

LABORATORY OF COMPUTER VISION AND MEDIA TECHNOLOGY

---

PH.D. DISSERTATION

COMPUTER VISION-BASED MOTION  
CAPTURE OF BODY LANGUAGE

APPLYING SPATIALLY-BASED PRUNING  
OF THE STATE-SPACE

THOMAS B. MOESLUND

DEPARTMENT OF HEALTH SCIENCE AND TECHNOLOGY

---

AALBORG UNIVERSITY 2003

## AUTHOR



### THOMAS BALTZER MOESLUND

Thomas Baltzer Moeslund received the M.Sc.EE. and the Ph.D. degrees in 1996 and 2003, respectively, both from Aalborg University, Denmark. The main topics of his Ph.D.-thesis are related to motion capture of human body language. Currently he is an assistant professor at Aalborg University working on computer vision for gesture recognition and object tracking. He is also involved in teaching activities and coordinates the M.Sc.EE. specialisation "Computer Vision and Graphics". Dr. Moeslund has published more than 30 journal and conference papers in computer vision, human computer interaction and related fields.

*Computer Vision-Based Motion  
Capture of Body Language  
Applying Spatially-Based Pruning  
of the State-Space*

A Ph.D. dissertation

by

Thomas B. Moeslund

Laboratory of Computer Vision and Media Technology

Department of Health Science and Technology

Aalborg University, Denmark

E-mail: [tbm@cvmt.dk](mailto:tbm@cvmt.dk)

URL: <http://www.cvmt.dk/~tbm>

October 2003

©Copyright 2003 by Thomas B. Moeslund



This dissertation was submitted in April 2003 to the Faculty of Engineering and Science, Aalborg University, Denmark, in partial fulfillment of the requirements for the Doctor of Philosophy degree.

While the first edition was approved, this second edition includes revisions in accordance with comments from the adjudication committee.

The following adjudication committee was appointed to evaluate the thesis. Note that the supervisor was a non-voting member of the committee.

**Professor Pascal Fua, Ph.D.**

Computer Vision Laboratory  
School of Computer and Communication Science  
Swiss Federal Institute of Technology (EPFL)  
Lausanne, Switzerland

**Reader Adrian Hilton, Ph.D.**

Centre for Vision, Speech and Signal Processing  
Department of Electronic Engineering  
University of Surrey  
Guildford, Surrey, GU2 7XH  
United Kingdom

**Associate Professor Claus B. Madsen, Ph.D. (committee chairman)**

Computer Vision and Media Technology Laboratory  
Department of Health Science and Technology  
Aalborg University  
Aalborg, Denmark

**Professor Erik Granum, Ph.D. (supervisor)**

Computer Vision and Media Technology Laboratory  
Department of Health Science and Technology  
Aalborg University  
Aalborg, Denmark

©All rights reserved. No part of this report may be reproduced, stored in a retrieval system, or transmitted, in any form by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the author.



# Abstract

Capturing the motion of a human body utilising computer vision is the focus of this thesis. Normally the capturing process is carried out by applying a priori knowledge, in the form of a geometrical model, i.e., applying a model-based approach. Different configurations of the model is synthesised and compared with the image data. The configuration most similar to the current image data defines the current state of the model, i.e., its pose. Each degree of freedom in the geometrical model is represented by one variable.

When first initialised this provides a very powerful pruning as long as the assumption of "smooth motion" is fulfilled. However, under practical circumstances the temporally-based pruning often breaks down and requires a complicated reinitialisation. Hence, alternative or supplementary methods of pruning that are independent of the temporal context are of interest. The purpose of this thesis is to investigate possibilities for exploiting spatial information to achieve a similar pruning effect. The context of the investigation into spatially-based pruning is to capture the 3D pose of a human arm given one static camera.

The thesis is divided into three parts. In the first part motion capture in general is described and a comprehensive survey of the relevant literature is presented. In the second part spatially-based pruning is applied to derive a more compact state-space representation of the arm by including low-level image features. Concretely it is shown how the primary degrees-of-freedom (DoF) in the shoulder and arm can be efficiently modelled by utilising the position of the hand in each individual image. This model is denoted the local screw axis model. Furthermore, this part also describes how to reduce the size of the state-space by introducing six spatially-based constraints. Together, the constraints and the local screw axis model reduces the size of the state-space significantly. In part three the spatially-based pruning is implemented in different systems in order to demonstrate its effect. Part three also shows how to apply Sequential Monte Carlo tracking to make an efficient search in the pruned state-space.

The primary findings are first of all the local screw axis model which allows the 12 primary DoF in the shoulder and arm to be modelled by just two DoF. Secondly, the six spatially-based constraints that allow for a pruning of the state-space of 97.3% in average. Both findings suggest that the proposed approach for spatially-based pruning is a realistic alternative for coping with the problems inherent in temporally-based pruning.



## Resume

Fokus for denne Ph.d.-afhandling er, at registrere bevægelserne af den menneskelige krop vha. computer vision. Registreringen foregår normalt ved at benytte a priori viden i form af en geometrisk model, dvs. en model-baseret tilgang. Forskellige konfigurationer af modellen sammenlignes med billede data. Den konfiguration der ligner billede data mest, definerer den nuværende tilstand for modellen.

Det største problem med en model-baseret tilgang er, at der potentielt kan forekomme et meget stort antal forskellige konfigurationer. For at imødegå dette problem kan den temporale kontekst benyttes til at reducere antallet af sandsynlige konfigurationer. Desværre har denne tilgang nogle u hensigtsmæssige konsekvenser, som gør at alternative tilgange bør findes.

I denne afhandling foreslåes det, at bruge den spatiale kontekst som et alternativ eller supplement til den temporale kontekst. Denne mulighed undersøges i sammenhæng med registreringen af 3D bevægelsen af en arm givet billeder fra et statisk kamera.

Afhandlingen består af tre dele. I den første del beskrives hvordan man kan registrere bevægelserne af den menneskelige krop. I den anden del beskrives hvordan den spatiale kontekst kan udnyttes til at lave en mere kompakt repræsentation af armen ved at inkludere information fra billeder. Konkret vises det hvordan de primære frihedsgrader i skulderen og armen kan modelleres effektivt ud fra kendskab til positionen af hånden i billederne. Denne fremgangsmåde benævnes "local screw axis model". Ydermere vises det i denne del hvordan antallet af mulige konfigurationer kan reduceres ved at udnytte den spatiale kontekst. Seks forskellige metoder præsenteres. I den tredje del demonstreres effekten af at benytte den spatiale kontekst. Dette gøres ved at beskrive tre systemer hvori konceptet er benyttet.

De vigtigste resultater i denne afhandling er først og fremmest "the local screw axis model" som gør det muligt, at modellere de 12 primære frihedsgrader i skulderen og armen vha. blot to parametre. Næstvigtigst er de seks forskellige metoder til reduktion af antallet af mulige konfigurationer. De muliggør en gennemsnitlig reduktion på 97.3%. Det konkluderes, at begge resultater viser, at den spatiale kontekst kan bruges til at undgå de u hensigtsmæssige konsekvenser som den temporal tilgang har.



## Preface

This thesis is the documentation of the primary research I have conducted in the period 1998 - 2003. In 1998 I enrolled as a Ph.D.-student at the Laboratory of Image Analysis (to become the Computer Vision and Media Technology Laboratory (CVMT) ). The Ph.D.-scholarship was a part of the national project "The Staging of Virtual Inhabited 3D Spaces" funded by the Danish National Research Councils. After the scholarship expired in 2001 I have worked as a research assistance, in industry, and since the summer of 2002 as an assistant professor at CVMT.

I would like to thank a number of people who have assisted me in various ways throughout the last five years. First of all I would like to thank Erik Granum for his effort in establishing the funding for my scholarship and for being my supervisor for the last five years.

I would like to thank the staff and students at CVMT in the last five years for making the lab. an inspiring and dynamic environment where I enjoy to work.

I would like to thank the following people for taking the time to read and provide useful comments on parts of the text in this thesis. Erik Granum, Claus B. Madsen, Moritz Störring, Hanne E. Andreasen, Lars Qvortrup, Lone Koefoed Hansen, Volker Krueger, and Lorna Herda.

I would like to thank the following people who, numerous times and without complaining, have helped me solving tedious practical computer problems. Moritz Störring, Jørgen Bjørnstrup, and Claus S. Andersen.

Finally I would like to thank my family, especially Hanne, for keeping the faith in me and for bearing with me when working long hours and often having my mind elsewhere. This would not have been possible without you!

Thomas B. Moeslund  
Aalborg, October 2003

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Model-Based Approaches . . . . .	2
1.2	The Focus of the Thesis . . . . .	2
1.3	The Outline of this Thesis . . . . .	3
1.4	The Publications of this Ph.D.-work . . . . .	7
<b>I</b>	<b>Human Motion Capture</b>	<b>11</b>
<b>2</b>	<b>Interacting with a Virtual World through Motion Capture</b>	<b>13</b>
2.1	Introduction . . . . .	15
2.2	Motion Capture . . . . .	15
2.3	Devices Used for Capturing Motion . . . . .	16
2.3.1	Active Sensing . . . . .	17
2.3.2	Passive Sensing . . . . .	19
2.3.3	Complexity of Different Devices . . . . .	20
2.4	Motion Capture Used in Control Applications . . . . .	21
2.4.1	Interacting with the Real World Through a MoCap System . . . . .	22
2.4.2	Interacting with a Virtual World Through a MoCap System . . . . .	23
2.5	Discussion . . . . .	27
	References . . . . .	28
<b>3</b>	<b>A Survey of Computer Vision-Based Human Motion Capture</b>	<b>31</b>
3.1	Introduction . . . . .	34
3.1.1	Application Areas . . . . .	34
3.1.2	Alternative Technologies for Motion Capture . . . . .	34
3.1.3	Content of this Paper . . . . .	35

---

3.2	Surveys and Taxonomies . . . . .	35
3.2.1	Previous Surveys . . . . .	36
3.2.2	A Taxonomy based on Functionalities . . . . .	36
3.2.3	Assumptions . . . . .	41
3.3	Initialisation . . . . .	43
3.4	Tracking . . . . .	44
3.4.1	Figure-Ground Segmentation . . . . .	45
3.4.2	Representation . . . . .	48
3.4.3	Tracking over Time . . . . .	50
3.5	Pose Estimation . . . . .	51
3.5.1	Model-Free . . . . .	51
3.5.2	Indirect Model Use . . . . .	53
3.5.3	Direct Model Use . . . . .	54
3.6	Recognition . . . . .	61
3.6.1	Static Recognition . . . . .	62
3.6.2	Dynamic Recognition . . . . .	62
3.7	Discussion . . . . .	64
3.7.1	Performance Characterisation . . . . .	64
3.7.2	State of the Art . . . . .	65
3.7.3	Future Directions . . . . .	67
3.8	Conclusion . . . . .	69
	References . . . . .	71

## II Spatially-Based Pruning of the State-Space 85

<b>4</b>	<b>Deriving a Compact State-Space Representation by Applying Low-Level Image Features</b>	<b>87</b>
4.1	Introduction . . . . .	89
4.1.1	Outline of this Chapter . . . . .	89
4.2	The DoF of the Arm . . . . .	89
4.3	Modelling the Gleno-Humeral Joint and the Elbow Joint . . . . .	92
4.3.1	Cartesian Coordinates . . . . .	94
4.3.2	Angle Representation . . . . .	94
4.3.3	Screw Axis Representation . . . . .	97

---

4.3.4	Approximated Screw Axis Representation . . . . .	99
4.3.5	The Sensorimotor Transformation Model . . . . .	99
4.4	The Local Screw Axis Model . . . . .	101
4.4.1	Defining the Local Screw Axis Model . . . . .	103
4.4.2	Eliminating the Effect of the Prismatic Joints . . . . .	105
4.4.3	Relating $\phi$ and $\Delta v$ . . . . .	106
4.4.4	Relating $\phi$ and $\Delta h$ . . . . .	110
4.4.5	Evaluating the Modelling Approach . . . . .	111
4.4.6	Relating the Position of the Hand and $\phi$ . . . . .	113
4.5	Discussion . . . . .	116
4.5.1	Summary . . . . .	116
4.5.2	Contributions . . . . .	118
4.5.3	Conclusion . . . . .	118
	References . . . . .	118
<b>5</b>	<b>Pruning the State-Space Representation using Extrinsic and Intrinsic Object Characteristics</b>	<b>123</b>
5.1	Introduction . . . . .	125
5.1.1	Extrinsic and Intrinsic Constraints . . . . .	125
5.1.2	Outline of this Chapter . . . . .	127
5.2	Pruning $H_z$ using a Distance Constraint . . . . .	127
5.2.1	General Pruning Effect . . . . .	128
5.2.2	Minimum, Maximum, and Average Pruning Effects . . . . .	128
5.2.3	Summary . . . . .	129
5.3	Pruning $H_z$ using Angle Constraints . . . . .	129
5.3.1	General Pruning Effect . . . . .	130
5.3.2	Minimum, Maximum, and Average Pruning Effects . . . . .	130
5.3.3	Summary . . . . .	131
5.4	Pruning $H_z$ using an Occlusion Constraint . . . . .	132
5.4.1	General Pruning Effect . . . . .	132
5.4.2	Minimum, Maximum, and Average Pruning Effects . . . . .	133
5.4.3	Summary . . . . .	133
5.5	Pruning $H_z$ using Temporal Constraints . . . . .	134

---

5.5.1	Finding the Minimum, Maximum and Average Displacements of $\theta_1$ and $\theta_4$ . . . . .	134
5.5.2	General Pruning Effect . . . . .	139
5.5.3	Minimum, Maximum, and Average Pruning Effects . . . . .	141
5.5.4	Summary . . . . .	144
5.6	Pruning $\alpha$ using Joint Angle Constraints . . . . .	145
5.6.1	Using $\theta_4$ as a Constraint . . . . .	145
5.6.2	Constraints using $\theta_1$ , $\theta_2$ , and $\theta_3$ . . . . .	146
5.6.3	Combined Pruning Effects for all Four Joint Angles . . . . .	154
5.6.4	Summary . . . . .	154
5.7	Pruning $\alpha$ using a Collision Constraint . . . . .	154
5.7.1	Classical Approach . . . . .	155
5.7.2	A More Efficient Approach . . . . .	156
5.7.3	Minimum, Maximum, and Average Pruning Effects . . . . .	158
5.7.4	Summary . . . . .	159
5.8	Overall Pruning Effects . . . . .	159
5.8.1	Overall Pruning Effects of $H_z$ . . . . .	160
5.8.2	Overall Pruning Effects of $\alpha$ . . . . .	163
5.8.3	Overall Combined Pruning Effects . . . . .	165
5.8.4	Temporal Aspects . . . . .	165
5.9	Discussion . . . . .	166
5.9.1	Assumptions . . . . .	167
5.9.2	Uncertainties . . . . .	168
5.9.3	Conclusion . . . . .	169
	References . . . . .	170

### III Applying the Spatially Pruned State-Space Representation in a Model-Based Framework 171

<b>6</b>	<b>Estimating the 3D Shoulder Position using Monocular Vision and a Detailed Shoulder Model</b> . . . . .	<b>173</b>
6.1	Introduction . . . . .	176
6.1.1	The Approach . . . . .	176
6.2	Estimating the Positions of the Hand . . . . .	177

---

6.3	Estimating the Ellipse . . . . .	177
6.3.1	Correcting the Measurements . . . . .	178
6.4	Estimating the 3D Shoulder Position . . . . .	179
6.4.1	Estimating the Intersection Plane . . . . .	181
6.4.2	Estimating the Circle Centre . . . . .	181
6.4.3	Finding the Correct Solution . . . . .	182
6.5	Tests and Results . . . . .	183
6.5.1	Testing the Ellipse Fitting Algorithm . . . . .	183
6.5.2	Testing the 3D Shoulder Position . . . . .	183
6.5.3	Testing on Real Data . . . . .	185
6.6	Discussion and Conclusion . . . . .	185
	References . . . . .	186
<b>7</b>	<b>3D Human Pose Estimation using 2D-Data and an Alternative Phase Space Representation</b>	<b>189</b>
7.1	Introduction . . . . .	192
7.2	The Approach . . . . .	192
7.2.1	Scenario and Model Complexity . . . . .	193
7.2.2	Content of the Paper . . . . .	193
7.3	Initialisation and Segmentation . . . . .	194
7.3.1	Initialisation . . . . .	194
7.3.2	Segmentation . . . . .	194
7.4	The Phase Space . . . . .	195
7.5	Pruning the Phase Space . . . . .	196
7.5.1	Constraints on $H_z$ . . . . .	196
7.5.2	Constraints on $(\alpha, H_z)$ . . . . .	197
7.6	Spatial Image Features . . . . .	197
7.6.1	Silhouette Matching . . . . .	198
7.6.2	Box Matching . . . . .	199
7.6.3	Combining the Two Methods . . . . .	201
7.7	Results . . . . .	202
7.8	Discussion . . . . .	202
7.9	Conclusion . . . . .	203
	References . . . . .	205

---

<b>8</b>	<b>Improving Sequential Monte Carlo Tracking by Bootstrapping</b>	<b>207</b>
8.1	Introduction . . . . .	210
8.1.1	Multi-Modal Distributions . . . . .	210
8.1.2	The Content of this Paper . . . . .	212
8.2	Prediction and Bootstrapping . . . . .	213
8.2.1	Bootstrapping . . . . .	214
8.3	Modelling the Arm - The State-Space Representation . . . . .	214
8.3.1	Bootstrapping the Screw Axis Representation . . . . .	215
8.4	Image- and Object Representations . . . . .	217
8.4.1	Estimating the Orientations in the Image . . . . .	218
8.5	Bootstrapping the SMC Algorithm . . . . .	220
8.5.1	Summary of Algorithm . . . . .	222
8.5.2	Issues in the SMC Algorithm . . . . .	223
8.6	Results . . . . .	226
8.7	Discussion and Conclusion . . . . .	228
8.7.1	Summary . . . . .	228
8.7.2	Discussion . . . . .	231
8.7.3	Conclusion . . . . .	234
	References . . . . .	234
<b>9</b>	<b>Conclusion</b>	<b>239</b>
9.1	Contributions . . . . .	240
9.2	Discussion . . . . .	241
9.2.1	The hypotheses . . . . .	241
9.2.2	Generality of the Approach . . . . .	242
9.2.3	Uncertainties . . . . .	242
9.2.4	Probabilistic Pruning . . . . .	242
9.3	Further Research Topics . . . . .	243
9.3.1	Static Torso . . . . .	243
9.3.2	The Shoulder Model . . . . .	243
9.3.3	Dependency between the Joint Angles . . . . .	244
9.3.4	Pruning $H_z$ using Temporal Constraints . . . . .	244
9.3.5	Stochastic Features in Bootstrapping . . . . .	244
	References . . . . .	245

---

<b>A Review of Additional Papers</b>	<b>247</b>
Additional References . . . . .	251
<b>B Converting Rodrigues' Formula into Matrix Form</b>	<b>257</b>
References . . . . .	258
<b>C Solution to the Transcendental Equation</b>	<b>259</b>