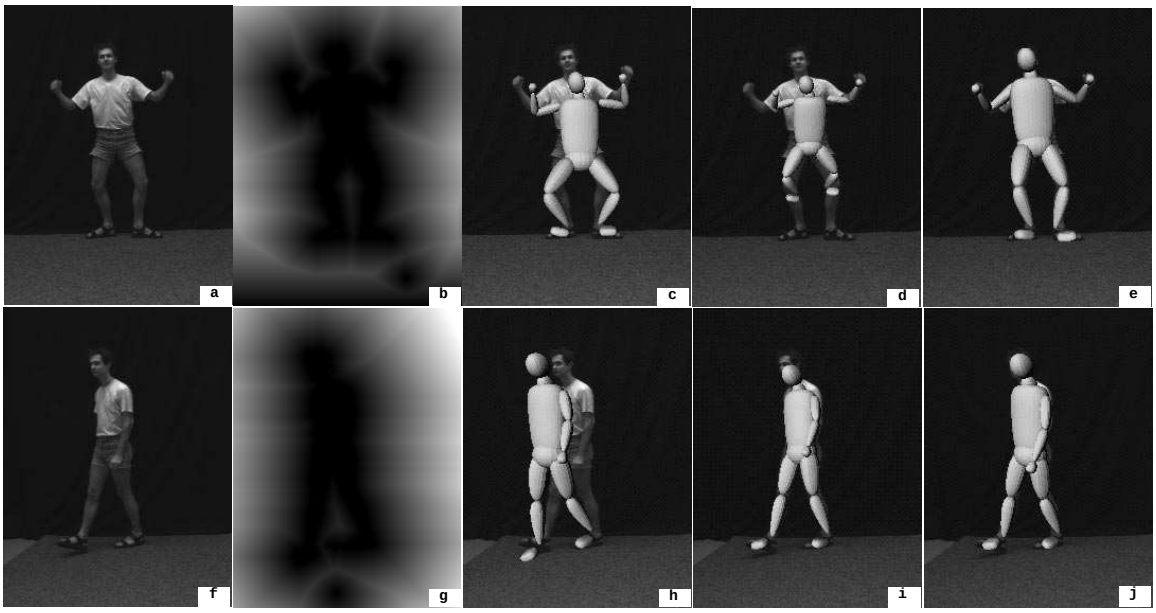


Figure 3.8: Model estimation based on various silhouette terms original images (a,f), their distance transforms (b,g), initial models (c,h), silhouette attraction term only (d,i), silhouette attraction and area overlap terms (e,j). Silhouette attraction term produces inconsistent likelihoods (d,i). The pair of attraction/explanation terms produces good fits and enforces consistency (e,j).

The cost term proposed here is based on a pair of sub-components, one which pushes the model inside the image silhouette, while the other maximizes the model-image silhouette area overlap. The cost term is global and consistent, in that it is not only enforcing the model remains within the image silhouette, but also demands that the image silhouette is entirely explained.

3.7.1 Silhouette-Model Area Overlap Term

The area of the predicted model can be computed from the model projected "triangulation" by summing over all "visible" triangles $t \in V_t$ (triangles having all the vertices $(x_i, y_i)_{i=1..3}$ visible).

$$S_t = \sum_{i=1}^{3} x_{i\odot 3}(y_{i+1\odot 3} - y_{i+2\odot 3}) \tag{3.21}$$

where $\odot$ describes the modulo operation, and the computation assumes the triangle vertices are sorted in counter-clockwise order to preserve positive area sign. In subsequent derivations we drop the modulo notation for simplicity.

Let S_g be the area of the target silhouette. The area alignment cost writes:

$$e_a = \frac{1}{2\sigma^2}(\sum_{t \in V_t} S_t - S_g)^2 \tag{3.22}$$

The gradient and Hessian for the area-based cost-term can subsequently be derived (by dropping the scaling term):

$$\mathbf{g}_a = \frac{\mathbf{d}e_a}{\mathbf{dx}} = (\sum_{t \in V_t} S_t - S_g) \sum_{t \in V_t} \frac{\mathbf{d}S_t}{\mathbf{dx}} \tag{3.23}$$

where:

$$\frac{\mathbf{d}S_t}{\mathbf{dx}} = \sum_{i=1}^{3} \frac{\mathbf{d}x_i}{\mathbf{dx}}^\top (y_{i+1} - y_{i+2}) + \sum_{i=1}^{3} x_i (\frac{\mathbf{d}y_{i+1}}{\mathbf{dx}} - \frac{\mathbf{d}y_{i+2}}{\mathbf{dx}})^\top \tag{3.24}$$

$$\mathbf{H}_a = \frac{\mathbf{d}^2 e_a}{\mathbf{dx}^2} = \sum_{t \in V_t} \frac{\mathbf{d}S_t}{\mathbf{dx}}^\top \frac{\mathbf{d}S_t}{\mathbf{dx}} + (\sum_{t \in V_t} S_t - S_g) \sum_{t \in V_t} \frac{\mathbf{d}^2 S_t}{\mathbf{dx}^2} \tag{3.25}$$

where:

$$\frac{\mathbf{d}^2 S_t}{\mathbf{dx}^2} = \sum_{i=1}^{3} \frac{\mathbf{d}^2 x_i}{\mathbf{dx}^2}(y_{i+1} - y_{i+2}) + \sum_{i=1}^{3} x_i (\frac{\mathbf{d}^2 y_{i+1}}{\mathbf{dx}^2} - \frac{\mathbf{d}^2 y_{i+2}}{\mathbf{dx}^2})^\top \tag{3.26}$$

$$+2 \sum_{i=1}^{3} \frac{\mathbf{d}x_i}{\mathbf{dx}}^\top (\frac{\mathbf{d}y_{i+1}}{\mathbf{dx}} - \frac{\mathbf{d}y_{i+2}}{\mathbf{dx}}) \tag{3.27}$$

One should notice that the individual partial derivatives $\frac{\mathbf{d}x_i}{\mathbf{dx}}$ and $\frac{\mathbf{d}y_i}{\mathbf{dx}}$ represent the columns of the individual Jacobian matrix evaluated at the corresponding prediction for the mesh node i, $\mathbf{r}_i(\mathbf{x}) = (x_i, y_i)$. In practice computing node visibility and area differences is rather fast calculation which we perform using the z-buffer.

3.7.2 Silhouette Attraction Term

This term pushes the model inside the image silhouette. Adding over all projected model nodes i the cost writes:

$$e_s = \frac{1}{2\sigma^2} \sum_i e_{s_i}(\mathbf{r}_i(\mathbf{x}), S_g) \tag{3.28}$$

The distance from a predicted model point $\mathbf{r}_i(\mathbf{x})$ to a given silhouette S_g can be written as a quadratic Chamfer distance, between model prediction $\mathbf{r}_i(\mathbf{x})$ and points $\mathbf{s}_i$ on the given silhouette S_g:

$$e_{s_i}(\mathbf{r}_i(\mathbf{x}), S_g) = \min_{s_i \in S_g} \parallel \mathbf{r}_i(\mathbf{x}) - \mathbf{s}_i \parallel \tag{3.29}$$

It is known that such Hausdorff distance is giving a good quadratic approximation ((Toyama and Blake, 2001)). The Chamfer distance can be computed fast by means of dynamic programming (Gavrila and Davis, 1996) (note that (Gavrila and Davis, 1996; Toyama and Blake, 2001) employed this distance for edges and in a discrete evaluation context). However, the distance lacks immediate continuous structure. Given a Chamfer silhouette image (see fig.3.8 b and fig.3.8 g), we build a 2-dimensional continuous potential surface $\mathbf{P_s}$ by fitting local quadric surfaces to 3x3 image patches. The gradient and Hessian of the corresponding cost term can therefore be derived from the model-image Jacobian, and the corresponding quadric terms:

$$\mathbf{g}_s = \sum_i \frac{\mathbf{dP_s}(\mathbf{r}_i(\mathbf{x}))}{\mathbf{dx}} = \sum_{i \in V} \mathbf{J}_i^\top \frac{\mathbf{dP_s}}{\mathbf{dr}_i} \tag{3.30}$$

$$\mathbf{H}_s = \sum_i \frac{\mathbf{d^2P_s}}{\mathbf{dx}^2} \approx \sum_{i \in V} \mathbf{J}_i^\top \frac{\mathbf{d}^2\mathbf{P_s}}{\mathbf{dr}_i^2} \mathbf{J}_i \tag{3.31}$$

Experiments : In fig.3.8 we show, for two image silhouettes, the initial image (fig.3.8 a and fig.3.8 f), the distance transform of the corresponding silhouette (fig.3.8 b and fig.3.8 g), the initial model configuration (fig.3.8 c and fig.3.8 h) and the fitting results obtained when using only the silhouette attraction term (fig.3.8 d and fig.3.8 i) and finally both the silhouette and the area overlap term (fig.3.8 e and fig.3.8 j). One can notice that, as expected, the silhouette attraction terms does not suffice for a good fit and any configuration that places the model inside the image silhouette can be potentially chosen. On the contrary, the area overlap term stabilizes the estimation and drives it towards rather satisfactory results. Note that the cost term has the desired properties of a wide attraction zone, being thus a good candidate for tracking applications where recovery from tracking failure is a highly desirable property.

Chapter 4

Covariance Scaled Sampling

In this chapter we present a method for recovering 3D human body motion from monocular video sequences, based on a robust image matching metric, incorporation of joint limits and non-self-intersection constraints, and a new sample-and-refine search strategy guided by rescaled cost-function covariances. Monocular 3D body tracking is challenging: besides the difficulty of matching an imperfect, highly flexible, self-occluding model to cluttered image features, realistic body models have at least 30 joint parameters subject to highly nonlinear physical constraints, and about a third of these degrees of freedom are essentially unobservable in any given monocular image. For image matching we use a carefully designed robust cost metric combining robust optical flow, edge energy, and motion boundaries. The nonlinearities and matching ambiguities make the parameter-space cost surface multi-modal, ill-conditioned and highly nonlinear, so searching it is difficult. We discuss the limitations of CONDENSATION-like samplers, and describe a novel hybrid search algorithm that combines inflated-covariance-scaled sampling and robust continuous optimization subject to physical constraints and model priors. Our experiments on challenging monocular sequences show that robust cost modeling, joint and self-intersection constraints, and informed sampling are all essential for reliable monocular 3D motion estimation.

Keywords: 3D human body tracking, particle filtering, high-dimensional search, constrained optimization, robust matching.

4.1 Introduction

Extracting 3D human motion from natural monocular video sequences poses difficult modeling and computation problems: (*i*) Even a minimal human model is very complex, with around 30 joint parameters and many more body shape ones, subject to highly nonlinear joint limits and non-self-intersection constraints. (*ii*) Unlike the 2D and the multi-camera 3D cases, about a third of the degrees of freedom are nearly unobservable in any given monocular image (depths, relative depths, also rotations of near-cylindrical limbs about their axes). (*iii*) Matching a complex, imperfectly

known, self-occluding model to a cluttered scene is inherently difficult. These difficulties interact strongly in practice — minor modeling or matching errors tend to lead to large compensatory biases in hard-to-estimate depth parameters, which in turn cause mis-prediction and tracking failure — and we believe that a successful monocular 3D body tracking system must pay attention to all of them.

To control correspondence errors, we use a carefully designed robust matching metric that combines robust optical flow, edge energy, and motion boundaries (§3.2). Our system includes full 3D occlusion prediction, and as far as we know it is the first to enforce both hard joint angle limits and body non-self-intersection constraints. These nonlinearities and ambiguities make the parameter-space cost function multi-modal, ill-conditioned and highly nonlinear, so some form of non-local search is required to optimize it. Existing approaches that we are aware of do not work well in this context (§4.1.1,4.1.2), so we introduce a novel hybrid search scheme that combines oversized, covariance-scaled sampling with robust local optimization subject to joint and non-self-intersection constraints (§4.2). We end with experimental results on some challenging monocular sequences, that illustrate the need for each of robust cost modeling, joint and self-intersection constraints, and well-controlled sampling plus local optimization.

4.1.1 High-Dimensional Search Strategies

Locating good poses in a high-dimensional body configuration space is intrinsically difficult. Three main classes of search strategies exist: **local descent** incrementally improves an existing estimate, *e.g.* using local Taylor models to predict good search directions (Bregler and Malik, 1998; Rehg and Kanade, 1995; Kakadiaris and Metaxas, 1996; Wachter and Nagel, 1999); **regular sampling** evaluates the cost function at a predefined pattern of points in (a slice of) parameter space, *e.g.* a local rectangular grid (Gavrila and Davis, 1996); and **stochastic sampling** generates random sampling points according to some hypothesis distribution encoding "good places to look" (Deutscher et al., 2000; Sidenbladh et al., 2000). Densely sampling the entire parameter space would in principle guarantee a good solution, but it is infeasible in more than 2–3 dimensions. In 30 dimensions any feasible sample must be extremely sparse, and hence likely to miss significant cost minima. Local descent does at least find a local minimum[1], but with multimodality there is no guarantee that the global one is found. Whichever method is used, effective focusing is the key to high-dimensional search. This is an active research area (Deutscher et al., 2000; Heap and Hogg, 1998; Cham and Rehg, 1999), but no existing method can guarantee a global minima.

During tracking, the search method is applied time-recursively, the starting point(s) for the current search being obtained from the optimized results at the previous time step, perhaps according to some noisy dynamical model. To the (often limited!) extent that the dynamical law and the image matching cost are statistically realistic models, Bayes-law propagation of a probability density for

[1] In 30-D, even starting with an exactly-known, spherically normalized Gaussian, samples typically lie $\sqrt{30} \sim 5.5\sigma$ from the optimum, and the probability of them falling closer than 3σ is negligible. However, along any individual direction, samples are unlikely to fall further than 2σ.

the true state is possible. For linearized monomodal dynamics and observation models under least squares / Gaussian noise, this leads to Extended Kalman Filtering. For likelihood-weighted random sampling under general multimodal dynamics and observation models, bootstrap filters (Gordon et al., 1993; Gordon and Salmond, 1995) or CONDENSATION (Isard and Blake, 1998) result. In either case, parameters such as dynamical and measurement noise levels must be carefully tuned for good performance. Visual tracking usually works in the 'shotgun in the dark' regime: observation likelihoods are quite sharply peaked but multimodal, so to avoid mistracking, the dynamical noise has to be turned up until it produces a scatter of samples large enough to reach typically-nearby peaks. In this regime there is negligible trajectory smoothing, so Kalman-style covariance updating is superfluous: the previous posterior determines the locations and importance-weightings of the search regions, the dynamical noise determines their breadth, and the observation likelihood determines the location and shape of the new posterior peak(s) within each region.

Many existing implementations use inflated dynamical noise for empirical search focusing in this way (Cham and Rehg, 1999; Heap and Hogg, 1998; Deutscher et al., 2000), but for body tracking we find that this produces very poorly shaped search regions that waste most of the sampling effort in unprofitable configurations. We believe that an efficient high-dimensional search strategy must adapt to the local shape of the cost surface, focusing its samples on the lowest-cost areas near (but not too close to) the current minimum. Rather than inflating the dynamical noise, we will argue that one should define the search region by using realistic dynamics, then modestly inflating the resulting *prior* (previous posterior + dynamics) covariance. Because this inflates the posterior uncertainty as well as the dynamical one, it produces much deeper sampling along the most uncertain directions, and hence reduces mistracking due to inadequate exploration of these hard-to-estimate parameters (here, poorly observable depth d.o.f.). The change may seem small, but it makes a huge difference in practice. Consider the 32 d.o.f. cost spectrum in fig. 4.6. For inflation large enough to double the sampling radius along the most uncertain direction (*e.g.*, for a modest search for local minima along this cost valley), the uniform dynamical noise method would produce a search volume 10^{54} times larger than that of our prior-based one, and an overwhelming fraction of its samples would have extremely high cost (see also fig. 4.1 on page 53). Such wastage factors are clearly untenable: in practice, inflated dynamical noise methods can not sample deeply enough along uncertain directions to find local minima lying along the floors of long narrow valleys in the cost surface.

4.1.2 Previous Work

Below we will compare our method to several existing ones. For a more consistent literature review and details, see chapter 2. 3D body tracking from monocular sequences is significantly harder than 2D (Cham and Rehg, 1999; Ju et al., 1996) or multi-camera 3D (Kakadiaris and Metaxas, 1996; Gavrila and Davis, 1996; Bregler and Malik, 1998; Delamarre and Faugeras, 1999) tracking, and surprisingly few works have addressed it (Deutscher et al., 2000; Sidenbladh et al., 2000; Wachter

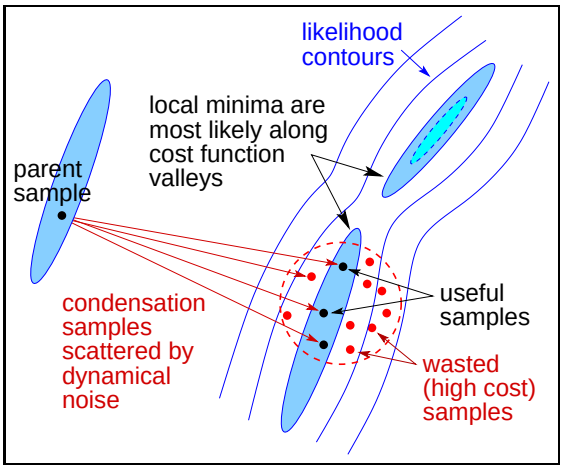


Figure 4.1: Why boosting the dynamical noise wastes samples ?

and Nagel, 1999; Gonglaves et al., 1995; Howe et al., 1999; Brand, 1999). The main additional difficulty is the omnipresence of depth ambiguities. Every limb or body segment lying near a frontoparallel plane has a first-order observability singularity. Small rotations towards or away from the camera leave the image unchanged to first order. Even for finite rotations there is a binary uncertainty, as towards-camera and away-from-camera rotations give very similar images. So even without image correspondence ambiguities, the matching cost function is always multimodal in parameter space. To successfully handle these difficulties, time integration or additional domain constraints such as joint limits and body parts non-intersection must be incorporated.

Deutscher *et al* uses a sophisticated 'annealed sampling' strategy to speed up CONDENSATION, but for his main sequence uses 3 cameras and a black background (Deutscher et al., 2000). Sidenbladh *et al* uses a similar importance sampling technique with a strong learned prior walking model to track a walking person in an outdoor sequence (Sidenbladh et al., 2000). Our current method uses no motion model (we optimize static poses), but it is true that when they hold, prior motion models are very effective tracking stabilizers. It is possible, but expensive, to track using a bank of motion models (Blake et al., 1999). Partitioned sampling (MacCormick and Isard, 2000) is another notable sampling technique for articulated models, under certain labeling assumptions (MacCormick and Isard, 2000; Deutscher et al., 2000).

Related particle filtering approaches have also addressed the some difficulties of discrete sampling that transfers the samples near the modes very slowly (Pitt and Shephard, 1997; Heap and Hogg, 1998; Cham and Rehg, 1999; Merwe et al., 2000; Choo and Fleet, 2001), especially in situations where the current observation likelihood peaks in the tail of the prior. The problem is much harder in high dimensions where a sample-based approximation of the posterior may require long sampling runs and could thus become prohibitively expensive. Therefore (Heap and Hogg, 1998; Cham and Rehg, 1999; Merwe et al., 2000) combine CONDENSATION style sampling with local optimization or Kalman filtering, while (Pitt and Shephard, 1997) samples using current observation

likelihood (and not the transition prior) while still relying on a purely discrete procedure. For visual tracking, Heap & Hogg and Cham & Rehg combine CONDENSATION-style sampling with local optimization, but they only consider the simpler case of 2D tracking (Heap and Hogg, 1998; Cham and Rehg, 1999). Cham & Rehg combine their heuristic 2D Scaled Prismatic Model (SPM) body representation with a first order motion model and a piecewise Gaussian resampling method for the CONDENSATION step. The Gaussian covariances are obtained from the Gauss-Newton approximation at the fitted optima, but the search region widths are controlled by the traditional method of adding a large dynamical noise. This appears to work reasonably well for 2D SPM tracking, which is essentially free of observability singularities. But we find (§4.4) that it can not handle the much less well-conditioned monocular 3D case. One questionable point in (Cham and Rehg, 1999) is the presence of closely-spaced minima with overlapping peaks, which motivated Cham & Rehg to introduce their piecewise Gaussian distribution model. We do not observe such overlaps, and we suspect that they were caused in (Cham and Rehg, 1999) by incomplete convergence in the optimizer, perhaps due to either over-loose convergence criteria or a noisy cost function (we took care to keep ours smooth).

Both Howe *et al* and Brand pose 3D estimation as a learning and inference problem, assuming that some form of 2D tracking (stick 2D model positions or silhouettes) is available over an entire time-series (Howe et al., 1999; Brand, 1999). Howe *et al* learn Gaussian distributions over short "snippets" of observed human motion trajectories, then use these as priors in an EM-based Bayesian MAP framework to estimate new motions (Howe et al., 1999). Brand learns a HMM with piecewise linear states and solves for the MAP estimate using an entropy minimization framework (Brand, 1999). As presented, these methods are basically monomodal so they can not accommodate multiple trajectory interpretations, and they also rely heavily on their learned-prior temporal models to stabilize the tracking. Nevertheless, they provide a powerful higher-level learning component that is complementary to the framework proposed in this paper.

4.1.3 Motivation of Approach

In order to perform monocular 3D tracking, here we advocate a balance between local and global search effort, using a combination of local optimization and carefully controlled sampling, *c.f.* (Heap and Hogg, 1998; Cham and Rehg, 1999; Merwe et al., 2000). Conventional global minimization aims to find the absolute global minimum, but it is important to realize that even if this were tractable, it would not, in general, suffice for problems with competing minima. In contrast, purely sampling-based methods aim at finding a full, 'true' underlying representation of the posterior distribution over model parameters, but it should be also understood that given the high-dimensionality of the problem and it's large number of minima, this is also a goal that is practically difficult to meet. Therefore, in many tracking or inference applications, one searches a good compromise [2], perhaps locating and tracking a set of minima that are 'representative' of the posterior

[2] See section §4.5 for a discussion on approximation accuracy as well as on optimization and sampling viewpoints.

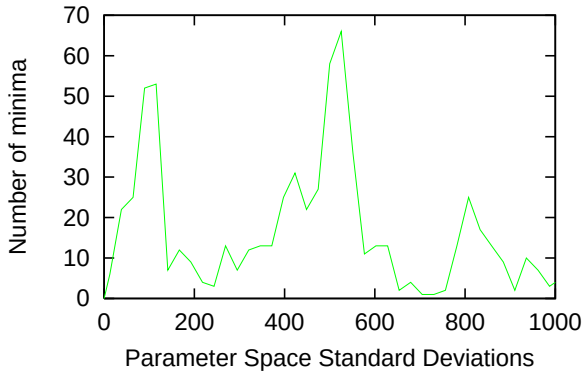


Figure 4.2: Typical parameter space minima distribution measured with respect to an arbitrary minimum. Notice that the minima are far from each other in parameter space so wide sampling is necessary to find them. However, boosting the dynamics by sampling from the transition prior (as in particle filtering, see fig. 4.1 on page 53) leads to inefficiencies.

likelihood encoded by the cost function, *i.e.* that capture most of its weight.

During tracking, it is necessary to both find and track local cost minima efficiently and escape from their basins of attraction [3] to discover other more remote ones. Reasons why these are difficult problems are: (*i*) The extent of the basin of attraction of each minimum is not known a-priori, but in general minima are far (many standard deviations) from each-other in parameter space (see fig. 4.2 on page 55 for typical distribution of inter-mode distances. See also chapter 7 for more exhaustive quantitative results). (*ii*) Distant sampling is therefore necessary to escape the basins and this leads to large volumes that need to be sampled efficiently. The problem is high-dimensional and a dense coverage is usually not feasible, so sparse sample proposals are needed. (*iii*) For ill-conditioned problems, the minimum shape is non-spherical and usually not aligned with the axes of the parameter space – instead it is highly elongated along directions corresponding to difficult to observe parameter combinations (see fig. 4.6 on page 62). Spherical sampling will either produce insufficient local scatter or lead to large wastage factors (see fig. 4.1 on page 53) so cost sensitive sampling is necessary (see fig. 4.3 on page 56). (*iv*) Even if a sample escapes out of the current basin of attraction, in high-dimensions, the probability of hitting the high-cost surroundings of another minimum is much larger than that of hitting its low cost core – This is due to the large increase of volume with radius, which makes the volume of the core of the minima tiny in comparison with their full basins of attraction (see fig. 4.4 on page 56). So local descent is needed to get back to lower cost regions (see for instance the large difference in medians between sampled and sampled then subsequently optimized configurations in table 7.1 on page 120). (*v*) Even for discrete sampling methods, high-cost samples are unlikely to be selected for resampling, unless the entire distribution is dominated by them – which will usually happen only if the track has already been lost.

[3] This is necessary while encountering temporal minima splittings (bifurcations) due to increased uncertainty.

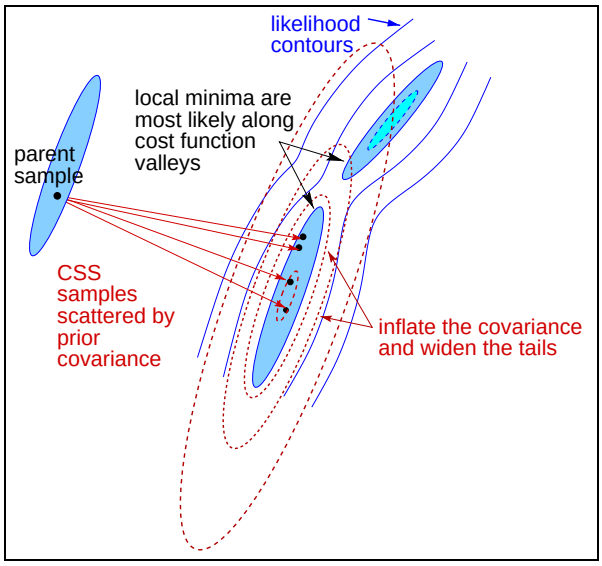


Figure 4.3: Sampling broadly on low-cost regions requires local cost covariance scaling.

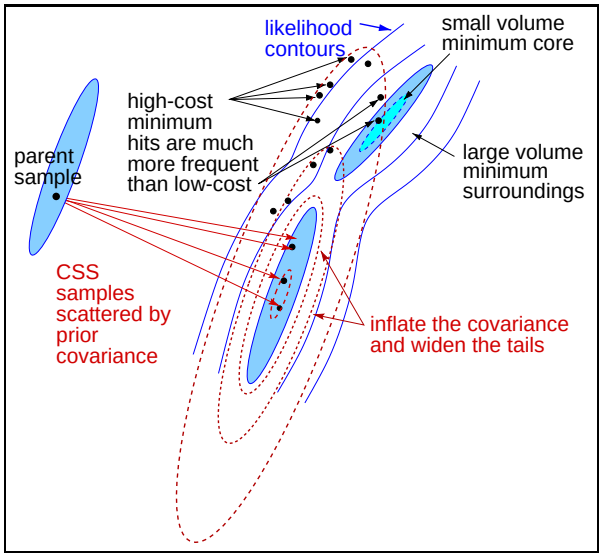


Figure 4.4: Why local optimization is necessary after sampling ?

We address the above issues by three mechanisms: (*i*) *wide tail sampling* in order to reduce trapping, (*ii*) *adaptive local cost covariance scaling* in order to avoid the sample wastage (*iii*) *local optimization (robust, constraint consistent)* in order to reduce the needle in haystack high-dimensional sampling effects.

4.1.4 Distribution Representation

We represent parameter space distributions as sets of separate modes $m_i \in \mathcal{M}$, each having an associated overall probability, mean and covariance matrix $m_i = (\mu_i, \Sigma_i, c_i)$. These can be viewed as Gaussian mixtures. Cham & Rehg also use multiple Gaussians, but they had to introduce a special piecewise representation as their modes seem to occur in clusters after optimization (Cham and Rehg, 1999). We believe that this is an artifact of their cost function design. In our case, as the modes are the result of robust continuous optimization, they are necessarily either separated or confounded. Our 3D monocular application also requires significantly more complex local optimization and sampling methods than (Cham and Rehg, 1999), as explained in §4.2.

4.1.5 Temporal Propagation

Equation 3.2 reflects the search for the model parameters in a static image, under likelihood terms and model priors but without a temporal or initialization prior. For temporal observations $\mathbf{R}_t = \{\bar{\mathbf{r}}_1, \bar{\mathbf{r}}_2 ..., \bar{\mathbf{r}}_t\}$, and sequence of states $\mathbf{X}_t = \{\mathbf{x}_1, \mathbf{x}_2 ..., \mathbf{x}_t\}$, the posterior distribution over model parameters becomes:

$$p(\mathbf{x}_t|\mathbf{R}_t) \;\propto\; p(\bar{\mathbf{r}}_t|\mathbf{x}_t)\, p(\mathbf{x}_t) \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1})\, p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}) \tag{4.1}$$

where $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ is a dynamical prior and $p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1})$ is the prior distribution from $t-1$. Together they form a (temporal) initialization prior $p(\mathbf{x}_t|\mathbf{R}_{t-1})$ for the static image search given by (3.2) on page 37 in the thesis.[4]

Proof. We can express the posterior over model parameters at current time step, by doing marginalization over parameters over the entire sequence up to the previous time step:

$$p(\mathbf{x}_t|\mathbf{R}_t) = \int_{\mathbf{X}_{t-1}} p(\mathbf{x}_t, \mathbf{X}_{t-1}|\mathbf{R}_t) \tag{4.2}$$

We introduce a first-order Markov assumptions that the state (model parameters) for the current time step depend only on the state at the previous time step. Thus, the equation above simplifies to:

$$p(\mathbf{x}_t|\mathbf{R}_t) = \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t, \mathbf{x}_{t-1}|\bar{\mathbf{r}}_t, \mathbf{R}_{t-1}) \tag{4.3}$$

Assuming also the independence of $\bar{\mathbf{r}}_t$ and $\mathbf{R}_{t-1}$ the equation (4.2) writes:

$$p(\mathbf{x}_t, \mathbf{x}_{t-1}|\bar{\mathbf{r}}_t, \mathbf{R}_{t-1}) = \frac{p(\bar{\mathbf{r}}_t, \mathbf{R}_{t-1}|\mathbf{x}_t, \mathbf{x}_{t-1}) p(\mathbf{x}_t, \mathbf{x}_{t-1})}{P(\mathbf{x}_t, \mathbf{x}_{t-1})} = \tag{4.4}$$

$$= \frac{p(\bar{\mathbf{r}}_t|\mathbf{x}_t, \mathbf{x}_{t-1}) p(\mathbf{R}_{t-1}|\mathbf{x}_t, \mathbf{x}_{t-1}) p(\mathbf{x}_t, \mathbf{x}_{t-1})}{p(\bar{\mathbf{r}}_t) p(\mathbf{R}_{t-1})} \tag{4.5}$$

[4] In practice, at any given time step we work on a negative log-likelihood 'energy' function that is essentially static, being based on both the current observation likelihood and the parameter space priors, as in (3.3) on page 37. The samples from the temporal prior $p(\mathbf{x}_t|\mathbf{R}_{t-1})$ are used as initialization seeds for local energy minimization. The different minima found will represent the components of the posterior mixture representation.

Given that $\mathbf{R}_t$ conditioned by $\mathbf{x}_t$ doesn't depend on $\mathbf{x}_{t-1}$ and $\mathbf{R}_{t-1}$ is independent as well of $\mathbf{x}_{t-1}$, the equation transforms to:

$$p(\mathbf{x}_t, \mathbf{x}_{t-1}|\bar{\mathbf{r}}_t, \mathbf{R}_{t-1}) = \frac{p(\bar{\mathbf{r}}_t|\mathbf{x}_t)p(\mathbf{R}_{t-1}|\mathbf{x}_{t-1})p(\mathbf{x}_t, \mathbf{x}_{t-1})}{p(\bar{\mathbf{r}}_t)p(\mathbf{R}_{t-1})} = \tag{4.6}$$

$$= \frac{p(\bar{\mathbf{r}}_t|\mathbf{x}_t)p(\mathbf{R}_{t-1}|\mathbf{x}_{t-1})p(\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1})}{p(\bar{\mathbf{r}}_t)p(\mathbf{R}_{t-1})} = \tag{4.7}$$

$$= \frac{p(\bar{\mathbf{r}}_t|\mathbf{x}_t)}{p(\bar{\mathbf{r}}_t)}p(\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})\frac{p(\mathbf{R}_{t-1}|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1})}{p(\mathbf{R}_{t-1})} \tag{4.8}$$

where in the equation above:

$$p(\mathbf{x}_t, \mathbf{x}_{t-1}) = p_s(\mathbf{x}_t)p_d(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}) = p(\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}) \tag{4.9}$$

Thus, we have factored $p(\mathbf{x}_t, \mathbf{x}_{t-1})$ into a static prior over model parameters p_s and a dynamic prior p_d conditioned by the previous state. In subsequent derivations, indices were dropped for notational simplicity.

Given that $1/p(\bar{\mathbf{r}}_t)$ is actually constant the equation (4.6) can be written by dropping the scaling factor:

$$p(\mathbf{x}_t, \mathbf{x}_{t-1}|\bar{\mathbf{r}}_t, \mathbf{R}_{t-1}) = \frac{p(\bar{\mathbf{r}}_t|\mathbf{x}_t)}{p(\bar{\mathbf{r}}_t)}p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}) \tag{4.10}$$

$$\propto p(\bar{\mathbf{r}}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}) \tag{4.11}$$

By substitution into equation (4.3) we obtain:

$$p(\mathbf{x}_t|\mathbf{R}_t) = \int_{\mathbf{x}_{t-1}} p(\bar{\mathbf{r}}_t|\mathbf{x}_t)p(\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}) \tag{4.12}$$

$$\propto p(\bar{\mathbf{r}}_t|\mathbf{x}_t)p(\mathbf{x}_t)\int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}) \tag{4.13}$$

4.2 Search Algorithm

Our parameter search technique combines robust constraint-consistent local optimization with a more global discrete sampling method.

4.2.1 Mode Seeking using Robust Constrained Continuous Optimization

The cost function is expressed as the negative log likelihood of probability. In order to optimize a sample $\mathbf{x}$ to find the center of its associated likelihood peak, we employ an iterative second order robust constrained local continuous optimization procedure. At each iteration, the log-likelihood

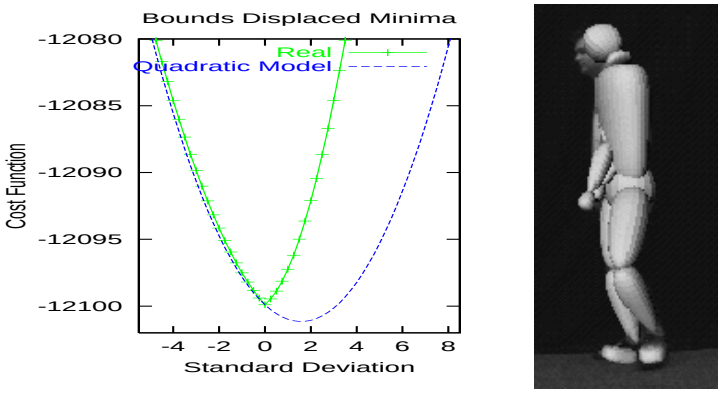


Figure 4.5: (a) Displaced minimum due to joint limits constraints, (b) Joint limits without body non-self-intersection constraints do not suffice for physical consistency.

gradients and Hessians corresponding to all observations are assembled, together with soft negative log-likelihood terms for prior contributions[5] from (3.3):

$$\mathbf{g} = \frac{\mathbf{d}f}{\mathbf{d}\mathbf{x}} = \mathbf{g}_o + \bigtriangledown f_{an} + \bigtriangledown f_s + \bigtriangledown f_{pa} \tag{4.14}$$

$$\mathbf{H} = \frac{\mathbf{d}^2 f}{\mathbf{d}\mathbf{x}^2} = \mathbf{H}_o + \bigtriangledown^2 f_{an} + \bigtriangledown^2 f_s + \bigtriangledown^2 f_{pa} \tag{4.15}$$

For local optimization we use a second order trust region method, where a descent direction is chosen by solving the regularized subproblem (Fletcher, 1987):

$$(\mathbf{H} + \lambda \mathbf{W})\delta\mathbf{x} = -\mathbf{g} \quad \text{subject to} \quad C_{bl} \cdot \mathbf{x} < 0 \tag{4.16}$$

where $\mathbf{W}$ is a symmetric positive-definite matrix and λ is a dynamically chosen weighting factor. Joint limits C_{bl} are handled as hard bound constraints in the optimizer, by projecting the gradient onto the current active constraint set. Adding joint constraints effectively changes the shape of the cost function, and hence the minimum reached. Fig. 4.5 plots a 1D slice through the constrained cost function together with a second order Taylor expansion of the unconstrained cost. Owing to the presence of the bounds, the cost gradient is nonzero (orthogonal to the active constraints) at the constrained minimum. Although the unconstrained cost function is smooth, the constrained one changes gradient abruptly when a constraint is hit because the active-set projection method changes the motion direction during the slice to maintain consistency with the constraints.

4.2.2 Covariance Scaled Sampling

Although representations based on propagating multiple modes, hypotheses or samples tend to increase the robustness of model estimation, the great difficulty with high-dimensional distributions

[5] 'Soft' means that these terms are part of the cost surface. This is in contrast with 'hard' joint angle limit priors that restrict the range of variation of their corresponding parameters.

is finding a sample-able proposal density that often hits their **typical sets** — the areas where most of their probability mass is concentrated. Here we develop a proposal density based on local parameter estimation uncertainties[6]. The local sample optimizations give us not only local modes, but also their (robust, constraint consistent) Hessians and hence estimates of the local posterior parameter estimation uncertainty at each mode[7].

The main insight is that alternative cost minima are most likely to occur along local valleys in the cost surface, *i.e.* along highly uncertain directions of the covariance. It is along these directions that cost-modeling imperfections and noise, and 3D nonlinearities and constraints, have the most likelihood of creating multiple minima, as the cost function is shallowest and the 3D movements are largest there. This is particularly true for monocular 3D estimation, where the covariance is unusually ill-conditioned owing to the many unobservable motion-in-depth d.o.f. Some examples of such multimodal behavior along high covariance eigen-directions are given in fig. 4.8. Also, it is seldom enough to sample at the scale of the estimated covariance. Samples at this scale almost always fall back into the same local minimum, and significantly deeper sampling is necessary to capture nearby but non-overlapping modes lying further up the valley[8]. Hence, we sample according to rescaled covariances, typically scaling by a factor of 10 or so. Finally, one can sample either randomly, or according to a regular pattern[9]. For the experiments showed here, we use random sampling using CSS with both Gaussian and heavy tail distributions. Fig. 4.7 summarizes the resulting covariance-scaled search method.

Given the explanations above, some comments are in order, since the sample generation process has both static and dynamic aspects, that we may need to address in the following practical situations:

(*i*) Generate fair samples from a prior for which the modes are known. This is a relatively simple problem. In our case, the Gaussian mixture[10] can be used as an importance sampling distri-

[6] Variable metric sampling methods exist in global optimization, in the context of continuous annealing (Vanderbilt and Louie, 1984) and have been applied by (Black, 1982) to low-dimensional ($2d$) optical flow computation.

[7] A sample is optimized to convergence to obtain the corresponding mode mean. The Hessian matrix at convergence point gives the principal curvature directions and magnitude around the mean and the inverse Hessian gives the covariance matrix, reflecting the cost local uncertainty structure. The Hessian is estimated by the algorithm §4.2.1 during optimization (using (4.14)), and the covariance is readily obtained from there.

[8] In part this is due to imperfect modeling, which easily creates biases greater than a few standard deviations, particularly in directions where the measurements are weak. Also, one case in which multiple modes are likely to lie close enough together in position and cost to cause confusion is when a single mode fragments due to smooth evolutions of the cost surface. In this case, generic singularity ('catastrophe') theory predicts that generically, exactly two modes will arise (bifurcation) and that they will initially move apart very rapidly (at a speed proportional to $1/\sqrt{t}$). Hence, it is easy for one mode to get lost if we sample too close to the one we are tracking.

[9] For efficiency purposes, an implementation could sample regularly, in fact only along lines corresponding to the lowest few covariance eigen-directions. Although this gives a very sparse sampling indeed, this is an avenue that can be explored in practice.

[10] There are at least two ways a mixture can be obtained. One is by using clustering techniques on a set of generated posterior samples like the ones obtained in a CONDENSATION step. This may lead to modes that are not necessarily separated, and might not actually reflect the true modes of the posterior because of sampling artifacts. A second possibility,

bution, and correction weighting can readily be performed. Being parametric, the mixture provides a compact, explicit multiple mode representation and accurate localization. These were the main rationales behind their introduction by (Heap and Hogg, 1998) and (Cham and Rehg, 1999) (section 5.1, page 5). Their sampling stage is based on the process model (with deterministic and large noise dynamics) and therefore similar to that of the CONDENSATION algorithm.

(*ii*) Recover new modes of a prior for which only *some* of the modes are known, under static image observations. Clearly, this is a significantly more difficult problem, due to the fact that, a-priori, the distribution of unknown modes is not available and neither one the boundaries (or of the saddles) basin of attraction of existing modes (in order to easily find their neighbors). Such a situation is typical in practice, as a full estimate of the posterior or even of all its peaks in rarely available in high-dimensions. But a set of initial modes may be known, perhaps as an incompletely recovered posterior from a prior tracking step or from a semi-automatic initialization process. Nevertheless, likelihood peaks are far from each other in parameter space (*e.g.* the static reflective ambiguities for human pose, or cascades of incorrect matches when a model limb is assigned to an incorrect image limb side). For visual examples, see fig. 5.6 at page 88 and fig. 5.8 on page 92. See also fig. 5.7 upper two rows, on page 91 for quantitative results on inter-minimum distance in parameter space and on the basin of attraction (saddle) positions. For minima distribution in parameter space and in cost, see fig. 5.7 upper two rows, on page 91, and fig. 7.3-7.10 on page 112. Consequently, given the distance between peaks, sampling purely based on the known (and potentially incomplete) prior is fatal, as the overwhelming majority of the samples will fall back into the modes that they arose from (see also chapter 6) leading to poor search diversity[11]. So broader sampling is necessary, but is also important to focus the samples in relatively low cost regions (see also fig. 4.1). To achieve this we propose to use the local cost surface to shape a broad sampling distribution. As expected on theoretical grounds, this turns out to give significantly improved results for CSS (for sample cost median, number of minima found, their cost) than competing methods based on either pure prior-based sampling or prior-based sampling plus spherical 'dynamical' noise (see table 7.1 on page 120).

(*iii*) Sample a prior under dynamic observations but without making restrictive assumptions on

followed here, is to optimize the samples locally. In this case the multiple modes found are true local peaks that are, necessarily, either separated or confounded.

[11]Certain metrics exist for assessing the efficiency of particle filters (Liu, 1996; MacCormick and Isard, 2000). The 'survival diagnostic' (also known as the 'effective sample size') intuitively indicates how many particles will survive a resampling operation – if the weights are very unbalanced, only few of the particles may survive, thus reducing search diversity. However, even if the the weights of the sample set are well balanced, this does not necessary imply that the real modes of the distribution are well explored. For example, the samples might lie (*e.g.* if the sampling radius is too small) approximately on the isocontours of a single mode or in a flat region. Another property of a particle set is the 'survival rate' which in some (rather limited) cases characterizes the ratio of the volume of support of the posterior to the volume of support of the prior. The intuition is that when this ratio is too low, the density estimation method may produce inaccurate estimates. Nevertheless, it can easily be seen that this ratio will be reasonably high when all of the samples are trapped in just one of the density's modes corresponding to a single hypothesis that evolves smoothly.

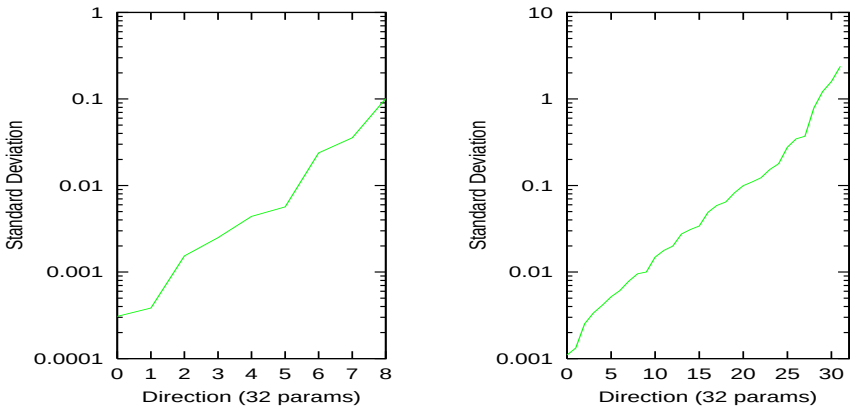


Figure 4.6: Typical covariance eigenvalue spectra plotted on a logarithmic scale, for a local minimum. $\sigma_{\max}/\sigma_{\min}$ is 350 for the 8 d.o.f. arm model, and 2000 for the 32 d.o.f. body one.

the motion of its peaks. In this case the modes from time $t-1$ are available, and it is critical that the sampling procedure covers the peaks of the observation density in the next time step t. This means that samples should be generated in the basins of attraction of the density peaks *after* they moved in between two temporal observations. In the absence of knowledge about the peaks' motion (*i.e.* known system dynamics), we exploit the local uncertainty structure in the distribution, and shape the search region based on it. Again, broader sampling is necessary, since the tracked object moves between observation frames. Also, as explained above, the mode tracking process is not one-to-one. New modes might emerge or split under the effect of increased uncertainty, and it is important that the sampling process does not miss such events by sampling too close to a given mode core, which may both move and split between two temporal observations. Our quantitative results in §4.4 directly support such findings *e.g.* for mode splitting reflecting bi-modality generated by locally planar versus in depth motion explanations (see below).

Our reason for *not* using specific motion models is that we want to track *general* human motion. Therefore, we cannot assume a mostly-known dynamics for the human subject (see for instance, the sequences given in fig. 4.10, fig. 4.11 and fig. 4.12). Consequently, no sophisticated form for the process model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ is used. For the experiments shown in the next section, we actually use trivial driftless diffusion dynamics, so that CSS only has to account for local uncertainty and cover moving peaks. But one could also use *e.g.* a constant velocity model. Note however, that even such weak models may turn out be misleading for tracking, especially when the subject suddenly changes trajectories, *e.g.* when encountering obstacles, doing unexpected motions like turning, or switching activities, *etc.* If the dynamics of the tracked subject was known a-priori (as in the case of people walking or running), one can use specifically learned motion models to stabilize tracking as in (Rohr, 1994; Deutscher et al., 2000; Sidenbladh et al., 2000).

To build up intuition about the shape of our cost surface, we studied it empirically by sampling along

From the 'old' mixture prior $p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}) = \sum_{i=1}^{K} \pi_i^{t-1} \mathcal{N}(\mu_i^{t-1}, \Sigma_i^{t-1})$, at time $t-1$, build 'new' mixture posterior $p(\mathbf{x}_t|\mathbf{R}_t) = \sum_{i=1}^{K} \pi_i^t \mathcal{N}(\mu_i^t, \Sigma_i^t)$, at time t, as follows:

---

1. Build covariance scaled proposal density $p_{t-1}^* = \sum_{i=1}^{K} \pi_i^{t-1} \mathcal{P}(\mu_i^{t-1}, \Sigma_i^{t-1})$. For the experiments we have used both Gaussian and Cauchy tails. The covariance scaled Gaussian component proposals are $\mathcal{P} = \mathcal{N}(\mu_i^{t-1}, s\Sigma_i^{t-1})$ with s=4–14 in our experiments. The covariance scaled Cauchy component proposals are defined as $\mathcal{P} = \mathcal{H}(\mu_i^{t-1}, \Sigma_i^{t-1})$, where $\mathcal{H}_{\mathbf{x}}(\mu, \Sigma) = \frac{1}{\pi|\Sigma|^{1/2}[1+(\mathbf{x}-\mu)\Sigma^{-1}(\mathbf{x}-\mu)^t]}$.

---

2. Generate components of the posterior at time t by sampling from p_{t-1}^* as follows. Iterate over $j = 1...N$ until the desired number of samples N are generated:
2.1. Choose component i from p_{t-1}^* with probability π_i^{t-1}.
2.2. Sample from $\mathcal{P}(\mu_i^{t-1}, \Sigma_i^{t-1})$ to obtain $\mathbf{s}_j$.
2.3. Optimize $\mathbf{s}_j$ over the observation likelihood at time t, $p(\mathbf{x}|\bar{\mathbf{r}}_t)$ defined by (3.2), using the local continuous optimization algorithm (§4.2.1). The result is the parameter space configuration at convergence μ_j^t, and the covariance matrix $\Sigma_j^t = \mathbf{H}(\mu_j^t)^{-1}$. If the μ_j^t mode has been previously found by a local descent process, discard it (For notational clarity, without any loss of generality, consider all the modes found are different).

---

3. Construct an un-pruned posterior for time t as: $p_t^u(\mathbf{x}_t|\mathbf{R}_t) = \sum_{j=1}^{N} \pi_j^t \mathcal{N}(\mu_j^t, \Sigma_j^t)$ where $\pi_j^t = \frac{p(\mu_j^t|\bar{\mathbf{r}}_t)}{\sum_{j=1}^{N} p(\mu_j^t|\bar{\mathbf{r}}_t)}$.

---

4. Prune the posterior p_t^u to keep the best K components with highest probability π_j^t (rename indices $j = 1...N$ into the set $k = 1...K$) and renormalize the distribution as follows: $p_t^p(\mathbf{x}_t|\mathbf{R}_t) = \sum_{k=1}^{K} \pi_k^t \mathcal{N}(\mu_k^t, \Sigma_k^t)$ where $\pi_k^t = \frac{p(\mu_k^t|\bar{\mathbf{r}}_t)}{\sum_{j=1}^{K} p(\mu_j^t|\bar{\mathbf{r}}_t)}$.

---

5. For each mixture component $j = 1...K$ in p_t^p, find the closest prior component i in $p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1})$, according to a Bhattacharyya distance $\mathcal{B}_{ij}(\mu_i^{t-1}, \Sigma_i^{t-1}, \mu_j^t, \Sigma_j^t) = (\mu_i^{t-1} - \mu_j^t)^T \left[\frac{\Sigma_i^{t-1}+\Sigma_j^t}{2}\right]^{-1} (\mu_i^{t-1} - \mu_j^t) + \frac{1}{2}\log \frac{\left|\frac{\Sigma_i^{t-1}+\Sigma_j^t}{2}\right|}{\sqrt{|\Sigma_i^{t-1}||\Sigma_j^t|}}$. Recompute $\pi_j^t = \pi_j^t * \pi_i^{t-1}$. Discard the component i of $p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1})$ from further consideration.

---

6. Compute the posterior mixture $p(\mathbf{x}_t|\mathbf{R}_t) = \sum_{k=1}^{K} \pi_k^t \mathcal{N}(\mu_k^t, \Sigma_k^t)$ where $\pi_k^t = \frac{\pi_k^t}{\sum_{j=1}^{K} \pi_j^t}$.

Figure 4.7: The steps of our covariance-scaled sampling algorithm.

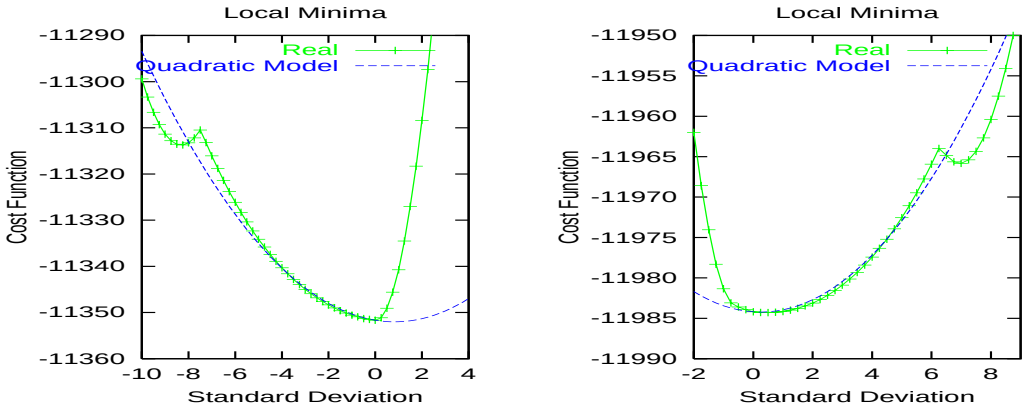


Figure 4.8: Multimodal behavior along highest uncertainty eigen-directions (0.8 and 0.9 s in cluttered body tracking sequence).

uncertain covariance directions (in fact eigenvectors of the covariance matrix), for various model configurations. With our carefully selected image descriptors, the cost surface is smooth apart from the apparent gradient discontinuities caused by active-set projection at joint constraint activation points. Hence, our local optimizer reliably finds a local minimum. We find that multiple modes do indeed occur for certain configurations, usually separated by cost barriers that a classical (uninflated) sampling strategy would have difficulty crossing. For example, fig. 4.8 shows the two most uncertain modes of the fig. 4.10 human tracking sequence at times 0.8 s and 0.9 s. (More precisely, the higher-cost slice-minima are not full-parameter-space minima, but they do lie in the attraction zones of ones). Secondary minima like those shown here occur rather often, typically for one of two reasons. The first is incorrect registration and partial loss of track when both edges of a limb model are attracted to the same image edge of the limb. This is particularly critical when there is imperfect body modeling and slightly mis-estimated depth. The second occurs when motion in-depth can be temporarily interpreted as a different quasi-planar motion. This does not immediately cause loss of registration but this eventually occurs when the uncertainty is "revealed" as a result of an "observable" motion. It is this second situation that occurs in fig. 4.8 (see also fig. 4.13). Identifying and tracking such ambiguous behaviors is critical, as incorrect quasi-planar depth interpretations quickly lead to tracking failure. Fig. 4.9a shows some typical cost eigendirection slices at much larger scales in parameter space. Note that we recover the expected robust shape of the distribution, with some but not too many spurious local minima. This is crucial for tracker robustness, given that, despite its power in representing and tracking multiple possible solutions, it still has limited resources that can be lost in uninformative, spurious minima. Consequently, the combination of our robust cost function design and informed search is likely to be computationally efficient.

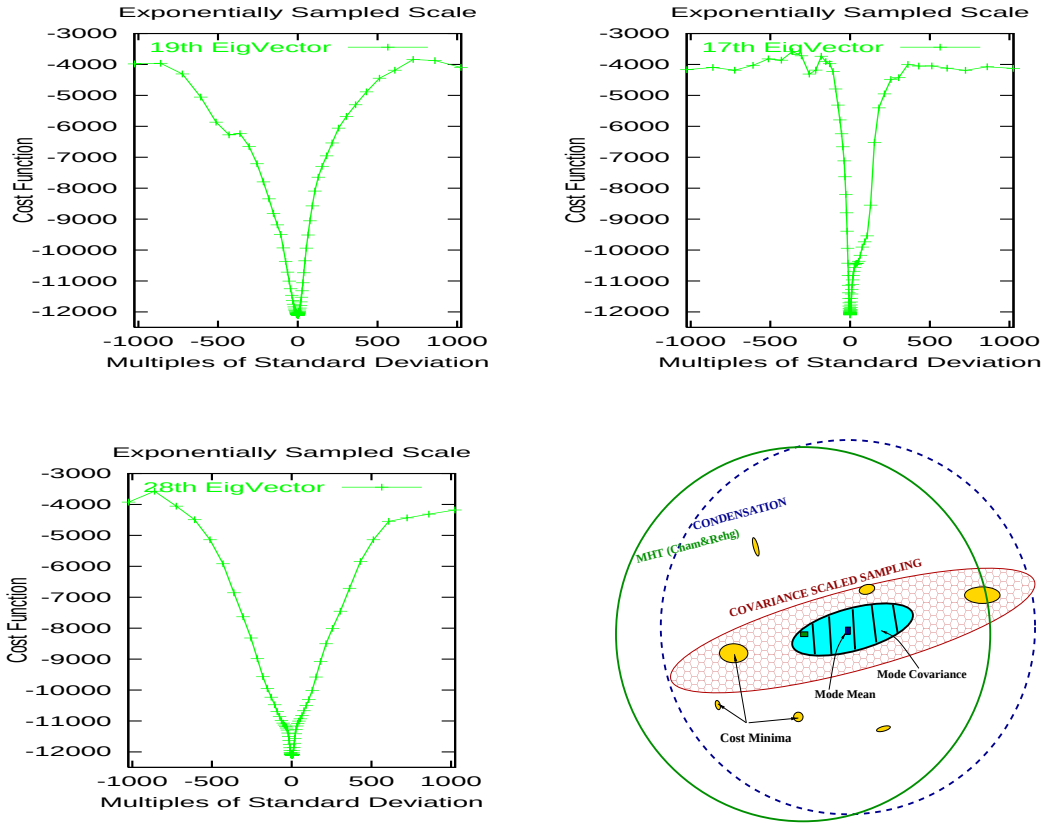


Figure 4.9: (a,b,c) Cost function slices at large scales, (d) Comparison of sampling methods: (1) CONDENSATION (dashed circle coverage) randomizes each sample by dynamic noise, (2) MHT ((Cham and Rehg, 1999), solid circle) samples within covariance support (dashed ellipse) and applies the same noise policy as (1), finally, our (3) *Covariance Scaled Sampling* (pattern ellipse) targets good cost minima (flat filled ellipses) by inflating or heavy tail sampling the local robust covariance estimation (dashed ellipse)).

4.3 Model Initialization

The visual tracking starts with a set of initial hypotheses resulting from a model initialization process. Correspondences need to be specified between model joint locations and approximate joint positions of the subject in the initial image. Given this input, we perform a consistent estimation of joint angles and body dimensions.

Previous approaches to model initialization from similar input and a single view, have not fully addressed the generality and consistency problems (Taylor, 2000; Barron and Kakadiaris, 2000), failing to enforce the joint limits constraints, and making assumptions regarding either restricted camera models or restricted human subject poses in the image, respectively. Alternative approaches based on approximate 2D joint recovery based on learned silhouette appearance (Rosales and

Sclaroff, 2000), might be used to bootstrap an algorithm like the one we propose.

Our initialization algorithm is a hierarchical three step process that follows a certain parameter estimation scheduling policy. Each stage of the estimation process is essentially based on the formulation described in §4.2.1.

Hard joint limits constraints are automatically enforced at all stages by the optimization procedure, and corresponding parameters on the left and right sides of the body are "mirrored", while we collect measurements from the entire body (see below). Initialization proceeds as follows. Firstly, we solve an estimation problem under the given 3D model to 2D image correspondences (by minimizing the projection error), and prior intervals on the internal proportions and sizes of the body parts (namely parameters $\mathbf{x}_a$, $\mathbf{x}_i$ and and some of $\mathbf{x}_d$). Secondly we optimize only over the remaining volumetric body sizes alone (limb crossections and their tapering parameters $\mathbf{x}_d$) while holding the other parameters fixed, using both the given correspondences and the local contour signal from image edges. Finally, we refine all model parameters ($\mathbf{x}$) based on similar image information as in the second stage. The covariance matrix corresponding to the final estimate is used to generate an initial set of hypotheses, which are propagated in time using the algorithm described in §4.2. While the process is heuristic, it gives a balance between stability and flexibility. In practice we find that enforcing the joint constraints, mirror information and prior bounds on the variation of body parameters gives far more stable and satisfactory results. However, in the monocular case, the initialization always remains ambiguous and highly uncertain in some parameter space directions, especially under 3D-2D joint correspondence data. In our case, we employ a suitable, coarse parameter initialization and use the above process for fine refinement, but if available, one can fuse pose information from multiple images.

4.4 Experiments

For the entire set of experiments, we use an edge and intensity based cost function and modeling based on priors and constraints as explained in chapter 3. We use Gaussian tails for CSS. A quantitative sampling evaluation of both Gaussian and Cauchy tails and different scalings appears in chapter 7 on page 108.

To illustrate our method we show results for an 8 second arm tracking sequence and two full body ones (3.5 s and 4 s). All of these sequences contain both self-occlusion and significant relative motion in depth. The first two (fig. 4.10) were shot at 25 frames (50 fields) per second against a cluttered, unevenly illuminated background. The third (fig. 4.12) is at 50 non-interlaced frames per second against a dark background, but involves a more complex model and motions. More detailed plates containing sequences with full images and tracking results can be found in (fig. 7.11, on page 125) and in (fig. 7.12 on page 126). In our unoptimized implementation, a 270 Mhz SGI O2 required about 5 s per field to process the arm experiment and 180 s per field for the full body ones, most of the time being spent evaluating the cost function. The figures overlay the current

best candidate model on the original images. We are also exploring various tracker components failure modes as follows. A *Gaussian single mode tracker* is a single hypothesis tracker doing local continuous optimization based on Gaussian error distributions and without enforcing any physical constraints. a *Robust single mode tracker* is an improved version of the above that uses robust error distributions. A *Robust single mode tracker with joint limits* additionally enforces physical constraints. For multimodal trackers, the sampling strategy could be either CONDENSATION-based or CSS-based as introduced in previous sections.

Cluttered background sequences: These sequences explore 3D estimation behavior with respect to image assignment and depth ambiguities, for a bending rotating arm under an 8 d.o.f. model and a pivoting full-body motion under a 30 d.o.f. one. They have cluttered backgrounds, specular lighting and loose fitting clothing. In the arm sequence, the deformations of the arm muscles are significant and other imperfections in our arm model are also apparent.

The *Gaussian single mode tracker* manages to track 2D frontoparallel motions in moderate clutter, although it gradually slips out of registration when the arm passes the strong edges of the white pillar (0.5 s and 2.2 s for the arm sequence and 0.3 s for the human body sequence). Any significant motion in depth is untrackable.

The *robust single mode tracker* tracks frontoparallel motions reasonably well even in clutter, but quickly loses track during in-depth motions, which it tends to misinterpret as frontoparallel ones. In the arm tracking sequence, shoulder motion towards the camera is misinterpreted as frontoparallel elbow motion, and the error persists until the upper bound of the elbow joint is hit at 2.6 s and tracking fails. In the full body sequence, the pivoting of the torso is underestimated, being partly interpreted as quasi-frontoparallel motion of the left shoulder and elbow joints. Despite the presence of anatomical joint constraints, the fist eventually collapses into the body if non-self-intersection constraints are not present.

The *robust joint-limit-consistent multi-mode tracker* correctly estimates the motion of the entire arm and body sequence. We retain just the 3 best modes for the arm sequence and the 7 best modes for the full human body sequence. As discussed in §4.2.2, multimodal behavior occurs mainly during significantly non-frontoparallel motions, between 2.2–4.0 s for the arm sequence, and over nearly the entire full body sequence (0.2–1.2 s). For the latter, the modes mainly reflect the ambiguity between true pivoting motion and its incorrect "frontoparallel explanation".

We also compared our sampling method with a 3D version of Cham & Rehg's MHT (Cham and Rehg, 1999) for the body turn sequence. Note that the original method is based on non-robust least-squares optimization and doesn't incorporate physical constraints or model priors. We used 10 modes to represent the distribution in our 30 d.o.f. 3D model, whereas (Cham and Rehg, 1999) used 10 for their 38 d.o.f. 2D SPM model. Our first set of experiments used a non-robust SSD image matching metric and a Levenberg-Marquardt routine for local sample optimization, as in (Cham and Rehg, 1999) (except that we use analytical Jacobians). With this cost function, we find that outliers cause large fluctuations, bias and frequent convergence to physically invalid configu-

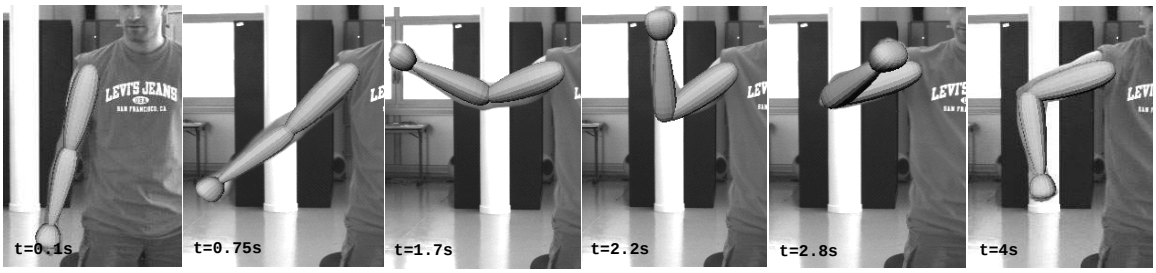


Figure 4.10: Arm tracking against a cluttered background.

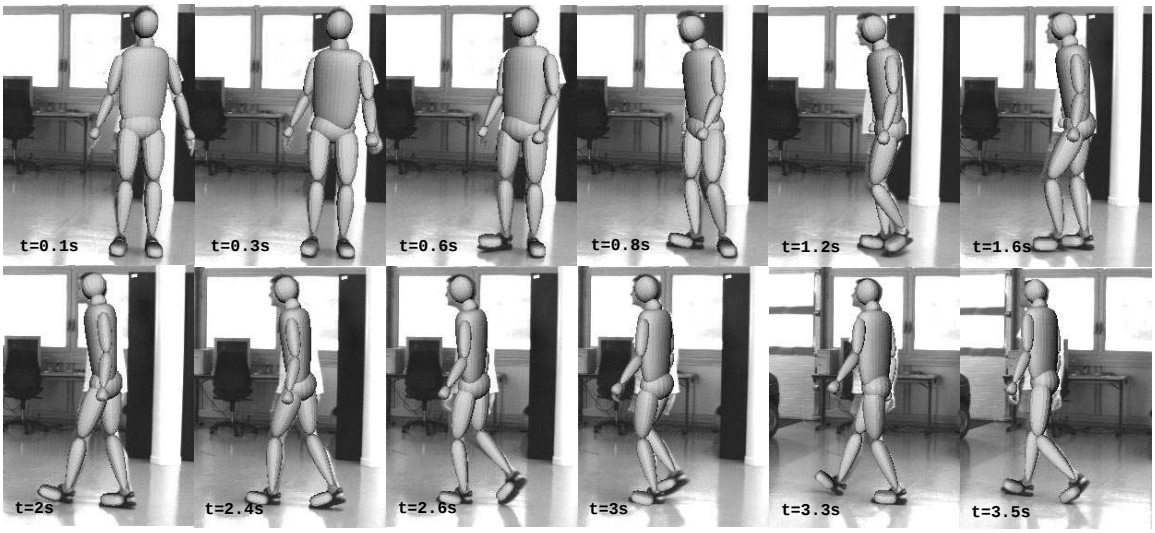


Figure 4.11: Human tracking against a cluttered background.

rations. Registration is lost early in the turn (0.5 s), as soon as the motion becomes significantly non-frontoparallel. Our second experiments used our robust cost function and optimizer, but still with MHT-style sampling. The track survived further into the turn, but was lost at 0.7 s when the depth variation became larger. As expected, we find that a dynamical noise large enough to provide usefully deep sampling along uncertain directions produces far too deep sampling along well-controlled ones, so that most of the samples are wasted on uninformative high-cost configurations. Similar arguments apply to standard CONDENSATION, as can be seen in the monocular 3D experiments of (Deutscher et al., 2000).

Black background sequence: In this experiment we focus on 3D errors, in particular depth ambiguities and the influence of physical constraints and parameter stabilization priors. We use an improved body model with 34 d.o.f. The four extra parameters control the left and right clavicle joints in the shoulder complex, which we find to be essential for following many arm motions. Snapshots from the full 4 s sequence are shown in fig. 4.12, and various failures modes in fig. 4.13.

The *Gaussian single mode tracker* manages to follow near-frontoparallel motions fairly reliably

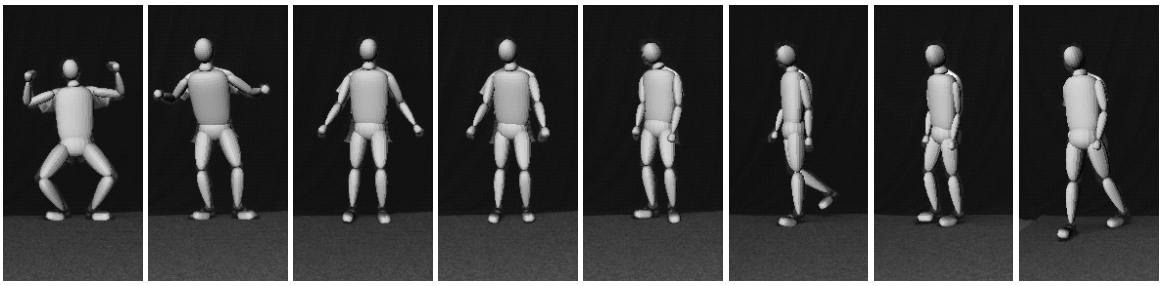


Figure 4.12: Human tracking under complex motion.

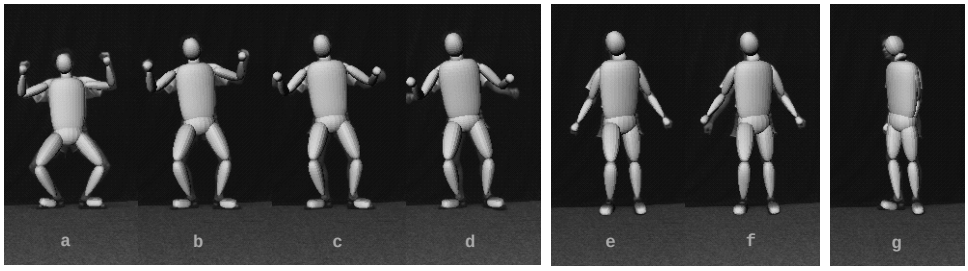


Figure 4.13: Failure modes of various components of the tracker (see text).

owing to the absence of clutter, but it eventually loses track after 0.5 s (fig. 4.13a–d). The *robust single mode tracker* tracks the non-frontoparallel motion somewhat longer (about 1 s), although it significantly mis-estimates the depth (fig. 4.13e–f — the right leg and shoulder are pushed much too far forward and the head is pushed forward to match subject contour, *c.f.* the "correct" pose in fig. 4.12). It eventually loses track during the turn. The *robust multi-mode tracker with joint-limits* is able to track quite well, but, as body non-self-intersection constraints are not enforced, the modes occasionally converge to physically infeasible configurations (fig. 4.13g) with terminal consequences for tracking. Finally, the *robust fully constrained multi-mode tracker* is able to deal with significantly more complex motions and tracks the full sequence without failure (fig. 4.12). Plates containing the full images and tracking results can be found in (fig. 7.11) on page 125 and (fig. 7.12) on page 126.

4.5 Limitations and Approximation Accuracy

In the previous section, we discussed the failure modes of some of the components of the CSS algorithm and showed their behavior in practical tracking contexts. Here we turn to more technical points. A more general discussion will appear in chapter 7.

The CSS algorithm involves both local continuous optimization and global covariance-scaled sampling. It therefore has a natural mechanism to trade-off speed and robustness. When tracking fails, both the number of modes used to represent the distribution and the number of samples

produced in the sampling stage can be increased. This may allow the tracking to survive through particularly difficult portions of the image sequence (although at an increased computational cost). In a weak, asymptotic sense, this is ensured by the sampling stage which guarantees that any region of the parameter space is eventually visited, and therefore that the basin of attraction of each mode is sampled. Consequently, when using local optimization, any peak of the distribution can be found with non-zero probability. It has been argued that mixed continuous/discrete trackers (Heap and Hogg, 1998; Cham and Rehg, 1999; Merwe et al., 2000) will 'diverge' if the visual information is ambiguous and converge to a 'best' mode when the target in the image is easily detectable. However, this kind of divergence is not that important here. We are working with likelihood surfaces that have multiple peaks with individual probabilities. Local optimization methods can converge to any of these peaks and sampling methods will eventually 'condense' near them if they use enough samples. Given the sampling stage, both methods have a chance of jumping between peaks (*i.e.* escaping spurious ones), although this may be a very slow process. The method presented in this chapter and the ones introduced later in the thesis are expressly designed to address the problems of efficient and more systematic multi-modal exploration. Note also that CSS can be viewed as an importance sampling distribution and correction weighting for fair sample generation can be readily performed with respect to the true prior.

A second issue concerns the algorithm's efficiency versus bias behavior. In tracking contexts, under the assumption of temporal coherency, one may want to confine to search in the neighborhood of the configurations propagated from the previous tracking time step. This can be done implicitly by designing a likelihood surface with strong local responses[12], or by tuning the search process for locality. In either case, a full 'objective' estimate of the posterior distribution is too expensive both for sampling and optimization algorithms. Nevertheless, restricting attention to quasi-local search or to artificially localized likelihood surfaces carries the risk of missing important peaks, or searching a density that becomes incorrectly peaked.

A third issue concerns the approximation accuracy of a Gaussian mixture for arbitrary multi-modal distributions. The mixture representation may lose precision away from the mode cores, and this may affect the accuracy of statistical calculations based on such an approximation. Nevertheless, for tracking or localization applications in vision, we are mostly interested in the modes themselves and less in the low-probability regions in their remote tails. Alternatively, sampling methods, being non-parametric, can in principle be more accurate there if they use enough samples. However sampling the tails is a fairly infrequent event in itself, since the samples tend to cluster towards the high probability regions near the modes' cores (see also chapter 6) [13]. In any case, the

[12]This can be for instance an optical flow correspondence field like the one we have described in §3.6.1. If used in connection with a Gaussian brightness error model, this can behave as a local prior, forcing local image velocity explanations and pruning away remote, potentially 'objective' multi-modality.

[13]Pure sample-based representations also provide little insight into the structure of the uncertainty and the degree of multi-modality of the likelihood surface. For multi-modal distributions, the sample mean becomes uninformative. Smoothing algorithms prove effective in disambiguating optimal temporal trajectories, but mostly for transient multi-

issue of approximation accuracy in low-probability regions is not a main concern here. Provided initial seeds are available in the individual mode's basins of attraction, sampling methods can generate fair samples from the modes and optimization methods can precisely identify their means and covariances by local descent. The two techniques can be use inter-changeably, depending on the application. It is however the process of finding the initial seeds for each mode that represents the major difficulty for high-dimensional multi-modal distributions.

A fourth and important practical issue concerns the properties of the likelihood function. For many complex models a good likelihood is difficult to build, and the one used may be a poor reflection of the desired observation density. In these situations, the strength of true and spurious responses is similar and this may affect the performance of the tracking algorithm, irrespective how much computational power is used. In such contexts, it can be very difficult to identify the true tracked trajectory in a temporal flow of spurious responses. This is a particularly complex problem, since many likelihoods commonly used in vision degrade un-gracefully under occlusion/disocclusion events and viewpoint change. We empirically find this is the most frequent reason for failure in our tracking system. At present, we do not have good mechanisms for detecting disocclusion events in complex backgrounds. The CSS algorithm has an elegant mechanism that accounts for high-uncertainty if particular degrees of freedom are not observed (like occluded limbs, *etc*.) and it will automatically sample broadly there. However, for sub-sequences with long occlusion events, it is often more likely to attach occluding limbs to segments in the background than to maintain them occluded. A more global silhouette or contour detector or higher-order matching consistency may help here. As an indication of the potential benefits of this, for our experiments we currently use silhouettes if available from background subtraction and motion segmentation, or the motion boundaries from the robust optical flow computation to weight the importance of contours. This controlled tracking drift in the sequences shown, but a more general and robust solution would be desirable. We defer further discussion to chapter 7, where we classify the properties of likelihood surfaces and discuss the limitations of the similarity measures we have employed in this thesis.

4.6 Conclusions

In this chapter, we have introduced a new method for monocular 3D human body tracking, based on optimizing a robust model-image matching cost metric combining robustly extracted edges, flow and motion boundaries, subject to 3D joint limits, non-self-intersection constraints, and model priors. Optimization is performed using Covariance Scaled Sampling, a novel high-dimensional search strategy based on sampling a hypothesis distribution followed by robust constraint-consistent local refinement to find a nearby cost minima. The hypothesis distribution is determined by combining the posterior at the previous time step (represented as a Gaussian mixture defined by the observed cost minima and their Hessians / covariances) and the assumed dynamics to find the current-

modality.

timestep prior, then inflating the prior covariances to sample more broadly. Our experiments on real sequences show that — as expected on theoretical grounds — this is significantly more effective than using inflated dynamical noise estimates as in previous approaches because it concentrates the samples on low-cost points near the current minima, rather than points that are simply nearby whatever their cost.

We shall compare stochastic and regular-sampling CSS, and variant covariance-scaled hypothesis generators such as longer-tailed or coreless distributions in chapter 7. It should also be possible to extend the benefits of CSS to CONDENSATION by using inflated (diluted weight) posteriors and dynamics for sample generation, then re-weighting the results, *c.f.* (Deutscher et al., 2000). Also, this tracking framework can be extended by incorporating better pose and motion priors.

Chapter 5

Building Deterministic Trajectories for Finding Nearby Minima

Getting trapped in suboptimal local minima is a perennial problem in model based vision, especially in applications like monocular human body tracking where complex nonlinear parametric models are repeatedly fitted to ambiguous image data. We show that the trapping problem can be attacked by building 'roadmaps' of nearby minima linked by *transition pathways* — paths leading over low 'cols' or 'passes' in the cost surface, found by locating the *transition state* (codimension-1 saddle point) at the top of the pass and then sliding downhill to the next minimum. We know of no previous vision or optimization work on numerical methods for locating transition states, but such methods do exist in computational chemistry, where transitions are critical for predicting reaction parameters. We present two families of methods, originally derived in chemistry, but here generalized, clarified and adapted to the needs of model based vision: *eigenvector tracking* is a modified form of damped Newton minimization, while *hypersurface sweeping* sweeps a moving hypersurface through the space, tracking minima within it. Experiments on the challenging problem of estimating 3D human pose from monocular images show that our algorithms find nearby transition states and minima very efficiently, but also underline the disturbingly large number of minima that exist in this and similar model based vision problems.

Keywords: Model based vision, global optimization, saddle points, 3D human tracking.

5.1 Introduction

Many visual modeling problems can be reduced to cost minimization in a high dimensional parameter space. Local minimization is usually feasible, but practical cost functions often have large numbers of local minima and it can be very difficult to ensure that the desired one is found. Exhaustive search rapidly becomes impracticable in more than 2–3 dimensions, so most global optimization methods focus on heuristics for finding 'good places to look next'. This includes both determinis-

tic techniques like branch-and-bound and pattern search, and stochastic importance samplers like simulated annealing, genetic algorithms and tabu search.

Unfortunately, global optimization remains expensive with any of these methods. In this chapter we develop an alternative strategy based on building 1-D 'roadmaps' of the salient minima, linked by paths passing over **saddle points**: stationary (zero gradient) points that have one or more negative curvature directions, so that they represent 'cols' rather than 'hollows' in the cost surface. We will restrict attention to **transition states** (saddles with just one negative curvature direction), as these give the minimum-peak-cost pathways between local minima. Our focus is on methods for finding the salient transitions surrounding an initial minimum. Given these, adjacent minima can be found simply by sliding downhill using local minimization.

Despite the omnipresence of local minima, we know of no previous vision or optimization work on systematic numerical algorithms for locating transition states. As far as we can judge, this was generally considered to be intractable. However such methods *do* exist in the computational chemistry / solid state physics community, where transitions are central to the theory of chemical reactions[1]. We will describe two families of transition-finding algorithms that have roots in computational chemistry: **eigenvector tracking** is a modified form of damped Newton minimization, while **hypersurface sweeping** sweeps a moving hypersurface through the space, tracking minima within it. These methods are potentially useful in almost any visual modeling problem where local minima cause difficulties. Examples include model based tracking, reconstruction under correspondence ambiguities, and various classes of camera pose and calibration problems. We present experimental results on monocular model based human pose estimation.

5.1.1 Literature Review

We start with a brief overview of the computational chemistry / solid state physics literature on locating transition states. This literature should be accessible to vision workers with high-school chemistry and a working knowledge of optimization. However the underlying ideas can be difficult to disentangle from chemistry-specific heuristics, and some papers are rather naive about numerical optimization issues. We therefore give a self-contained treatment and generalization of two of the most promising approaches below, in numerical analysts language.

A transition state is a local minimum along its $n-1$ positive curvature directions in parameter space, but a local maximum along its remaining negative curvature one. So transition state search methods often reduce to a series of $(n-1)$-D minimizations, while moving or maximizing along the remaining direction. The main differences lie in the methods of choosing the directions to use.

[1] Atomic assemblies can be modeled in terms of the potential energy induced by interactions among their atoms, *i.e.* by an energy function defined over the high-dimensional configuration space of the atoms' relative positions. A typical assembly spends most of its time near an energy minimum (a stable or quasi-stable state), but thermal perturbations may sometimes cause it to cross a transition state to an adjacent minimum (a chemical reaction). The energy of the lowest transition joining two minima determines the likelihood of such a perturbation, and hence the reaction pathway and rate.

Eigenvector tracking methods (Crippen and Scheraga, 1971; Hilderbrandt, 1977; Cerjan and Miller, 1981; Wales, 1989; Wales and Walsh, 1996; Munro and Wales, 1999; Henkelman and Jonsson, 1999; Simons et al., 1983; Nichols et al., 1990; Helgaker, 1991; Culot et al., 1992; Bofill, 1994) are modified Newton minimizers designed to increase the cost along one of the curvature eigendirections, rather than reducing it along all eigendirections. If the lowest curvature direction is chosen they attempt to find the lowest gradient path to a transition state by walking along the 'floor' of the local cost 'valley'. However this behavior can not be guaranteed (Jorgensen et al., 1988; Sun and Ruedenberg, 1993; Jensen, 1995) and valleys need not even lead to saddle points: they might be 'blind', with the transition states located off to one side. An early method (Hilderbrandt, 1977) used explicit Newton minimization in the $(n-1)$-D space obtained by eliminating the coordinate with the largest overlap with the desired up-hill direction. Later quasi-Newton methods use Lagrange multipliers (Cerjan and Miller, 1981; Wales, 1989; Wales and Walsh, 1996) or shifted Hessian eigenvalues (Simons et al., 1983; Nichols et al., 1990; Helgaker, 1991; Culot et al., 1992; Bofill, 1994) to ensure that the cost function is increased along the chosen 'uphill eigenvector' direction while being minimized in all orthogonal ones. Maintaining a consistent direction to follow can be delicate and several competing methods exist, including using a fixed eigenvector index (Hilderbrandt, 1977; Cerjan and Miller, 1981; Wales, 1989; Wales and Walsh, 1996; Munro and Wales, 1999; Henkelman and Jonsson, 1999) and attempting to track corresponding eigenvectors from step to step (Simons et al., 1983; Nichols et al., 1990; Helgaker, 1991; Culot et al., 1992; Bofill, 1994). Eigenvector tracking can be motivated as a 'virtual cost minimization' obtained by inverting the sign of the negative Hessian eigenvalue and the corresponding gradient component (Mousseau and Berkema, 1998; Helgaker, 1991). This gives an intuitive algebraic analogy with minimization, but none of its convergence guarantees as the virtual cost function changes at each step.

Trajectory methods in global optimization (Fiodorova, 1978; Griewank, 1981; Branin and Hoo, 1972) are based on a similar idea of tracing routes through stationary points of a high-dimensional cost function. Some methods are based either on sequences of descents and ascents along eigenvectors (Fiodorova, 1978), but they only give a 2D formulation, with no obvious extension to high-dimensions and no clear ascent heuristic. Golf methods (Griewank, 1981), aim at exploring different minima and equilibrate at a certain energy level (assumed known a-priori). Branin type methods (Branin and Hoo, 1972; Anderson and Walsh, 1986) are based on solving differential equations derived from the gradient stationarity condition (essentially forms of Newton update), but managing singularities and bifurcations (Newton leaves) during the iteration is still an unsolved problem.

Constraint based methods (Crippen and Scheraga, 1971; Abashkin and Russo, 1994; Abashkin et al., 1994; Barkema, 1996; Mousseau and Berkema, 1998; Henkelman and Jonsson, 1999) aim to use some form of constrained optimization to guarantee more systematic global progress towards a transition state. Crippen & Sheraga's early method (Crippen and Scheraga, 1971) builds an uphill path by minimizing in the orthogonal hyperplane of a ray emanating from the initial minimum and passing through the current configuration. Mousseau & Barkema use a similar but less rigorous

technique based on changing the gradient sign in one direction followed by conjugate root-finding in the other directions (Mousseau and Berkema, 1998). Barkema (Barkema, 1996) uses a biased repulsive spherical potential and optimizes subject to this soft constraint. New minima are found but the method does not attempt to pass exactly through a saddle point. Abashkin & Russo (Abashkin and Russo, 1994; Abashkin et al., 1994) minimize on successively larger radius hyperspheres centered at a minimum, and also include a method for refining approximately located saddle points. The use of hyperspheres forces the search to move initially along the valley floor of the cost surface (Jensen, 1995), so usually at most two distinct saddles can be found. Below we show how to steer the initial search along any desired direction using ellipsoidal surfaces. There are also stochastic search methods designed to find transition states and they are introduced and analyzed in chapter 6.

Relevant constrained approaches in global optimization are known as *penalty methods* (filled function, tunneling) and attempt to change the optimized cost function in order to prevent multiple determination of the same local minimum, and force the discovery of new ones (Goldstein and Price, 1971; Levy and Montalvo, 1985; Ge, 1987). Effectively, the methods perform a sequence of local optimizations on an augmented cost surface containing increasing radii repellers centered in the known minima, until new ones are found (this is necessary given that the extent of the basins of attraction of the minima are not generally known). However, none of these methods appears to give a general treatment for high-dimensions, nor do they define a repeller functional that is well adapted to the local shape of the original cost surface, which is critical in ensuring search diversity. Moreover, strategies more efficient than local descent may be desirable when increasing repeller's radius or detecting transitions to new minima basins.

5.2 Algorithms for Finding Transition States

Many efficient methods exist for finding local minima of smooth high dimensional cost surfaces. Minimization allows strong theoretical guarantees as the reduction in the function value provides a clear criterion for monitoring progress. For example, for a bounded-below function in a bounded search region, any method that ensures 'sufficient decrease' in the function at each step is 'globally convergent' to some local minimum (Fletcher, 1987). Finding saddle points is much harder as there is no universal progress criterion and no obvious analogue of a 'downhill' direction. Newton-style iterations provide rapid *local* convergence near the saddle, but it is not so obvious how to find sufficiently nearby starting points. We will consider several methods that extend the convergence zone. We are mainly interested in saddles as starting points for finding adjacent minima, so we will focus on methods that can be started from a minimum and tuned to find nearby transition states. (Efficient 'rubber band relaxation' methods also exist for finding the transition state(s) linking two given minima (Sevick et al., 1993)).

5.2.1 Newton Methods

Let $f(\mathbf{x})$ be the cost function being optimized over its n-D parameter vector $\mathbf{x}$, $\mathbf{g} \equiv \frac{\partial f}{\partial \mathbf{x}}$ be the function's **gradient** and $\mathbf{H} \equiv \frac{\partial^2 f}{\partial \mathbf{x}^2}$ be its **Hessian**. We seek **transition states**, stationary points $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ at which the Hessian has one negative and $n-1$ positive eigenvalues. If there is a stationary point at $\mathbf{x}+\boldsymbol{\delta}\mathbf{x}$, a first order Taylor approximation at $\mathbf{x}$ gives:

$$\mathbf{0} = \mathbf{g}(\mathbf{x}+\boldsymbol{\delta}\mathbf{x}) \approx \mathbf{g}(\mathbf{x}) + \mathbf{H}\,\boldsymbol{\delta}\mathbf{x} \tag{5.1}$$

Solving this linear system for $\boldsymbol{\delta}\mathbf{x}$ and iterating to refine the approximation gives the **Newton iteration**:

$$\mathbf{x} \leftarrow \mathbf{x}+\boldsymbol{\delta}\mathbf{x} \qquad \text{with update} \qquad \boldsymbol{\delta}\mathbf{x} = -\mathbf{H}^{-1}\mathbf{g} \tag{5.2}$$

When started sufficiently close to any regular[2] stationary point, Newton's method converges to it, but how close you need to be is a delicate point in practice.

For Newton-based minimization, convergence can be globalized by adding suitable damping to shorten the step and stabilize the iteration. The standard methods use the **damped Newton** update:

$$\boldsymbol{\delta}\mathbf{x} = -(\mathbf{H}+\lambda\mathbf{D})^{-1}\mathbf{g} \tag{5.3}$$

where $\mathbf{D}$ is a positive diagonal matrix (often the identity). The damping factor $\lambda > 0$ is manipulated by the algorithm to ensure stable and reliable progress downhill towards the minimum. Damping can be viewed as Newton's method applied to a modified local model for f, whose gradient at $\mathbf{x}$ is unchanged but whose curvature is steepened to $\mathbf{H}+\lambda\mathbf{D}$.

Similarly, to reduce the step and stabilize the iteration near a saddle point, negative curvatures must be made more negative, and positive ones more positive. In a Hessian eigenbasis $\mathbf{H} = \mathbf{V}\,\mathbf{E}\,\mathbf{V}^\top$, where $\mathbf{E} = \mathrm{diag}(\lambda_1, ..., \lambda_n)$ are the eigenvalues of $\mathbf{H}$ and the columns of $\mathbf{V}$ are its eigenvectors, the undamped Newton update becomes:

$$\boldsymbol{\delta}\mathbf{x} = -\mathbf{V}\,(\bar{g}_1/\lambda_1, \ldots, \bar{g}_n/\lambda_n)^\top \tag{5.4}$$

where $\bar{g}_i \equiv (\mathbf{V}^\top\mathbf{g})_i$ are the eigen-components of the gradient. Damping can be introduced by replacing this with[3]:

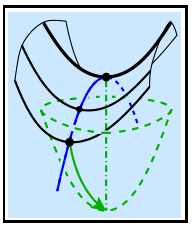
$$\boldsymbol{\delta}\mathbf{x} = -\mathbf{V}\,\mathbf{u}(\lambda), \quad \mathbf{u}(\lambda) \equiv \left(\frac{\bar{g}_1}{\lambda_1+\sigma_1\lambda}, ..., \frac{\bar{g}_n}{\lambda_n+\sigma_n\lambda}\right)^\top = \left(\frac{\sigma_1\bar{g}_1}{\sigma_i\lambda_1+\lambda}, ..., \frac{\sigma_n\bar{g}_n}{\sigma_n\lambda_n+\lambda}\right)^\top \tag{5.5}$$

where $\sigma_i = \pm 1$ is a desired sign pattern for the λ_i. Damping $\lambda > \max_i(-\sigma_i\lambda_i, 0)$ ensures that the denominators are positive, so that the iteration moves uphill to a maximum along the eigendirections with $\sigma_i = -1$ and downhill to a minimum along the others. At each step this can be viewed

[2] 'Regular' means that $\mathbf{H}$ is nonsingular and 2^{nd} order Taylor expansion converges.

[3] There is nothing absolute about eigenvalues! Affine changes of coordinates leave the original Newton method unchanged but produce essentially inequivalent eigen-decompositions and dampings.

as the minimization of a virtual local function with curvatures $\sigma_i\lambda_i$ and sign-flipped gradients $\sigma_i\bar{g}_i$. But the model changes at each step so none of the usual convergence guarantees of well-damped minimization apply: f itself is *not* minimized.



As in minimization, λ must be varied to ensure smooth progress. There are two main strategies for this: **Levenberg-Marquardt** methods manipulate λ directly, while the more sophisticated **trust region** ones maintain a local region of supposed-'trustworthy' points and choose λ to ensure that the step stays within it, for instance $\|\boldsymbol{\delta}\mathbf{x}(\lambda)\| = \|\mathbf{u}(\lambda)\| \lesssim r$ where r is a desired 'trust radius'. (Such a λ can be found efficiently with a simple 1-D Newton iteration started at large λ (Fletcher, 1987)). In both cases, convergence criteria and model accuracy metrics such as the relative f-prediction error:

$$\beta \;=\; \left| \frac{f(\mathbf{x}+\boldsymbol{\delta}\mathbf{x}) - f(\mathbf{x})}{\mathbf{g}^\top\boldsymbol{\delta}\mathbf{x} + \boldsymbol{\delta}\mathbf{x}^\top\mathbf{H}\boldsymbol{\delta}\mathbf{x}/2} - 1 \right| \tag{5.6}$$

are monitored, and the damping is increased (larger λ or shorter r) if the accuracy is low, decreased if it is high, and left unchanged if it is intermediate (*e.g.*, by scaling λ or r up or down by fixed constants).

As in minimization, if the exact Hessian is unavailable, quasi-Newton approximations based on previously computed gradients can be used. Positive definiteness is not required so update rules such as Powell's are preferred (Bofill, 1994):

$$\mathbf{H} \leftarrow \mathbf{H} - \frac{\boldsymbol{\delta}\mathbf{x}^\top\boldsymbol{\xi}}{\|\boldsymbol{\delta}\mathbf{x}\|^4}\boldsymbol{\delta}\mathbf{x}\,\boldsymbol{\delta}\mathbf{x}^\top + \frac{\boldsymbol{\xi}\,\boldsymbol{\delta}\mathbf{x}^\top + \boldsymbol{\delta}\mathbf{x}\,\boldsymbol{\xi}^\top}{\|\boldsymbol{\delta}\mathbf{x}\|^2} \tag{5.7}$$

where:

$$\boldsymbol{\xi} = \mathbf{g}(\mathbf{x}+\boldsymbol{\delta}\mathbf{x}) - \mathbf{g}(\mathbf{x}) - \mathbf{H}(\mathbf{x})\boldsymbol{\delta}\mathbf{x} \tag{5.8}$$

5.2.2 Eigenvector Tracking

Now consider the choice of the signs σ_i. Roughly speaking, the damped iteration moves uphill to a maximum along directions with $\sigma_i = -1$ and downhill to a minimum along directions with $\sigma_i = +1$, *i.e.* it tries to find a stationary point whose principal curvatures λ_i have signs σ_i. To find minima we need $\sigma_i = +1$, and for transition states exactly one σ_i should be made negative[4]. The question is, which one.

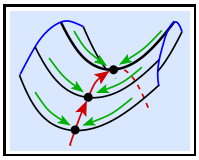
This question is thornier than it may seem. To ensure continued progress we need to track and modify "the same" eigenvector(s) at each step. Unfortunately, there is no globally well defined correspondence rule linking eigenvectors at different points, especially given that the trajectory followed depends strongly on the eigenvector(s) chosen. So in practice we must resort to one of several imperfect correspondence heuristics. For transition state searches we need only track

[4]This holds irrespective of the λ_i and $\bar{g}_i$ at the *current* state, which affect only the damping required for stability.

one eigenvector (the one that is given $\sigma_i = -1$), so we will concentrate on this case. A simple approach would be to choose a fixed direction in space (perhaps the initial eigendirection) and take the eigenvector with maximal projection along this direction. But many such directions are possible and the most interesting saddles may happen not to have negative curvature along the particular direction chosen. Alternatively, we can try to track a given eigenvector as it changes. The problem is that globally, eigendirections are by no means stable. Eigenvalues change as we move about the space, but generically (in codimension 1) they never cross. When they approach one another, the eigenbasis of their 2D subspace becomes ill-conditioned and slews around through roughly 90° to avoid the collision. Seen from a large enough scale, the eigenvalues do seem to cross with more or less constant eigenvectors, but on a finer scale there is no crossing, only a smooth but rapid change of eigendirection that is difficult to track accurately. Whichever of the two behaviors is desired, it is difficult to choose a step length that reliably ensures it, so the numerical behavior of eigenvector-tracking methods is often somewhat erratic. In fact, the imprecise coarse scale view is probably the desired one: if we are tracking a large eigenvalue and hoping to reduce it to something negative, it will have to "pass through" each of the smaller eigenvalues. Tracking at too fine a scale is fatal as it (correctly) prevents such crossings, instead making the method veer off at right angles to the desired trajectory.

Even without these problems there would be no guarantee that a saddle point of the desired signature was found (*e.g.* the trajectory could diverge to infinity). Also, as with other damped Newton methods, the whole process is strongly dependent on the affine coordinate system used. Nevertheless, eigenvector tracking is relatively lightweight, simple to implement, and it often works well in practice.

5.2.3 Hypersurface Sweeping

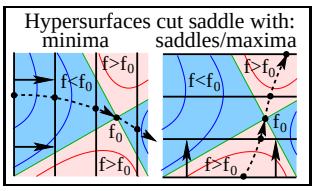


Eigenvector trackers do not enforce any notion of global progress, so they could sometimes behave erratically, *e.g.* cycling or stalling. To prevent this we can take a more global approach to the '$(n-1)$-D minimization and 1D maximization' required for transition state search. 'Hypersurface sweeping' approaches sweep an $(n-1)$-D hypersurface across the parameter space — typically a moving hyperplane or an expanding hyper-ellipsoid centered at the initial minimum — tracking local minima within the hypersurface and looking for temporal maxima in their function values. The intuition is that as the hypersurface expands towards a transition state, and assuming that it approaches along its negative curvature direction, the $(n-1)$-D minimization forces the hypersurface-minimum to move along the lowest path leading up to the saddle's 'col', and the 1-D maximization detects the moment at which the col is crossed. The method can not stall or cycle as the hypersurface sweeps through each point in the space exactly once.

The moving hypersurface can be defined either implicitly as the level sets $c(\mathbf{x}) = t$ of some function $c(\mathbf{x})$ on the parameter space (a linear form for hyperplanes, a quadratic one for hyper-

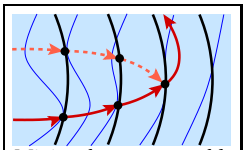
ellipsoids...), or explicitly in terms of a local parameterization $\mathbf{x} = \mathbf{x}(\mathbf{y}, t)$ for some hypersurface-t-parameterizing $(n-1)$-D vector $\mathbf{y}$. The minimum-tracking problem becomes:

$$\text{local\_max}_t \; f_c(t) \qquad \text{where} \qquad f_c(t) \equiv \begin{cases} \text{local\_min}_{c(\mathbf{x})=t} \, f(\mathbf{x}) \\ \text{local\_min}_{\mathbf{y}} \, f(\mathbf{x}(\mathbf{y}, t)) \end{cases} \tag{5.9}$$



Different local minima on the hypersurface typically lead to different transition states. To find the lowest cost transition we would in principle have to track every minimum. More seriously, transitions that are cut by the hypersurfaces in negative curvature directions are missed: they appear as saddle points or local maxima within the hypersurface, and so can not be found by tracking only minima. Nor do local maxima of $f_c(t)$ always indicate true transition states. At any hypersurface-stationary point, f's isosurface is necessarily tangent to the local hypersurface:

$$\mathbf{g} \propto \frac{\partial c}{\partial \mathbf{x}} \qquad \text{or} \qquad \mathbf{g}^\top \frac{\partial \mathbf{x}}{\partial \mathbf{y}} = \mathbf{0} \tag{5.10}$$



The point is a hypersurface-minimum, -saddle, or -maximum respectively as the cost isosurface has higher/mixed/lower signed curvature than the local hypersurface (*i.e.* as the isosurface is locally inside/mixed/outside the hypersurface). At points where the moving hypersurface transitions from being outside to being mixed w.r.t. the local isosurface, the minimum being tracked abruptly disappears and the solution drops away to some other local minimum on the hypersurface, causing an abrupt 'sawtooth' maximum in $f_c(t)$ (generically, the hypersurface-minimum collides with a hypersurface-saddle and is annihilated). The search can continue from there, but it is important to verify that the maxima found really are saddle points.

As any given family of hypersurfaces is necessarily blind to some saddle orientations, it is wise to try a range of different families. Hyperplanes search preferentially along a fixed direction whereas hyper-ellipsoids can find saddles lying in any direction. The initial direction of the minimum trajectory is determined by the hyperplane normal or the ellipsoid shape. Near a minimum $\mathbf{x}_0$ with Hessian $\mathbf{H}_0$, consider the ellipsoids $c(\mathbf{x}) = (\mathbf{x}-\mathbf{x}_0)^\top \mathbf{A} (\mathbf{x}-\mathbf{x}_0) = t$, where $\mathbf{A}$ is some positive definite matrix. To second order, $f(\mathbf{x})$ generically has exactly two local minima on an infinitesimal ellipsoid $c(\mathbf{x}) = t$: the $\pm$ directions of the smallest eigenvector of the matrix pencil[5] $\mathbf{A}+\lambda\,\mathbf{H}$. For most $\mathbf{A}$ there are thus only two possible initial trajectories for the moving minimum, and so at most two first saddles will be found. To find additional saddles we need to modify $\mathbf{A}$. We can enforce any desired initial direction $\mathbf{u}$ by taking the 'neutral' search ellipsoids $\mathbf{A} = \mathbf{H}$ (on which f is constant to second order, so that all initial directions are equally good) and flattening them slightly relative to the cost isosurfaces in the $\mathbf{u}$ direction. *E.g.*, to satisfy the Lagrange

[5] Numerically, these can be found by generalized eigen-decomposition of $(\mathbf{A}, \mathbf{H})$, or standard eigen-decomposition of $\mathbf{L}^{-\top}\mathbf{H}\mathbf{L}^{-1}$ where $\mathbf{L}\mathbf{L}^\top$ is the Cholesky decomposition of $\mathbf{A}$.

multiplier condition for a constrained minimum:

$$\frac{\boldsymbol{\partial} c}{\boldsymbol{\partial} \mathbf{x}} \propto \frac{\boldsymbol{\partial} f}{\boldsymbol{\partial} \mathbf{x}} \tag{5.11}$$

we can take:

$$\mathbf{A} = \mathbf{H} + \mu \frac{\mathbf{g}\,\mathbf{g}^\top}{\mathbf{u}^\top \mathbf{g}} \tag{5.12}$$

where $\mathbf{g} = \mathbf{H}\,\mathbf{u}$ is the cost gradient (and hence isosurface normal) at displacements along $\mathbf{u}$ and μ is a positive constant, say $\mu \sim 0.1$ for mild flattening. Similarly, for hyperplanes $c(\mathbf{x}) = \mathbf{n}^\top(\mathbf{x}-\mathbf{x}_0) = t$ with normal $\mathbf{n}$, the initial minimum direction is $\mathbf{u} = \pm\mathbf{H}^{-1}\,\mathbf{n}$, so to search in direction $\mathbf{u}$ we need to take $\mathbf{n} = \mathbf{H}\,\mathbf{u}$.

The minimum tracking process is a fairly straightforward application of constrained optimization, but for completeness we summarize the equations needed in the following section.

Summary: None of the current methods are foolproof. Damped Newton iteration is useful for refining estimated saddles but its convergence domain is too limited for general use. Eigenvector tracking extends the convergence domain but it is theoretically less sound (or at least, highly dependent on the step size and 'same eigenvector' heuristics). Hypersurface sweeping is better founded and provides at least weak guarantees of global progress, but it is more complex to implement and no single sweep finds all saddle points.

5.2.4 Hypersurface Sweeping Equations

Here we summarize the equations needed to implement hypersurface sweeping for both implicit and parametric hypersurfaces. For the implicit approach, let $\mathbf{g}_c \equiv \frac{\boldsymbol{\partial} c}{\boldsymbol{\partial} \mathbf{x}}$ and also $\mathbf{H}_c \equiv \frac{\boldsymbol{\partial}^2 c}{\boldsymbol{\partial} \mathbf{x}^2}$. The hypersurface constraint is enforced with a Lagrange multiplier λ, solving:

$$\frac{\boldsymbol{\partial}}{\boldsymbol{\partial} \mathbf{x}}\,(f + \lambda\, c) \;=\; \mathbf{g} + \lambda\,\mathbf{g}_c \;=\; \mathbf{0} \qquad \text{subject to} \qquad c = t \tag{5.13}$$

If we are currently at $(\mathbf{x}, \lambda)$, second order Taylor expansion of these equations for a constrained minimum at $(\mathbf{x}+\boldsymbol{\delta}\mathbf{x}, \lambda+\delta\lambda)$ gives the standard **sequential quadratic programming** update rule for $(\boldsymbol{\delta}\mathbf{x}, \delta\lambda)$:

$$\begin{pmatrix} \mathbf{H}_\lambda & \mathbf{g}_c \\ \mathbf{g}_c^\top & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\delta}\mathbf{x} \\ \delta\lambda \end{pmatrix} \;=\; -\begin{pmatrix} \mathbf{g} + \lambda\,\mathbf{g}_c \\ c - t \end{pmatrix} \qquad \text{where} \qquad \mathbf{H}_\lambda \;\equiv\; \mathbf{H} + \lambda\,\mathbf{H}_c \tag{5.14}$$

(The $\lambda\,\mathbf{H}_c$ term in the Hessian is often dropped for simplicity. This slows the convergence but still gives correct results). Similarly, in the parametric approach let $\mathbf{J} \equiv \frac{\boldsymbol{\partial}}{\boldsymbol{\partial} \mathbf{y}}\mathbf{x}(\mathbf{y}, t)$. The chain rule gives the reduced gradient $\mathbf{g}_\mathbf{y}$ and Hessian $\mathbf{H}_\mathbf{y}$:

$$\mathbf{g}_\mathbf{y} = \mathbf{J}\,\mathbf{g} \tag{5.15}$$

$$\mathbf{H}_\mathbf{y} = \mathbf{J}\,\mathbf{H}\,\mathbf{J}^\top + (\tfrac{\boldsymbol{\partial}}{\boldsymbol{\partial} \mathbf{y}}\mathbf{J})\,\mathbf{g} \tag{5.16}$$

These can be used directly in the Newton update rule $\boldsymbol{\delta}\mathbf{y} = -\mathbf{H}_{\mathbf{y}}^{-1}\,\mathbf{g}_{\mathbf{y}}$. In particular, if we eliminate one $\mathbf{x}$-variable — say x_n so that $\mathbf{y} = (x_1, ..., x_{n-1})$ and $x_n = x_n(\mathbf{y}, t)$ — we have:

$$\mathbf{J} = \left(\mathbf{I} \,|\, \tfrac{\boldsymbol{\partial} x_n}{\boldsymbol{\partial}\mathbf{y}}\right), \qquad \mathbf{g}_{\mathbf{y}} \;=\; \tfrac{\boldsymbol{\partial} f}{\boldsymbol{\partial}\mathbf{y}} + g_n\,\tfrac{\boldsymbol{\partial} x_n}{\boldsymbol{\partial}\mathbf{y}}, \qquad \mathbf{H}_{\mathbf{y}} \;=\; \mathbf{J}\,\mathbf{H}\,\mathbf{J}^\top + g_n\,\tfrac{\boldsymbol{\partial} x_n}{\boldsymbol{\partial}\mathbf{y}} \tag{5.17}$$

To save optimization work and for convergence testing and step length control, it is useful to be able to extrapolate the position and value of the next minimum from existing values. This can be done, *e.g.*, by linear extrapolation from two previous positions, or analytically by solving the constrained minimum state update equations $(\mathbf{g}{+}(\lambda{+}\delta\lambda)\mathbf{g}_c)(\mathbf{x}{+}\boldsymbol{\delta}\mathbf{x}) = \mathbf{0}$ or $\mathbf{g}_{\mathbf{y}}(\mathbf{y}{+}\boldsymbol{\delta}\mathbf{y}, t{+}\delta t) = \mathbf{0}$ to first order, assuming that $\mathbf{x}, t$ is already a minimum and $t \to t{+}\delta t$:

$$(\boldsymbol{\delta}\mathbf{x},\ \delta\lambda) \;=\; \tfrac{\delta t}{\mathbf{g}_c^\top\,\mathbf{H}_\lambda^{-1}\,\mathbf{g}_c}\left(\mathbf{H}_\lambda^{-1}\,\mathbf{g}_c,\ -1\right) \tag{5.18}$$

$$\boldsymbol{\delta}\mathbf{x} \;=\; \mathbf{J}\,\boldsymbol{\delta}\mathbf{y} + \tfrac{\partial \mathbf{x}}{\partial t}\,\delta t, \qquad \boldsymbol{\delta}\mathbf{y} \;=\; -\mathbf{H}_{\mathbf{y}}^{-1}\left(\tfrac{\partial \mathbf{J}}{\partial t}\,\mathbf{g} + \mathbf{J}\,\mathbf{H}\,\tfrac{\partial \mathbf{x}}{\partial t}\right)\,\delta t \tag{5.19}$$

Taylor expansion of $f(\mathbf{x}{+}\boldsymbol{\delta}\mathbf{x})$ then gives:

$$f_c(t{+}\delta t) \;\approx\; f_c(t) + f_c'\,\delta t + \frac{1}{2} f_c''\,\delta t^2 \tag{5.20}$$

with $f_c' \;=\; \mathbf{g}\,\tfrac{\boldsymbol{\delta}\mathbf{x}}{\delta t}$ and $f_c'' \;=\; \tfrac{\boldsymbol{\delta}\mathbf{x}}{\delta t}^\top\,\mathbf{H}\,\tfrac{\boldsymbol{\delta}\mathbf{x}}{\delta t}$. For step length control, we can either fix δt and solve for $\boldsymbol{\delta}\mathbf{x}$ or $\boldsymbol{\delta}\mathbf{y}$ (and hence $\mathbf{x}{+}\boldsymbol{\delta}\mathbf{x} \equiv \mathbf{x}(\mathbf{y}{+}\boldsymbol{\delta}\mathbf{y}, t{+}\delta t)$), or fix a desired trust region for $\boldsymbol{\delta}\mathbf{x}$ or $\boldsymbol{\delta}\mathbf{y}$ and work backwards to find a δt giving a step within it.

5.2.5 Implementation Details

We have tested several variants of each of the above methods. In the experiments below we focus on just two, which are summarized in fig. 5.1:

Hyper-ellipsoid sweeping: We start at a local minimum and use centered, curvature-eigenbasis-aligned ellipsoidal hypersurfaces flattened along one eigendirection, say the e^{th}. This restricts the initial search to an eigendirection (the e^{th}). This limitation could easily be removed, but gives a convenient, not-too-large set of directions to try. All calculations are performed in eigen-coordinates and the minimum is tracked using variable elimination (5.17) on x_e. In eigen-coordinates, the on-hypersurface constraint becomes:

$$\sum_i (x_i/\sigma_i')^2 = t^2 \tag{5.21}$$

where the σ_i' are the principal standard deviations, except that the e^{th} (eliminated) one is shrunk by say 20%. Solving for x_e gives:

$$x_e(\mathbf{y}, t) = \pm\sigma_e'(t^2 - \sum_{i \neq e} (x_i/\sigma_i')^2)^{1/2} \tag{5.22}$$

Hyper-ellipsoid Sweeping Transition State Search

1. Initialization

Given initial minimum $\mathbf{x}_0$ with Hessian $\mathbf{H}$, eigen-decompose $\mathbf{H}$ to $(\lambda_i, \mathbf{v}_i)$ with principal radii $\sigma_i = 1/\sqrt{\lambda_i}$. Choose an initial search eigen-direction e. Shrink σ_e by say 20% and prepare to eliminate x_e. Set initial step $\mathbf{x}_1 = \mathbf{x}_0 + t_1 \sigma_e \mathbf{v}_e$ where t_1 is say 3. Go to step 2.B.

2. Loop, Updating Hypersurface and Minimizing

A. k=k+1. Estimate an initial $\mathbf{x}_k$ by linear extrapolation to the trust radius. Compute the resulting t_k.

B. Minimize f on the t_k ellipsoid to get $f_c(t_k)$: $\mathbf{y}_k = \arg\min_{\mathbf{y}} f(\mathbf{x}_k(\mathbf{y}, t_k))$.

C. Compute $f_c' = \frac{\partial}{\partial t} f_c(t_k)$. If $f_c' < \epsilon$ we are near or past saddle: go to step 3.A. Otherwise go to step 2.A.

3. Line Search for Transition State Refinement

A. If $|f_c'| < \epsilon$, exit.

B. k=k+1. Estimate t_{saddle} by linear interpolation of last two f_c' values.

B. Optimize $\mathbf{y}_k$ as in step 2.B and go to step 3.A.

Eigenvector Tracking Transition State Search

Initialization

Set starting point $\mathbf{x}_0$, initial tracking direction $\mathbf{t}$ and initial trust radius r.

Eigenvector Tracking Loop

A. At $\mathbf{x}_k$, find $f_k, \mathbf{g}_k, \mathbf{H}_k$, the Hessian eigen-decomposition $(\lambda_i, \mathbf{v}_i)$ and eigen-basis gradient $\bar{g}_k$. Set n_- to the number of negative eigenvalues. If the problem has active internal constraints, project $\mathbf{t}$ onto the constraint surface.

B. If $k > 0$ choose $e = \max_i |\mathbf{v}_i^\top \mathbf{t}|$. Set $\alpha = |\mathbf{v}_e^\top \mathbf{t}|$ and $\mathbf{t} = \mathbf{v}_e$.

C. If $\lambda_e > 0$ set $\bar{g}_e = -\bar{g}_e$. Take an undamped Newton step if n_-=1 and $\|\boldsymbol{\delta}\mathbf{x}\| \leq r$. Otherwise take a damped one (5.5) with λ chosen so that $\|\boldsymbol{\delta}\mathbf{x}\| \leq r$.

D. Find the new f and the modeling error β (5.6). If $\beta > 0.3$ (say) or $\alpha < 0.7$ (say), shrink the trust radius r by say 50%. Otherwise, if $\beta < 0.2$ (say) and we took a damped step, grow r by say 40%.

E. If $\beta \geq 1$ go to step C. If $\|\mathbf{g}_k\| < \epsilon$ return success if n_-=1, failure otherwise. Otherwise, go to step A.

Figure 5.1: Our ellipsoid sweeping and eigenvector tracking algorithms for transition state search.

where we take $\mathbf{y} = (x_1, ..., x_{e-1}, x_{e+1}, ..., x_n)$. Derivatives are easily found. At each time step we

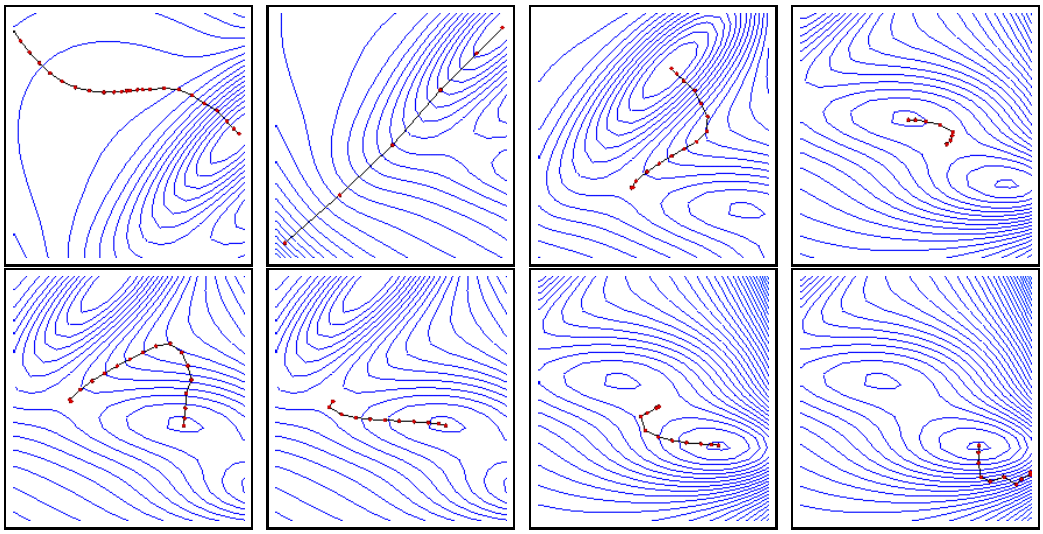


Figure 5.2: Trajectories for the eigenvector following on the Müller cost surface for trajectories started in different minima, along different principal curvature directions.

predict the new minimum by linear extrapolation from the previous two:

$$\mathbf{x} = \mathbf{x}_k + r \frac{\mathbf{x}_k - \mathbf{x}_{k-1}}{\|\mathbf{x}_k - \mathbf{x}_{k-1}\|} \tag{5.23}$$

where r is a trust region radius for $\boldsymbol{\delta}\mathbf{x}$, then solve for the corresponding t_{k+1} using the ellipsoid constraint.

Eigenvector tracker: We use the damped Newton saddle step (5.5), moving away from the minimum by reversing the sign of the gradient in the tracked eigendirection if this has positive curvature. The damping $\lambda > 0$ is controlled to keep the step within a trust radius r and to dominate any undesired negative eigenvalues. The trust radius is set by monitoring the accuracy (5.6) of the local model for f.

In some of our target applications, the underlying problem has bound constraints that must be maintained. For hypersurface sweeping this just adds additional constraints to the within-hypersurface minimizations. For eigenvector following, our trust region step routine uses a projection strategy to handle constraints on $\mathbf{x}$, and also projects the eigenvector-tracking direction $\mathbf{t}$ along the constraints to ensure stability.

5.3 Globalization

The above algorithms can be used to find the unknown first order saddle points surrounding a known minimum (*e.g.* found from a random initialization by local optimization using §4.2.1). From the detected saddles, we can slide downhill using local optimization to identify neighboring minima. It is therefore immediately clear how such an algorithm can be turned into a quasi-global optimizer.

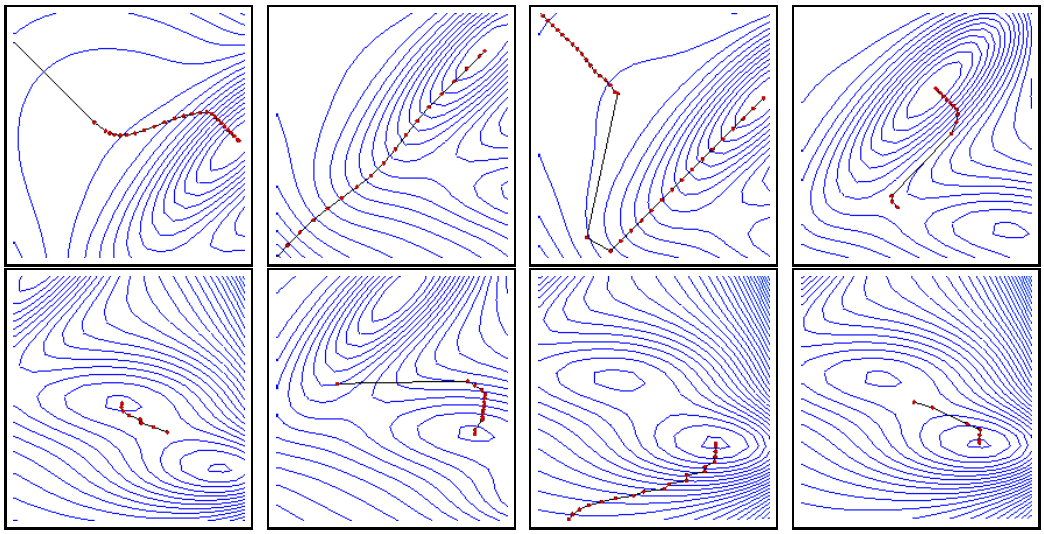


Figure 5.3: Trajectories for the hypersurface sweeping on Müller cost surface for trajectories started in different minima, along different principal curvature directions.

Start with one or perhaps several initial minima in a 'working queue', initialize saddle search trajectories from there, detect new saddles and their corresponding minima, and add any newly discovered minima to the 'working queue'. The algorithm can run until the working set becomes empty (all discovered minima have been processed) or until a given computational resource is exhausted (*e.g.* a number of function evaluations). The algorithm has very useful local to global search properties, because the search proceeds by expanding locally: the neighbors of a starting minimum are found, then their neighbors, *etc*. This behavior qualifies the method not only for static initialization problems, but also for tracking applications, where multi-modality in the cost surface often arises around the configurations propagated from the previous time step. It is important to use local search algorithms that efficiently locate multiple minima surrounding a set of given ones. Note that simply sampling randomly around such known configurations is not likely to be efficient computationally. Even neighboring minima are often separated by large distances in parameter space (see fig. 5.7, especially first two rows, also chapter 7 and the quantitative experiments therein) and covering such regions with a large radius noise sphere, and sampling this exhaustively is simply intractable. It is not only the volume of the space that is an issue here, but rather the small relative volumes of the minima's low-cost cores relative to their high-cost surrounding neighborhoods, given that the volume increases very rapidly with radius in high dimensions. Consequently the chances of samples hitting low cost region directly are very small. A learned dynamical model might be useful to help focus the samples near interesting low-cost regions, but sufficiently precise dynamical predictions are rarely available when tracking *general motion*. Also, for static multi-modal search problems,

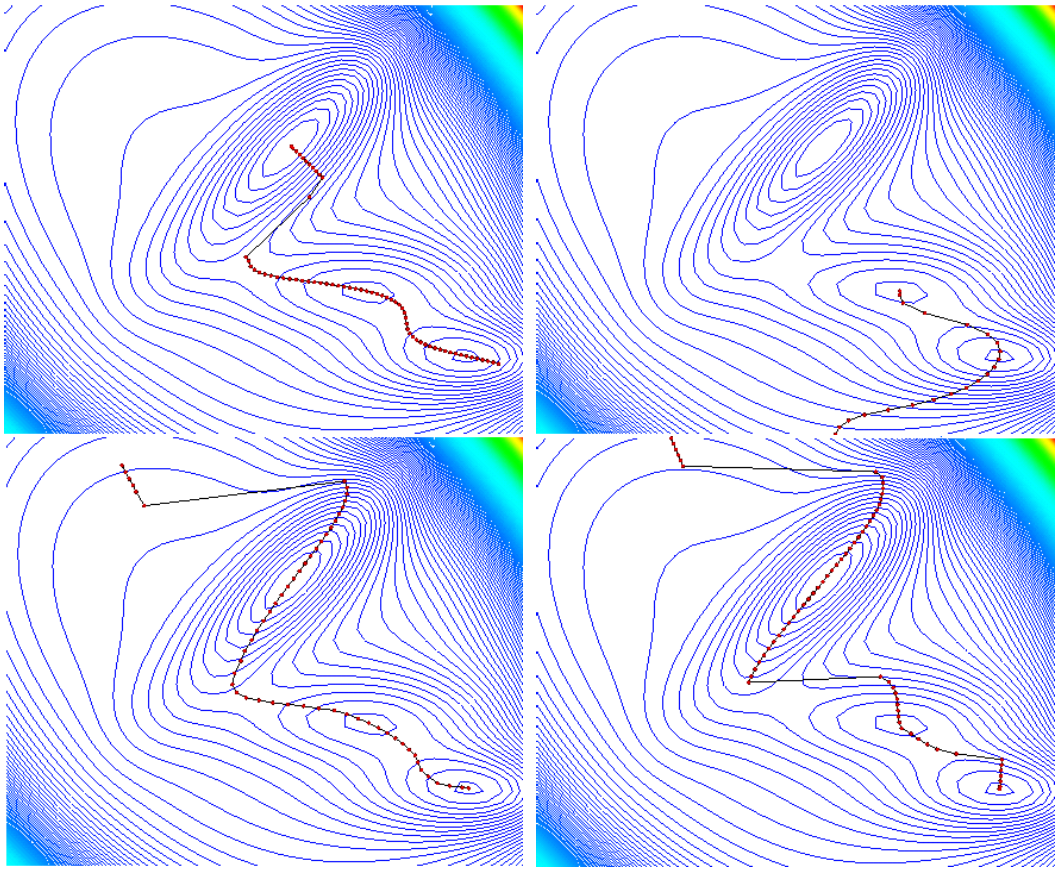


Figure 5.4: Trajectories for the hypersurface sweeping on Müller cost surface for trajectories started in different minima, but not stopped after the first saddle detection. Notice that sometimes multiple saddles and minima are found.

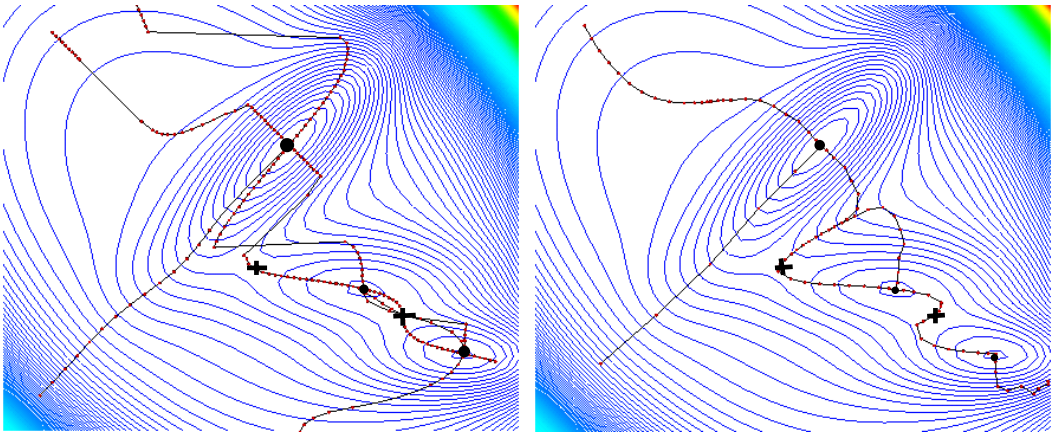


Figure 5.5: Trajectories for the hyper-ellipsoid sweeping (left) and eigenvector following (right) algorithms on the Müller cost surface, initialized along the $\pm$ eigendirections of the 3 minima.

there is no obvious general learned dynamics that can easily jump between minima[6].

5.4 Human Domain Modeling

We briefly review here the model, image features and priors we use for the experiments. For details, see chapter 3.

Representation: The 3D body model used in the human pose and motion estimation experiments here consists of a kinematic 'skeleton' of articulated joints controlled by angular joint parameters, covered by a 'flesh' built from superquadric ellipsoids with additional global deformations. For the experiments here we estimate typically 35 joint parameters.

Observation Likelihood: Robust model-to-image matching cost metrics are evaluated for each predicted image feature, and the results are summed over all observations to produce the image contribution to the parameter space cost function. We use a robust combination of extracted-feature-based metrics and intensity-based ones such as optical flow and robustified normalized edge energy. We also give results for a simpler likelihood designed for model initialization, based on squared distances between reprojected model joints and their specified image positions.

Priors and Constraints: The model incorporates both hard constraints (for joint angle limits) and soft priors (penalties for anthropometric model proportions, collision avoidance between body parts, and stabilization of useful but hard-to-estimate model parameters). In the experiments below we use mainly joint angle limits and body part non-interpenetration constraints.

Estimation: We apply Bayes rule and maximize the total posterior probability to give locally MAP parameter estimates as in (3.2) in chapter 3. For temporal multiple-hypothesis tracking we use the propagation law given by (4.1) in chapter 4.

5.5 Experiments

We illustrate our transition state search algorithms on a 2 d.o.f. toy problem, and on 3D human pose and motion estimation from monocular images.

The Müller Potential: This simple analytical 2D cost function given in appendix B is often used to illustrate transition state methods in chemistry. Fig. 5.5 shows its 3 minima and 2 saddles (the black dots and crosses) and plots the trajectories of the two methods starting from each minimum, along different principal curvature directions. The hypersurface sweeping algorithm is run for extended trajectories through several saddles and minima (left plot). In this simple example, a single sweep started at the top left minimum successfully finds all of the other minima and transition states.

Articulated 3D Human Motion Estimation: There is a large literature on human motion tracking but relatively little work on the thorny issue of local minima in the difficult 3D-from-monocular

[6] On occasion, importance samplers based on learned static parameter space priors or low-level image cues may be available to improve sample focusing.

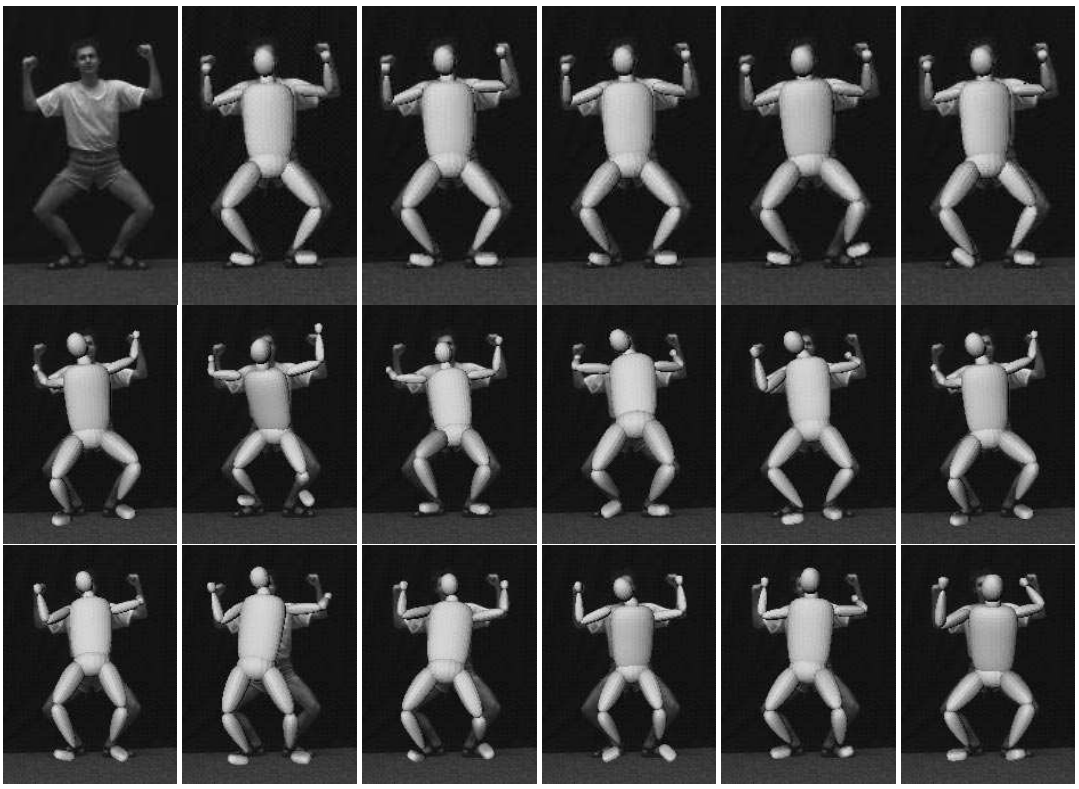


Figure 5.6: Minima of image-based cost functions. Top row: contour and optical flow likelihood. Middle and bottom rows: silhouette and edge likelihood. The model is initialized in one minimum (first row, second figure), and search trajectories for other minima (along different principal curvature directions) are initiated from there. The other figures show some other minima found, that correspond to incorrect contour assignments or to configurations where the intensity robustifiers turn off (see text).

case. Cham & Rehg combine local optimization and CONDENSATION sampling for 2D tracking (Cham and Rehg, 1999). Deutscher *et al* use an annealed sampling method and multiple cameras to widen the search and limit the presence of spurious minima (Deutscher et al., 2000). Sidenbladh *et al* use particle filtering with importance sampling based on a learned walking model to focus the search in the neighborhood of known trajectory pathways (Sidenbladh et al., 2000). Sminchisescu & Triggs combine robust constraint-consistent local continuous optimization with a covariance-scaled sampling method that focuses samples in high uncertain parameter space regions likely to have low cost, inside or beyond the current minimum ((Sminchisescu and Triggs, 2001b), see also chapter 4). All of these works note the difficulty of the multiple-minimum problem and attempt to develop techniques or constraints (on the scene, motion, number of cameras or background) to tackle it.

Here we show examples from a set of experiments with a 32 d.o.f. articulated full-body model, including pose estimation and tracking in monocular images using cost surfaces based on different

combinations of image cues. The figures show examples of minima found in likelihood models based on image contours and optical flow (fig. 5.6, top row), contours and silhouette-image data (fig. 5.6, middle and bottom rows), and model-to-image joint correspondences (fig. 5.8). In all cases, the model has been initialized in one of the minima (the algorithm in §4.2.1 has been run to convergence for the different cost surfaces mentioned above) and subsequently, saddle point search trajectories were initiated from that minimum, along different principal curvature directions. For each of the first-order saddles found, local optimization using the algorithm §4.2.1, has been performed in order to locate the neighboring minima. It is straightforward to see how the algorithm can be globalized: initialize the current model configuration in one minimum, perform saddle searches along different curvature directions, detect new minima and restart the search from there, until no new minima are found or a given computational limit is reached (see §5.3 on page 84 for a discussion).

Fig. 5.7 attempts to capture some more quantitative information about the methods, here for the joint correspondence cost function. The first row displays the parameter space and cost distances of the 56 minima found during a set of 64 constrained searches (the $\pm$ directions of the 32 eigenvectors of an initial minimum, distances being measured w.r.t. this minimum, in radians and meters for the parameter space). The second row again shows parameter space distances, but now measured in standard deviations and for saddles rather than minima, for the same frontal view and for a slightly more side-on one (fig. 5.8, respectively top and bottom). The plots reveal the structure of the cost surface, with nearby saddles at 4–8 standard deviations and progressively more remote ones at 20–50, 80–100 and 150–200 standard deviations. It follows that no multiple-minimum exploration algorithm can afford to search only within the 'natural' covariance scale of its current minima: significantly deeper sampling is needed to capture even nearby minima (as previously noted, *e.g.* by (Sminchisescu and Triggs, 2001a,b)).

The last two rows of fig. 5.7 show some sample cost profiles for typical runs of the eigenvector following (row 3) and constrained hyper-surface (row 4) saddle search methods. In the eigenvector method, it is preferable to represent the joint limits using a 'hard' active set strategy (row 3 right) rather than soft constraints (row 3 left): the stiff 'cost walls' induced by the soft constraints tend to force the eigenvector follower into head-on collision with the wall, with the cost climbing rapidly to infinity. The active set strategy avoids this problem at the price of more frequent direction changes as the joint limits switch on and off (row 3 right). The hyper-ellipsoid method (row 4) produces trajectories that do not require special joint limit processing, but its cost profiles have characteristic sawtooth edges (row 4 right) associated with sudden state readjustments on the hypersphere at points where the tracked minimum becomes locally unstable.[7]

[7] Note the different reasons for instability in the generated trajectories for the two methods: the eigenvector tracking deals with projecting the current trajectory step onto the constraint surface, while for the hypersurface sweeping algorithm, the instabilities have to do with the relative curvatures of the hypersurface ellipsoid with respect to the cost isosurface – the constraints being already part of the cost surface, they do not play a particular role in emergent instabilities there.

Fig. 5.6 shows minima for costs based on various combinations of image cues. In the first row the minima correspond to a small interframe motion, using contour and robust optical flow information. This case has relatively few, but closely spaced local minima owing to the smoothing/quadratic effect of the flow. (Remoter minima do still exist at points where the robust contributions of sets of flow measurements turn off, particularly when these coincide with incorrect edge assignments). The second and third rows show minima arising from a silhouette and edge based cost function. The minima shown include 'reflective' (depth-related) ambiguities, incorrect edge assignments and singular 'inside-silhouette' configurations (which can be alleviated to some extent by augmenting the likelihood term as in (Sminchisescu, 2002)).

Finally, fig. 5.8 shows depth ambiguities for articulated 3D frontal and side poses, under model to image joint correspondences. The arm-shoulder complex is very flexible and therefore tends to induce more minima than the legs. We also find that side views tend to generate fewer minima than frontal ones, perhaps due to presence of body-part non-self-intersection and joint constraints that render many 'purely reflective' minima infeasible.

5.6 Conclusions and Open Research Directions

In this chapter, we have described two families of deterministic-optimization-based algorithms for finding 'transition states' (saddle points with 1 negative eigenvalue) in high-dimensional multimodal cost surfaces. These allow us to build topological 'roadmaps' of the nearby local minima and the transition states that lead to them. The methods are based on ones developed in computational chemistry, but here generalized, clarified and adapted for use in computational vision. Experiments on the difficult problem of articulated 3D human pose from monocular images show that our algorithms can stably and efficiently recover large numbers of transition states and minima, but also serve to underline the very large numbers of minima that exist in this problem.

The methods are potentially useful for other multimodal, non-linear problems in vision, including structure from motion or shape from shading. The methods can be also used to quantify the degrees of ambiguity of different cost functions. This could allow, longer term, to design better cost functions based on higher-level features and groupings.

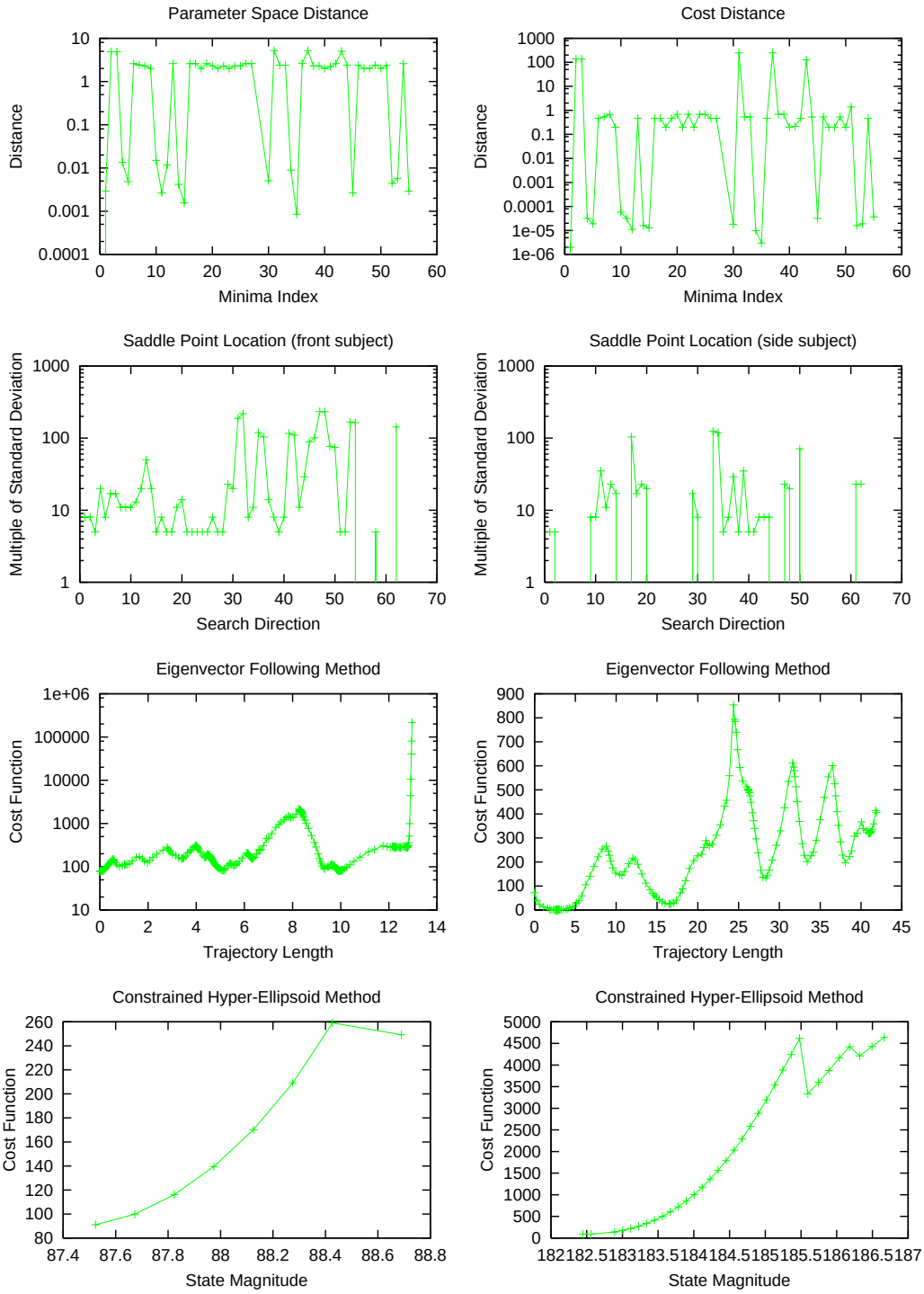


Figure 5.7: Transition state and minimum location algorithms for ± 32-eigendirection search trials. Top row: parameter space distance and cost difference between initial and current minimum. Second row: saddle point distances in standard deviations for a frontal and a partly side-on 3D pose. Third and fourth rows: cost profiles for different trajectories and constraints (see text).

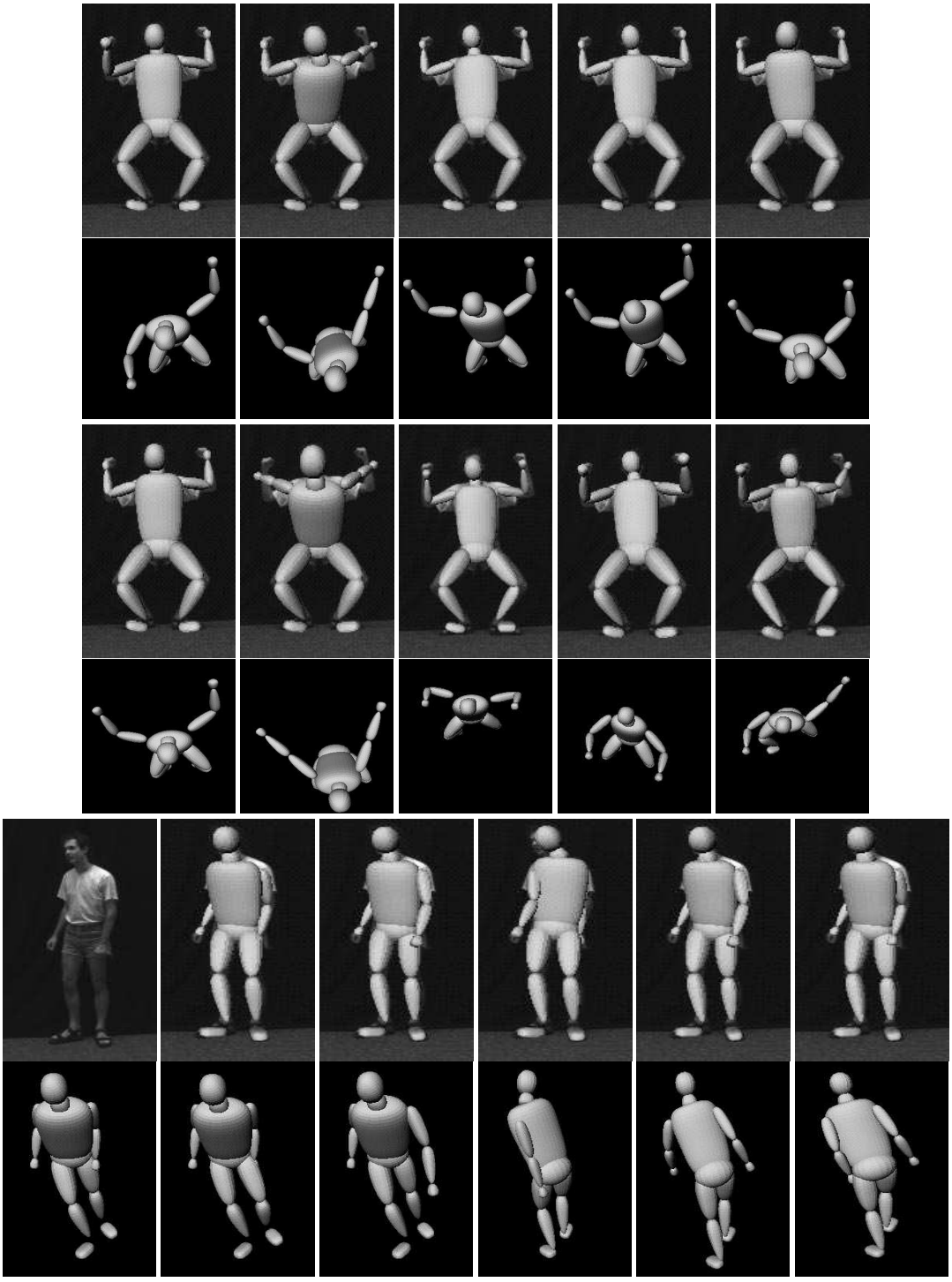


Figure 5.8: 'Reflective' kinematic ambiguities under the model/image joint correspondence cost function. The model is initialized in one minimum (second row, leftmost figure), and search trajectories for other minima (along different principal curvature directions) are initiated from there. The images show some of the new minima found. Each pair of rows displays the original image overlayed with the projected model, and the 3D model position seen from a fixed synthetic overhead camera. Note the pronounced forwards-backwards character of these reflective minima, and the large parameter space distances that often separate of them.

Chapter 6

Hyperdynamic Importance Sampling

Sequential random sampling (‘Markov Chain Monte-Carlo’) is a popular strategy for many vision problems involving multimodal distributions over high-dimensional parameter spaces. It applies both to *importance sampling* (where one wants to sample points according to their ‘importance’ for some calculation, but otherwise fairly) and to *global optimization* (where one wants to find good minima, or at least good starting points for local minimization, regardless of fairness). Unfortunately, most sequential samplers are very prone to becoming ‘trapped’ for long periods in unrepresentative local minima, which leads to biased or highly variable estimates. We present a general strategy for reducing MCMC trapping that generalizes Voter’s ‘hyperdynamic sampling’ from computational chemistry. The local gradient and curvature of the input distribution are used to construct an adaptive importance sampler that focuses samples on low cost negative curvature regions likely to contain ‘transition states’ — codimension-1 saddle points representing ‘mountain passes’ connecting adjacent cost basins. This substantially accelerates inter-basin transition rates while still preserving correct relative transition probabilities. Experimental tests on the difficult problem of 3D articulated human pose estimation from monocular images show significantly enhanced minimum exploration.

Keywords: Hyperdynamics, Markov-chain Monte Carlo, importance sampling, global optimization, human tracking.

6.1 Introduction

Many vision problems can be formulated either as global minimizations of highly non-convex cost functions with many minima, or as statistical inferences based on fair sampling or expectation-value integrals over highly multi-modal distributions. Importance sampling is a promising approach for such applications, particularly when combined with sequential (‘Markov Chain Monte-Carlo’), layered or annealed samplers (Forsyth et al., 2001; Choo and Fleet, 2001; Deutscher et al., 2000), optionally punctuated with bursts of local optimization (Heap and Hogg, 1998; Cham and Rehg, 1999;

Sminchisescu and Triggs, 2001b). Sampling methods are flexible, but they tend to be computationally expensive for a given level of accuracy. In particular, when used on multi-modal cost surfaces, current sequential samplers are very prone to becoming trapped for long periods in cost basins containing unrepresentative local minima. This 'trapping' or 'poor mixing' leads to biased or highly variable estimates whose character is at best quasi-local rather than global. Trapping times are typically exponential in a (large) scale parameter, so 'buying a faster computer' helps little. Current samplers are myopic mainly because they consider only the size of the integrand being evaluated or the lowness of the cost being optimized when judging 'importance'. *For efficient global estimates, it is also critically 'important' to include an effective strategy for reducing trapping,* e.g. *by explicitly devoting some fraction of the samples to moving between cost basins.*

This chapter describes a method for reducing trapping by 'boosting' the dynamics of the sequential sampler. Our approach is based on Voter's 'hyperdynamics' (Voter, 1997a,b), which was originally developed in computational chemistry to accelerate the estimation of transition rates between different atomic arrangements in atom-level simulations of molecules and solids. There, the dynamics is basically a thermally-driven random walk of a point in the configuration space of the combined atomic coordinates, subject to an effective energy potential that models the combined inter-atomic interactions. The configuration-space potential is often highly multimodal, corresponding to different large-scale configurations of the molecule being simulated. Trapping is a significant problem, especially as the fine-scale dynamics must use quite short time-steps to ensure accurate physical modeling. Mixing times of 10^6–10^9 or more steps are common. In our target applications in vision the sampler need not satisfy such strict physical constraints, but trapping remains a key problem.

Hyperdynamics reduces trapping by boosting the number of samples that fall near 'transition states' — low lying saddle points that the system would typically pass through if it were moving thermally between adjacent energy basins. It does this by modifying the cost function, adding a term based on the gradient and curvature of the original potential that raises the cost near the cores of the local potential basins to reduce trapping there, while leaving the cost intact in regions where the original potential has the negative curvature eigenvalue and low gradient characteristic of transition neighborhoods. Hyperdynamics can be viewed as a generalized form of MCMC importance sampling whose importance measure considers the gradient and curvature as well as the values of the original cost function. The key point is not the specific form adopted for the potential, but rather the refined notion of 'importance': deliberately adding samples to speed mixing and hence reduce global bias ('finite sample effects'), even though the added samples are not directly 'important' for the calculation being performed.

Another general approach to multi-modal optimization is *annealing* — initially sampling with a reduced sensitivity to the underlying cost ('higher temperature'), then progressively increasing the sensitivity to focus samples on lower cost regions. Annealing has been used many times in

vision and elsewhere[1], *e.g.* (Neal, 1998, 2001; Deutscher et al., 2000), but although it works well in many applications, it has important limitations as a general method for reducing trapping. The main problem is that it samples indiscriminately within a certain energy band, regardless of whether the points sampled are likely to lead out of the basin towards another minimum, or whether they simply lead further up an ever-increasing potential wall. In many applications, and especially in high-dimensional or ill-conditioned ones, the cost surface has relatively narrow 'corridors' connecting adjacent basins, and it is important to steer the samples towards these using local information about how the cost appears to be changing. Hyperdynamics is a first attempt at doing this. In fact, these methods are complementary: it may be possible to speed up hyperdynamics by annealing its modified potential, but we will not investigate this here.

6.1.1 What is a Good Multiple-Mode Sampling Function ?

'The curse of dimensionality' causes many difficulties in high-dimensional search. In stochastic methods, long sampling runs are often needed to hit the distribution's 'typical set' — the areas where most of the probability mass is concentrated. In sequential samplers this is due to the inherently local nature of the sampling process, which tends to become 'trapped' in individual modes, moving between them only very infrequently. More generally, choosing an importance sampling distribution is a compromise between tractable sampleability and efficient focusing of the sampling resources towards 'good places to look'.

There are at least three issues in the design of a good multi-modal sampler: (*i*) *Approximation accuracy*: in high dimensions, when the original distribution is complex and highly multi-modal (as is the case in vision), finding a good approximating function can be very difficult, thus limiting the applicability of the method. It is therefore appealing to look for ways of using a modified version of the original distribution, as for instance in annealing methods (Neal, 1998, 2001; Deutscher et al., 2000). (*ii*)*Trapping*: even when the approximation is locally accurate (*e.g.* by sampling the original distribution, thus avoiding any sample-weighting artifacts), most sampling procedures tend to get caught in the mode(s) closest to the starting point of sampling. Very long runs are needed to sample infrequent inter-mode transition events that lie far out in the tails of the modal distributions, but that can make a huge difference to the overall results. (*iii*)*Biased transition rates*: annealing changes not only the absolute inter-mode transition rates (thus reducing trapping), but also their relative sizes (Sorensen and Voter, 2000). So there is no guarantee that the modes are visited with the correct relative probabilities implied by the dynamics on the original cost surface. This may seem irrelevant if the aim is simply to discover 'all good modes' or 'the best mode', but the levels of annealing needed to make difficult transitions frequent can very significantly increase the number of modes and the state space volume that are available to be visited, and thus cause the vast bulk

[1] Raising the temperature is often unacceptable in chemistry applications of hyperdynamics, as it may significantly change the problem. *E.g.*, the solid being simulated might melt...

of the samples to be wasted in fruitless regions[2]. This is especially important in applications like tracking, where only the neighboring modes that are separated from the current one by the lowest energy barriers need to be recovered.

To summarize, for complex high dimensional problems, finding good, sampleable approximating distributions is hard, so it is useful to look at sequential samplers based on distributions derived from the original one. There is a trade-off between sampling for local computational accuracy, which requires samples in 'important' regions, usually mode cores, and sampling for good mixing, which requires not only more frequent samples in the tails of the distribution, but also that these should be focused on regions likely to lead to inter-modal transitions. Defining such regions is delicate in practice, but it is clear that steering samples towards regions with low gradient and negative curvatures should increase the likelihood of finding transition states (saddle points with one negative curvature direction) relative to purely cost-based methods such as annealing.

6.1.2 Related Work

Now we briefly summarize some relevant work on high-dimensional search, especially in the domain of human modeling and estimation. Cham & Rehg perform 2D tracking with scaled prismatic models. Their method combines a least squares intensity-based cost function, particle filtering with dynamical noise style sampling, and local optimization of a mixture of Gaussians state probability representation (Cham and Rehg, 1999). Deutscher *et al* track 3D body motion using a multi-camera silhouette-and-edge based likelihood function and annealed sampling within a temporal particle filtering framework (Deutscher et al., 2000). Their sampling procedure resembles one used by Neal, but Neal also includes an additional importance sampling correction designed to improve mixing (Neal, 1998, 2001). Sidenbladh *et al* use an intensity based cost function and particle filtering with importance sampling based on a learned dynamical model to track a 3D model of a walking person in an image sequence (Sidenbladh et al., 2000). Choo & Fleet combine particle filtering and hybrid Monte Carlo sampling to estimate 3D human motion, using a cost function based on joint re-projection error given input from motion capture data (Choo and Fleet, 2001). Sminchisescu & Triggs recover articulated 3D motion from monocular image sequences using an edge and intensity based cost function, with a combination of robust constraint-consistent local optimization and 'oversized' covariance scaled sampling to focus samples on probable low-cost regions (Sminchisescu and Triggs, 2001b).

Hyperdynamics uses stochastic dynamics with cost gradient based sampling as in (Forsyth et al., 2001; Neal, 1993; Choo and Fleet, 2001), but 'boosts' the dynamics with a novel importance sampler constructed from the original probability surface using local gradient and curvature

[2]There is an analogy with the chemist's melting solid, liquids being regions of state space with huge numbers of small interconnected minima and saddles, while solids have fewer, or at least more clearly defined, minima. Also remember that state space volume increases very rapidly with sampling radius in high dimensions, so dense, distant sampling is simply infeasible.

information. All of the annealing methods try to increase transition rates by sampling a modified distribution, but only the one given here specifically focuses samples on regions likely to contain transition states. There are also deterministic local-optimization-based methods designed to find transition states and they are presented and studied in chapter 5.

6.2 Sampling and Transition State Theory

6.2.1 Importance Sampling

Importance sampling works as follows. Suppose that we are interested in quantities depending on the distribution of some quantity $\mathbf{x}$, whose probability density is proportional to $f(\mathbf{x})$. Suppose that it is feasible to evaluate $f(\mathbf{x})$ pointwise, but that we are not able to sample directly from the distribution it defines, but only from an approximating distribution with density $f_b(\mathbf{x})$. We will base our estimates on a sample of N independent points, $\mathbf{x}_1, ..., \mathbf{x}_N$ drawn from $f_b(\mathbf{x})$. The expectation value of some quantity $V(\mathbf{x})$ with respect to $f(\mathbf{x})$ can then be estimated as $\bar{V} = \sum_{i=1}^{N} w_i V(\mathbf{x}_i) / \sum_{i=1}^{N} w_i$, where the **importance weighting** of $\mathbf{x}_i$ is $w_i = f(\mathbf{x}_i)/f_b(\mathbf{x}_i)$ (this assumes that $f_b(\mathbf{x}) \neq 0$ whenever $f(\mathbf{x}) \neq 0$). It can be proved that the importance sampled estimator converges to the mean value of V as N increases, but it is difficult to assess how reliable the estimate $\bar{V}$ is in practice. Two issues affect this accuracy: the variability of the importance weights due to deviations between $f(\mathbf{x})$ and $f_b(\mathbf{x})$, and statistical fluctuations caused by the improbability of sampling infrequent events in the tails of the distribution, especially if these are critical for estimating $\bar{V}$.

6.2.2 Stochastic Dynamics

Various methods are available for speeding up sampling. Here we use a stochastic dynamics method on the potential surface defined by our cost function (the negative log-likelihood of the state probability given the observations, $f(\mathbf{x}) = -\log p(\mathbf{x}|\cdot)$). Canonical samples from $f(\mathbf{x})$ can be obtained by simulating the phase space dynamics defined by the Hamiltonian function:

$$H(\mathbf{x}, \mathbf{p}) = f(\mathbf{x}) + K(\mathbf{p}) \tag{6.1}$$

where $K(\mathbf{p}) = \mathbf{p}^\top \mathbf{p}/2$ is the kinetic energy, and $\mathbf{p}$ is the momentum variable. Averages of variables V over the canonical ensemble can be computed by using classical 2N-dimensional phase-space integrals:

$$\langle V \rangle \;=\; \frac{\int\int V(\mathbf{x},\mathbf{p}) e^{-\alpha f(\mathbf{x})} e^{-\alpha K(\mathbf{p})} \mathbf{dx dp}}{\int\int e^{-\alpha f(\mathbf{x})} e^{-\alpha K(\mathbf{p})} \mathbf{dx dp}} \tag{6.2}$$

where $\alpha = 1/T$ is the temperature constant. Dynamics (and hence sampling) is done by locally integrating the Hamilton equations:

$$\frac{\mathbf{dx}}{dt} \;=\; \mathbf{p} \qquad \text{and} \qquad \frac{\mathbf{dp}}{dt} \;=\; -\frac{df(\mathbf{x})}{\mathbf{dx}} \tag{6.3}$$

using a Langevin Monte Carlo type integration/rejection scheme that is guaranteed to perform sampling from the canonical distribution over phase-space:

$$\mathbf{x}_{i+1} \;=\; \mathbf{x}_i - \frac{\Delta t_{sd}^2}{2}\,\frac{df(\mathbf{x})}{\mathbf{dx}} + \Delta t_{sd}\mathbf{n}_i \tag{6.4}$$

where $\mathbf{n}_i$ is a vector of independently chosen Gaussian variables with zero mean and unit variance, and Δt_{sd} is the stochastic dynamics integration step. Compared to so called 'hybrid' methods, the Langevin method can be used with a larger step size and this is advantageous for our problem, where the step calculations are relatively expensive (see (Neal, 1993) and its references for a more complete discussion of the relative advantages of hybrid and Langevin Monte Carlo methods)[3]. For physical dynamics t represents the physical time, while for statistical calculations it simply represents the number of steps performed since the start of the simulation. The simulation time is used in §6.3 below to estimate the acceleration of infrequent events produced by the proposed biased potential.

6.2.3 Transition State Theory

Continuing the statistical mechanics analogy begun in the previous section, the behavior of the physical system can be characterized by long periods of 'vibration' within one 'state' (energy basin), followed by infrequent transitions to other states via saddle points. In the 'transition state theory' (TST) approximation, the transition rates between states are computed using the sample flux through the *dividing surface* separating them. For a given state S, this is the $N-1$ dimensional surface separating the state S from its neighbors. The rate of escape from state S is:

$$k_{S\rightarrow}^{tst} \;=\; \langle\, |\nu_S|\, \delta_S(\mathbf{x}) \,\rangle_S \tag{6.5}$$

where $\delta_s(\mathbf{x})$ is a Dirac delta function positioned on the dividing surface of S and ν_s is the velocity normal to this surface. Crossings of the dividing surface correspond to true state change events, and we assume that the system loses all memory of this transition before the next event.

6.3 Accelerating Transition State Sampling

In this section we explain how to transform the original potential in order to accelerate the transitions between minima. This means that more minima are found when sampling the modified potential (for the same number of samples and the same level of accuracy), than the original one.

[3]Note that the momenta are only represented implicitly in the Langevin formulation: there is no need to update their values after each leapfrog step as they are immediately replaced by new ones drawn from the canonical distribution at the start of each iteration. If approximate cost Hessian information is also available, the gradient in (6.4) can be projected onto the Hessian eigen-basis and its components weighted by the local eigen-curvatures to give an effective 'Newton-like' step. We use such steps near saddle points, where the hyperdynamic bias potential is essentially zero, to avoid the inefficiencies of random walk behavior there.

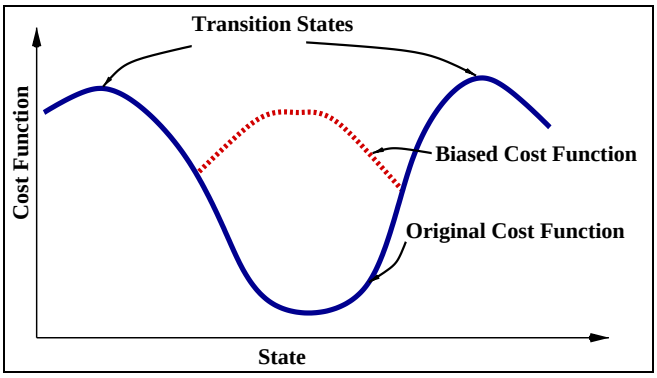


Figure 6.1: The original cost function and the bias added for hyperdynamics. The bias prevents long-trapping in the minimum by raising the cost there, but leaves it unchanged in the transition neighborhoods.

We describe the general properties that such a transformation should obey and we subsequently derive the quantitative results for the expected inter-minimum transition acceleration. A precise functional form that approximates the required transformation properties is then given in §6.4.

According to the transition state theory formalism presented in the previous section, the TST rate can be evaluated as follows, using (6.5) and (6.2):

$$k^{tst}_{S\rightarrow} \;=\; \frac{\iint\, |\nu_S|\, \delta_S(\mathbf{x})\, e^{-\alpha f(\mathbf{x})}\, e^{-\alpha K(\mathbf{p})}\, \mathbf{dx}\, \mathbf{dp}}{\iint\, e^{-\alpha f(\mathbf{x})}\, e^{-\alpha K(\mathbf{p})}\, \mathbf{dx}\, \mathbf{dp}} \tag{6.6}$$

Now consider adding a positive bias or boost cost $f_b(\mathbf{x})$ (with a corresponding 'biased' state S_b) to the original cost $f(\mathbf{x})$, with the further property that $f_b(\mathbf{x}) = 0$ whenever $\delta_S(\mathbf{x}) \neq 0$, *i.e.* the potential is unchanged in the transition state regions. The TST rate becomes:

$$k^{tst}_{S\rightarrow} \;=\; \frac{\iint\, |\nu_S|\, \delta_S(\mathbf{x})\, e^{-\alpha[f(\mathbf{x})+f_b(x)]}\, e^{\alpha f_b(\mathbf{x})}\, e^{-\alpha K(\mathbf{p})}\, \mathbf{dx}\, \mathbf{dp}}{\iint\, e^{-\alpha f(\mathbf{x})}\, e^{-\alpha K(\mathbf{p})}\, \mathbf{dx}\, \mathbf{dp}} \tag{6.7}$$

$$= \;\frac{\left\langle\, |\nu_S|\, \delta_S(\mathbf{x})\, e^{\alpha f_b(\mathbf{x})}\,\right\rangle_{S_b}}{\left\langle\, e^{\alpha f_b(\mathbf{x})}\,\right\rangle_{S_b}} \;=\; \frac{\left\langle\, |\nu_S|\, \delta_S(\mathbf{x})\,\right\rangle_{S_b}}{\left\langle\, e^{\alpha f_b(\mathbf{x})}\,\right\rangle_{S_b}} \tag{6.8}$$

The boost term increases every escape rate from state S as the cost well is made shallower, but it leaves the *ratios* of escape rates from S, S_b to other states S_1, S_2 invariant:

$$\frac{k^{tst}_{S\rightarrow S_1}}{k^{tst}_{S\rightarrow S_2}} \;=\; \frac{k^{tst}_{S_b\rightarrow S_1}}{k^{tst}_{S_b\rightarrow S_2}} \tag{6.9}$$

This holds because all escape rates from S all have the partition function of S as denominator, and replacing this with the partition function of S_b leaves their ratios unchanged. Concretely, suppose that during N_t steps of classical dynamics simulation on the biased cost surface, we encounter N_e escape attempts over the dividing surface. For the computation, let us also assume that the simulation is artificially confined to the basin of state S by reflecting boundaries. (This does not

happen in real simulations: it is used here only to estimate the 'biased boost time'). The TST escape rate from state S can be estimated simply as the ratio of the number of escape attempts to the total trajectory length: $k_S^{tst} = N_e/(N_t \Delta t_{sd})$. Consequently, the mean escape time (inverse transition rate) from state S can be estimated from (6.7) as:

$$\tau_{esc}^S = \frac{1}{k_{S\rightarrow}^{tst}} = \frac{\langle\, e^{\alpha f_b(\mathbf{x})} \,\rangle_{S_b}}{\langle\, |\nu_S|\, \delta_S(\mathbf{x}) \,\rangle_{S_b}} = \frac{\frac{1}{N_t}\sum_{i=1}^{N_t} e^{\alpha f_b(\mathbf{x}_i)}}{N_e/(N_t\, \Delta t_{sd})} = \frac{1}{N_e}\sum_{i=1}^{N_t} \Delta t_{sd}\, e^{\alpha f_b(\mathbf{x}_i)} \tag{6.10}$$

The effective simulation time boost achieved in step i thus becomes simply:

$$\Delta t_{b_i} = \Delta t_{sd} e^{\alpha f_b(\mathbf{x}_i)} \tag{6.11}$$

The dynamical evolution of the system from state to state is still correct, but it works in a distorted time scale that depends exponentially on the bias potential. As the system passes through regions with high f_b, its equivalent time Δt_b increases rapidly as it would originally have tended to linger in these regions (or more precisely to return to them often on the average) owing to their low original cost. Conversely, in zones with small f_b the equivalent time progress at the standard stochastic dynamics rate. Of course, in reality the simulation's integration time step and hence its sampling coarseness are the same as they were in the unboosted simulation. The boosting time (6.11) just gives an intuition for how much time an unaccelerated sampler would probably have wasted making 'uninteresting' samples near the cost minimum. But that is largely the point: the wastage factors are astronomical in practice — unboosted samplers can not escape from local minima.

Depending on the application, the modified potential can be used in different ways. One possibility is as an importance sampler. In this case correction weighting can be readily performed, but it is likely that the samples in the mode cores will have low weights, as the 'bias' was purposively designed to avoid long trapping there and to focus more samples in the transition regions. Asymptotically, the importance sampling will be correct anyway, but in practical situations, some improvements are immediately possible. (*i*) If fair samples of the original distribution are of interest, the 'biased' sampler could provide initial seeds for a subsequent classical stochastic dynamics or random sampling stage that operates on the original cost. The advantage of using such seeds is that substantially more modes are discovered, as previously proved in this section. (*ii*) If the precise localization of the modes is important as for instance in global optimization, the 'biased' sampler provides seeds for subsequent local optimization on the original cost. Whichever of the two behaviors above is desired, sampling the modified potential offers a fast and principled way to improve the multimodal exploration of complex high-dimensional distributions – the only fundamental difficulty encountered when either sampling them or seeking their multiple peaks.

6.4 The Biased Cost

The main requirements on the bias potential are that it should be zero on all dividing surfaces, that it should not introduce new sub-wells with escape times comparable to the main escape time from

the original cost well, and that its definition should not require prior knowledge of the cost wells or saddle points (if we knew these we could avoid trapping much more efficiently by including explicit well-jumping samples). For sampling, the most 'important' regions of the cost surface are minima, where the Hessian matrix $\mathbf{H}$ has strictly positive eigenvalues, and transition states, where it has exactly one negative eigenvalue $e_1 < 0$. The gradient vector vanishes in both cases. The rigorous definition of the TST boundary is necessarily global[4], but locally near a transition state the boundary contains the state itself and adjacent points where the Hessian has a negative eigenvalue and vanishing gradient component along the corresponding eigenvector:

$$g_{p1} \;=\; \mathbf{V}_1^{\top}\mathbf{g} \;=\; 0 \qquad \text{and} \qquad e_1 < 0 \tag{6.12}$$

where $\mathbf{g}$ is the gradient vector and $\mathbf{V}_1$ is the first Hessian eigenvector. Voter (Voter, 1997a,b) therefore advocates the following bias cost for hyperdynamics:

$$f_b \;=\; \frac{h_b}{2}\left[\,1 + \frac{e_1}{\sqrt{e_1^2 + g_{p1}^2/d^2}}\,\right] \tag{6.13}$$

where h_b is a constant controlling the strength of the bias and d is a length scale (*e.g.* an estimate of the typical nearest-neighbor distance between minima, if this is available). In the neighborhood of any first-order saddle point, (6.12) gives a very good approximation to the true TST dividing surface. Far away from the saddle, the approximation may not hold. For instance, the equation (6.12) may be satisfied in regions internal to a minimum that do not represent state boundaries or for some parts of the TST surface, the Hessian may have no negative eigenvalues and the gradient may not be zero along the lowest Hessian eigenvector (but some higher one). However, the low cost regions of the dividing surface are the most likely inter-minimum transition neighborhoods, so if we can define a bias potential that is zero in the most important parts of the dividing surface (*i.e.* near saddle points), we then have a useful approximation.

Increasing h_b increases the bias and hence the nominal boosting. In principle it is even permissible to raise the cost of a minimum above the level of its surrounding transition states. However, there is a risk that doing so will entirely block the sampling pathways through and around the minimum, thus causing the system to become trapped in a newly created well at one end of the old one. Hence, it is usually safer to select a more moderate boosting. Note, however, that independent of the choice of h_b, the bias potential has the desirable property that automatically decreases (or vanishes) in regions that have the structure of transition neighborhoods.

One difficulty with Voter's potential (6.13) is that direct differentiation of it for gradient-based dynamics requires third order derivatives of $f(\mathbf{x})$. However an inexpensive numerical estimation method based on first order derivatives was proposed in (Voter, 1997b). For completeness we

[4]The basin of state S can be defined as the set of configurations from which gradient descent minimization leads to the minimum S. This basin is surrounded by an $(n-1)$-D hypersurface, outside of which local descent leads to states other than S.

summarize this in the next section. These calculations are more complex than those needed for standard gradient based stochastic simulation, but we will see that the bias provides a degree of acceleration that often pays-off in practice.

6.5 Estimating the Gradient of the Bias Potential

Many efficient sampling or optimization methods require the gradient of the cost to be sampled or optimized. Direct calculation of the gradient of Voter's potential (6.13) requires third order derivatives of $f(\mathbf{x})$, but an inexpensive numerical estimation method based on first order derivatives was proposed in (Voter, 1997b). An eigenvalue can be computed by numerical approximation along it's corresponding eigenvector direction $\mathbf{s}$:

$$e(\mathbf{s}) \;=\; [f(\mathbf{x}+\eta\mathbf{s}) + f(\mathbf{x}-\eta\mathbf{s}) - 2f(\mathbf{x})]/\eta^2 \tag{6.14}$$

The eigenvector direction can be estimated numerically using any gradient descent method, based on a random initialization $\mathbf{s}$ or on the one from the previous dynamics step, using:

$$\frac{de}{d\mathbf{s}} \;=\; [\mathbf{g}(\mathbf{x}+\eta\mathbf{s}) - \mathbf{g}(\mathbf{x}-\eta\mathbf{s})]/\eta \tag{6.15}$$

The lowest eigenvector obtained from the minimization (6.15) is then used to compute the corresponding eigenvalue via (6.14). The procedure can be repeated for higher eigenvalue-eigenvector pairs by maintaining orthogonality with previous directions. The derivative of the projected gradient g_{1p} can then be obtained by applying the minimization to the matrices $\mathbf{H} + \lambda\,\mathbf{g}\,\mathbf{g}^\top$ and $\mathbf{H} - \lambda\,\mathbf{g}\,\mathbf{g}^\top$. One thus minimizes:

$$\frac{de_i}{\mathbf{dx}} \;=\; \left\{[\mathbf{g}(\mathbf{x}+\eta\mathbf{s}) + \mathbf{g}(\mathbf{x}-\eta\mathbf{s}) - 2\mathbf{g}(\mathbf{x})]/\eta^2\right\}_{\mathbf{s}=\mathbf{s}_i} \tag{6.16}$$

where:

$$e_{\pm\lambda} \;=\; e(\mathbf{s}) \pm \lambda\Big[\frac{f(\mathbf{x}+\eta\mathbf{s}) - f(\mathbf{x}-\eta\mathbf{s})}{2\eta}\Big]^2 \tag{6.17}$$

A good approximation to g_{p1} can be obtained from (Voter, 1997b):

$$g_{p1} \;=\; \frac{1}{2\lambda}(e_{+\lambda} - e_{-\lambda}), \qquad \text{and} \qquad \frac{dg_{p1}}{\mathbf{dx}} \;=\; \frac{1}{2\lambda}\left(\frac{de_{+\lambda}}{\mathbf{dx}} - \frac{de_{-\lambda}}{\mathbf{dx}}\right) \tag{6.18}$$

6.6 Human Domain Modeling

This section briefly describes the exact humanoid visual tracking models and features used in the hyperdynamic boosting experiments. For more details see chapter 3.

Representation: The body model used for experiments contains kinematic skeletons of articulated joints controlled by angular joint parameters, and surface coverage built from superquadric ellipsoids with additional global deformations. The typical models used here have about 30-35 joint parameters that we estimate.

Estimation: We aim for a probabilistic interpretation and optimal estimates of the model parameters by maximizing the total probability according to Bayes rule as in (3.2) in chapter 3. Discretizing the continuous problem, we build the potential energy function $f(\mathbf{x})$ in (6.1) as the negative log-likelihood for the total posterior probability as in (3.3) in chapter 3.

Observation Likelihood: In the below experiments we actually only used a very simple Gaussian likelihood based on given model-to-image joint correspondences. The negative log-likelihood for the observations is just the sum of squared model joint re-projection errors. Our full tracking system uses this cost function only for initialization, but it still provides an interesting (and difficult to handle) degree of multimodality owing to the kinematic complexity of the human model and the large number of parameters that are unobservable in a singular monocular image. In practice we find that globalizing the search is at least as important for initialization as for tracking, and this cost function is significantly cheaper to evaluate than our full image based one, allowing more extensive sampling experiments.

Priors and Constraints: Both hard and soft priors are used for the experiments. We use anthropometric priors on model proportions, parameter stabilizers for hard to estimate but useful modeling parameters, terms for collision avoidance between body parts, and joint angle limits.

6.7 Experiments and Results

In this section we illustrate the hyperdynamics method on a toy problem involving a two-dimensional multi-modal cost surface, and on the problem of initial pose estimation for an articulated 3D human model based on given joint-to-image correspondences. In both cases we compare the method with standard stochastic dynamics on the original cost surface. The parameters of the two methods (temperature, integration step, number of simulation steps, *etc.*) are identical, except that hyperdynamics requires values for the two additional parameters h_b and d that control the properties of the bias potential (6.13).

6.7.1 The Müller Cost Surface

Müller's Potential, given in appendix B, is a simple 2D analytic cost function with three local minima, and two saddle points.

Fig. 6.2 (right) shows the result of standard stochastic dynamic sampling on the original cost surface. Despite 6000 simulation steps at a reasonable step size $\Delta t_{sd} = 0.01$, only the basin of the starting minimum is sampled extensively, and no successful escape has yet taken place. Fig. 6.3 shows two hyperdynamics runs with parameters set for moderate boosting. Note the reduced emphasis on sampling in the core of the minimum — in fact the minimum is replaced by a set of higher energy ones — and the fact that the runs escape the initial basin. In the right hand plot there is a clear focusing of samples in the region corresponding to the saddle point linking the two adjacent minima M_1 and M_2. Finally, fig. 6.4 shows results for more aggressive bias potentials that cause

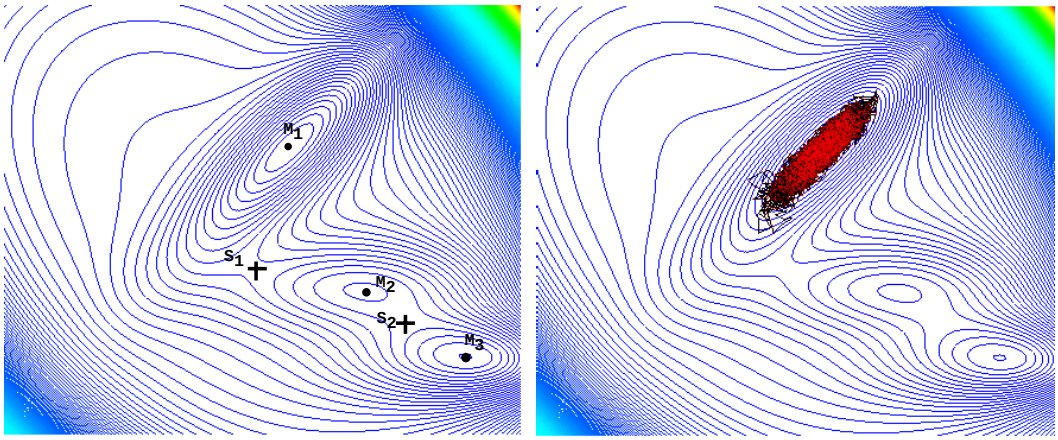


Figure 6.2: The Müller Potential (left) and a standard stochastic dynamics gradient sampling simulation (right) that gets trapped in the basin of the starting minimum.

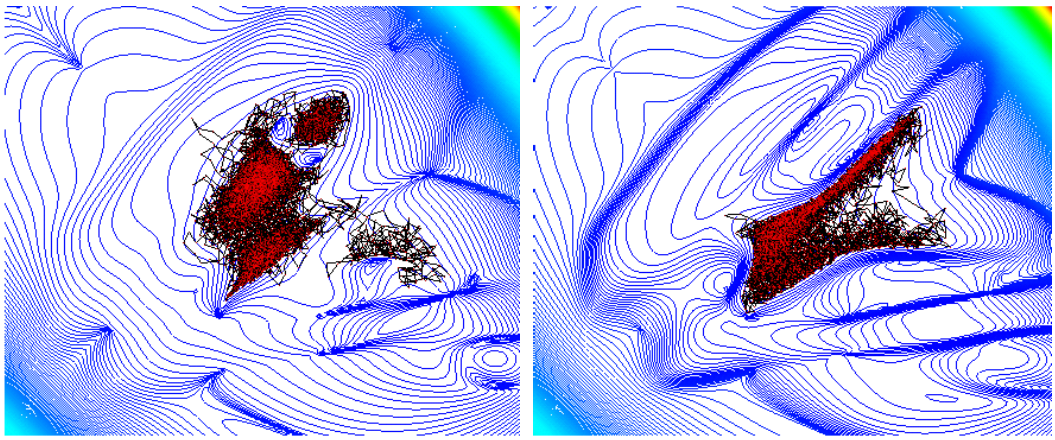


Figure 6.3: Hyperdynamic sampling with $h_b = 150, d = 0.1$ and $h_b = 200, d = 0.5$.

the basins of all three minima to be visited, with strong focusing of samples on the inter-minimum transition regions. The bias here turns the lowest positive curvature region of the initial minimum into a local maximum.

The plots also show that the Voter potential is somewhat 'untidy', with complicated local steps and ridges. Near the hypersurfaces where the first Hessian eigenvalue e_1 passes down through zero, the bias jumps from h_b to 0 with an abruptness that increases as the length scale d increases (sic) or the gradient projection g_{p1} decreases, owing to the $e_1/\sqrt{e_1^2 + g_{p1}^2/d^2}$ term in (6.13). A small d makes these $e_1 = 0$ transitions smoother, but increases the suddenness of ridges in the potential that occur on hypersurfaces where g_{1p} passes through zero.

Fig. 6.5 plots the simulation boosting time for two bias potentials. The left plot has a milder potential that simply encourages exploration of saddle points, while the right plot has a more aggressive one that is able to explore and jump between individual modes more rapidly. (Note the

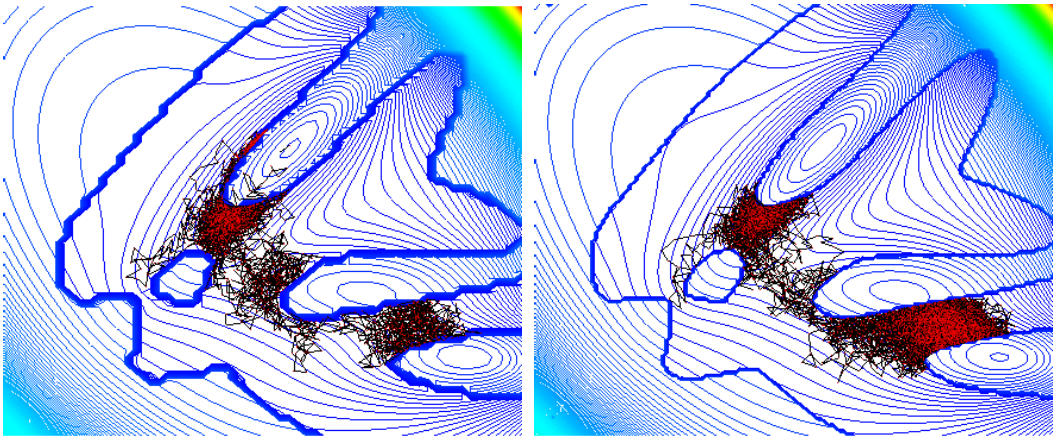


Figure 6.4: Hyperdynamic sampling with $h_b = 300, d = 10$ and $h_b = 400, d = 100$.

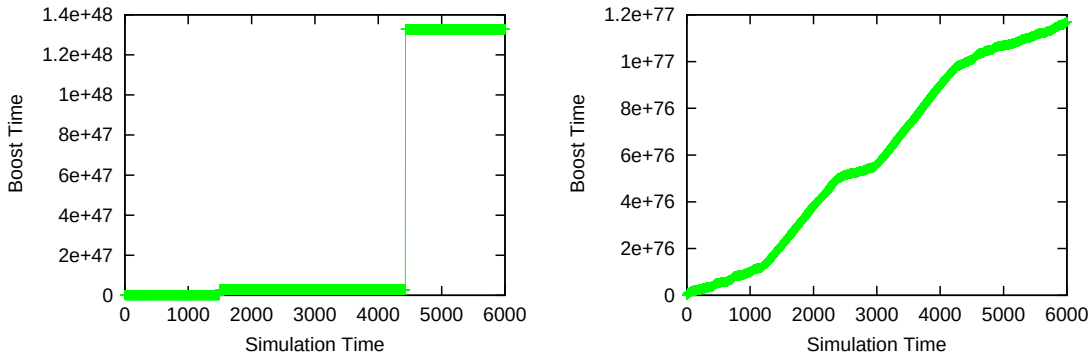


Figure 6.5: Effective boost times for mild (left) and more aggressive (right) bias potentials.

very large and very different sizes of the boosting time scales in these plots).

6.7.2 Monocular 3D Pose Estimation

Now we explore the potential of the hyperdynamics method for monocular 3D human pose estimation under model to image joint correspondences. This problem is well adapted to illustrating the algorithm, as its cost surface is highly multimodal. Of the 32 kinematic model d.o.f., about 10 are subject to 'reflective' kinematic ambiguities (forwards *vs*. backwards slant in depth), which potentially creates around $2^{10} = 1024$ local minima in the cost surface (Lee and Chen, 1985), although some of these are not physically feasible and are automatically pruned during the simulation (see below). Indeed, we find that it is very difficult to ensure initialization to the 'correct' pose with this kind of data.

The simulation enforces joint limit constraints using reflective boundary conditions, *i.e*. by reversing the sign of the particle's normal momentum when it hits a joint limit. We found that this gives an improved sampling acceptance rate compared to simply projecting the proposed config-

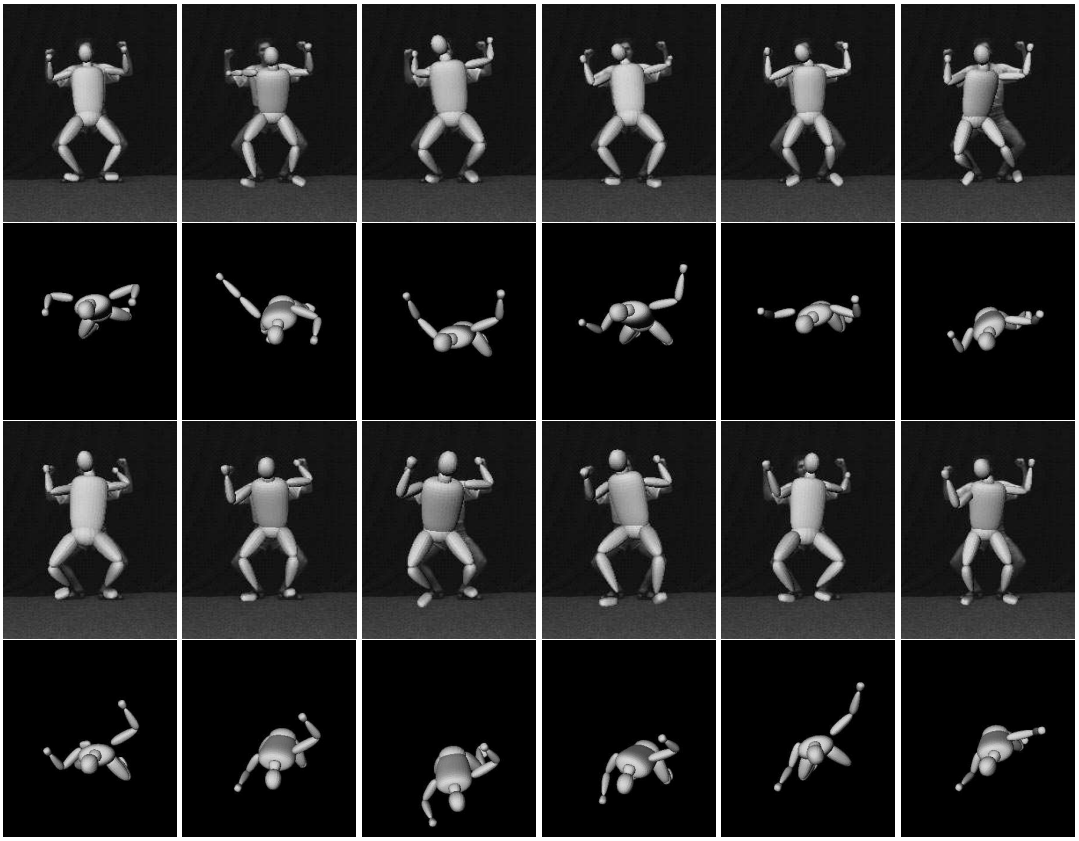


Figure 6.6: Human poses sampled using hyperdynamics on a cost surface based on given model-to-image joint correspondences, seen from the camera viewpoint and from above. Hyperdynamics finds a variety of different poses including well separated reflective ambiguities (which, as expected, all look similar from the camera viewpoint). In contrast, standard stochastic dynamics (on the same underlying cost surface with identical parameters) essentially remains trapped in the original starting mode even after 8000 simulation steps (fig. 6.8).

uration back into the constraint surface, as the latter leads to cascades of rejected moves until the momentum direction gradually swings around.

We ran the simulation for 8000 steps with $\Delta t_{sd} = 0.01$, both on the original cost surface (fig. 6.8) and on the boosted one (fig. 6.6). It is easy to see that the original sampler gets trapped in the starting mode, and wastes all of its samples exploring it repeatedly. Conversely, the boosted hyperdynamics method escapes from the starting mode relatively quickly, and subsequently explores many of the minima resulting from the depth reflection ambiguities.

Fig. 6.7 plots the estimated boosting times for two different bias potentials, $h_b = 200, d = 2$, and $h_b = 400, d = 20$. The computed mean state variance of the original estimator was 4.10^{-6}, compared to 7.10^{-6} for the boosted one.

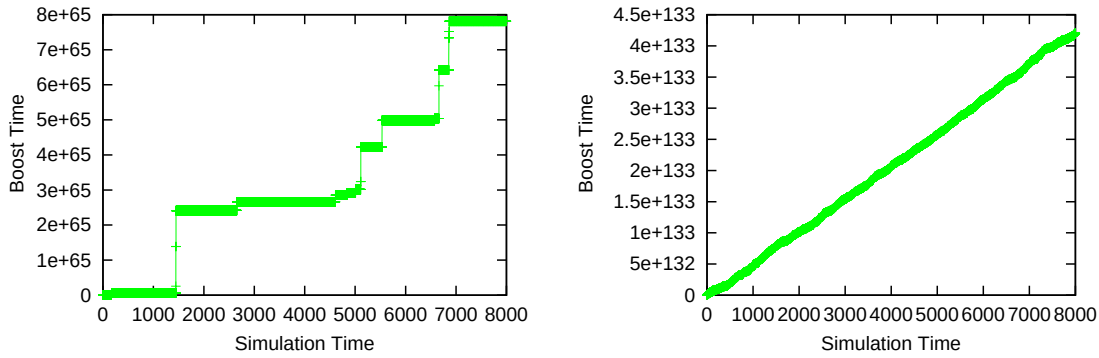


Figure 6.7: Boosting times for human pose experiments, with mild (left) and strong (right) bias.

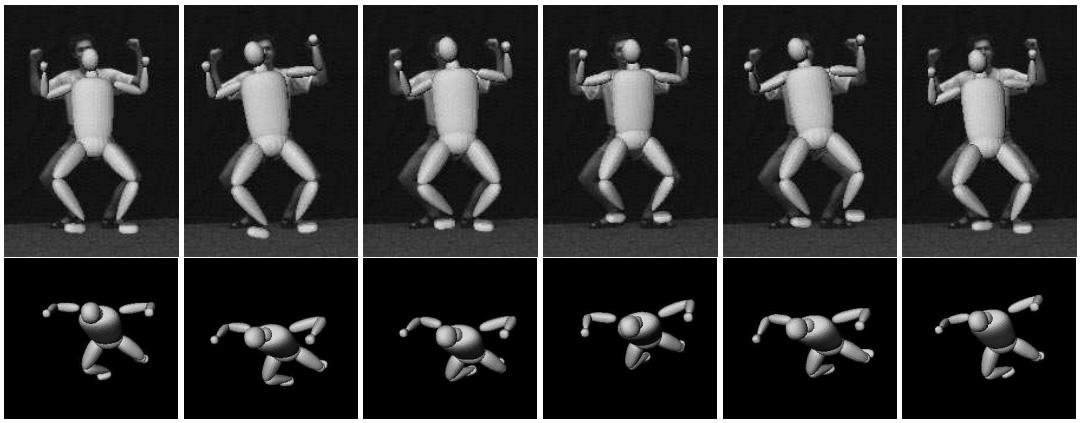


Figure 6.8: Stochastic dynamics on the original cost surface leads to "trapping" in the starting mode.

6.8 Conclusions and Open Research Directions

In this chapter, we underlined the fact that for global investigation of strongly multimodal high dimensional cost functions, importance samplers need to devote some of their samples to reducing trapping in local minima, rather than focusing only on performing their target computation. With this in mind, we presented an MCMC sampler designed to accelerate the exploration of different minima, based on the 'hyperdynamics' method from computational chemistry. It uses local cost gradients and curvatures to construct a modified cost function that focuses samples towards regions with low gradient and at least one negative curvature, which are likely to contain the transition states (low cost saddle points with one negative curvature direction) of the original cost. Our experimental results demonstrate that the method significantly improves inter-minimum exploration behavior in the problem of monocular articulated 3D human pose estimation.

An interesting research direction would be the derivation and investigation of alternative, computationally more efficient biased sampling distributions.

Chapter 7

A View of the Search Problem

This chapter presents an overview of the search problem from the point of view of the factors that affect its difficulty. We study the impact of distribution shape on sampling efficiency, defining several classes of search problems, features influencing search complexity and possible solution techniques. Based on these criteria, we analyze the properties of the likelihood functions used throughout this thesis. We also give suggestions for different search methods that may prove effective in various problem contexts.

7.1 The Choice of Sampling Distributions

We run illustrative experiments in order to study the behavior of different sampling regimes, and their efficiency in locating minima or low-cost parameter space regions. We are interested in how the sampling distribution (as characterized by the shape of its core and the width of its tails) impacts the sampling efficiency. For the study here we use a simple, static, but highly multi-modal likelihood surface based on 3D joint to image correspondences for a 34 d.o.f. articulated model. We run experiments involving Covariance Scaled Sampling (CSS) and Spherical Sampling (SS) for both Gaussian and heavy tail (Cauchy) distributions. For fairer comparisons, we keep the volume of the cores of the distributions constant, *i.e*. the volume of the covariance ellipsoid is always equal to the volume of the corresponding sphere. This is achieved simply by taking the radius R of the sphere to be $R = \sqrt[n]{\lambda_1 ... \lambda_n}$, where λ_i are the eigenvalues of the covariance ellipsoid. The standard deviation σ for a corresponding Gaussian distribution will then be, $\sigma = 1/\sqrt{R}$, *etc*. We report on experiments for Gaussians at scalings 1,4,8,16 σ as well as for a Cauchy distribution with scaling 1. During sampling, we preserve the physical constraints, so samples that are not admissible are projected onto the constraint surface. This can lead to samples that are no longer, *i.e*. Gaussian even though they were originally generated from such a distribution. The samples are subsequently locally optimized, under physical constraints, using the method described in §4.2.1. We report of the number of minima found by each method, the median for distances and standard deviations in parameter space, as well as the costs. These are given in table 7.1. Distributions (histograms)

of the number of samples and minima with parameter space distance, standard deviation and cost, are also shown for different scalings and distribution type in fig. 7.1-7.10. 'Unoptimized' means the histograms correspond to sampled configurations whereas 'Optimized' stands for histograms of samples that have been optimized locally. Note that the distributions of optimized samples are highly multimodal, reflecting the large number of true minima that exist in the problem. Note also the significantly larger number of minima found by CSS with respect to SS, and also that the samples are placed in much lower cost regions by CSS than by SS. One can also see the large cost differences between un-optimized and optimized samples. Also, at a first glance it may appear that the SS generates lower-cost minima than CSS. This is simply because CSS searches more thoroughly. Firstly, the CSS is naturally able to sample further away due to its variable scaling, and this can lead to the detection of higher cost minima which can be expected further from the target. Secondly, CSS detects significantly more minima that SS, and this influences the computation of the median (some of the extra minima, being further away are likely to have higher cost).

Notice the substantially improved sampling efficiency of CSS versus SS, and that sampling from a heavy tailed distribution gives a better balance between locality and globality than the use of pure Gaussian tails. However, although distant parameter space Cauchy tailed samples produce a more thorough global search, a different, quasi-local behavior may be desired in some applications. For instance, in tracking, one may look for low-cost samples in spatially nearby (ideally adjacent) minima, while for initialization one can be interested in more global solutions. In either case, we recommend CSS with Gaussian or heavy tails as an effective sampling distribution, both for locating larger numbers of minima, and for generating samples with good cost within a given search volume. For ill-conditioned problems like 3D tracking, where the cost surface evolves dynamically, sampling relatively distantly using the local covariance generates hypotheses along the highly uncertain and difficult to estimate directions of the parameter space, and therefore increases the chance of not losing representative minima when optimizing locally in the next time step.

In table 7.2 on page 120 we show results that quantify the relative efficiency of each of the methods we have proposed for the same likelihood cost surface based on model-image joint correspondence. The results are given in terms of the number of minima found per unit of work of local descent (*i.e.* one local optimization work). Notice the superior performance of ET/HS methods and that also CSS and HDIS are still very competitive. Nevertheless, the results should be interpreted with moderation as the situation could change on different cost surfaces with different properties and minimum structure. For instance a surface with more local attraction properties might render RS techniques even more inefficient while a surface with few and narrow cost basins might be very efficiently searched by HDIS (probably better than CSS). Alternatively for surfaces with many minima that are fairly unstructured RS might perform as well as any other method. A more general discussion will appear in the next few sections.

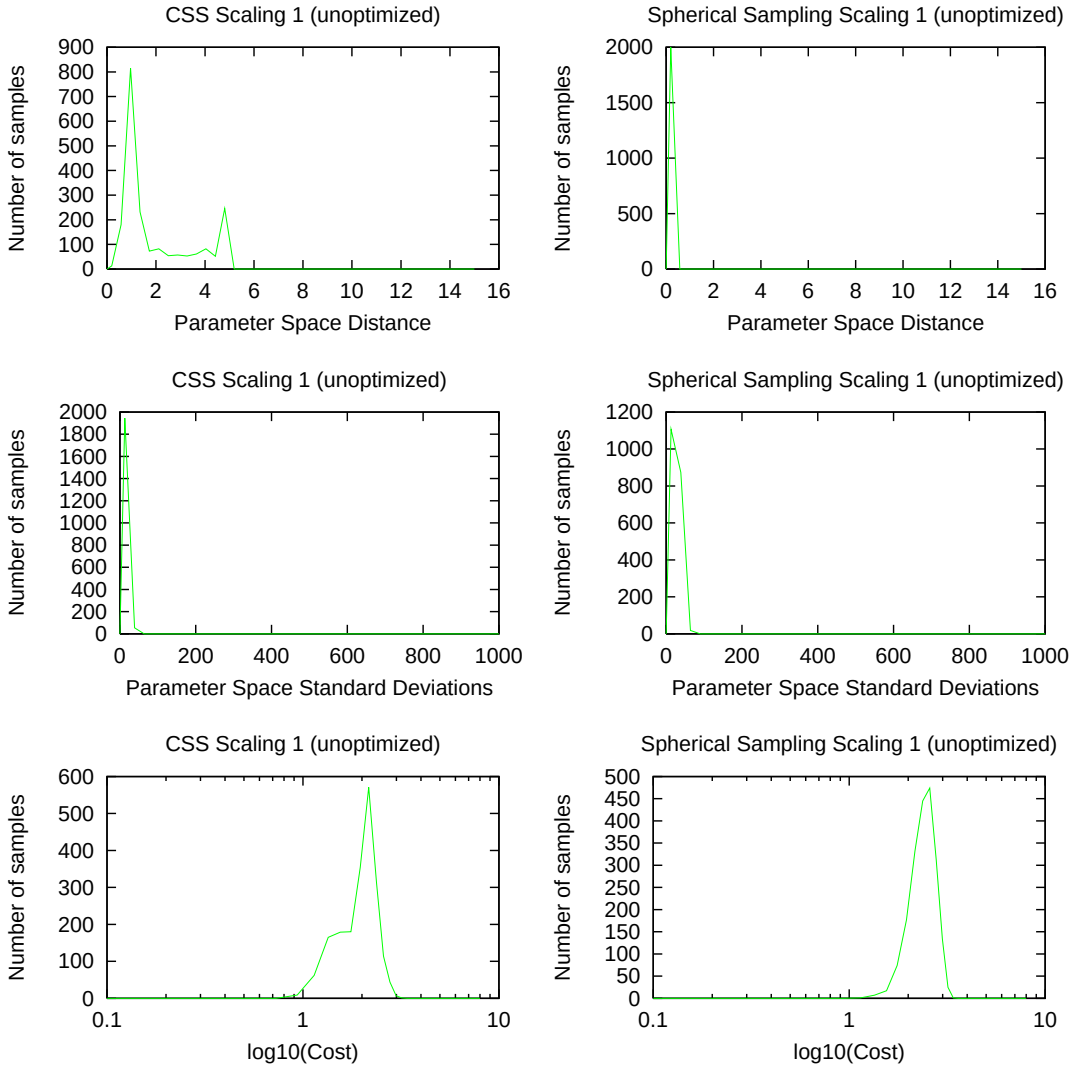


Figure 7.1: Unoptimized Spherical and Covariance Scaled Sampling (CSS) for the same space volume (scaling factor is 1).

7.2 Search Problems, Likelihood Structures and Solution Techniques

Analyzing the relations between the structure of the likelihood surface, the complexity of the search algorithm and the possible solution techniques, is a complex task that may be worth a thesis in its own right. In the remainder part of this chapter we briefly suggest a few possible ways of classifying, decomposing and attacking these problems, based on our experience.

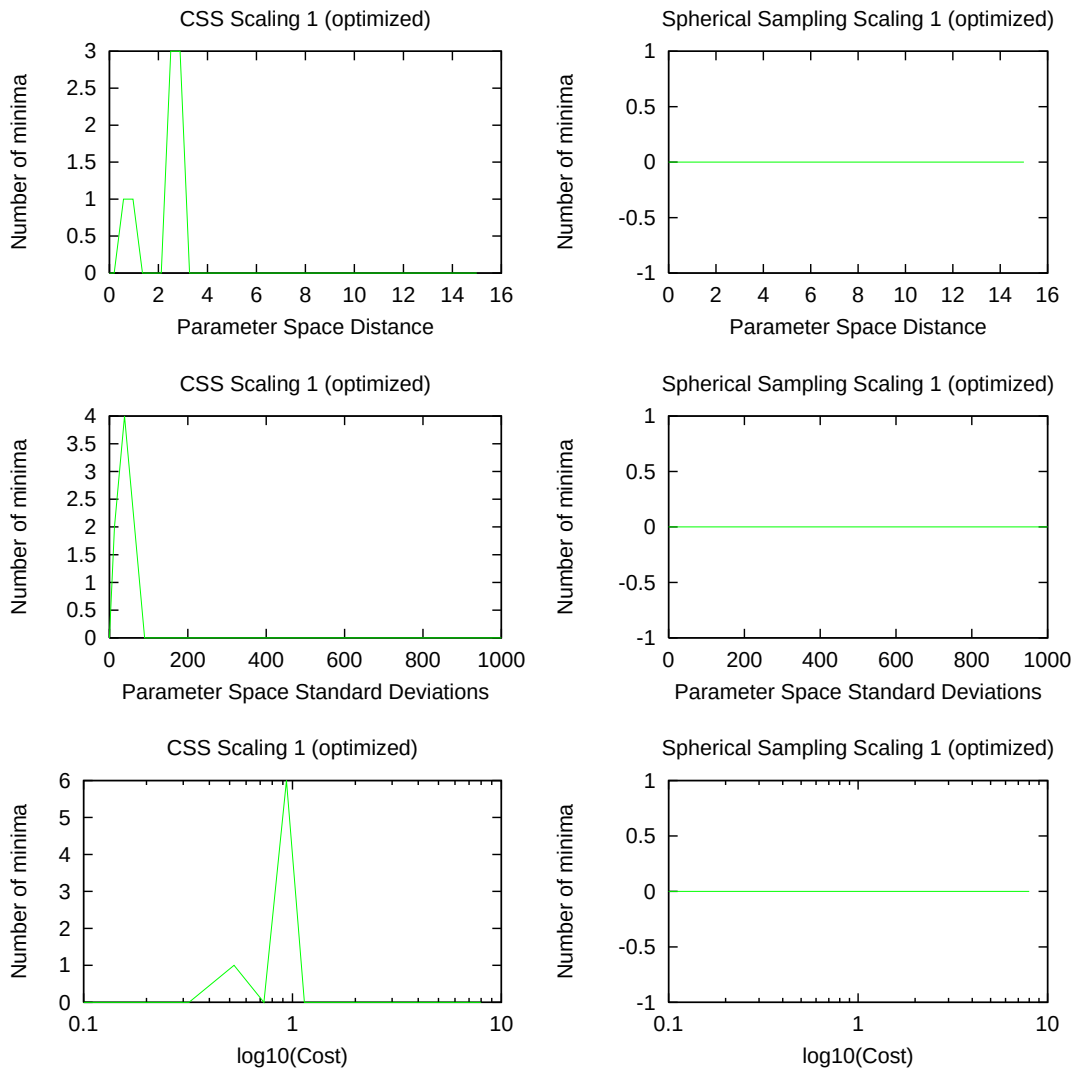


Figure 7.2: Optimized Spherical and Covariance Scaled Sampling (CSS) for the same space volume (scaling factor is 1).

7.2.1 Factors Influencing Search Complexity

In this section we consider the main factors that influence the expense and success rate of search algorithms for non-convex, multiple-minima error surfaces. These could have one or several global minima as well as other local minima. Conventional global minimization aims to find the absolute global minimum, but it is important to realize that even if this were tractable, it would not, in general, suffice. In particular, in many tracking or inference applications, one searches for a set of minima that are 'representative' of the posterior likelihood encoded by the cost function, *i.e.* that

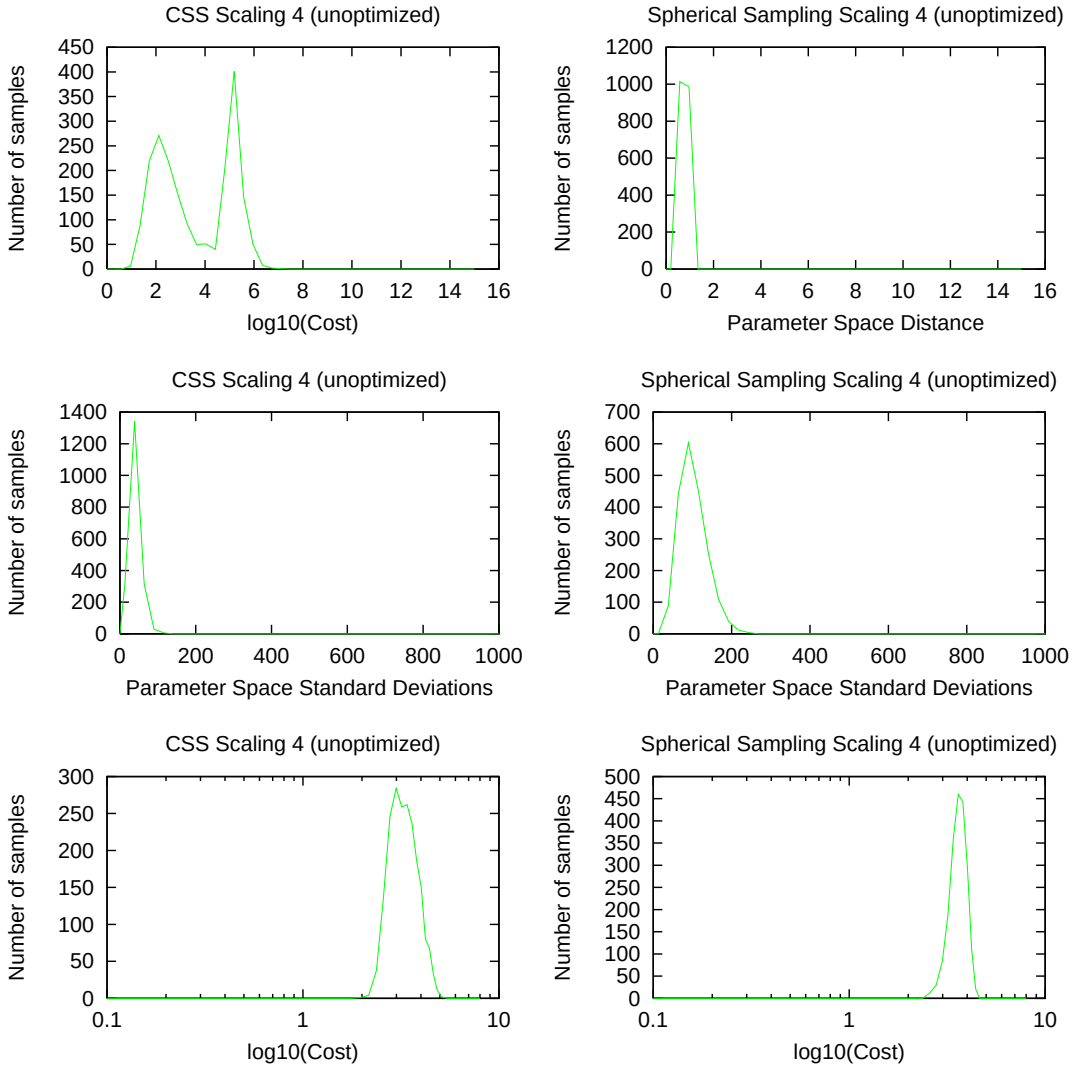


Figure 7.3: Unoptimized Spherical and Covariance Scaled Sampling (CSS) for the same space volume (scaling factor is 4).

capture most of its weight[1].

Cost evaluation and the volume of the 'typical set': Many complex cost functions have most of their representative minima concentrated in relatively narrow regions of parameter space. The volumes of the low-lying basins of attraction relative to the volume of the full parameter space are important determinants of search complexity. For instance, suppose we are interested in a set k of global minima with basins of attraction of relative volume p_i, and that N_i random function evaluations are available to search the i^{th} minimum. Then the probability of missing at least one

[1]Such minima are potentially global in a moderate neighborhood of their mean value. A related discussion in the context of global optimization can be found in (Törn et al., 1999).

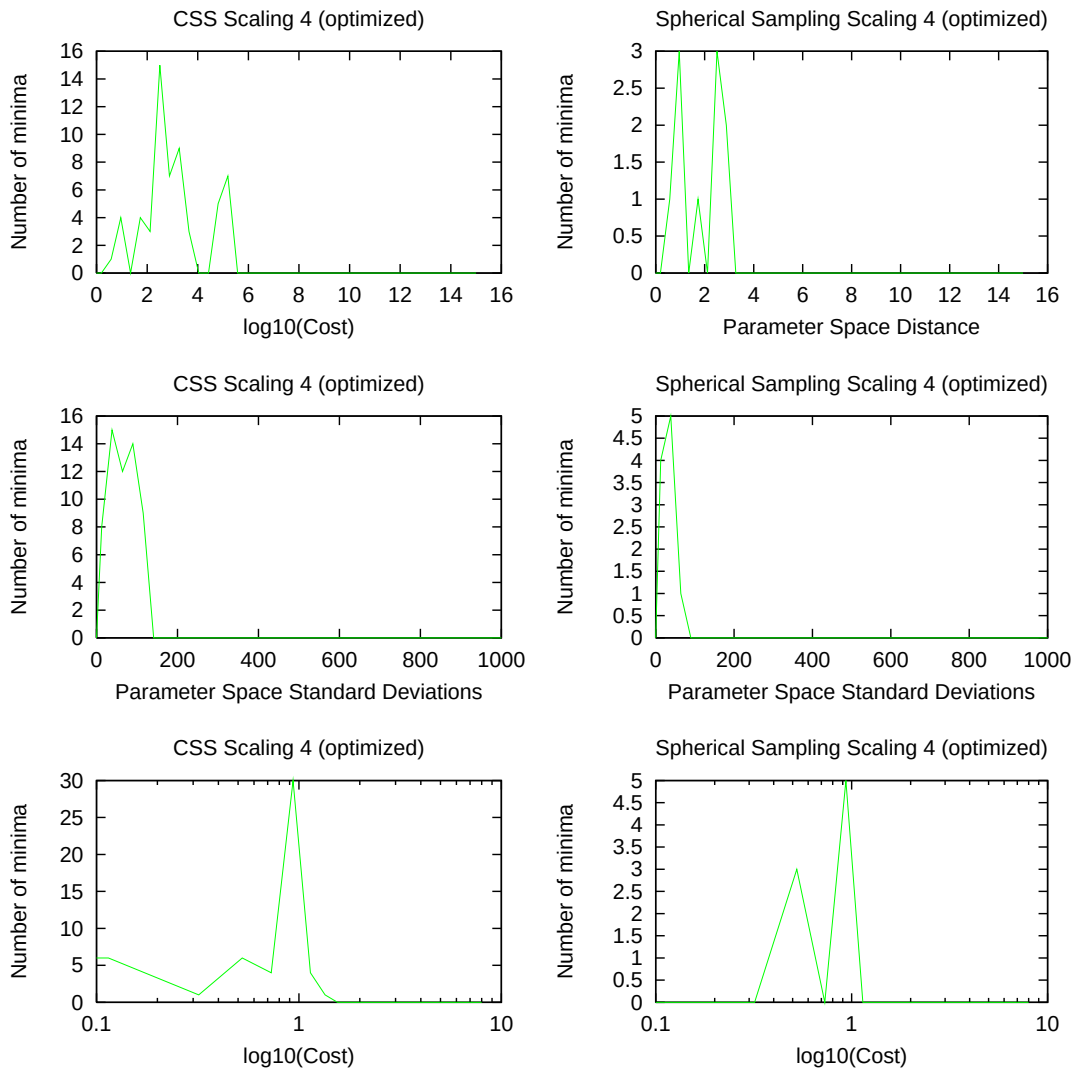


Figure 7.4: Optimized Spherical and Covariance Scaled Sampling (CSS) for the same space volume (scaling factor is 4).

minimum is: $\prod_{i=1}^{k}(1-p_i)^{N_i}$. For moderate N_i and large volume parameter spaces, the above probability of a global set miss can be very high[2].

The complexity of an individual cost evaluation also plays a role here. For the human tracking problem, this involves the 3D articulated mesh construction, occlusion prediction and computing feature prediction error on a subset of model points.

The clustering of minima: If any given minima is typically surrounded by other nearby minima,

[2]To see how the volume of the parameter space impacts the miss probability for even a single minima, consider an N-dimensional box type parameter space $[0,1]^N$ and suppose it is scaled, say twice, in each dimension so it becomes $[0,2]^N$. Then the probability of missing one single minimum decreases from $1-p_i$ to $1-p_i/2^n$.

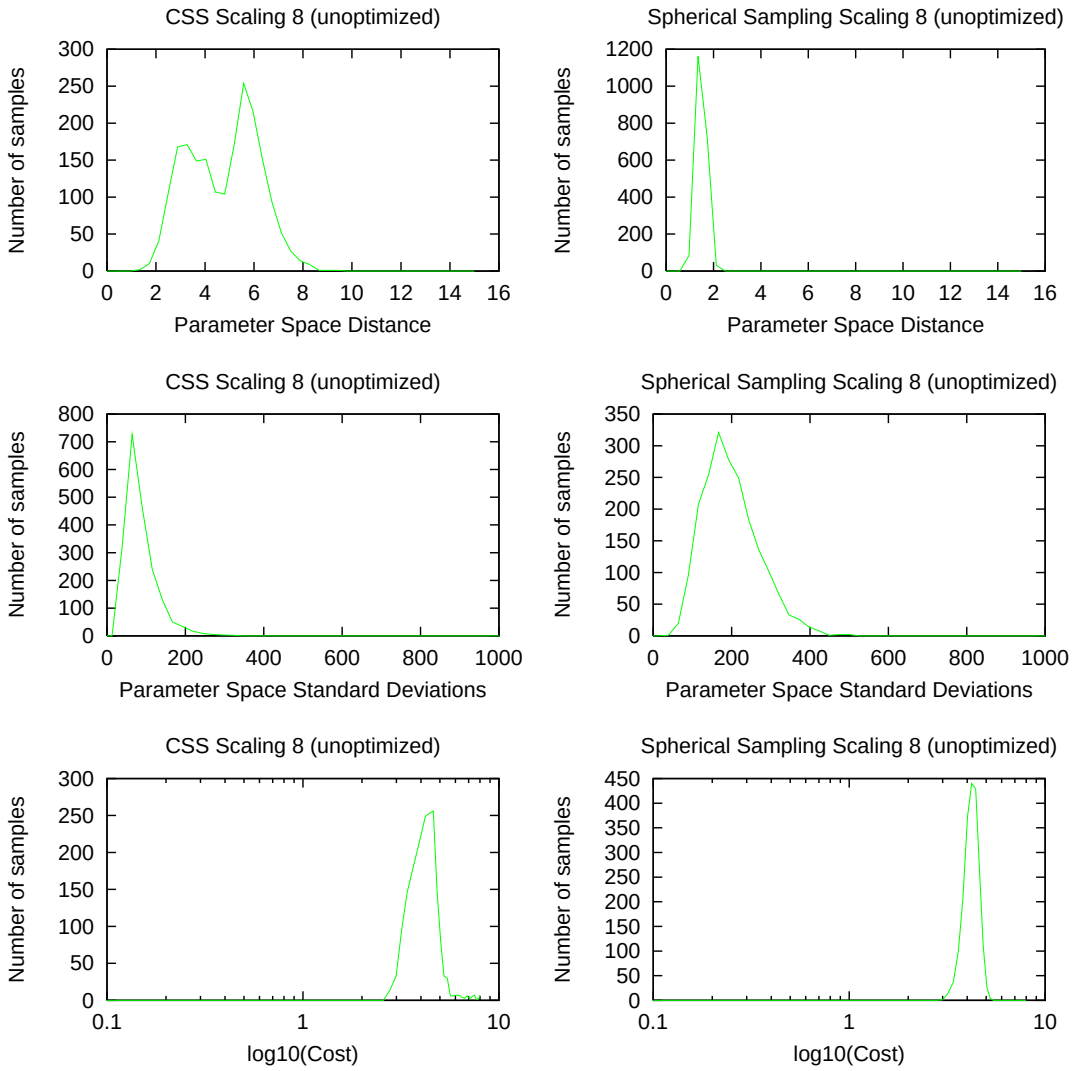


Figure 7.5: Unoptimized Spherical and Covariance Scaled Sampling (CSS) for the same space volume (scaling factor is 8).

adaptive quasi-local search can lead to the discovery of increasingly better minima, eventually including the lowest one. Isolated minima are the opposite case, where the position of a minimum give little information about others, but there is also little local ambiguity. In vision, isolated minima can arise from clutter or modeling artifacts, while clustered ones mostly arise close to the true target due to ambiguities in model to image matching.

The number of local minima and the attraction zones: The number of local minima is another good indicator of the difficulty of the search problem. Clearly, with many minima, the attraction zones of individual energy basins must become smaller and the search process can easily be distracted by the presence of such spurious minima.

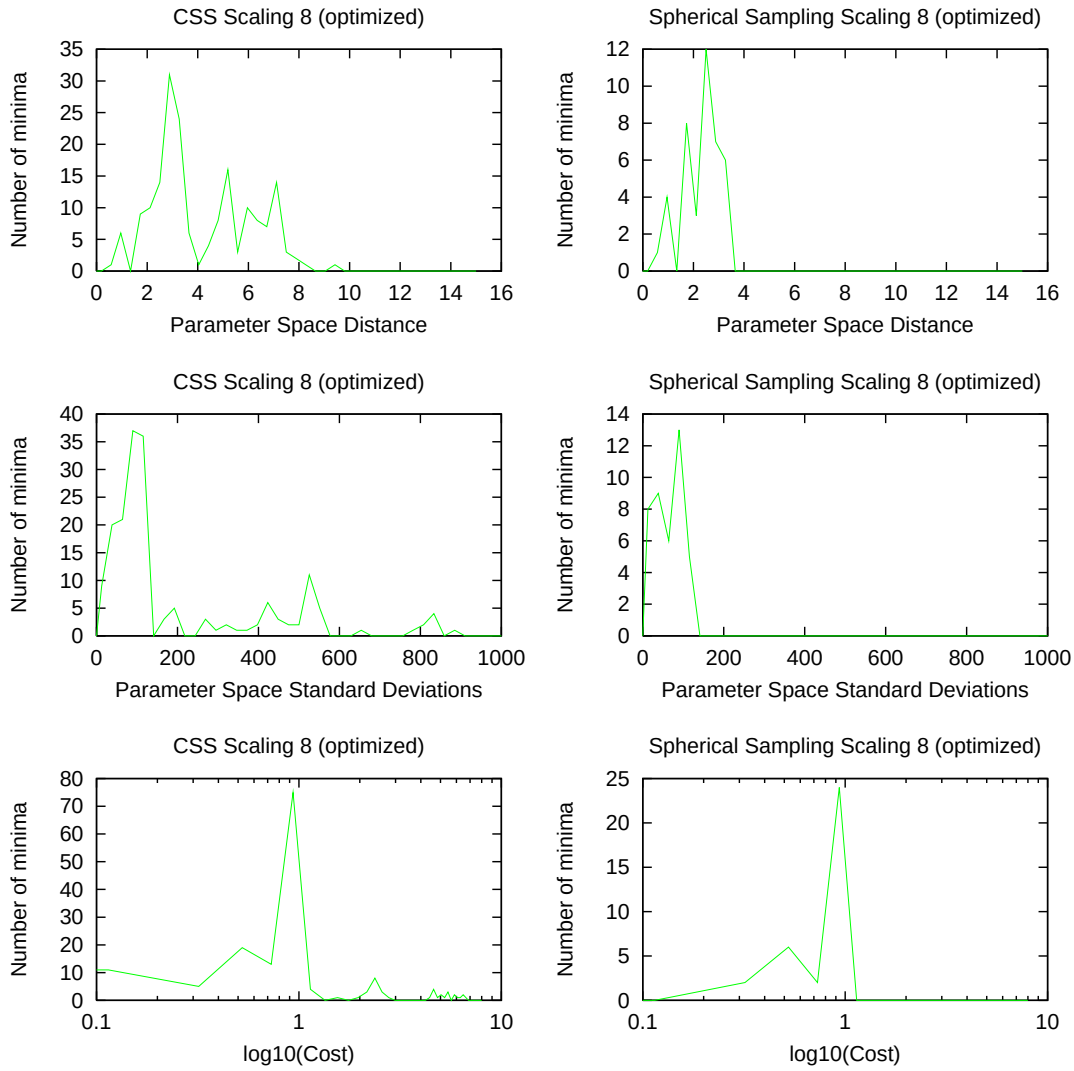


Figure 7.6: Optimized Spherical and Covariance Scaled Sampling (CSS) for the same space volume (scaling factor is 8).

7.2.2 The Likelihood Structure

In this section we comment on the likelihood functions developed during this thesis, in particular on their regions of attraction, globality of response and minimum structure.

The contour likelihood is constructed from the individual edge responses in the neighborhood of the model prediction. The resulting error surface has a local attraction zone but there are often secondary minima nearby, generated by other locally possible correspondences. Although these are regularized by model coherencies, spurious minima corresponding to partially incorrect assignments do exist, as showed in chapter 3. Away from the target, in regions where the local model-based search doesn't detect *any* globally human-like structures, the likelihood can still have sharp

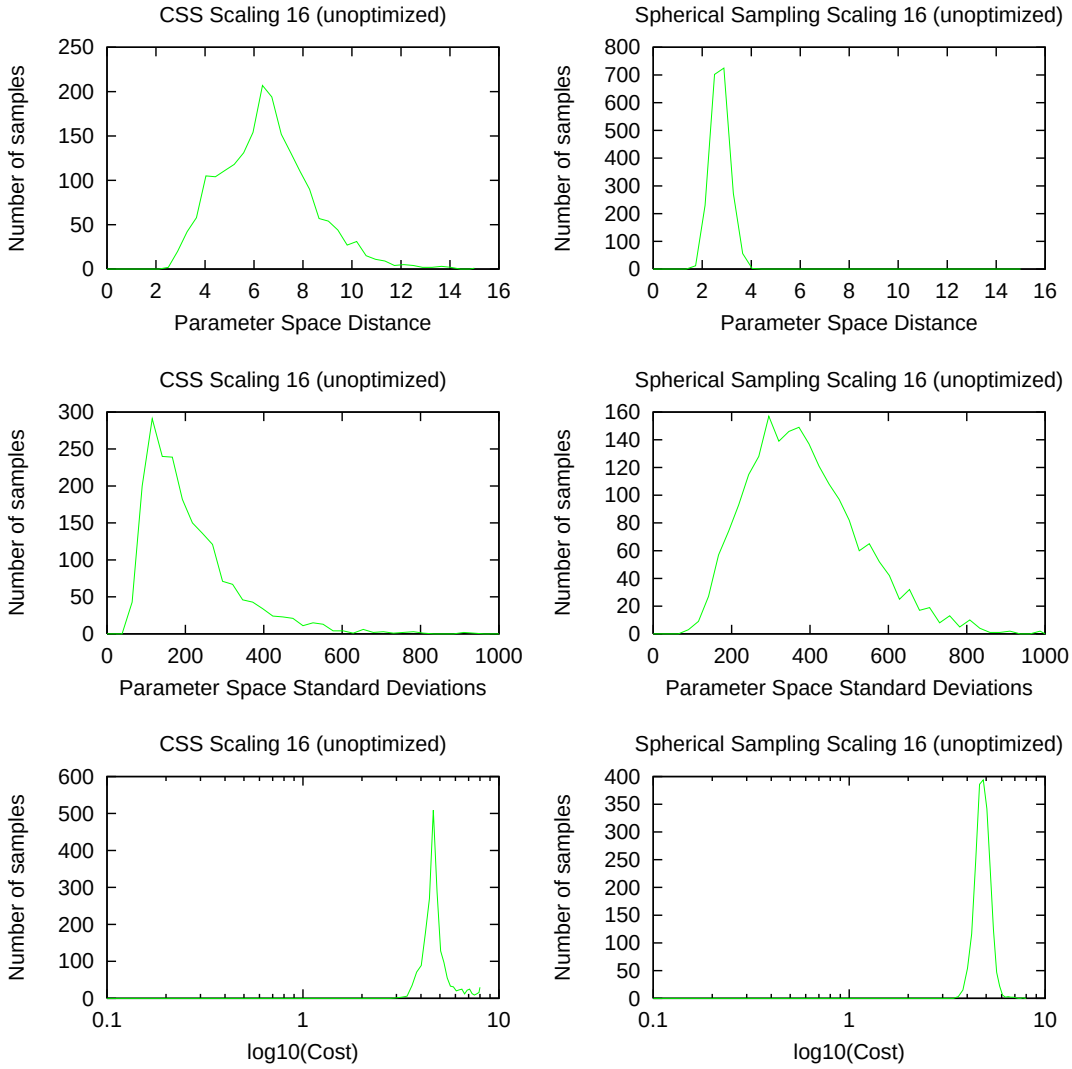


Figure 7.7: Unoptimized Spherical and Covariance Scaled Sampling (CSS) for the same space volume (scaling factor is 16).

but isolated peaks corresponding to alignments with spurious contours or linear coherencies in the background. The frequency of these false alarms depends strongly on both the properties of the background – there are many false alarms in cluttered scenes and few in controlled environments with uniform backgrounds – and the degree to which the feature detector enforces more global coherence – for example a ribbon or blob detector would give fewer false alarms than a simple local edge detector, but probably also more misses of true features. Our experience suggests that local edge likelihoods may suffice for tracking but not for initialization, and this is the context where we have used it[3].

[3]We don't rule out the usage of contour likelihoods for object localization or initialization *in general*. We only

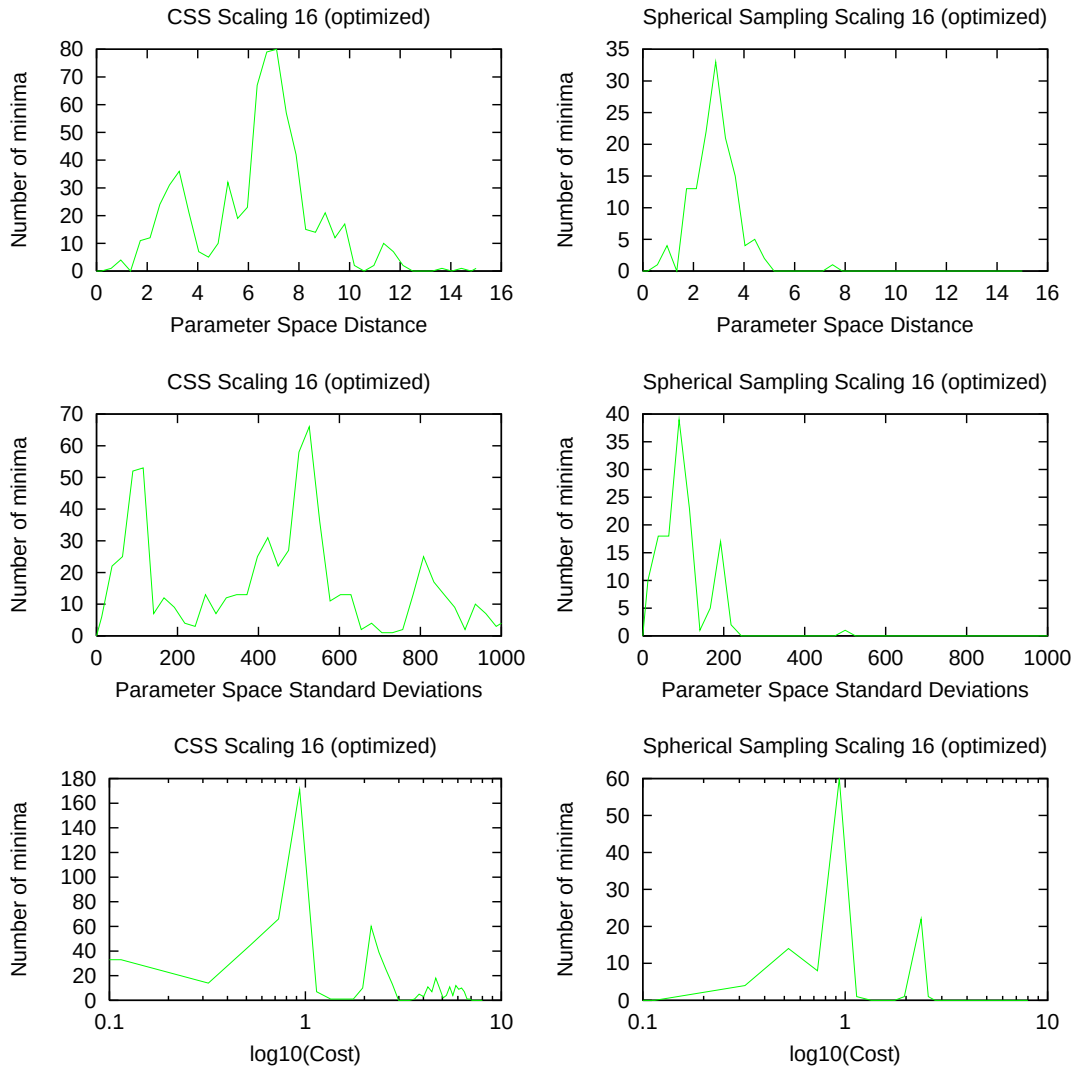


Figure 7.8: Optimized Spherical and Covariance Scaled Sampling (CSS) for the same space volume (scaling factor is 16).

The model-based intensity matching likelihood has many properties in common with the edge-based one. It is meaningful close to the target, but also has a fairly small attraction zone. The effect of clutter close to the target is less obvious than in the case of contours, but multi-modality is still possible, especially for objects with uniform of repetitive textures. Far from the true configuration, the intensity responses are still strongly dependent on the structure of the background, with many small isolated peaks in cluttered regions and noisy flatness in uniform ones. Again, we feel that intensity matching is also more useful for tracking than for initialization. As above, representations that use intensity cues to detect limbs and assemble lower-dimensional 2D cardboard models are

discourage their use with high-dimensional parameter spaces like the ones we deal with here.

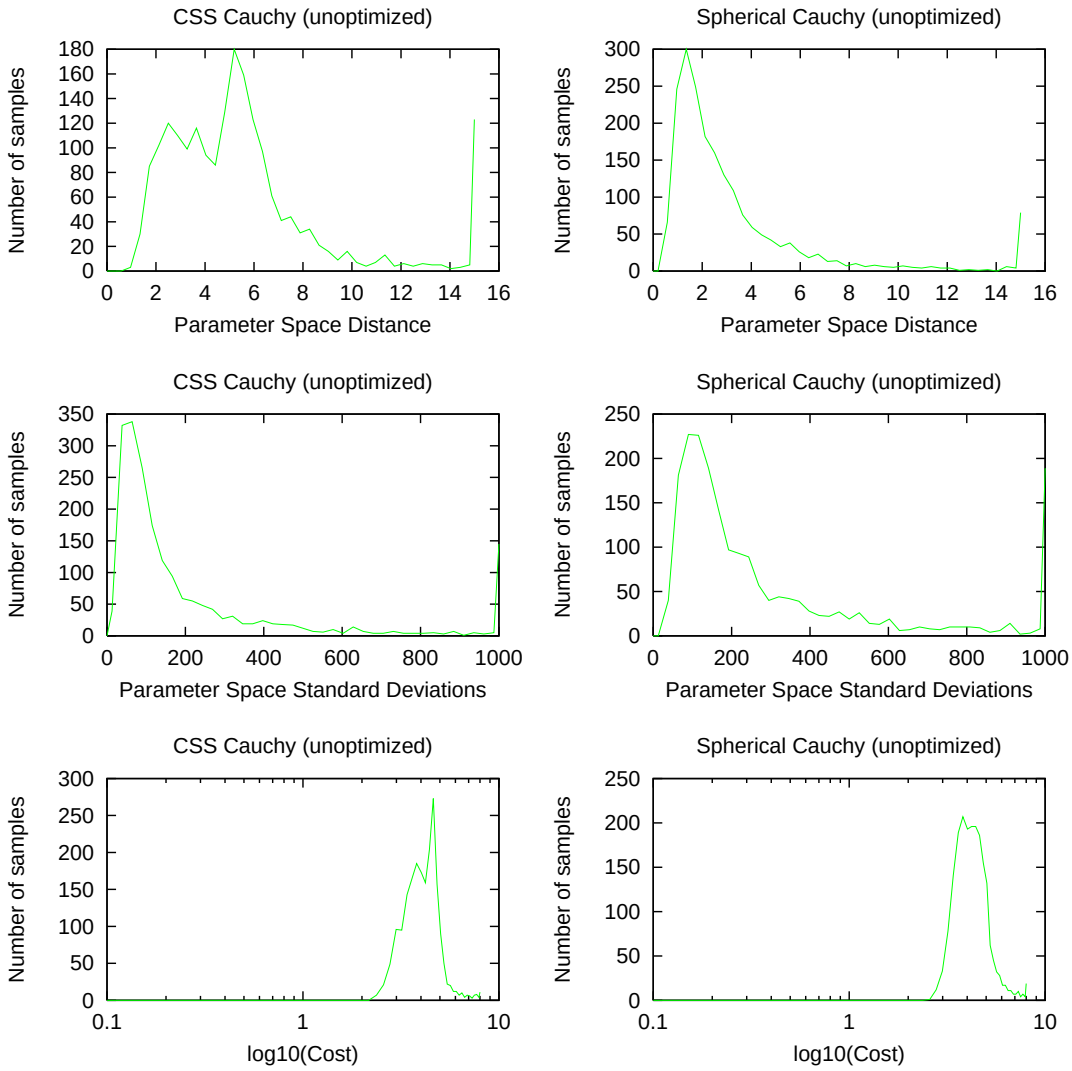


Figure 7.9: Unoptimized Spherical and Covariance Scaled Sampling (CSS) for Cauchy heavy tail distributions (same core volume, unscaled).

still useful. See the concluding chapter for a discussion.

If unambiguous silhouettes are available (*e.g.* in blue screen applications or with efficient background subtraction, contour tracking or motion segmentation), the silhouette likelihood has attractive global response properties and a large basin of attraction. There are very few minima far from the true target. Close to the target, it has a complex topology with multiple minima and embedded structure. Some of the minima are inherited from model projection/observability ambiguities, *e.g.* rotating the person back to front or reflective ambiguities for limbs, *etc.* Another problem with the silhouettes is that they are not very discriminating especially when the limbs are well inside the silhouette. Hence, by itself, the silhouette likelihood often leads to singular cost regions caused by

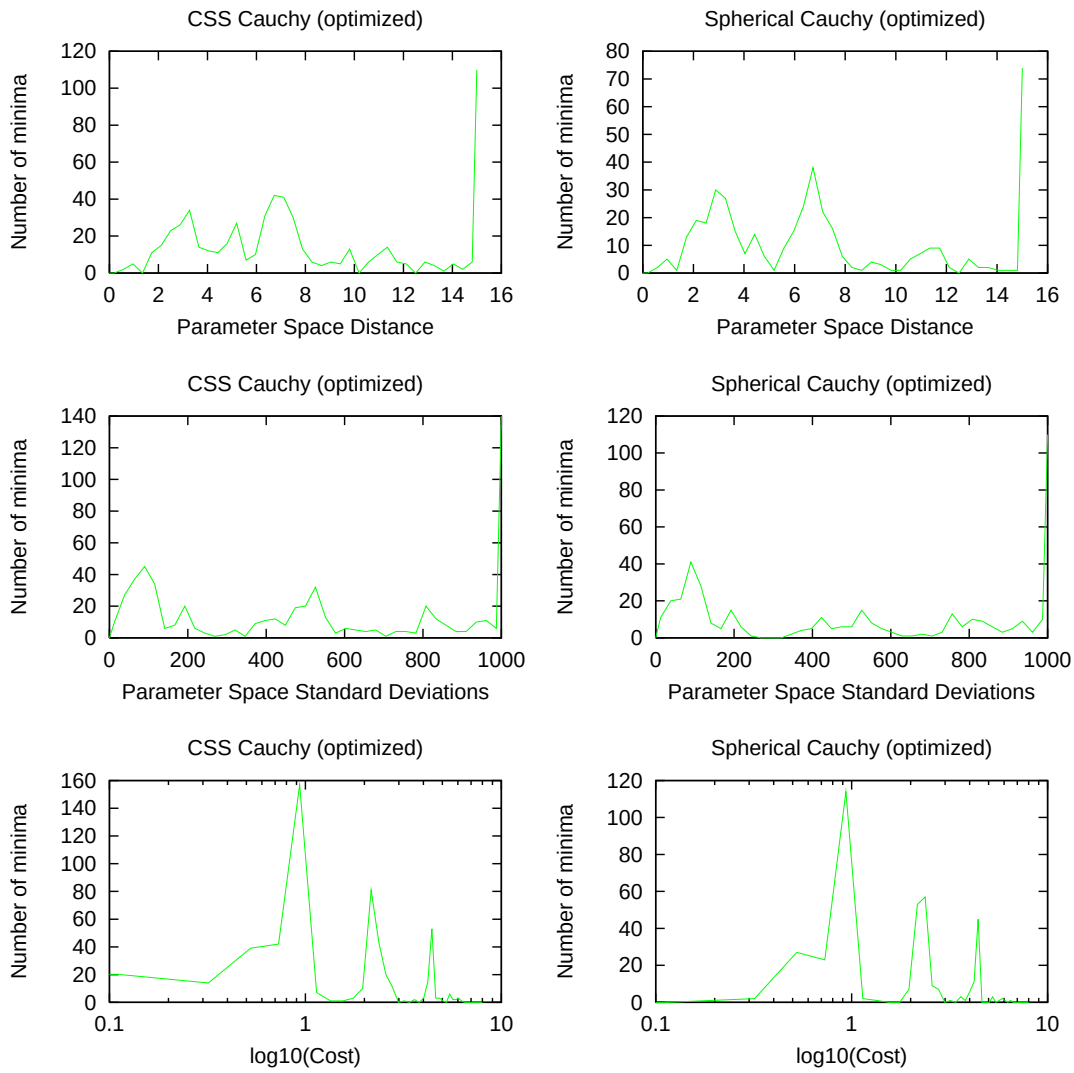


Figure 7.10: Optimized Spherical and Covariance Scaled Sampling (CSS) for Cauchy heavy tail distributions (same core volume, unscaled).

parameters that can't be properly observed[4]. This is less problematic with multi-camera systems, but with monocular ones the silhouette likelihood should be used either in conjunction with the edge or intensity one, or stabilized with priors on difficult to observe parameters, as explained in chapter 3.

The joint-center based likelihood used for initialization has the desirable global response properties of the silhouette likelihood and additional discriminative power owing to the precise 3D to 2D joint correspondence information. However, in the monocular case, it still retains multiple re-

[4]This is mostly true in the monocular case. For multiple-camera settings, the situation could improve, but still axial limb rotations can pass unobserved, *etc*.

METHOD	SCALE	NUMBER OF MINIMA	PARAMETER DISTANCE MEDIAN		STANDARD DEVIATIONS MEDIAN		COST MEDIAN	
			UNOPT	OPT	UNOPT	OPT	UNOPT	OPT
CSS	1	8	1.148	2.55242	10.9351	47.6042	116.951	8.49689
CSS	4	59	3.21239	2.9474	35.2918	55.3163	1995.12	6.98109
CSS	8	180	4.969	3.34661	75.1119	109.813	16200.8	7.09866
CSS	16	667	6.42423	6.72099	177.111	465.889	45444.1	8.69589
CSS	1/HT	580	5.05363	6.93626	106.631	517.387	15247.7	8.72421
SS	1	0	0.199367	-	24.5274	-	273.509	-
SS	4	11	0.767306	2.04928	96.1519	39.0745	4291.12	6.28014
SS	8	42	1.47262	2.54884	188.157	56.8268	16856.1	6.96481
SS	16	135	2.71954	2.8494	367.746	87.8533	63591.4	8.69588
SS	1/HT	232	2.18611	6.54748	178.647	535.999	18173	17.8807

Table 7.1: Quantitative results for minima distribution. Note substantially increased number of minima found by the covariance scaled sampling (CSS) versus the spherical sampling methods (SS), as well as the fact that CSS places samples on significantly lower cost regions.

METHOD	MINIMA LOCATION EFFICIENCY
RS+LD	0,0675
CSS+LD	0,3335
HDIS	0,2226
ET/HS	0,7741

Table 7.2: Relative performance of different methods in locating multiple minima. The efficiency is given in new minima found per 1 local optimization work or equivalent. 'CSS'=Covariance Scaled Sampling+local descent, 'SS'=spherical sampling+local descent, 'ET'=Eigenvector tracking, 'HS'=Hypersurface sweeping, 'HDIS'=Hyperdynamic Importance Sampling+local descent.

flective ambiguities owing to the loss of depth information. It has a clustered minima structure in the neighborhood of its global minima and it does still have a small number of secondary minima produced by kinematic or joint limit constraints[5] and 'crossing of the springs' that attract the human puppet joints towards the image joints.

The properties of the different likelihood types are summarized in table 7.3.

[5]However, it is still the case that the use of physical constraints helps more than it hurts here. Most of the time the joint positioning in the image is likely to be noisy and this can easily lead to unconstrained estimates that are not admissible. Even under exact joint correspondence information, the ambiguity of the model-image mapping can lead to incorrect fits.

LIKELIHOOD	ATTRACTION	TOPOLOGY OF MINIMA		NUMBER OF MINIMA		TYPICAL SET
		Close	Far	Close	Far	MISS
Edge (cb)	Local	Clustered	Isolated	Medium	Large	High
Edge (ub)					Small	Average
Intensity (cb)	Local	Clustered	Isolated	Medium	Large	High
Intensity (ub)					Small	Average
Silhouette	Global	Clustered	Isolated	Medium	Small	Average
Joint	Global	Clustered	Clustered	Large	Medium	Average

Table 7.3: Likelihood surface and their properties. 'Far' means away from the true response, 'Close' means near the true response, 'cb' stands for cluttered background and 'ub' stands for uniform background.

Limitations

It is apparent from the previous section that each of our likelihood terms has important limitations: some have only narrow zones of attraction, while others lead to ambiguities even under correct sets of correspondences. Ultimately, it is clear that our models of object similarity are still weak. They are based on over-local feature detection and recognition models that lack the the higher-level feedback that would allow incoherent configurations to be pruned away during the matching process. Furthermore, although we know that we need "semantically interpretable" minima, our likelihood functions have relatively poor invariance under semantic level transformations such as occlusion and viewpoint changes. Tracking failures can often be attributed to the existence of too much local ambiguity in the cost surface. Any multiple hypothesis tracker necessarily has only limited resources and if these are saturated, important peaks are sure to be lost. In most trackers, these problems are greatly exacerbated by poor mixing (difficulty of jumping) between nearby minima, so that temporary tracking failures can not be recovered. For these reasons, it is important to prune the search space and exploit as much as possible the structure of the problem, even though this is non-trivial in many applications. For instance, in problems with structured internal ambiguities such as kinematic reflection ones, it may be possible to compute well controlled jumps between modes to promote mixing or even to work initially with just one representation of one ambiguity, *i.e.* a 'folded' parameter space.

Now, we turn to reviewing the solution strategies that can be used with such complicated likelihood structures.

7.2.3 Solution Strategies

A balance between global and local search effort is necessary in many optimization problems including the monocular human tracking one. The dimension of the parameter space prohibits exhaustive search but, once a set of minima has been found, one can attempt to conserve effort by quasi-local search near these and perhaps also near other, globally different configurations that are known to give similar observables. We thus feel that three components are important in the design of an effective search algorithm: (*i*) A global representation involving estimation based on multiple hypotheses (*e.g*. samples or minima) (*ii*) An effective local descent method (LD), that could be either continuous local optimization (LO) or stochastic descent methods based on Markov Chains or controlled random search (CRS), (*iii*) An effective strategy for escaping minima and more global exploration. Depending on the problem complexity this might be random sampling (RS), annealed sampling (AS), our covariance-scaled sampling (CSS), or one of the saddle point search methods we have proposed: eigenvector tracking (ET), hypersurface sweeping (HS) or hyperdynamic importance sampling (HDIS).

Different solution strategies are preferable for different classes of problems. the results are summarized in table 7.4. The '*' signs show when would a certain solution strategy is desirable or '**' essential for success in the corresponding problem class.

7.2.4 Problem Classes

While we certainly feel that many vision problems are hard, we haven't yet come to decompose their difficulties in detail. In this section, we attempt to crudely classify the difficulties of different search problems and give suggestions on how to combine solution strategies for particular problems. We judge problem difficulty according to the probability of missing the 'typical set' of the distribution[6], the clustering of minima and the embedding of the global ones, and the number of local minima present.

Convex Problems

Convex problems have an essentially unique global solution with a basin of attraction covering the whole parameter space, so they need neither a global technique, nor escape strategies. The (*) in the column corresponding to the global search in table 7.4 only indicates that multiple starting points may have to be used for local descent, as in many practical contexts, we don't know the problem is unimodal apriori. Both continuous and stochastic local descent techniques are usable here.

[6]By 'typical set' we understand the set of significant energy basins around the important minima. These could also include other local minima, *etc*.

Easy Problems

The easy problems are characterized by a low probability of missing the attraction zones of the interesting minima. This means that either their corresponding basins of attraction are large or that sufficient samples can be generated in order to reduce the probability of a total miss. In such contexts, a good strategy could be to sample points globally and randomly and to do local optimization from there. For stochastic descent methods that do not converge to the exact minimum, clustering may be required to identify the different minima [7]. Annealed sampling can be an effective method for problems with few global minima each having a large attraction zone. Several runs can be started at different initialization points to capture all of the minima. For problems with many minima and smaller basins of attraction, more effective escape strategies are needed.

COMPLEXITY	TYPICAL SET MISS	CLUSTERING	MINIMA NUMBER	STRATEGY TYPE			METHOD
				Global Search	Local Descent	Escape Strategy	
Convex	0	No	1	(*)	**	-	LD
Easy	Low	Any	Small	*	*	-	RS+LD/AS/CRS
	Low	Any	Large	*	*	(*)	CSS/RS+LD
Moderate	High	Clustered	Small	*	**	*	AS/CSS/HS/HDIS
	High	Clustered	Large	*	**	**	CSS/ET
Hard	High	Isolated	Small	*	*	**	HDIS/ET
	High	Isolated	Large	*	**	**	ET/HS

Table 7.4: Classes of search problems, the solution strategies and methods suggested for use with each one (see text).

Moderately Difficult Problems

These problems are characterized by clustered minima with high probability of missing their regions of attraction, either because these volumes are small or because the cost function is so expensive to evaluate that the number of available evaluations very limited. The local search needs to be globalized and both local descent and escape strategies are useful here. For small or moderate numbers of local minima and ill-conditioning, or more widely separated ones, we appreciate CSS, HS or even HDIS can be effective methods. For larger numbers of minima, more systematic search procedures like HS or ET may be desirable.

[7] For continuous methods the clustering results are implicit, since running the optimizer to convergence guarantees separated (or confounded) minima.

Hard Problems

These are characterized by a high probability of missing the regions of attraction of the good minima which may be concentrated in very narrow regions of space. In such contexts, one may need to rely almost entirely on random sampling the space, if no signal for driving a more adaptive method is available. Nevertheless, for continuous or differentiable problems with a small number of minima, HDIS may be the preferred method. The additional effort involved in sampling or optimizing its bias potential would be rewarded because the method focuses the search in the *very few* transition neighborhoods that can lead to global minima. In such situations, an aggressive bias potential may be required as the saddle neighborhoods can have very large volumes indeed. Alternatively a moderate boosting potential can be used to focus the samples on transition neighborhoods and a lightweight saddle refinement method like ET can be initialized on negative curvature and proceed from there. For problems with a large number of isolated minima, the use of ET or HS techniques may still be effective as they allow systematic search along a large variety of directions in space. Additionally, being Hessian based, ET/HS are fairly insensitive to the shape of cost surface and can efficiently adapt their trajectories to steep, low-curvature or ill-conditioned regions.

7.2.5 Convergence

Every optimization or sampling method faces the delicate issue of convergence or stopping conditions. Sometimes these are based on theoretical convergence properties, but for most real optimization or sampling applications, the convergence properties are usually probabilistic, meaning that some method will find all the quasi-global minima or sample all the important modes with a probability that approaches one as the run-time increases. Finite convergence would mean that points in the regions of attraction of the quasi-global minima (for optimization methods), or in those of all minima (for sampling methods) could be found in a finite number of steps. This is usually not possible except in some controlled scenarios where we know bounds on global minima, or some combinatorial argument provides information on the number of peaks (*e.g*. for model-image feature assignment problems in vision, but also there are often nonlinearities that complicate the issue, *etc*.). Note that it is simple to modify a heuristic search method to run indefinitely and converge with probability one simply by adding a random sampling element that is applied, perhaps with an exponentially decreasing probability in order not to degrade the efficiency of the method (*e.g*. consider the following: any time the number of function evaluation reaches an upper bound, say U_b, sample a point at random, include it in the working sample set and increase U_b). Nevertheless, the heuristic and the ‘convergent’ methods would be almost identical in practical applications.

Figure 7.11: Clutter human tracking sequence detailed results for the CSS algorithm in §4.4, chapter 4.

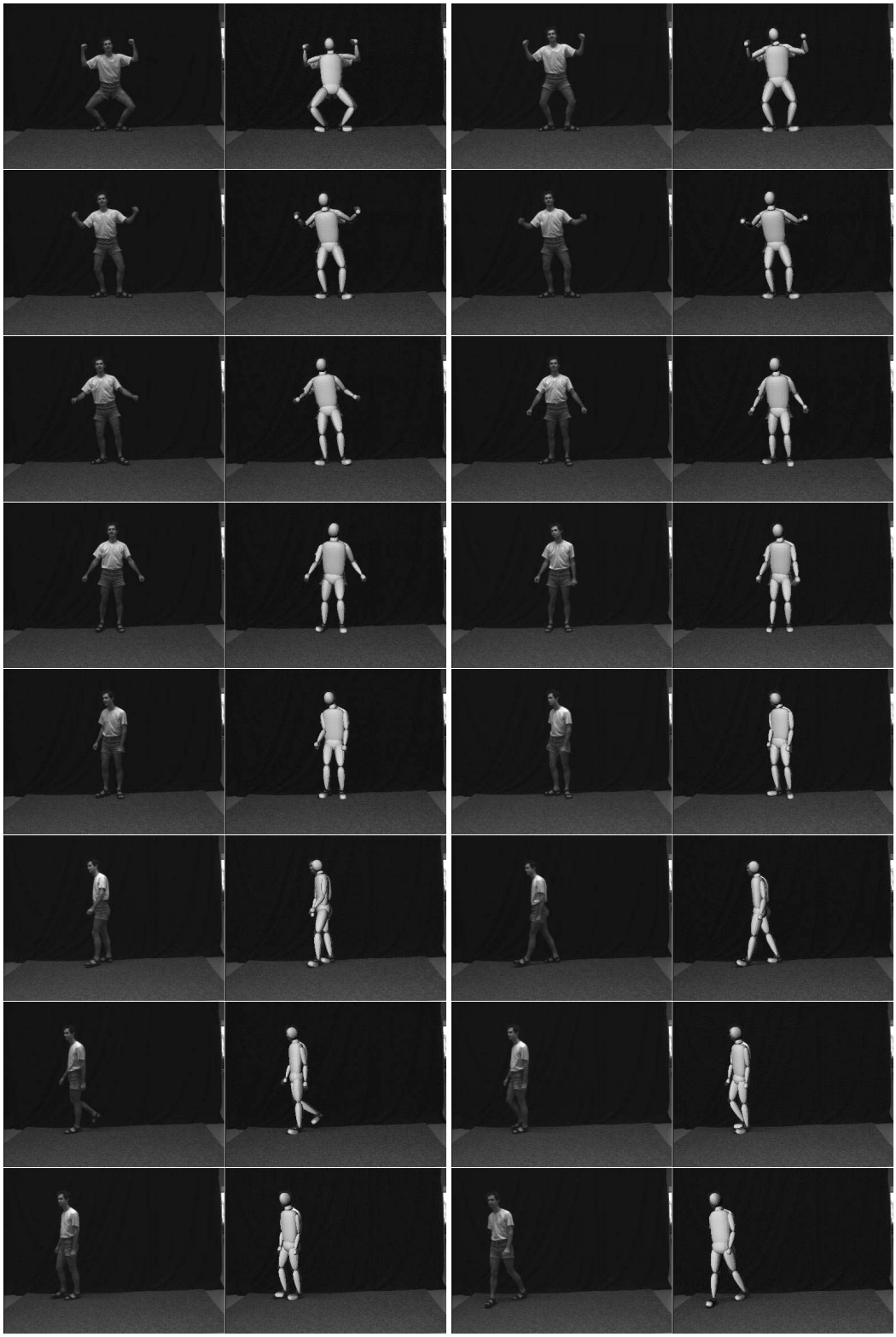


Figure 7.12: Complex motion tracking sequence detailed results for the CSS algorithm in §4.4, chapter 4.

Chapter 8

Incremental Model-Based Estimation Using Geometric Consistency Constraints

In this chapter, we present a model-based framework for incremental object shape estimation and tracking in monocular image sequences. Parametric structure and motion estimation approaches usually assume a fixed class of shape representations (splines, deformable superquadrics, *etc*.) that are initialized prior to tracking. Since the model shape coverage is fixed a-priori, incremental discovery of structure is decoupled from tracking, thereby limiting both processes in their scope and robustness. We describe a parametric model-based tracking framework that supports the automatic detection and integration of geometric primitives (lines and points) incrementally during tracking. Such primitives are not explicitly captured in the initial model, but they are moving consistently with its image motion. The consistency tests used to reveal new structure are based on trinocular relationships between geometric primitives. The method allows us to increase both the model's scope and, ultimately, its higher-level shape coverage. It improves tracking robustness and accuracy by directly employing the newly recovered features in both model forward prediction and reconstruction. The formulation represents a step towards automating model shape estimation and tracking, since it allows weaker assumptions on the availability of a prior shape representation. The method is also robust and unbiased. We demonstrate the proposed approach in two separate image-based tracking domains, each one involving complex 3D object structure and motion.

Keywords: Model-based vision, geometric constraints, object tracking, model grouping, bundle adjustment.

8.1 Introduction

Model-based estimation offers a powerful framework for recovering the object shapes and motions by fitting reduced degree of freedom models to image observations. The quest for robustness and accuracy during model estimation naturally leads to the search for geometrically and photometrically derived image features (cues), for efficient model parameterizations, and for ways to combine them in a consistent manner. Many different parameterizations and combinations of cues have been suggested, but most approaches do assume *a fixed and known* initial model *shape representation*. In this chapter, we present a technique that augments the model incrementally during tracking, by means of geometric consistency tests. These tests are used to combine heterogeneous information, relating the initial model's rigid parameters to independently tracked features in the image. The goal is a robust and flexible model accommodating both high-level and low-level representation detail whose scope that increases dynamically and automatically during the tracking process. *The technique therefore aims to bridge the gap between top-down model-based estimation techniques and bottom-up, feature-based reconstruction, by relaxing some constraints on each side. The complete model representation is no longer fully known a-priori, while knowledge of the motion of the features is acquired during reconstruction, once their identity as parts of the model is established*. The final model is a mixture of low-level features (*e.g*. lines, points) and parameterized shapes, and generalizes the shape coverage of previously-used parametric models. We show how the proposed framework naturally accommodates rigid and non-rigid high-level parametric shape representations and their associated constraints, and various model discretizations such as points and lines, by combining the simplicity of linear reconstruction methods with the refinement and bias removal of non-linear robust estimation methods.

We conclude with experiments involving objects with complex shape and motion in monocular video sequences, and show that the method is able to recover new model structure efficiently. Quantitative results support the claim that the additionally recovered structure significantly improves the accuracy and speed of the tracking process, providing important additional constraints, especially during difficult to estimate towards camera motions.

8.1.1 Relation to Previous Work

Many model-based techniques exist for representing and tracking objects: (*i*) CAD based methods (Lowe, 1987; Armstrong and Zisserman, 1995; Drummond and Cipolla, 2000) assume precise, off-line constructed rigid object models with motion estimated based on image measurements (this is done using mainly least-square techniques). Nevertheless, despite their practical effectiveness such methods are fundamentally limited to tracking known objects having fixed, non-adaptive shapes. (*ii*) Physics-based or regularization frameworks (Terzopoulos and Metaxas, 1991; Kass et al., 1988; Pentland and Horowitz, 1991; Pentland and Sclaroff, 1991; Fua and Leclerc, 1996) employ either reduced d.o.f. object parameterizations like splines or superquadrics, or point based representations

with regularized physical properties to extract and track non-rigid shape and motion. From an optimization perspective, the 'deformable' approaches are variational methods based on quadratic energy functions. They and are typically solved using integration schemes based on gradient descent to find local minima[1]. (*iii*) parametric models (Lowe, 1991; McReynolds and Lowe, 1997) are specifically built to represent certain classes of objects, but without 'physical' analogy. Apart from formulation detail, they have similar flexibility and mathematical treatment as the physics-based parametric methods: *i.e*. their structural and rigid parameters are estimated iteratively using non-linear techniques.

In this work, we shall rely on such flexible parametric models, as basic representational primitives. Nevertheless, as powerful as these techniques are, they have two important limitations:

(*i*) *Fixed Representation*: The common assumption is that the model representation is fixed and known a-priori, sometimes imposing a heavy burden on the model initialization/recovery process (Dickinson et al., 1992; Dickinson and Metaxas, 1994). Furthermore, a representational "gap" exists between the coarse, high-level parametric shapes used to model the objects, and low-level features like points, lines, corners or curved contours that can be detected in the image. It is not obvious how to bridge this gap, for example to represent object markings, discontinuities or other fine surface detail, through the inclusion of other basic geometric primitives, *e.g*., lines or planes, *etc*. It is also known (Seitz, 1998) that the diversity of parameterizations corresponding to different features at different abstraction levels usually leads to difficulties when integrating those features within a single representation or optimization procedure. Our goal here is to provide a representation that is flexible, can be estimated jointly, provides higher-level abstraction and low-level image coverage and furthermore that can be augmented as we track. The approach we follow is to combine model-based estimation techniques and feature-based reconstruction methods.

Pure feature based structure and motion estimation techniques differ in the types of correspondences (2-D to 2-D, 2-D to 3-D, or 3-D to 3-D) and features (lines, points, or corners), that they assume available simultaneously – see (Huang and Netravali, 1994) for a review. Incremental algorithms (Crowley et al., 1992; Vieville and Faugeras, 1990) reconstruct features as they become available on a per-frame basis. Batch approaches rely on non-linear least-squares techniques and assume correspondences over an entire image sequence, but do tolerate missing data. Such methods are based on the inversion of the forward model (Szeliski and Kang, 1994; Azarbayejani and Pentland, 1995; Taylor and Kriegman, 1996) and can be related to the classical relative orientation techniques (Horn, 1990).

(*ii*) *Estimation and Dimensionality*: The additional flexibility of an object representation that can adapt or deform during tracking comes at the expense of an increased number of parameters to estimate. To avoid singularities or ill-conditioning, it is important to extract complementary image cues that can induce good local minima in parameter space (corresponding to true object localization

[1]Regularization comes either implicitly from a reduced parameterization or explicitly from 'physical' properties like stiffness or damping.

in the image) and thus stabilize the estimation process. Approaches to constraint integration in a model-based framework have used contours and stereo (Terzopoulos et al., 1988), shading and stereo (Fua and Leclerc, 1996), contours and optical flow (DeCarlo and Metaxas, 1996, 1999), and shading (Samaras and Metaxas, 1998) with good results. Beyond the particular choice of sources of information to use, these approaches differ in the way they fuse information sources. Some combine the information in a symmetric manner weighting it statistically (*e.g.soft* constraints). Others favor a particular hierarchical constraint satisfaction order with an exact policy, such that inconsistent contributions to the solution from constraints further down in the hierarchy are pruned away by constraints higher-up (*hard* constraints).

The formulation we propose is based on a robust model-based probabilistic tracking process as presented in chapter 3 and 4. In the extension described here, we assume that the initial model is incomplete, and recover additional structure through the use of geometric consistency tests embedded in a tracking framework. Specifically, we integrate observations in the form of new line features discovered in the image as moving consistently with the model, with observations in the form of point features derived from a discretization of the model.

The line feature consistency tests used are based on those used in separate, bottom-up structure and motion estimation under 2-D to 2-D line correspondences in the Euclidean calibrated case (Huang and Netravali, 1994; Mitiche and Aggarwal, 1986; Yen and Huang, 1983). A later body of research in the calibrated, uncalibrated, and projective reconstruction cases (Spestakis and Aloimonos, 1991; Shashua, 1995; Hartley, 1997) derived the trilinear constraints between points and lines in 3 views, and proposed linear, bottom-up, and model-free rigid reconstruction methods (see (Hartley and Zisserman, 2000) for a comprehensive review). However, unlike these approaches, we assume an incomplete parametric (adaptive/deformable) and perhaps non-rigid articulated Euclidean model and we do not solve for the rigid parameters. On the contrary, given the model estimated rigid parameters, we only test whether the consistency conditions are indeed satisfied for arbitrarily independent tracked lines in at least three frames. Once such lines have been identified, they are reconstructed, and included in the model representation and used in the model forward prediction. Their image contributions are fused together with point-based contour and intensity observations, to construct an augmented cost function (see section 8.3.5).

8.2 Model Representation and Estimation

The proposed method is based on a MAP estimation framework as described in chapter 4 and an object representation based on superquadrics with parametric global deformations[2], as described in chapter 3. For more robust global estimates, a multiple hypothesis framework can be used, but we

[2]In one of the experiments, we actually use an articulated chain consisting of 3 such parametric shapes, a hierarchical representation similar to the one used for human body modeling.

shall not discuss this here[3]. We employ contour and intensity image observations for the basic cost function design for tracking.

8.3 Line Feature Formulation

The formulation of the tracking process in the previous sections is based on a robust estimation method where the model parameters are constrained by contour and intensity observations in the image. These observations are localized in the neighborhood of predictions for the already-known model parts. In this section, we extend this formulation by integrating new line features into the model – features that are not part of the initial model, but for which evidence exists in the image. We begin by tracking a minimal model in a sequence of images. Incrementally over time we: 1) identify line features moving consistently with the model, and 2) augment the model with these consistent features to improve its tracking. At each iteration, the model increases in scope to cover more of the image features, resulting in more robust tracking of the object.

The approach involves image-level and model-level processes. For the model, we use contour and intensity observations to estimate its rigid and non-rigid parameters. Independently, we use purely image-based techniques to detect and track lines in a sequence of frames. Such lines are called *image-tracked lines* **(ITL)**. Then, we decide whether an ITL represents a line belonging to the object (but not present in its model) by means of two geometric consistency tests, derived from tracking the ITL in at least three successive frames. The lines that pass this test are called *consistent image-tracked lines* **(CITL)**. We robustly recover their structure in terms of an underlying parameterization in a model-centered frame and we predict how a they will appear in subsequent images, based on the current estimate of the model's rigid motion. We call these predictions *model-predicted lines* **(MPL)**. We use the error between a CITL and a MPL to define additional image alignment cost terms. Their robust gradient and Hessian are estimated and combined with the ones derived from contour and intensity measurements. The initially reconstructed lines are then re-estimated jointly with the entire model structure and motion, to remove bias effects. The pipeline of the estimation process is depicted in fig. 8.1 on page 132.

8.3.1 Line parameterization

In the following treatment, we denote a line in 3-D by lower-case letters, $l_i(i = 1..n)$, and its corresponding projected line (or segment) in the image plane by capital letters, $L_i(i = 1..n)$. A 3-D line is parameterized by a unit vector **v**, representing one of two possible directions on the 3-D line l, and a vector **d** terminating on l and perpendicular to it (see fig. 8.2a). This line representation

[3]For the clarity of the mathematical treatment we present only the core of the method that involves detecting and reconstructing new un-modeled image features, and building modified cost functions that include them. We then perform robust continuous MAP estimates for model parameters. For more robust solutions in practical applications, using the multiple hypotheses framework in chapter 4 should be straightforward.

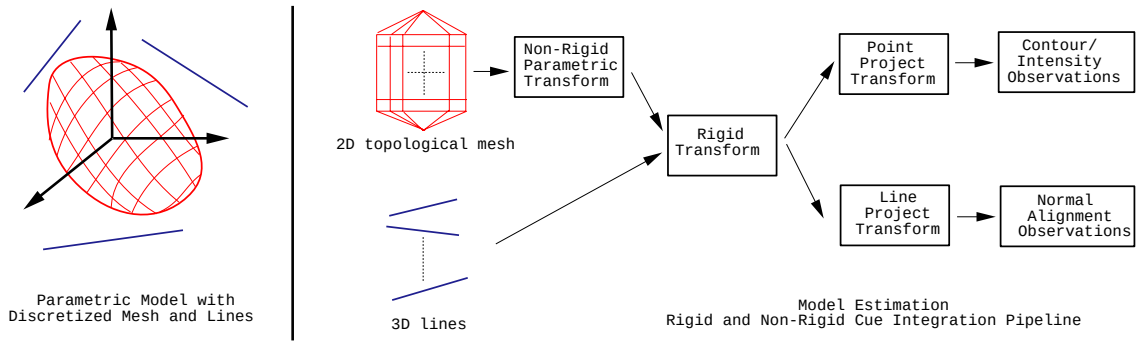
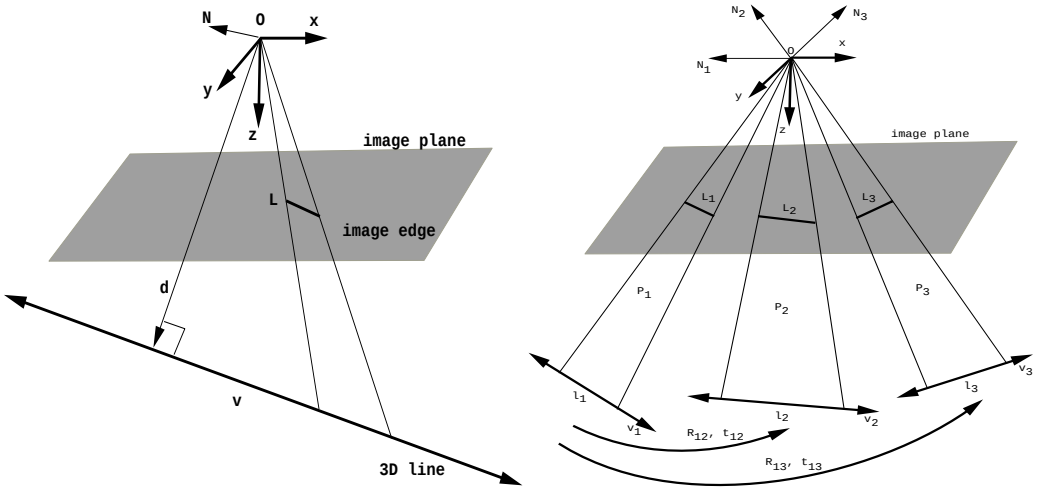


Figure 8.1: Estimation Pipeline

forms a 6-dimensional parameter space with 4 degrees of freedom. Consequently, the underlying relation between the line primitives can be identified as a 4-dimensional manifold embedded in the abstract parameter space, and any line can be identified with a point (actually two) on this manifold (Kanantani, 1996; Taylor and Kriegman, 1996).



(a) Line Pameterization (b) Line Motion in Three Frames

Figure 8.2: Line Representation and Motion

The 3-D line, l, and the optical center of the camera determine a plane (the line's interpretation plane) with normal, $\mathbf{N} = (N_x, N_y, N_z)^\top$. This plane intersects the image plane, defined by the equation $z = f$ (f being the camera focal length) at line L. The equation of line L in the image plane can be written as:

$$N_x X + N_y Y + N_z f = 0 \tag{8.1}$$

The above relation allows the normal of the plane containing a 3-D line in space to be recovered

from the image plane equation of its projected (observed) line, or segment. In vector notation, given a plane, $\mathbf{P} = (N_x, N_y, N_z, 0)^\top = (\mathbf{N}^\top, 0)^\top$, and any point belonging to the interpretation plane of a given line, $\mathbf{X} = (X, Y, Z, 1)^\top$, we can write the plane equation as $\mathbf{P}^\top \mathbf{X} = \mathbf{0}$.

8.3.2 Model-Based Consistency Tests

The standard formulations for recovering motion and structure using 2-D to 2-D line correspondences (e.g., (Huang and Netravali, 1994; Yen and Huang, 1983; Mitiche and Aggarwal, 1986)) rely on three frames and at least six line correspondences (although no formal proof is yet available, (Huang and Netravali, 1994)) for uniquely recovering the structure and motion of a rigid object[4]. For the two view case, the resulting system of equations does not constrain the motion at all – there is a consistent structure for any image lines and any motion.

Consider now the motion of a 3-D line, l, in three successive frames ($l_i, i = 1, 2, 3$, with direction support $\mathbf{v_i}, i = 1, 2, 3$) and it's corresponding image projected lines ($L_i, i = 1, 2, 3$). The motion between frames 1 and 2 is described by the translation and rotation, $\mathbf{t}_{12}$ and $\mathbf{R}_{12}$, and for frames 1 and 3, by $\mathbf{t}_{13}$ and $\mathbf{R}_{13}$, respectively. The corresponding normals for the interpretation planes, P_1, P_2, P_3, determined by the line L and the center of projection in the three frames, are $\mathbf{N_1}, \mathbf{N_2}, \mathbf{N_3}$, respectively (see fig. 8.2b).

The following two relations can be derived either geometrically or algebraically, from the quantities presented above (Huang and Netravali, 1994; Mitiche and Aggarwal, 1986; Yen and Huang, 1983):

$$\mathbf{N_1} \cdot (\mathbf{R}_{12}^{-1}\mathbf{N_2} \times \mathbf{R}_{13}^{-1}\mathbf{N_3}) = 0 \tag{8.2}$$

$$-\mathbf{t}_{12} \cdot (\mathbf{R}_{12}\mathbf{N_1}) = \frac{\|\mathbf{N_2} \times \mathbf{R}_{12}\mathbf{N_1}\|}{\|\mathbf{N_2} \times \mathbf{R}_{23}^{-1}\mathbf{N_3}\|} \cdot \mathbf{R}_{23}^{-1}\mathbf{t}_{23} \cdot \mathbf{R}_{23}^{-1}\mathbf{N_3} \tag{8.3}$$

In this model-based formulation, we are not interested in solving for the rotation and translation in a bottom-up manner, but given a model with *known* motion and some *independent* ITLs, to *verify* whether those lines are moving consistently with the model (CITLs). More specifically, given an ITL in three frames (that is, knowing $\mathbf{N_1}$, $\mathbf{N_2}$ and $\mathbf{N_3}$) as well as the motion of the model (that is, $\mathbf{R}_{12}$, $\mathbf{t}_{12}$ and $\mathbf{R}_{13}$, $\mathbf{t}_{13}$), the equations (8.2) and (8.3) are used to test whether the motion of the line is consistent with the model's motion. If so, we hypothesize that the line is part of the object and therefore should be added to the model. It can be proven that the above two relations are necessary and sufficient conditions for consistency[5]. In practice, consistency relations are never satisfied exactly, due to slight errors in the image tracked lines and errors in the estimated model rigid parameters, so a threshold must be chosen. At present, we have selected it based on the

[4] Within a scale factor for translation and structure parameters.

[5] As mentioned before, a 3D line has 4 intrinsic degrees of freedom while a projected image line has just 2. Measurements are collected in 3 frames, so this will determine 3x2-4=2 independent relations.

covariance of both model rigid parameters estimation and the image tracked lines, respectively, but we acknowledge that more work needs to be done here.

8.3.3 Robust Unbiased Model-Based Structure Recovery

Once a moving image line has been assigned to the model through the above consistency tests, the next step is to recover its structure, i.e., the vector pair $(\mathbf{v}, \mathbf{d})$ in a model-centered coordinate system. In order to increase the robustness of the recovery process, one can use as many line correspondences in as many frames (at least two) as are available. The process can be formulated as follows: all interpretation planes for the line correspondences in a camera frame are transformed to a common, model-centered coordinate frame. Each line, l_i, having the interpretation plane, $\mathbf{P_i}$, is subject to the displacement, $\mathbf{D_c^{-1}D_i^{-1}}$, where:

$$\mathbf{D_c} = \begin{bmatrix} \mathbf{R}_c & \mathbf{t}_c \\ 0 & 1 \end{bmatrix} \tag{8.4}$$

is the displacement corresponding to the camera, and $\mathbf{D_i}$ is the displacement corresponding to the model (in the world coordinate system) in frame i. Then, the equation of the plane in the object-centered frame is: $\mathbf{P_i}^\top \cdot \mathbf{D_c^{-1}D_i^{-1}} \cdot \mathbf{X} = \mathbf{0}$. By stacking together the equations for corresponding lines, we obtain:

$$\mathbf{A} \cdot \mathbf{X} = \begin{bmatrix} \mathbf{P_1}^\top \cdot \mathbf{D_c^{-1}D_1^{-1}} \\ \mathbf{P_2}^\top \cdot \mathbf{D_c^{-1}D_1^{-1}} \\ \cdots \\ \cdots \\ \cdots \\ \mathbf{P_k}^\top \cdot \mathbf{D_c^{-1}D_k^{-1}} \end{bmatrix} \cdot \mathbf{X} = \mathbf{0} \tag{8.5}$$

Since all the planes should intersect in a common line, the above [k x 4] matrix $\mathbf{A}$ should have rank 2. Any point $\mathbf{p}$ on the intersecting line can be written as a linear combination of the singular vectors corresponding to the 2 smallest singular values of the matrix $\mathbf{A}$:

$$\mathbf{p} = a \cdot \mathbf{X_{s1}} + b \cdot \mathbf{X_{s2}} \tag{8.6}$$

The corresponding line representation can be recovered as:

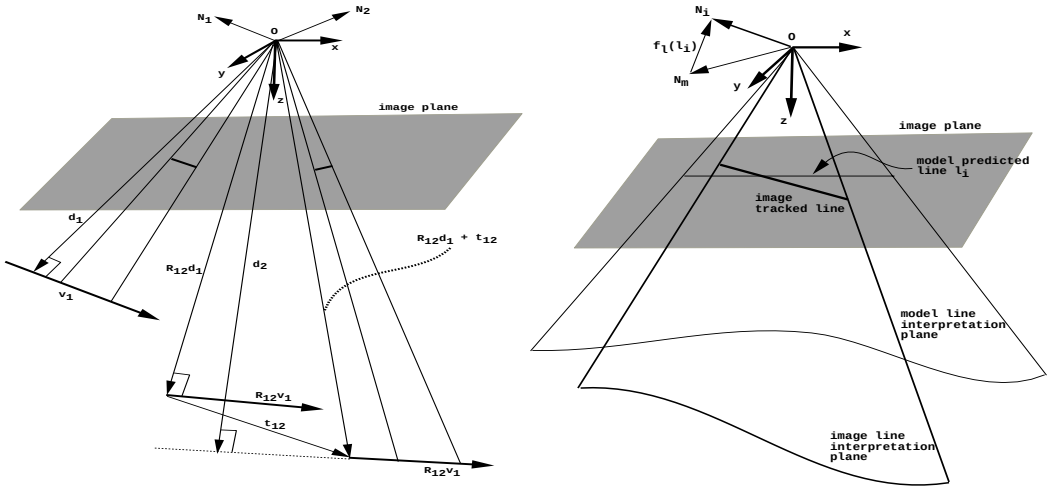
$$\mathbf{v} = \mathbf{X_{s1}} - \mathbf{X_{s2}} \qquad \mathbf{d} = (\mathbf{I} - \frac{\mathbf{v} \cdot \mathbf{v}^\top}{\| \mathbf{v} \|^2}) \cdot \mathbf{X_{s1}} \tag{8.7}$$

The stability of the reconstruction can be verified in terms of the ratio of the 2nd and 3rd singular values of $\mathbf{A}$ (remember that in the noise free case the last two singular values should be zero), being satisfactory when this ratio is high. The above linear method, although robust, might be prone to bias in the initial line parameter estimates, due to fixed rigid displacements. However, we work in

a non-linear parametric model framework, in which *both* model structure *and* its motion are jointly estimated based on robust and statistically meaningful error norms. Consequently, any initial linear bias is automatically removed during estimation in subsequent frames.

8.3.4 Forward Model Line Prediction

Once consistent lines (CITL) have been identified, recovered, and effectively added to the model, they are used to improve the tracking of the augmented model by imposing further constraints on its alignment with the data. The approach proposed here is to define alignment residuals between a CITL in the image and the projection (MPL) of its corresponding (new) model line, as predicted under the forward motion of the model.



(a) Forward Line Transfer (b) Alignment Error

Figure 8.3: Lines transfer and alignment

Consider the two frame case, as illustrated in fig. 8.3a. Given $(\mathbf{N_1}, \mathbf{v_1}, \mathbf{d_1})$, one can obtain $(\mathbf{N_2}, \mathbf{v_2}, \mathbf{d_2})$ by geometric means as follows:

$$\mathbf{v_2} = \mathbf{R}_{12}\mathbf{v_1} \tag{8.8}$$

$$\mathbf{d_2} = (\mathbf{R}_{12}\mathbf{d_1} + \mathbf{t_{12}}) - \mathbf{v_2}((\mathbf{R}_{12}\mathbf{d_1} + \mathbf{t_{12}}) \cdot \mathbf{v_2}) \tag{8.9}$$

$$\mathbf{N_2} = \frac{\mathbf{v_2} \times \mathbf{d_2}}{\| \mathbf{v_2} \times \mathbf{d_2} \|} \tag{8.10}$$

As mentioned in the introductory section, $\mathbf{N_2}$ fully identifies the MPL in the second frame.

The Jacobian matrix associated with the mapping between the adaptive line structure after a rigid motion has been derived analytically (but it is omitted here as such derivations are bulky

and non-informative). Given a line representation, $\mathbf{u}_l = (\mathbf{v}^\top, \mathbf{d}^\top)^\top$, and the rigid displacement parameterization $\mathbf{x}_D$, of the model, we assemble the parameters of interest[6] as: $\mathbf{x}_l = (\mathbf{u}_l^\top, \mathbf{x}_D^\top)^\top$. The corresponding Jacobian matrix of the line parameters with respect to the model parameters of interest is a [6x13] matrix:

$$\mathbf{J}_{\mathbf{u}_l} = \frac{\mathbf{d}\mathbf{u}_l}{\mathbf{d}\mathbf{x}_l} \tag{8.11}$$

8.3.5 Image Line Observations

Given an MPL, identified by equation (8.1), and a CITL, identified by its endpoints $\mathbf{P_1} = (X_1, Y_1)$ and $\mathbf{P_2} = (X_2, Y_2)$, we define 2-D residuals to compensate for their alignment error. Since the CITL and MPL each define an interpretation plane, we define the line alignment error in terms of the alignment of their corresponding interpretation planes (see fig. 8.3b). Given $\mathbf{N_i}$, the normal of the CITL interpretation plane, and $\mathbf{N_m}$, the normal of the MPL interpretation plane, the cost $f_{l_i}(\mathbf{x})$ corresponding to the alignment error $\Delta\mathbf{N}_i(\mathbf{x}) = \mathbf{N}_i - \mathbf{N}_m$, can be written as:

$$f_{l_i}(\mathbf{x}) = \frac{1}{2}\rho_i(\Delta\mathbf{N}_i(\mathbf{x})\mathbf{W}_i\Delta\mathbf{N}_i(\mathbf{x})^\top) \tag{8.12}$$

Notice that a final transformation maps a line representation to an observable "normal" on which the line alignment is actually performed. The corresponding transformation is given by equation (8.10), and involves the computation of another [3x6] Jacobian matrix:

$$\mathbf{J}_N = \frac{\mathbf{d}\mathbf{N}}{\mathbf{d}\mathbf{u}_l} \tag{8.13}$$

The Jacobian $\mathbf{J}_l = \frac{\mathbf{dN}}{\mathbf{dx}_l}$ corresponding to the mapping between the line representation in the model frame and the observation (the normal of the corresponding line interpretation plane) is computed via the chain rule in terms of Jacobians (8.11) and (8.13).

The corresponding gradient and Hessians for contour, intensity and line observations can now be derived using (3.8) and (3.15):

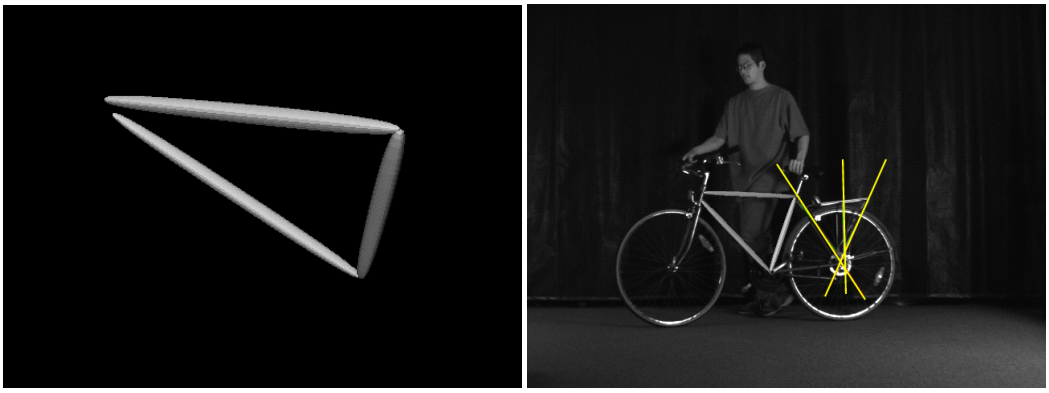
$$\mathbf{g} = \mathbf{g}_c + \mathbf{g}_I + \sum_i \mathbf{J}_{l_i}^\top \rho_i' \mathbf{W}_i \Delta\mathbf{N}_i \tag{8.14}$$

$$\mathbf{H} \approx \mathbf{H}_c + \mathbf{H}_I + \sum_i \mathbf{J}_{l_i}^\top (\rho_i{}' \mathbf{W}_i + 2\rho_i'' (\mathbf{W}_i\Delta\mathbf{N}_i)(\mathbf{W}_i\Delta\mathbf{N}_i)^\top)\mathbf{J}_{li} \tag{8.15}$$

8.4 Experiments

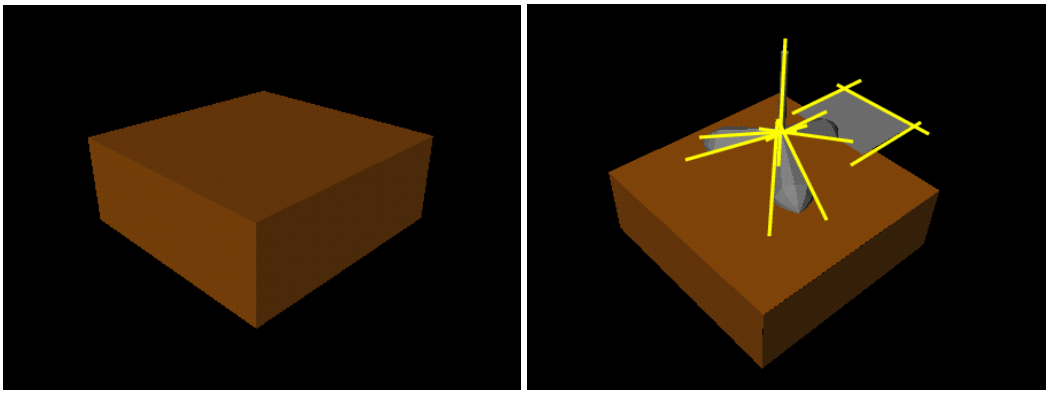
The experiments consist of two sequences, each containing of 4 seconds of video (200 frames recorded at 50 fps) of a moving bike and of a space robotics end-effector grapple fixture. Both

[6] These are parameters that are estimated with contributions from image line observations, namely 7 parameters to represent rigid displacements using quaternions for rotations and 6 parameters corresponding to the 3D line parameterization in the reference frame.



(a) (b)

Figure 8.4: Initial and Reconstructed Model Tracking



(a) (b)

Figure 8.5: Initial and reconstructed model, with new superquadrics parts fitted

sequences involve significant translational and rotational motion in the camera frame. Part of the bike frame is modeled and tracked by means of a model consisting of 3 parts (fig. 8.4a), while the grapple fixture consists of only a square parallelepiped (fig. 8.5a). Newly recovered models are displayed in fig. 8.4b and fig. 8.5b, in which additional parametric shapes have been fitted based on 3-D reconstructed model lines, their covariance, and their corresponding image contours (an algorithm for such higher-level shape-recovery and grouping is beyond the scope of this work, and is subject to further investigation). Note that tracking based only on a rigid model consisting of the square parallelepiped in fig. 8.8 fails. This is due to an incorrect monocular shape and pose initialization that appeared to be correct during the initial part of the sequence. Nevertheless, it turns out that the model shape recovery based on only one image was inaccurate and, as expected,

the non-adaptive model gradually drifted and ultimately failed to track rigid motion in frame 208, and this can be clearly observed in fig. 8.8.

For the adaptive tracking case, the models have been manually initialized (as before) in the first frame of the sequence, and are displayed in subsequent frames aligned (overlaid and rendered in flat-shaded gray for the bike and wireframe (brown on color plates) for the grapple fixture, respectively) with the part of the object they model (for tracking results, see fig. 8.7 and fig. 8.9). Additional lines, not part of the initial models, are tracked through the sequence by means of an independent gradient-based line tracker, involving interest point-based tracking, followed by line fitting in each frame. Lines that are determined to be CITLs are displayed in green and are visible mostly on color plates. The model reconstructed/predicted lines are displayed in light yellow. The consistency test is evaluated, using a threshold $\tau = 0.05$, based on the lines present in frames (20, 40, 60) for the bike sequence. For the space shuttle sequences, lines are incrementally tracked and recovered (some are not visible initially) and the consistency tests are performed several times for different lines, in the frames (20, 40, 60), (120, 140, 160) and (160, 175, 190).

It is important that the estimation of the models' rigid motion is accurate as it affects both the validity of consistency tests and the quality of the line reconstruction. In practice, since the models are initialized manually, there is always uncertainty associated with their initialization. For this reason, we do not consistency test and reconstruct lines immediately after model initialization, but rather allow a slight delay (around 20 frames in our experiments), such that the model locks onto the data.

Any initial *bias* in the linear reconstruction is subsequently eliminated through the re-estimation of the line representation jointly with all the rigid and non-rigid model parameters in the non-linear framework used.

The least-squares-based reconstruction for the bike sequence is based on 12 frames within the interval, 20-60, while for the grapple fixture, we use 12 frames in 20-60, 120-160, 160-190, respectively. In the experiments we did, this provided sufficiently important motion and sufficient additional lines for accurate, constrained reconstruction. The stability of the reconstruction is checked, as explained in §8.3.3 by the ratio of the 2nd and 3rd singular values associated with the $\mathbf{A}$ matrix. We thus employ a principled criteria for selecting sufficient lines and corresponding informative displacements for the reconstruction. In both sequences, the frames prior to new line reconstruction are tracked using the prior model along with contour and intensity-based image residuals, while the rest of the sequence (once the CITLs have been reconstructed) is tracked using the enhanced model (consisting of both the initial model *and* the reconstructed lines, and using *both* contour, intensity and alignment residuals between CITLs and MPLs).

Although no lines are reconstructed until frame 60 in the bike sequence, their re-projection is displayed over the entire sequence (note that since they are reconstructed in a model-centered coordinate frame, they can be transferred backwards, since the inter-frame motion of the model has already been estimated). For the grapple fixture sequence, we do not plot the lines over the entire

sequence in order to show how they are incrementally tracked and recovered. Notice that for both sequences, the reconstructed lines' projections are correct.

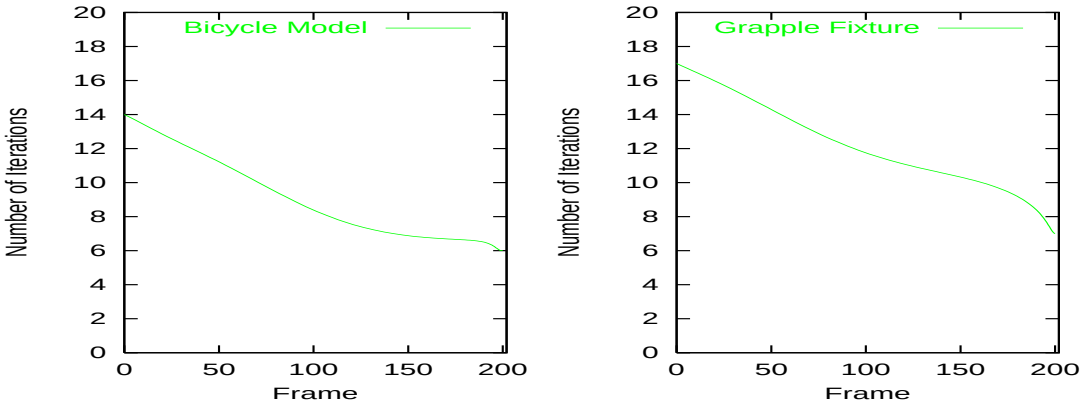


Figure 8.6: Number of iterations per frame (Bezier interpolation) decreases due to the addition of line constraints.

The tracking results emphasize increased stability, as the line features are reconstructed and integrated into the tracking process as additional cues, especially during difficult to track towards camera motion of the grapple fixture. The number of iterations in the robust non-linear refinement loop decreases with the integration of additional line cues (see fig. 8.6 for plots). The average per-node error decreased from 0.9 (initially) to 0.4 pixels (frame 60) in the bike tracking sequence, and from 1.2 pixels (initially) to 0.8 (frame 60), 0.6 (frame 160) and 0.2 (frame 190) in the grapple fixture sequence, respectively, reflecting improved accuracy due to the integration of line soft constraints in the model estimation process.

8.5 Conclusions

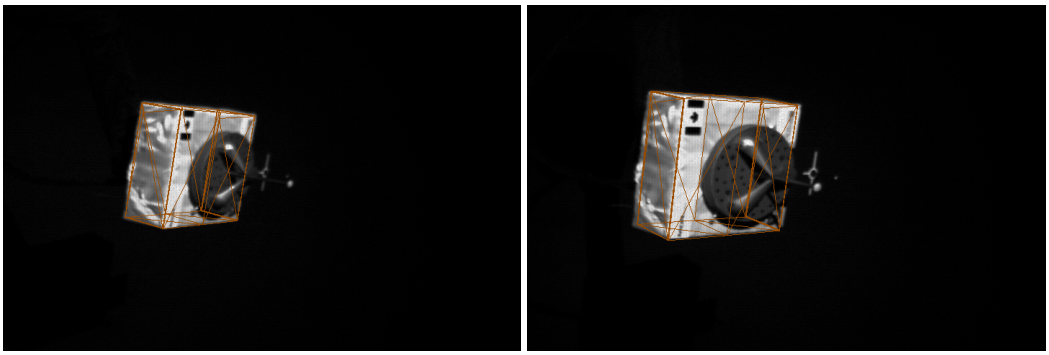
In this chapter, we have presented a framework for incremental model acquisition and tracking using parametric models. We have relaxed the constraint that the representation of the model has to be fully known a-priori, and have enhanced its basic discretization structure in terms of points and lines while preserving higher-level shape information, in terms of parametric shapes. This has allowed us to formulate geometric constraints for feature consistency, reconstruction, and tracking via cue integration in a flexible manner through the integration of top-down, unbiased and robust non-linear model-based estimation techniques and bottom-up, linear feature reconstruction techniques. Our experimental results show that the proposed method is able to deal with objects with complex shape and motion, and to track and incrementally recover structure with improved accuracy. Potential research directions may include more rigorous quantitative evaluation of the reconstruction and tracking improvement results, as well as grouping and high-level model abstraction.

(a) frame 0 (b) frame 40

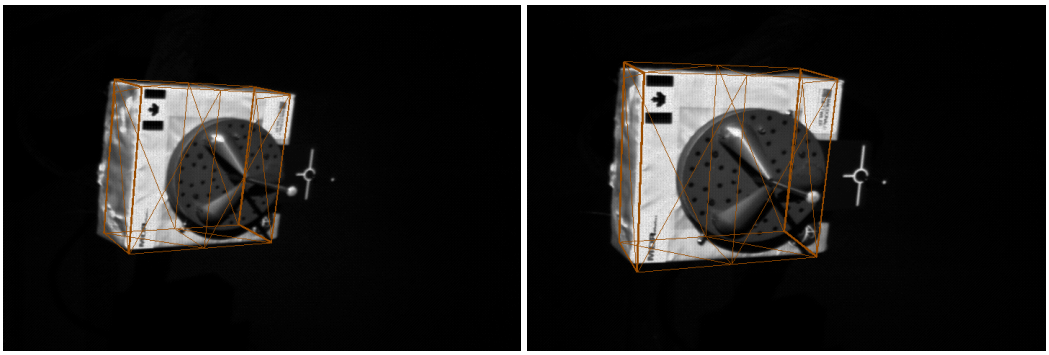
(c) frame 80 (d) frame 130

(e) frame 170 (f) frame 200

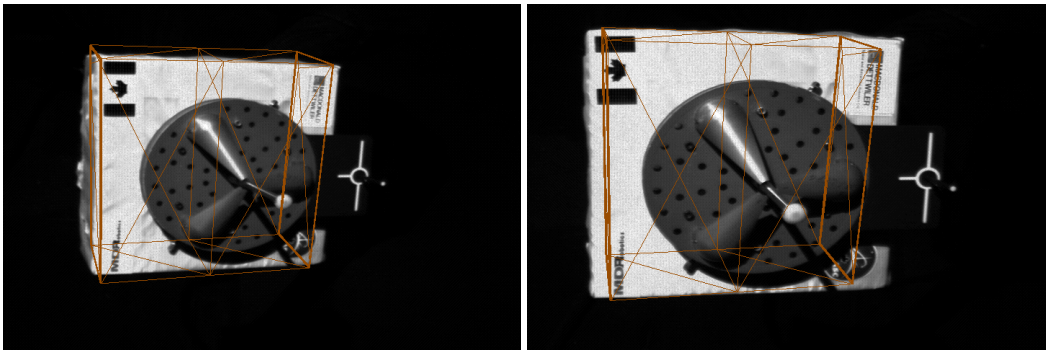
Figure 8.7: Bike Model Tracking (the model frame in grey) with MPLs (light yellow). CITLs (green) are also visible on color plates.



(a) frame 60 (b) frame 100

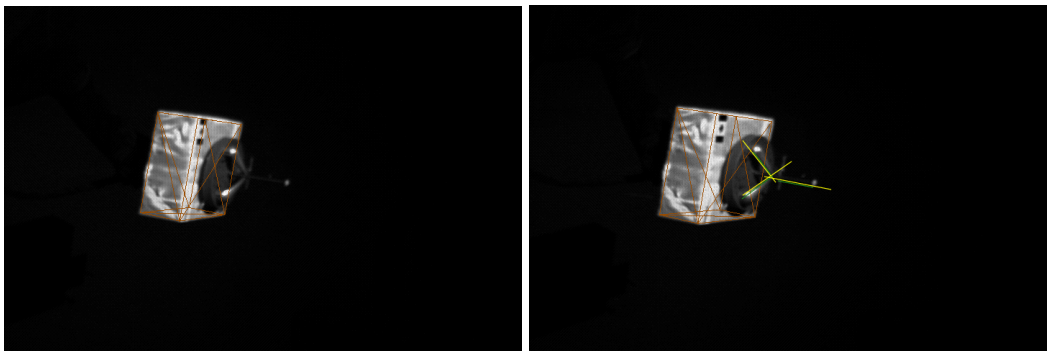


(c) frame 130 (d) frame 160



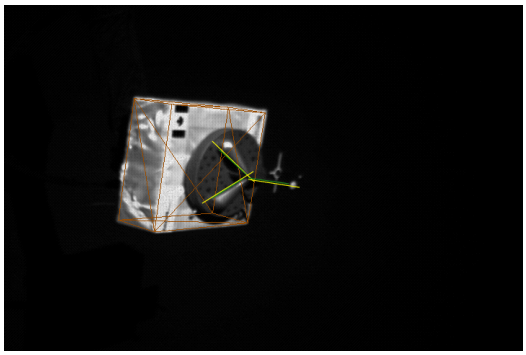
(e) frame 180 (f) frame 208

Figure 8.8: Using an uncertainly initialized rigid and incomplete model (wireframe, in brown) leads to tracking failure.

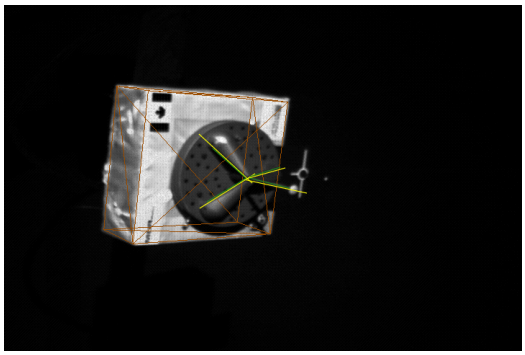


(a) frame 0, no reconstructed lines

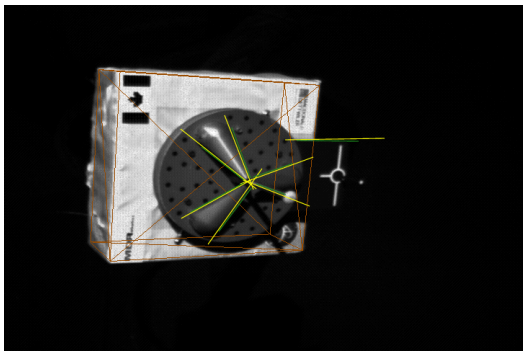
(b) frame 40, 3 reconstructed lines



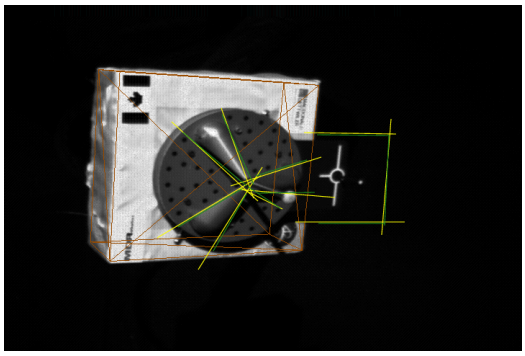
(c) frame 80, 3 reconstructed lines



(d) frame 120, 4 reconstructed lines



(e) frame 160, 7 reconstructed lines



(f) frame 190, 10 reconstructed lines

Figure 8.9: Grapple fixture model tracking (the grapple fixture model in wireframe, in brown) with MPLs (light yellow). CITLs (green) are also visible on color plates.

Chapter 9

Perspectives and Open Problems

> It's not the consequence that makes a problem important, it is that you have a reasonable attack.
>
> Richard HAMMING
> *You and Your Research*

This chapter concludes the thesis with an overview of open research directions, not only for human tracking, but for model-based vision in general. We believe that vision systems with predictable behavior can only emerge through a flexible combination of learned and preprogrammed representations, coherent modeling and efficient estimation methods.

9.1 Initialization Methods

Many estimation problems in vision, and particularly the high-dimensional ones we have addressed in this thesis, are non-linear and highly non-convex. It is therefore necessary to efficiently localize multiple minima and if required to track them through time, and this often requires favorable initial estimates within their basins of attraction. As we explained in the previous chapter, building likelihoods that are well behaved (coherent, with non-singular peaks having large attraction zones, *etc*.) is difficult for human tracking and similar model-based vision applications. It is therefore important to derive flexible methods that can provide object localization, and thus allow both initialization and recovery from tracking failure. Several directions could be explored here: (*i*) RANSAC-like combinatorial fitting schemes loop over possible sub-sets of model-image assignments and provide the set of estimated model parameters with the largest consensus among the choices tried. The first difficulty is that for multimodal problems, there are often multiple solutions even under an exact set of assignments, so fast non-linear fitting is non-trivial (*e.g*. the pose recovery from joint correspondences discussed in previous chapters, *etc*.). The second problem is that consensus methods tend

to be computationally expensive and not very reliable for the type of spaces encountered in human modeling and tracking[1]. (*ii*) Multi-resolution (coarse to fine, hierarchical) initialization schemes may also prove effective and play an important role. For instance, suppose that potential 2D human responses are detected using a SVM as in (Papageorgiu and Poggio, 1999) followed by subsequent approximate body labeling and extraction of a rough 2D cardboard model using body parts adjacency constraints as in (Ioffe and Forsyth, 2001; Ronfard et al., 2002). Finally, suppose that a higher granularity cost surface is built in terms of model to image limb correspondences. In this way, the computations remain efficient at each processing stage (*e.g.* for low dimensional representations like 2D human detectors, exhaustive search over image is still feasible, and so is body part labeling and 2D skeleton construction using dynamic programming, *etc.*).

9.2 Modeling Highly Non-Linear Manifolds

Modeling complex and highly variable objects like articulated or deformable structures (*i.e.* muscles, clothing) remains a major challenge. It is very difficult to hand-craft representations that are intuitive, that provide good coverage and binding with the images and that remain efficient to estimate. 'Smart' highly-structured generative models, that incorporate both pixel and shape layers, are desirable. In fact, the process of recovering model representations can be re-formulated as a large-scale learning process that estimates a manifold with intrinsic dimension that can synthesize or explain the variability of structures present in the application domain. Ideally we want to learn the topology and properties of such a space in terms of its geometric, photometric and dynamic features. Approaches based on such ideas have recently emerged for low and intermediate level vision (*e.g.* FRAME, (Zhu, 1999)) but they are still extremely complex computationally and it is not clear to what extent estimation methods based on pure MCMC can be effective here. One therefore has to use either sufficient statistics to constrain the model, or efficient search to identify alternative good-cost configurations for weak models. In other words any local minimum might be good for structured, well-constrained models, but several competing minima should be found when estimating weak models.

Another open question is how to find a middle-ground between pre-programmed and learned representations in order to reduce the estimation complexity and preserve their invariance properties[2]. An interesting step in this direction for 2D models is (Toyama and Blake, 2001).

[1] In order to diminish the search complexity one may be able to construct cost surfaces that are assembled from higher-level features which have a coarser minimum granularity and a more tractable, easier to enumerate, combinatorial assignment structure (see below).

[2] Being purely example-based, many manifold representations are not-invariant to simple transformations like translation or rotation in the input set. To achieve a limited invariance, replicated training sets (rotated, scaled, *etc.*) are often generated in the learning phase.

9.3 Quasi-Global Optimization Algorithms

The optimization algorithms proposed in this thesis represent a first step towards the systematic quasi-global exploration of the likelihood surfaces encountered in vision problems. Nevertheless, we believe that improved sampling based optimization and approximation schemes can be derived, that exploit the structure of the likelihood surface to accelerate the search process. Many vision problems (e.g. tracking) have a strong temporal coherence structure. Multi-modality is still an important issue, but temporal coherence partially simplifies the problem by allowing the search to focus on spatially nearby minima. It is therefore important to be able to efficiently locate minima neighboring the current one(s).

9.4 Coherent, High Level Likelihood Construction

Despite our efforts, modeling image similarity and matching model features to image ones, remains a major challenge. It would therefore be useful to be able to build globally more consistent likelihood surfaces by finding tractable approximations that account for the higher order couplings of the configurations involved in the data association process. The aim is a cost surface with a moderate number of 'interpretable' minima that can be searched efficiently. Most cost surfaces used today can potentially have exponentially many minima in the number of model-image correspondence choices and do not explicitly enforce high-level coherence.

A different approach would be to learn visual descriptors and fuse likelihoods for object localization and classification. The main requirements here are the efficient learning of descriptors from labeled and unlabeled data and good inter-class discrimination while retaining invariance to changes of viewpoint, occlusion, lighting and intra-class variation. It might be also useful to learn patterns of higher level feature couplings (*e.g.* by means of auto-associator networks that achieve 'pattern completion' on low level input features).

9.5 Prior Distributions and Model Selection

Many inverse problems in vision have parameters that are badly controlled by the observations, so a prior distribution or a simplified model are needed to stabilize the solution. Nevertheless, obtaining a representative prior can be difficult, and there is a risk that it will introduce a bias and increase the connectivity of the problem and thus its difficulty. Alternatively, model selection or constraint enforcement techniques do exist, but automatically deciding when to use them and comparing results obtained with different models is not always easy, and further developments of such methods are needed.

Priors can be also used to initialize non-linear algorithms, by learning 'soft' distributions over inverse image to model mappings.

9.6 Combining Visual Cues and Adaptive Attention Mechanisms

Many frameworks, including the Bayesian one, accommodate the fusion of different cues or cost terms (*e.g*. edge, shading, stereo or motion). But deciding which cues are reliable, using them adaptively and combining their (sometimes conflicting) results is still an unsolved problem. Usually, the contributions are fused based on inverse covariances, but, as in the case of prior distributions, their simultaneous usage can introduce additional difficulties by increasing the degree of multi-modality or ambiguity in the cost surface.

9.7 Evaluating and Balancing Bayesian Actors

Many different modeling and estimation frameworks exist in vision (Bayesian, pattern theory, energetic) and as explained in the introduction, they can be re-phrased in terms of similarity measures, priors and search. Nevertheless, it remains an open problem how these three essential components should be combined in each specific context, as most of the time their relative importance is only known qualitatively. For instance, if we want to localize an object we don't usually know what prior knowledge use for its structure or motion, what image descriptors to use for the likelihood construction, how to set suitable search bounds in a high-dimensional object parameter space and finally under what conditions the object is detectable or not. Advances on this front would be not only of theoretical interest, but also practically useful in the performance evaluation of vision algorithms.

9.8 Estimation and On-line Learning in Graphical Models

A long term challenge in vision is to be able to extract structure starting from bare images, in a purposive fashion, based only on a very weakly pre-programmed architecture. Ultimately, we want to learn on-line both the hidden structure of the models and their parameters, in an unified, flexible and automatic estimation process. Traditionally, models are built partially or totally off-line, based on heuristic domain intuitions and their structure is fixed a-priori. This always limits their adaptability to real data and their robustness in specific application contexts. Conversely, 'flexible' approach greatly augments the difficulty of the estimation problem in terms of the number of parameters local minima, and different model types (*e.g*. a mixed continuous/discrete composition, *etc*.). It is therefore not at all obvious what search algorithms and image constraints are needed to make on-line model estimation feasible. Recent variational methods are based on structural approximations obtained by decoupling (some of) the dependencies between the variables of a structured (*e.g*. graphical) model and exploitation of the simplified structure in order to compute a local minimum. Nevertheless, it is currently unclear how much errors are induced by the approximations and finding only a local minimum on such under-constrained cost surfaces is not likely to be very satisfactory for consistency or robustness.

9.9 Conclusions

The goal of this thesis was to investigate the feasibility of three-dimensional human motion reconstruction from monocular video sequences. We proposed a modeling framework able to deal with the high dimensionality and physical constraints of human bodies, and also several estimation methods for recovering the distribution of human poses over time sequences. We obtained results on tracking unconstrained motions in clutter and we also derived three deterministic and stochastic multiple minimum location and sampling algorithms that allow temporal tracking and analysis of the multimodal high-dimensional human pose probability surface: **Covariance Scaled Sampling**, **Eigenvector Tracking** and **Hypersurface Sweeping** and **Hyperdynamic Importance Sampling**. The search methods give general, principled approaches to the exploration of the non-convex error surfaces so often encountered in computational vision problems.

We believe that this research represents a step towards robust and efficient computational vision systems for interpreting human motion and for a broad range of related model-based vision applications.

Appendix A

Superquadrics and Deformations

Basic superquadric ellipsoid representation:

$$\mathbf{s}(\mathbf{q_s}, \mathbf{u}) = a \begin{pmatrix} a_x {C_u}^{\epsilon_1} {C_v}^{\epsilon_2} \\ a_y {C_u}^{\epsilon_1} {S_v}^{\epsilon_2} \\ a_z {S_u}^{\epsilon_1} \end{pmatrix}, \tag{A.1}$$

where:

- $\mathbf{q_s} = (a_x, a_y, a_z, \epsilon_1, \epsilon_2)$, $\mathbf{u} = (u, v)$
- $u \in [-\pi/2, \pi/2], v \in [-\pi, \pi]$
- ${S_w}^{\epsilon} = sgn(\sin w)|\sin w|^{\epsilon}$, ${C_w}^{\epsilon} = sgn(\cos w)|\cos w|^{\epsilon}$
- $0 \leq a_x, a_y, a_z \leq 1$ are aspect ratio parameters, a is a scale parameter
- $\epsilon_1, \epsilon_2 \geq 0$ are "squareness" parameters

Consider parameterized deformation functions of the form:

$$\mathbf{T} : \mathbf{R^q} \times \mathbf{R^3} \mapsto \mathbf{R^3} \tag{A.2}$$

$$\mathbf{x}(\mathbf{q}; \mathbf{u}) = \mathbf{T}(\mathbf{q_T}; \mathbf{s}(\mathbf{q}, \mathbf{u})) \tag{A.3}$$

Tapering and bending deformations:

$$\mathbf{s}(\mathbf{q_{tb}}, \mathbf{e}) = \begin{pmatrix} (\frac{t_1 e_3}{a_0 a_3 w} + 1)e_1 + b_1 cos(\frac{e_3 + b_2}{a_0 a_3 w}\pi b_3) \\ (\frac{t_2 e_3}{a_0 a_3 w} + 1)e_2 \\ e_3 \end{pmatrix} \tag{A.4}$$

where:

- $-1 \leq t_1, t_2 \leq 1$ are tapering parameters in principal axis 1, 2
- b_1 defines the bending magnitude
- $-1 \leq b_2 \leq 1$ defines location on axis 3 where bending is applied
- $0 \leq b_3 \leq 1$ defines the region of influence of bending

Appendix B

The Müller Potential

Müller's Potential (fig. B.1) is a simple 2D analytic cost function with three local minima M_1, M_2, M_3, and two saddle points S_1, S_2, which is often used in the computational chemistry literature to illustrate transition state search methods.

It has the form:

$$V(x,y) = \sum_{i=1}^{4} A_i \, e^{a_i(x-x_i)^2 + b_i(x-x_i)(y-y_i) + c_i(y-y_i)^2} \tag{B.1}$$

where $\mathbf{A} = (-200, -100, -170, 15)$, $\mathbf{a} = (-1, -1, -6.5, 0.7)$, $\mathbf{b} = (0, 0, 11, 0.6)$, $\mathbf{c} = (-10, -10, -6.5, 0.7)$, $\mathbf{x} = (1, 0, -0.5, -1)$, $\mathbf{y} = (0, 0.5, 1.5, 1)$. The inter-minimum distance is of order 1 length unit, and the transition states are around 100–150 energy units above the lowest minimum.

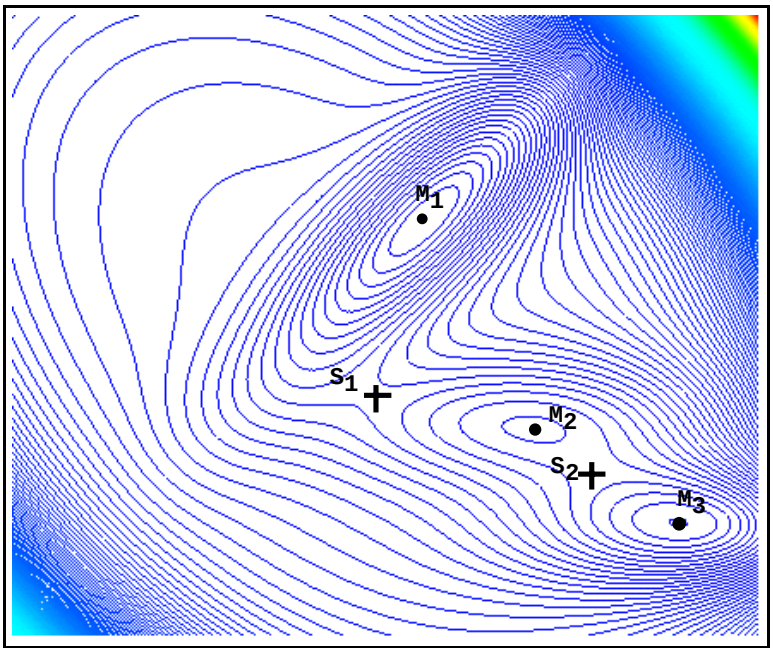


Figure B.1: The Isocontours of the Müller Potential

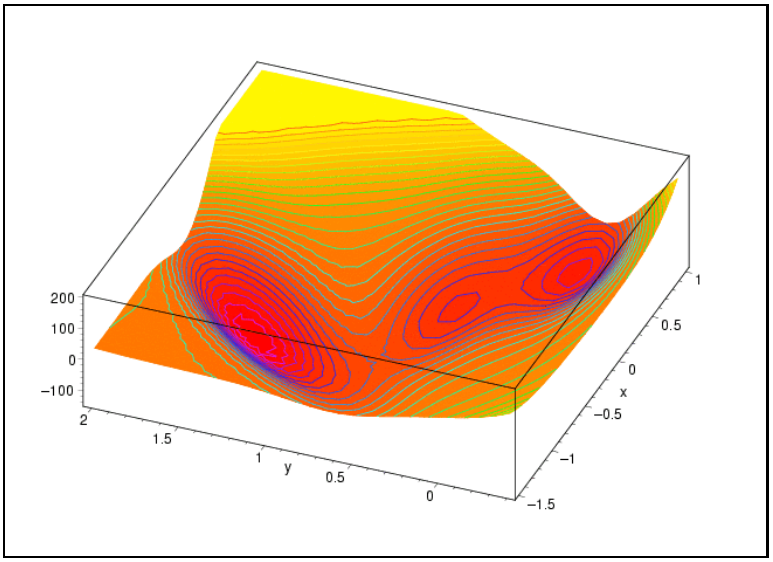


Figure B.2: 3D view of the Müller Potential

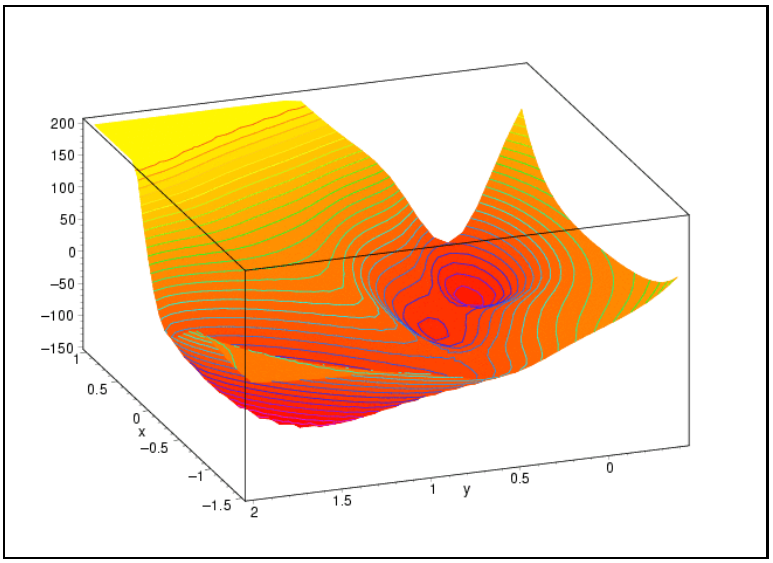


Figure B.3: 3D view of the Müller Potential

Appendix C

Implementations and Vision System Design

In this appendix we briefly describe the control and communication structure we propose in order to model computer vision applications of the type studied in this thesis. Our application model is based on optimal state estimation ideas. Such a framework allows modeling and experimenting with an important class of vision techniques including continuous optimization, sampling, or dynamical systems methods.

C.1 Introduction

The design of systems for computer vision tasks like reconstruction or tracking poses several major difficulties. First, such systems tend to be large and to involve methodologies from various areas, such as object modeling, optimization, control theory, statistics and computer graphics. Secondly, a clear control and communication structure is necessary in order to structure complete classes of vision application in terms of the above methodologies.

In order to fulfill such demands, we have developed an application model based on optimal state estimation concepts. It consists of:

- a *model* characterized by its state.
- a generative *transformation* which predicts discretized model features in the observation space, based on a current state configuration.
- an *association* of predicted and extracted features in the observation space to evaluate a configuration cost.
- a *control strategy* that updates the model state such that the evaluation cost meets an optimality criterium.

Such an application structuring can cover a wide range of techniques such as deformable models and dynamical systems, continuous optimization methods and sampling methods, as well as combinations of them (hybrid methods). The framework is flexible and extensible as it does not make assumptions about the ways the model is represented or discretized, the type and number of extracted features (cues), how the association is done, and what is the underlying strategy used to estimate and evolve the model state. More details are given in the next sections.

C.2 Application Model

In a generic state estimation approach, one derives optimal estimates of a model's parameters based on a (possibly temporal) sequence of observations. In detail, this involves the following elements (see also fig. C.1 which depicts these elements and the data streams between them):

1. a "representational" model *discretization* in a spatial domain.

2. a composite (generally non-linear) *generalized transformation* (T), parameterized in terms of the current model configuration (typically an instance of the model state) which generates predictions in the observation space for points in the discretized domain. This item and the previous one form the *model representation*.

3. a sequence of (possibly temporal) observations or extracted *features*.

4. a way to associate model predictions to extracted features to evaluate a *configuration cost* or higher order operators associated with it (their computation typically needs equivalent quantities of T).

5. a *strategy* to evolve the model state based on the evaluation of configuration costs or higher-order operators associated to it, such as its gradient or Hessian, in order to match an optimality criterium (see fig. C.1).

6. several *user interaction* and *visualization* components responsible for the application building and scene exploration, including interactive model-data couplings, 2D and 3D renderings, as well as classical graphics user interfaces (GUI).

C.2.1 Configuration Evaluator

The basic task of the configuration evaluator is to compute the cost corresponding to a model state or higher order operators (gradient or Hessian) associated with it. Their evaluation is typically performed in terms of equivalent quantities computed for the model/data direct generalized transform (T).

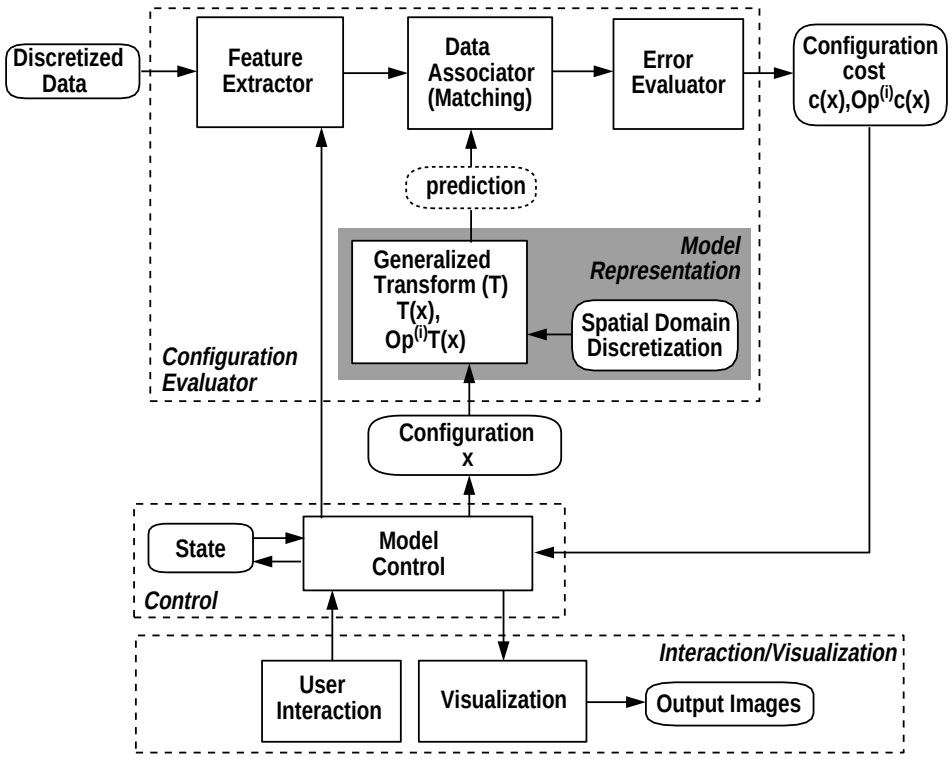


Figure C.1: Vision application model

Model Representation

The model is spatially discretized in a domain Ω. For any point $\mathbf{u} \in \Omega$, we can compute a prediction at the current configuration $\mathbf{x}$ in the observation space $\mathbf{r} = \mathbf{T}(\mathbf{x}, \mathbf{u})$. The Jacobian matrix of this direct mapping makes the connection between differential quantities in parameter and observation spaces and it can also be used to compute gradients or approximate Hessian matrices. We compute the gradient and Hessian analytically and in a modular fashion by back-propagation on the direct transformation chain.

Feature Extraction, Data Association and Error Evaluation

These components extract the various types of information used in vision applications. Typical datasets include 2D image data, namely edge, contour or silhouette information, optical flow information, or 3D range data obtained, for instance, from a multi-camera stereo reconstruction system.

A subsequent data associator or matching stage establishes correspondences between model predictions and data features. The matching is either explicit, as in the case of an optical flow module or a 3D model to range features, or implicit, when every predicted model feature has already a computed cost on a potential surface (see also chapter 3).

C.2.2 State Representation and Control

The state representation and control components constitute the core of the application model. Various state representations correspond to various control strategies as follows:

- *continuous* ones (deformable models, continuous optimization) which generally assume unimodal state representation and evaluation of the cost function and its higher order operators (gradient, Hessian, etc.);
- *discrete* ones which only involve cost function evaluations and sampling methods for focusing the search effort;
- *hybrid* strategies that involve a combination of the first two.

Regardless of the state representation and control employed, there is a unique interface between these components and the configuration evaluator (Fig. C.1).

Dynamical Systems and Deformable Models

A deformable model estimates the state of the system by numerical integrating a synthetic dynamical system that encodes the fitting error. The Lagrangian equations governing its evolution can be written as:

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{D}\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{f}_{\mathbf{x}}, \mathbf{f}_{\mathbf{x}} = \int \mathbf{L}^{\top}\boldsymbol{\delta}\mathbf{r} \tag{C.1}$$

where $\mathbf{M} = \delta \int \mathbf{J}^{\top}\mathbf{J}$, $\mathbf{D} = \gamma \int \mathbf{J}^{\top}\mathbf{J}$, and $\mathbf{K} = \mathbf{diag}(k_{si})$, with δ and γ being tuning parameters, k_{si} being the stiffness associated with the parameter i, $\mathbf{f}_{\mathbf{x}}$ are generalized "forces", acting on the state parameters and $\boldsymbol{\delta}\mathbf{r}$ is an "image force" in the observation space (typically an $L2$ norm).

Continuous Optimization Based Methods

Optimization based methods perform non-linear estimation using the direct generalized transform gradient and Hessian. Such gradient descent or Newton optimizers have been discussed extensively in chapter 3, 4 and 5. See there for the derivations of cost functions, gradient and Hessian approximations corresponding to different image features and error norms.

Sampling Methods

Sampling methods, are non-parametric techniques that propagate the entire distribution, over time, as a set of hypotheses, or samples, with their associated probability. In each frame, the distribution over parameters is recomputed (i.e. re-sampled and re-weighted) based on new image observations. Initially (Isard and Blake, 1998) these methods didn't have a continuous component, so estimating

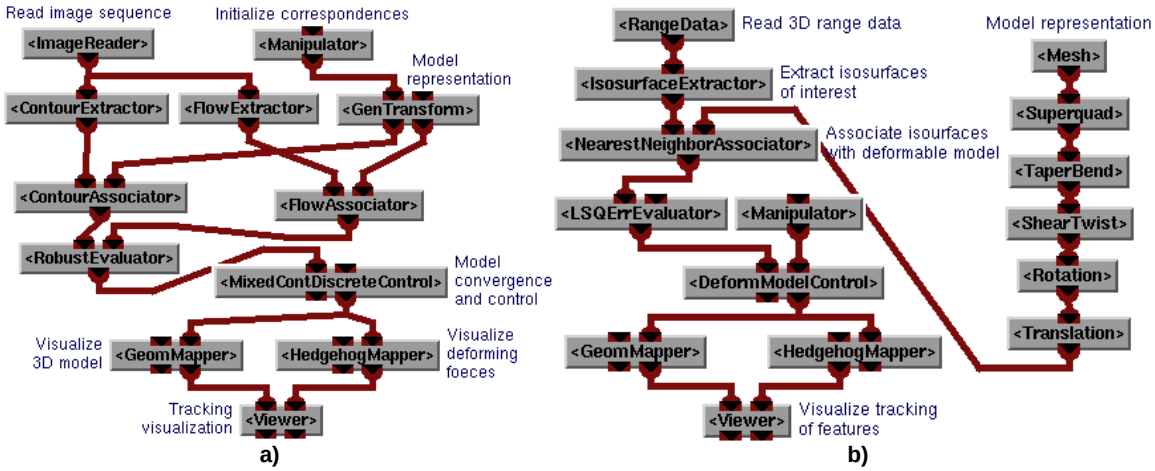


Figure C.2: Tracking networks for human motion (a) and 3D range data (b)

the distribution would (in our framework) require only evaluating configurations, but no higher-order operators. The computational burden lies in the strategies for sampling the distribution in order to locate its typical sets. Known strategies include importance, partitioned (Gibbs) or annealing based sampling methods (Neal, 1993; Deutscher et al., 2000). More recent approaches (Choo and Fleet, 2001; Sminchisescu and Triggs, 2002b) supplementary use gradient information to generate moves. We have developed an uniform interface such that various methods can be parameterized by the sample generation strategy.

Hybrid Continuous/Discrete Methods

Hybrid continuous/discrete methods have been proposed more recently (Heap and Hogg, 1998; Cham and Rehg, 1999; Forsyth et al., 2001; Sminchisescu and Triggs, 2001a,b) in order to combine generic robustness features of sample-based techniques with the local accuracy of continuous optimization ones. Depending upon the approach these methods propagate the distribution either non-parametrically as a set of samples or parametrically using a kernel mixture.

C.2.3 Interaction/Visualization

The interaction and visualization modules are responsible for scene exploration and manipulation, parameter monitoring and control, and interactive application building. We point out that for interactive applications one usually needs dual control: simultaneous user driven and application driven. Such a constraint imposes particular demands on the system design, namely *consistency* and *control* issues: user input has to be consistently integrated and propagated into the application data structures (even during application driven execution), while the structuring of individual application components has to allow external interaction during the execution of their different operations. We address these problems by means of automatically enforcing data-flow consistency and by using

object-orientation in order to design components with open, non-encapsulated, control.

Appendix D

Other Scientific Activities

- **Invited Talks**
  - INRIA Sophia-Antipolis (Oddysée/Robotvis), July 2002.
  - Xerox Research Center Europe, June 2002.
  - Brown University, USA, April 2002.
  - Swiss Federal Institute of Technology (EPFL), Switzerland, March 2001 and 2002.
  - Microsoft Research, Seattle, USA, February 2001 and Cambridge, UK, January 2002.
  - University of Pennsylvania, USA, June 2000.
  - Rutgers University, USA, June 1999 and April 2002.
  - Eindhoven University of Technology, The Netherlands, April 1999.
- **Collaborations**
  - Scientific collaborator in the European project VIBES (Video Browsing, Exploration and Structuring) (INRIA, Oxford, Weizmann, KTH, Leuven, EPFL).
  - Collaboration with MacDonald Dettwiler Space and Advanced Robotics Ltd. Canada and University of Toronto, on complex object tracking in space (project for the international space shuttle).
- **Reviewer (IEEE PAMI, IEEE IP, SIGGRAPH, ICCV, CVPR, ECCV, ICPR)**
  - IEEE Transactions on Pattern Analysis and Machine Intelligence/Image Processing
  - SIGGRAPH International Conference on Computer Graphics
  - International Conference on Computer Vision
  - International Conference on Computer Vision and Pattern Recognition
  - European Conference on Computer Vision
  - International Conference on Pattern Recognition

Bibliography

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.

Y. Abashkin, N. Russo, and M. Toscano. Transition States and Energy Barriers from Density Functional Studies: Representative Isomerization Reactions. *International Journal of Quantum Chemistry*, 1994.

J. Aggarwal and Q. Cai. Human Motion Analysis: A Review. *Computer Vision and Image Understanding*, 73(3):428–440, 1999.

N. Anderson and G. R. Walsh. A Graphical Method for a Class of Branin Trajectories. *Journal of Optimization Theory and Applications*, 1986.

M. Armstrong and A. Zisserman. Robust Object Tracking. In *Asian Conference on Computer Vision*, 1995.

A. Azarbayejani and A. Pentland. Recursive Estimation of Motion, Structure and Focal Length. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995.

Y. Bar-Shalom and T. Fortman. *Tracking and Data Association*. Academic Press, 1988.

G. T. Barkema. Event-Based Relaxation of Continuous Disordered Systems. *Physical Review Letters*, 77(21), 1996.

A. Barr. Global and Local Deformations of Solid Primitives. *Computer Graphics*, 18:21–30, 1984.

C. Barron and I. Kakadiaris. Estimating Anthropometry and Pose from a Single Image. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 669–676, 2000.

M. Bertero, T. Poggio, and V. Torre. Ill-posed Problems in Early Vision. *Proc. of IEEE*, 1988.

M. Black. *Robust Incremental Optical Flow*. PhD thesis, Yale University, 1982.

M. Black and P. Anandan. The Robust Estimation of Multiple Motions: Parametric and Piecewise Smooth Flow Fields. *Computer Vision and Image Understanding*, 6(1):57–92, 1996.

M. Black and A. Rangarajan. On the Unification of Line Processes, Outlier Rejection, and Robust Statistics with Applications in Early Vision. *International Journal of Computer Vision*, 19(1): 57–92, July 1996.

A. Blake and M. Isard. *Active Contours*. Springer, 2000.

A. Blake, B. North, and M. Isard. Learning Multi-Class Dynamics. *ANIPS 11*, pages 389–395, 1999.

A. Blake and A. Zisserman. *Visual Reconstruction*. MIT Press, 1987.

J. M. Bofill. Updated Hessian Matrix and the Restricted Step Method for Locating Transition Structures. *Journal of Computational Chemistry*, 15(1):1–11, 1994.

M. Brand. Shadow Puppetry. In *IEEE International Conference on Computer Vision*, pages 1237–1244, 1999.

F. Branin and S. Hoo. A Method for Finding Multiple Extrema of a Function of n Variables. *Numerical Methods of Nonlinear Optimization*, 1972.

C. Bregler and J. Malik. Tracking People with Twists and Exponential Maps. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 1998.

R. Carceroni and K. Kutulakos. Multi-View Scene Capture by Surfel Sampling: From Video Streams to Non-Rigid 3D Motion, Shape and Reflectance. In *ICCV*, 2001.

C. J. Cerjan and W. H. Miller. On Finding Transition States. *J. Chem. Phys.*, 75(6), 1981.

T. Cham and J. Rehg. A Multiple Hypothesis Approach to Figure Tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 239–245, 1999.

A. Chiusso, R. Brockett, and S. Soatto. Optimal Structure from Motion: Local Ambiguities and Global Estimates. *International Journal of Computer Vision*, 39(3):195–228, 2000.

K. Choo and D. Fleet. People Tracking Using Hybrid Monte Carlo Filtering. In *IEEE International Conference on Computer Vision*, 2001.

D. Comaniciu, V. Ramesh, and P. Meer. Real-time Tracking of Non-rigid Objects using Mean Shift. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2000.

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.

J. Crowley, P. Stelmaszyk, T. Skordas, and P.Pugget. Measurements and Integration of 3-D Structures by Tracking Edge Lines. *International Journal of Computer Vision*, 1992.

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.

D. DeCarlo and D. Metaxas. The Integration of Optical Flow and Deformable Models with Applications to Human Face Shape and Motion Estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 1996.

D. DeCarlo and D. Metaxas. Combining Information in Deformable Models Using Hard Constraints. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 1999.

Q. Delamarre and O. Faugeras. 3D Articulated Models and Multi-View Tracking with Silhouettes. In *IEEE International Conference on Computer Vision*, 1999.

J. Deutscher, A. Blake, and I. Reid. Articulated Body Motion Capture by Annealed Particle Filtering. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2000.

J. Deutscher, A. Davidson, and I. Reid. Articulated Partitioning of High Dimensional Search Spacs associated with Articulated Body Motion Capture. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2001.

J. Deutscher, B. North, B. Bascle, and A. Blake. Tracking through Singularities and Discontinuities by Random Sampling. In *IEEE International Conference on Computer Vision*, pages 1144–1149, 1999.

S. Dickinson and D. Metaxas. Integrating Qualitative and Quantitative Shape Recovery. *International Journal of Computer Vision*, 1994.

S. Dickinson, A. Pentland, and A. Rosenfeld. Shape Recovery using Distributed Aspect Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1992.

D. DiFranco, T. Cham, and J. Rehg. Reconstruction of 3-d figure motion from 2-d correspondences. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2001.

T. Drummond and R. Cipolla. Real-time Tracking of Multiple Articulated Structures in Multiple Views. In *European Conference on Computer Vision*, 2000.

T. Drummond and R. Cipolla. Real-time Tracking of Highly Articulated Structures in the Presence of Noisy Measurements. In *IEEE International Conference on Computer Vision*, 2001.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

A. Efros and T. Leung. Texture Synthesis by Non-Parametric Sampling. In *IEEE International Conference on Computer Vision*, 1999.

I. Fiodorova. Search for the Global Optimum of Multiextremal Problems. *Optimal Decision Theory*, 1978.

R. Fletcher. Practical Methods of Optimization. In *John Wiley*, 1987.

D. Forsyth, J. Haddon, and S. Ioffe. The Joy of Sampling. *International Journal of Computer Vision*, 41:109–134, 2001.

P. Fua and G. Leclerc. Taking Advantage of Image-Based and Geometry-based Constraints to Recover 3-D Surfaces. *Computer Vision and Image Understanding*, 1996.

D. Gavrila. The Visual Analysis of Human Movement: A Survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.

D. Gavrila and L. Davis. 3-D Model Based Tracking of Humans in Action:A Multiview Approach. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 73–80, 1996.

R. Ge. The theory of the Filled Function Method for Finding a Global Minimizer of a Nonlinearly Constrained Minimization Problem. *J. of Comp. Math.*, 1987.

D. Geiger, A. Gupta, L. Costa, and J. Vlontzos. Dynamic Programming for Detecting, Tracking and Matching Deformable Contours. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(3), 1995.

D. Geman and S. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1984.

A. Goldstein and J. Price. On Descent from Local Minima. *Mathematics of Computation*, 1971.

L. Gonglaves, E. Bernardo, E. Ursella, and P. Perona. Monocular Tracking of the human arm in 3D. In *IEEE International Conference on Computer Vision*, pages 764–770, 1995.

N. Gordon and D. Salmond. Bayesian State Estimation for Tracking and Guidance Using the Bootstrap Filter. *Journal of Guidance, Control and Dynamics*, 1995.

N. Gordon, D. Salmond, and A. Smith. Novel Approach to Non-linear/Non-Gaussian State Estimation. *IEE Proc. F*, 1993.

A. Griewank. Generalized Descent for Global Optimization. *Journal of Optimization Theory and Applications*, 1981.

Group. *Specifications for a Standard Humanoid*. Hanim-Humanoid Animation Working Group, http://www.h-anim.org/Specifications/H-Anim1.1/, 2002.

G. Hager and P. Belhumeur. Efficient Region Tracking With Parametric Models of Geometry and Illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998.

I. Haritaoglu, D. Harwood, and L. Davis. W4: Who, When, Where, What: A Real Time System for Detecting and Tracking People. In *FGR*, 1998.

R. Hartley. Lines and Points in Three Views and the Trifocal Tensor. *International Journal of Computer Vision*, 1997.

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

T. Heap and D. Hogg. Wormholes in Shape Space: Tracking Through Discontinuities Changes in Shape. In *IEEE International Conference on Computer Vision*, pages 334–349, 1998.

T. Helgaker. Transition-State Optimizations by Trust-Region Image Minimization. *Chemical Physics Letters*, 182(5), 1991.

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.

R. L. Hilderbrandt. Application of Newton-Raphson Optimization Techniques in Molecular Mechanics Calculations. *Computers & Chemistry*, 1:179–186, 1977.

K. B. Horn. Relative Orientation. *International Journal of Computer Vision*, 1990.

N. Howe, M. Leventon, and W. Freeman. Bayesian Reconstruction of 3D Human Motion from Single-Camera Video. *Neural Information Processing Systems*, 1999.

T. S. Huang and A. N. Netravali. Motion and Structure from Feature Correspondences: A Review. *Proc. of the IEEE*, 1994.

S. Ioffe and D. Forsyth. Human Tracking with Mixtures of Trees. In *IEEE International Conference on Computer Vision*, volume I, pages 690–695, 2001.

M. Isard and A. Blake. CONDENSATION – Conditional Density Propagation for Visual Tracking. *International Journal of Computer Vision*, 1998.

M. Isard and J. MacCormick. BRaMBLe: A Bayesian Multiple-blob Tracker. In *IEEE International Conference on Computer Vision*, 2001.

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.

P. Jorgensen, H. J. A. Jensen, and T. Helgaker. A Gradient Extremal Walking Algorithm. *Theoretica Chimica Acta*, 73:55–65, 1988.

S. Ju, M. Black, and Y. Yacoob. Cardboard People: A Parameterized Model of Articulated Motion. In *2nd Int.Conf.on Automatic Face and Gesture Recognition*, pages 38–44, October 1996.

I. Kakadiaris and D. Metaxas. 3D Human Body Model Acquisition from Multiple Views. In *IEEE International Conference on Computer Vision*, pages 618–623, 1995.

I. Kakadiaris and D. Metaxas. Model-Based Estimation of 3D Human Motion with Occlusion Prediction Based on Active Multi-Viewpoint Selection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 81–87, 1996.

K. Kanantani. *Statistical Optimization for Geometric Computation: Theory and Practice*. Elsevier, 1996.

M. Kass, A. Witkin, and D. Terzopoulos. Snakes:Active Contour Models. *International Journal of Computer Vision*, 1988.

O. King and D. Forsyth. How does CONDENSATION Behave with a Finite Number of Samples? In *European Conference on Computer Vision*, pages 695–709, 2000.

H. J. Lee and Z. Chen. Determination of 3D Human Body Postures from a Single View. *Computer Vision, Graphics and Image Processing*, 30:148–168, 1985.

M. Leventon and W. Freeman. Bayesian Estimation of 3-d Human Motion from an Image Sequence. Technical Report TR-98-06, MERL, 1998.

A. Levy and A. Montalvo. The Tunneling Algorithm for the Global Minimization of Functions. *SIAM J. of Stat. Comp.*, 1985.

D. Liebowitz and S. Carlson. Uncalibrated Motion Capture Exploiting Articulated Structure Constraints. In *IEEE International Conference on Computer Vision*, 2001.

J. Liu. Metropolized Independent Sampling with Comparisons to Rejection Sampling and Importance Sampling. *Statistics and Computing*, 6, 1996.

T. L. Liu and D. Geiger. Visual Deconstruction: Reconstructing Articulated Objects. In *Venice Workshop*, 1998.

D. Lowe. Three-dimensional Object Recognition from Single Two-dimensional Images. *Artificial Intelligence*, 31(3), 1987.

D. Lowe. Fitting Parameterized Three-dimensional Models to Images . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1991.

J. MacCormick and A. Blake. A Probabilistic Contour Discriminant for Object Localisation. In *IEEE International Conference on Computer Vision*, 1998.

J. MacCormick and M. Isard. Partitioned Sampling, Articulated Objects, and Interface-Quality Hand Tracker. In *European Conference on Computer Vision*, volume 2, pages 3–19, 2000.

D. P. McReynolds and D. Lowe. Rigidity Checking of 3D point Correspondences under Perspective Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997.

R. Merwe, A. Doucet, N. Freitas, and E. Wan. The Unscented Particle Filter. Technical Report CUED/F-INFENG/TR 380, Cambridge University, Department of Engineering, May 2000.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, 21(6):1087–1092, 1953.

A. Mitiche and J. Aggarwal. Line-based Computation of Structure and Motion using Angular Invariance. In *IEEE Workshop on Motion*, 1986.

T. Moeslund and E. Granum. A Survey of Computer Vision-Based Human Motion Capture. *Computer Vision and Image Understanding*, 81:231–268, 2001.

D. Morris and J. Rehg. Singularity Analysis for Articulated Object Tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 289–296, 1998.

N. Mousseau and G. T. Berkema. Traveling Through Potential Energy Lanscapes of Desordered Materials: The Activation-Relaxation Technique. *Physical Review E*, 57(2), 1998.

D. Mumford. Elastica and Computer Vision. In *Algebraic Geometry and its Applications*. Springer-Verlag, 1994.

D. Mumford. Pattern Theory: A Unifying Perspective. In D. Knill and W. Richards, editors, *Perception as Bayesian Inference*. Cambridge University Press, 1996.

L. J. Munro and D. J. Wales. Defect Migration in Crystalline Silicon. *Physical Review B*, 59(6): 3969–3980, 1999.

R. Neal. Probabilistic Inference Using Markov Chain Monte Carlo. Technical Report CRG-TR-93-1, University of Toronto, 1993.

R. Neal. Annealed Importance Sampling. Technical Report 9805, Department of Statistics, University of Toronto, 1998.

R. Neal. Annealed Importance Sampling. *Statistics and Computing*, 11:125–139, 2001.

J. Nichols, H. Taylor, P. Schmidt, and J. Simons. Walking on Potential Energy Surfaces. *J. Chem. Phys.*, 92(1), 1990.

B. North and A. Blake. Learning Dynamical Models by Expectation Maximization. In *IEEE International Conference on Computer Vision*, 1998.

J. Oliensis. The Error Surface for Structure from Motion. Technical report, NECI, 2001.

C. Papageorgiu and T. Poggio. Trainable Pedestrian Detection. In *International Conference on Image Processing*, 1999.

V. Pavlovic, J. Rehg, T. Cham, and K. Murphy. A Dynamic Bayesian Approach to Figure Tracking using Learned Dynamical Models. In *IEEE International Conference on Computer Vision*, 2001.

A. Pentland and B. Horowitz. Recovery of Non-rigid Motion and Structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1991.

A. Pentland and S. Sclaroff. Closed Form Solutions for Physically-based Shape Modeling and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1991.

P. Perez, A. Blake, and M. Gangnet. JetStream:Probabilistic Contour Extraction with Particles. In *IEEE International Conference on Computer Vision*, 2001.

M. Pitt and N. Shephard. Filtering via Simulation: Auxiliary Particle Filter. *Journal of the American Statistical Association*, 1997.

R. Plankers. *Human Body Modeling from Video Sequences*. PhD thesis, EPFL, 2001.

R. Plankers and P. Fua. Articulated Soft Objects for Video-Based Body Modeling. In *IEEE International Conference on Computer Vision*, pages 394–401, 2001.

J. Rehg and T. Kanade. Model-Based Tracking of Self Occluding Articulated Objects. In *IEEE International Conference on Computer Vision*, pages 612–617, 1995.

K. Rohr. Towards Model-based Recognition of Human Movements in Image Sequences. *Computer Vision, Graphics and Image Processing*, 59(I):94–115, 1994.

R. Ronfard, C. Schmid, and B. Triggs. Learning to Parse Pictures of People. In *European Conference on Computer Vision*, 2002.

R. Rosales and S. Sclaroff. Inferring Body Pose without Tracking Body Parts. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 721–727, 2000.

D. Samaras and D. Metaxas. Incorporating Illumination Constraints in Deformable Models. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 1998.

S. Seitz. Implicit Scene Reconstruction from Probability Density Functions. In *DARPA Image Understanding Workshop*, 1998.

E. M. Sevick, A. T. Bell, and D. N. Theodorou. A Chain of States Method for Investigating Infrequent Event Processes Occuring in Multistate, Multidimensional Systems. *J. Chem. Phys.*, 98 (4), 1993.

A. Shashua. Algebraic Functions for Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995.

H. Sidenbladh and M. Black. Learning Image Statistics for Bayesian Tracking. In *IEEE International Conference on Computer Vision*, 2001.

H. Sidenbladh, M. Black, and D. Fleet. Stochastic Tracking of 3D Human Figures Using 2D Image Motion. In *European Conference on Computer Vision*, 2000.

H. Sidenbladh, M. Black, and L. Sigal. Implicit Probabilistic Models of Human Motion for Synthesis and Tracking. In *European Conference on Computer Vision*, 2002.

J. Simons, P. Jorgensen, H. Taylor, and J. Ozmen. Walking on Potential Energy Surfaces. *J. Phys. Chem.*, 87:2745–2753, 1983.

C. Sminchisescu. Consistency and Coupling in Human Model Likelihoods. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 27–32, Washington D.C., 2002.

C. Sminchisescu, D. Metaxas, and S. Dickinson. Improving the Scope of Deformable Model Shape and Motion Estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 485–492, Hawaii, 2001a.

C. Sminchisescu, D. Metaxas, and S. Dickinson. Incremental Model-Based Estimation Using Geometric Consistency Constraints. Technical Report 4209, INRIA, June 2001b.

C. Sminchisescu and A. Telea. A Framework for Generic State Estimation in Computer Vision Applications. In S. Verlag, editor, *International Conference for Computer Vision, ICVS International Workshop on Computer Vision Systems*, pages 21–34, Vancouver, 2001.

C. Sminchisescu and A. Telea. Human Pose Estimation from Silhouettes. A Consistent Approach Using Distance Level Sets. In *WSCG International Conference for Computer Graphics, Visualization and Computer Vision*, Czech Republic, 2002.

C. Sminchisescu and B. Triggs. A Robust Multiple Hypothesis Approach to Monocular Human Motion Tracking. Technical Report RR-4208, INRIA, 2001a.

C. Sminchisescu and B. Triggs. Covariance-Scaled Sampling for Monocular 3D Body Tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 447–454, Hawaii, 2001b.

C. Sminchisescu and B. Triggs. Building Roadmaps of Local Minima of Visual Models. In *European Conference on Computer Vision*, volume 1, pages 566–582, Copenhagen, 2002a.

C. Sminchisescu and B. Triggs. Hyperdynamics Importance Sampling. In *European Conference on Computer Vision*, volume 1, pages 769–783, Copenhagen, 2002b.

M. R. Sorensen and A. F. Voter. Temperature-Accelerated Dynamics for Simulation of Infrequent Events. *J. Chem. Phys.*, 112(21):9599–9606, 2000.

M. Spestakis and J. Aloimonos. A Multi-frame Approach to Visual Motion Perception. *International Journal of Computer Vision*, 1991.

J. Sullivan, A. Blake, M. Isard, and J. MacCormick. Object Localization by Bayesian Correlation. In *IEEE International Conference on Computer Vision*, 1999.

J. Q. Sun and K. Ruedenberg. Gradient Extremals and Stepest Descend Lines on Potential Energy Surfaces. *J. Chem. Phys.*, 98(12), 1993.

R. Szeliski and S. B. Kang. Recovery 3-D Shape and Motion from Image Streams using Non-linear Least-squares. Technical report, DEC TR, 1994.

C. Taylor and D. Kriegman. Structure and Motion from Line Segments in Multiple Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1996.

C. J. Taylor. Reconstruction of Articulated Objects from Point Correspondences in a Single Uncalibrated Image. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 677–684, 2000.

D. Terzopoulos and D. Metaxas. Dynamic 3D models with Local and Global deformations: Deformable Superquadrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1991.

D. Terzopoulos, A. Witkin, and M. Kass. Constraints on Deformable Models: Recovering 3-D Shape and Non-rigid Motion. *A.I.*, 36(1), 1988.

A. Törn, M. Ali, and S. Viitanen. Stochastic Global Optimization: Problem Classes and Solution Techniques. *Journal of Global Optimization*, 14, 1999.

K. Toyama and A. Blake. Probabilistic Tracking in a Metric Space. In *IEEE International Conference on Computer Vision*, 2001.

B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle Adjustment - A Modern Synthesis. In Springer-Verlag, editor, *Vision Algorithms: Theory and Practice*, 2000.

D. Vanderbilt and S. G. Louie. A Monte Carlo Simulated Annealing Approach over Continuous Variables. *J. Comp. Physics*, 56, 1984.

T. Vieville and O. Faugeras. Feed-forward Recovery of Motion and Structure From a Sequence of 2-D Line Matches. In *IEEE International Conference on Computer Vision*, 1990.

G. H. Vineyard. Frequency factors and Isotope Effects in Solid State Rate Processes. *J. Phys. Chem. Solids*, 3:121–127, 1957.

A. F. Voter. A Method for Accelerating the Molecular Dynamics Simulation of Infrequent Events. *J. Chem. Phys.*, 106(11):4665–4677, 1997a.

A. F. Voter. Hyperdynamics: Accelerated Molecular Dynamics of Infrequent Events. *Physical Review Letters*, 78(20):3908–3911, 1997b.

S. Wachter and H. Nagel. Tracking Persons in Monocular Image Sequences. *Computer Vision and Image Understanding*, 74(3):174–192, 1999.

D. J. Wales. Finding Saddle Points for Clusters. *J. Chem. Phys.*, 91(11), 1989.

D. J. Wales and T. R. Walsh. Theoretical Study of the Water Pentamer. *J. Chem. Phys.*, 105(16), 1996.

L. Wei and M. Levoy. Fast Texture Synthesis using Tree-structured Vector Quantization. In *SIGGRAPH*, 2000.

C. Wren and A. Pentland. DYNAMAN: A Recursive Model of Human Motion. Technical Report TR-451, MIT Media Lab, 1998.

B. Yen and T. Huang. Determining 3-D Motion and Structure of a Rigid Body Using Straight line Correspondences. In *ASI NATO Series*, 1983.

S. C. Zhu. Embedding Gestalt Laws in Markov Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(11), 1999.

Index