Statistical Models for Human Body Pose Estimation from Videos

A dissertation submitted to the

ETH ZURICH

for the degree of

Doctor of Sciences

presented by

Tobias Jaeggli
MSc, Université de Fribourg
born July 26, 1976
citizen of Winterthur ZH

accepted on the recommendation of

Prof. Dr. Luc Van Gool, examiner
Prof. Dr. Cristian Sminchisescu, co-examiner

2008
Abstract

To investigate the task of multidimensional continuous inference from video sequences on a concrete example application, we focus on the problem of articulated 3D human tracking from monocular video. This is an interesting topic because of its relevance for biological vision systems, as well as its many applications in various domains. Estimating body pose and motion of humans is a challenging task, with difficulties such as self-occlusions and ambiguities. To account for unresolvable uncertainties of the visual analysis of such footage, we formulate the task as a probabilistic inference problem. The pose estimation and tracking algorithms are based on statistical models that can be automatically learned from a set of example data. Thanks to this architecture, the proposed approaches remain general and can be tailored to a specific task by the choice of training data sets. Prior knowledge can be provided in a flexible and theoretically well-motivated way. First, we propose an approach that is based on a model of the joint probability distribution of body pose and the corresponding human shape, as it can be observed in video images. Both body pose and shape are treated as multivariate random variables, by choosing suitable representations. The statistical model uses a mixture of Gaussian distributions to approximate the density, which enables efficient discriminative inference of body poses from shape descriptors. When additionally taking the unknown image locations of the persons into account, the posterior distributions become non-parametric. Therefore, a hybrid inference scheme based on a Rao-Blackwellised particle filter combines parametric inference with sample based inference. A second approach is based on a generative predictive model of human shape, using nonlinear regression. To enable efficient learning and sample based inference, a low-dimensional embedding of human locomotion is determined, with a nonlinear dynamical model. This method is implemented using Locally Linear Embedding, and Relevance Vector Machines for sparse nonlinear regression. We also propose an integrated formulation of the model, fully based on Gaussian Process regression. The resulting tracking algorithms are tested on realistic video sequences with low resolution and image noise. We present extensions of the framework, for simultaneously tracking multiple persons that occlude each other, and for recognising the performed activity along with the pose estimation.
Zusammenfassung

Acknowledgements

I would like to thank the following people and institutions for their help and support.

Esther Koller-Meier • Luc Van Gool • Cristian Sminchisescu • Eidgenössische Technische Hochschule Zürich ETH • Swiss National Science Foundation (SNF), for support through the Project PICSEL and the SNF NCCR IM2 • The Commission of the European Union, for support through the Project DIRAC (IST-027787) • Stephan Gammeter • Bastian Leibe • Andreas Ess • Konrad Schindler • Daniel Roth, Martin Von Siebenthal, Raphael Hoever, Robert McGregor, Philippe Cattin, for performing on the treadmill • and all my current and former colleagues at BIWI • Barbara Widmer • Vreni Vogt • Christina Krueger • Thomas Koninckx • Geert Caenen • Rik Fransens • Joris Vanden Wyngaerd • and my former colleagues at VISICS • Beni Rohrer • My parents • Monica Jaeggli
# Contents

List of Abbreviations xi

Notation xiii

1 Introduction
1.1 Ambiguous and Uncertain Inference from Images . . . . . . . . 1
1.2 Monocular Human Pose Estimation . . . . . . . . . . . . . . 1
1.3 Contributions . . . . . . . . . . . . . . . . . . . . . . . . . . . 2
1.4 Outline of the Thesis . . . . . . . . . . . . . . . . . . . . . . . 3

2 Inference from Partial Measurements 5
2.1 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . 5
2.2 Related Work: Partial Observations . . . . . . . . . . . . . . 6
2.3 Inferring Missing Observations . . . . . . . . . . . . . . . . . 8
   2.3.1 Partial Observations as Affine Linear Subspaces . . . . . 8
   2.3.2 Combining Observation with Gaussian Prior Model . . . . 9
   2.3.3 Unknown Alignment Transformations . . . . . . . . . . . 10
   2.3.4 Extension to Gaussian Mixture Models . . . . . . . . . . 11
   2.3.5 Uncertainty in Observations . . . . . . . . . . . . . . . 12
   2.3.6 Working on a Low-Dimensional Subspace . . . . . . . . 13
2.4 Example Applications . . . . . . . . . . . . . . . . . . . . . . 15
   2.4.1 3D Face Reconstruction . . . . . . . . . . . . . . . . . 15
   2.4.2 Body Pose Estimation . . . . . . . . . . . . . . . . . . 17

3 Statistical Approach to Tracking and Pose Estimation 21
3.1 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . 21
3.2 Related Work . . . . . . . . . . . . . . . . . . . . . . . . . . . 25
   3.2.1 Articulated Tracking . . . . . . . . . . . . . . . . . . . 26
   3.2.2 Learning Based Tracking . . . . . . . . . . . . . . . . . 29
   3.2.3 Multi-Person Articulated Tracking . . . . . . . . . . . 35
   3.2.4 Activity and Motion Recognition . . . . . . . . . . . . 36
   3.2.5 Learning Methods . . . . . . . . . . . . . . . . . . . . . 37
3.3 Proposed Overall Framework for Tracking and Pose Estimation 39
### Contents

3.4 Representations ........................................... 41  
3.4.1 Body Pose Representations ............................... 42  
3.4.2 Shape Representations .................................... 45  
3.5 Training Data Sets .......................................... 49  
3.5.1 LocoETH Dataset ........................................... 49  
3.5.2 CMU MocapData ............................................ 56

4 Combining Discriminative and Generative Inference in a Rao-Blackwellised Particle Filter 57  
4.1 Introduction .................................................. 57  
4.2 Modelling and Learning ...................................... 59  
4.3 Bayesian Recursive Filtering ................................. 60  
4.3.1 Monte Carlo integration .................................... 62  
4.3.2 Analytical Pose Inference .................................. 63  
4.3.3 Rao-Blackwellised particle filter ......................... 64  
4.4 Implementation and Experimentation ....................... 67  
4.4.1 Representations and Dimensionality Reduction ......... 67  
4.4.2 Dataset and Training ...................................... 68  
4.4.3 Tracking Experiments ..................................... 68  
4.5 Discussion ..................................................... 76

5 Generative Regression-Based Model for Pose Estimation 79  
5.1 Introduction .................................................. 79  
5.2 Modelling and Learning ...................................... 81  
5.2.1 Relevance Vector Machines ............................... 82  
5.2.2 Low-dimensional Pose Manifold ......................... 83  
5.2.3 Temporal Prior ............................................. 85  
5.2.4 Appearance Model ........................................ 86  
5.3 Tracking and Pose Inference ................................. 87  
5.3.1 On-Line Sample-Based Tracking ......................... 87  
5.3.2 Off-Line Global Optimisation ............................. 90  
5.4 Experiments .................................................. 93  
5.4.1 Empirical Ambiguity Analysis ............................ 93  
5.4.2 Tracking Results .......................................... 95  
5.5 Discussion ..................................................... 102

6 Articulated Multi-Body Tracking Based on Gaussian Processes105  
6.1 Introduction .................................................. 105  
6.2 Modelling with Gaussian Processes ......................... 107  
6.2.1 Gaussian Process Regression ............................. 107  
6.2.2 Pose, Appearance and Dynamics ......................... 108  
6.2.3 Learning the Embedding ................................ 110
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3</td>
<td>Integrated Multi-Person Tracking System</td>
<td>111</td>
</tr>
<tr>
<td>6.3.1</td>
<td>System Overview</td>
<td>111</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Multi-Body Tracker</td>
<td>112</td>
</tr>
<tr>
<td>6.3.3</td>
<td>Articulated Tracking</td>
<td>113</td>
</tr>
<tr>
<td>6.3.4</td>
<td>Guided Adaptive Segmentation</td>
<td>114</td>
</tr>
<tr>
<td>6.4</td>
<td>Experiments</td>
<td>115</td>
</tr>
<tr>
<td>6.5</td>
<td>Discussion</td>
<td>118</td>
</tr>
<tr>
<td>7</td>
<td>Inferring Activity Categories</td>
<td>121</td>
</tr>
<tr>
<td>7.1</td>
<td>Modelling and Learning</td>
<td>121</td>
</tr>
<tr>
<td>7.2</td>
<td>Tracking with Activity Recognition</td>
<td>122</td>
</tr>
<tr>
<td>7.3</td>
<td>Experiments</td>
<td>123</td>
</tr>
<tr>
<td>7.4</td>
<td>Discussion</td>
<td>125</td>
</tr>
<tr>
<td>8</td>
<td>Conclusion</td>
<td>129</td>
</tr>
<tr>
<td>8.1</td>
<td>Summary and Comparison of Frameworks</td>
<td>129</td>
</tr>
<tr>
<td>8.2</td>
<td>Perspectives</td>
<td>132</td>
</tr>
<tr>
<td>A</td>
<td>Equivalence of Gaussian Section and Gaussian Conditioning</td>
<td>135</td>
</tr>
<tr>
<td>Bibliography</td>
<td>137</td>
<td></td>
</tr>
<tr>
<td>Curriculum vitae</td>
<td>151</td>
<td></td>
</tr>
</tbody>
</table>
### List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPCA</td>
<td>Binary Principal Component Analysis</td>
</tr>
<tr>
<td>(k)CCA</td>
<td>(kernel) Canonical Correlation Analysis</td>
</tr>
<tr>
<td>DOF</td>
<td>Degree of Freedom</td>
</tr>
<tr>
<td>DV</td>
<td>Digital Video</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation - Maximisation</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian Mixture Model</td>
</tr>
<tr>
<td>GP</td>
<td>Gaussian Processes</td>
</tr>
<tr>
<td>GPDM</td>
<td>Gaussian Process Dynamical Model</td>
</tr>
<tr>
<td>GPLVM</td>
<td>Gaussian Process Latent Variable Model</td>
</tr>
<tr>
<td>LLE</td>
<td>Locally Linear Embedding</td>
</tr>
<tr>
<td>MAP</td>
<td>maximum a posteriori</td>
</tr>
<tr>
<td>MCMC</td>
<td>Markov Chain Monte-Carlo</td>
</tr>
<tr>
<td>MLR</td>
<td>Multivariate Linear Regression</td>
</tr>
<tr>
<td>MoCap</td>
<td>Motion Capture</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PC</td>
<td>Principal Component</td>
</tr>
<tr>
<td>pdf</td>
<td>probability density function/probability distribution</td>
</tr>
<tr>
<td>PLS</td>
<td>Partial Least Squares</td>
</tr>
<tr>
<td>RBF</td>
<td>Radial Basis Function</td>
</tr>
<tr>
<td>RBPF</td>
<td>Rao-Blackwellised Particle Filter</td>
</tr>
<tr>
<td>RVM</td>
<td>Relevance Vector Machine</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
</tbody>
</table>
Notation

\( X \) a random variable, possibly multivariate
\( x \) a vector
\( X \) a matrix, with rows \( x_i, i \in \{1, \ldots, N\} \), and \( N \) the number of rows
\( x_t \) state variable \( x \) at discrete time step \( t \in \{1, \ldots, T\} \), where \( T \) is the number of time steps
\( x_{1:t} \) accumulated state variables \( x_t \) from timestep 1 to \( t \), i.e. \( \{x_j, j \in \{1, \ldots, t\}\} \)
\( \mathcal{N}(x; \mu, \Sigma) \) a multivariate normal distribution with mean \( \mu \) and covariance matrix \( \Sigma \)

Terminology

The expressions \textit{Body Pose Estimation}, \textit{Articulated Tracking} and \textit{3D Body Tracking}, at the core of the presented work, are used interchangeably within this text. Similarly, \textit{body pose}, \textit{posture}, \textit{articulation} or \textit{configuration} are treated as synonyms. Another central concept, \textit{appearance descriptors}, will also be referred to as \textit{appearance representations}, \textit{image/shape descriptors} or \textit{(image) features}. 
1

Introduction

1.1 Ambiguous and Uncertain Inference from Images

For many machine vision tasks it is not possible to determine a single result with confidence. Often, multiple interpretations of the input images are possible, or the result remains uncertain in some aspects while the uncertainty of other aspects can be removed by analysing the images. In some cases, a temporal accumulation of image information can help disambiguation, but the state estimates are still temporarily uncertain. These challenges can be dealt with by formulating the image analysis tasks as probabilistic inference problems. This framework allows for handling remaining uncertainties and multimodalities of the estimation. Furthermore, it readily accommodates statistical models that encode prior knowledge about the problem domain, which is usually essential for machine vision tasks. While prior knowledge in the form of a basic understanding of the problem at hand is necessary for model design, many interesting statistical properties can automatically be estimated from example data using machine learning techniques. The statistical framework thus offers a principled way to incorporate prior knowledge and to express uncertain and ambiguous posterior beliefs.

In this thesis, statistical modelling and inference approaches are investigated, in a general manner, and mainly for the application to concrete computer vision tasks. We formulate a general framework that deals with partial observability and the estimation of unknown quantities from incomplete observations. These tools are applicable to a broad variety of tasks, a few of which are sketched in chapter 2.

1.2 Monocular Human Pose Estimation

The subsequent chapters 3 to 7 focus on a specific task - tracking and pose estimation of humans - and in particular the use of statistical models in this
context. Pose estimation is a challenging task with many potential applications in the domains of entertainment industry, surveillance, human-machine interfaces or medical diagnosis. Furthermore it is of great scientific interest, and suffers from many difficulties such as estimation ambiguities, occlusions and high-dimensional representation spaces. These difficulties can be approached by providing prior knowledge about body motion in some form. In contrast to previous work, the proposed approaches rely on learned models rather than geometrical models of the human body. These are trained on real data of human motion and corresponding synthetic appearance data; or real appearance data if available. The proposed approaches include discriminative and generative inference strategies, using linear or nonlinear statistical tools. Furthermore, parametric and sample-based representations of the multimodal posterior probability density functions are presented, compared and combined.

1.3 Contributions

The following main contributions are made in this thesis:

- The 2D location tracking and the pose estimation are treated as a two interleaved process that closely collaborate.
- We investigate different ways to model appearance, pose and dynamics of human motion, and exploit these models for tracking and pose estimation.
- A combination of analytical and sample-based inference is proposed for the simultaneous estimation of bounding box location and body articulation.
- A generative pipeline with a non-linearly embedded manifold, a dynamical model and a learned appearance model.
- Activity recognition is performed by estimating an additional activity label along with the articulated tracking.
- An extension to multi-object articulated tracking featuring global occlusion reasoning, and integrating different image cues.
- We present results on challenging real traffic sequences with low-resolution imagery and noise.
1.4 Outline of the Thesis

In chapter 2, a probabilistic framework is introduced, that allows for inference from partial measurements. The proposed set of tools is applicable in many situations where only a subset of modelled features are observable, while others are unseen, *e.g.* due to occlusions.

Chapters 3, 4, 5 and 6 build a main block of the thesis, concerned with articulated tracking and body pose estimation. First, in chapter 3, the problem is stated and the difficulties of the task as well as related work are discussed. A general framework for learning based pose estimation and tracking is introduced. Finally, an overview of representations for body articulations and image appearance are given, followed by a description of the data sets that were used for training the statistical models.

Then, in chapters 4, 5 and 6, three actual realisations and implementations of the tracking framework are presented, with different modelling approaches and based on various machine learning techniques. Chapter 4 deals with a modelling approach that is based on linear tools. For tracking, analytical and sample-based inference are combined in a Rao-Blackwellised Particle Filter. In chapter 5, a generative modelling approach uses nonlinear techniques for manifold learning and regression-based appearance prediction. Chapter 6 recasts the generative approach in a Gaussian Process framework, allowing for a more compact formulation of the learning task. The algorithm is then integrated into a multi-person tracking system that combines occlusion-reasoning and different image cues.

In chapter 7, we present an extension that performs activity recognition in addition to articulated tracking, by using a model-switching strategy.

Chapter 8 concludes the thesis with a comparison of the different proposed tracking approaches, a summary, and a discussion of possible continuative research directions.
2

Inference from Partial Measurements

2.1 Problem Statement

The data that can be retrieved from images or video sequences are often incomplete. Computer vision techniques that try to track and understand the motion of articulated objects have to deal with parts of the objects being invisible to the camera due to occlusions. For instance if a moving human body is observed, only a few points are visible during the whole sequence. Other points are only temporarily visible or stay completely occluded. Given a model of how humans move or perform specific actions, likely positions of the remaining body parts can be predicted. In some cases it may even be possible to recognise the activity of the subject, even if only a small subset of the body parts at a given time can be observed. Also, some observations may be more informative than others, i.e. reduce the uncertainty about how the subject moves in a more effective way and tell us more about the performed activity.

Similar problem settings occur in many different domains, related to computer vision, or statistical modelling and learning in general. Often the set of observed and unobserved features dynamically varies over time, while for other applications, it is clear beforehand which of the variables can be observed, and which can not. For example, when observing a 3D scene with a single camera, depth information is lost during the image formation process. It lies thus in the nature of the problem, that part of the information is missing and has to be made up for in some other way than directly from the images.

In this chapter, a probabilistic framework is derived, that allows to determine the remaining uncertainty given a set of observations. A parametric statistical model which has been learned from training data is used to predict occluded or missing features. The resulting distribution over the model parameters can be used for further inference steps, in substitution of a single noiseless measurement.
The key of the proposed approach is the representation of missing information as subspaces or with degenerate Gaussian distributions. Unobserved or missing variables are treated the same as uncertainty that is caused by unknown alignment transformations, such as an unknown scaling or offset of the measurements. Previous authors have often estimated those alignment operations in a preprocessing step. While the mathematics of the inference procedure are relatively simple, the topic has rarely been investigated in a general, comprehensive and flexible manner, which we try to do here. Many authors have described application specific and less general methods, of which many are in fact equivalent.

In this chapter, related work is reviewed first. Then, in section 2.3 the theoretical framework is introduced, starting with the basic technique, and followed by extensions that allow for more complex models, unknown alignment transformations, noisy measurements, and dimensionality reduction. After the general theoretical outline, section 2.4 sketches applications of the technique to face reconstruction and the analysis of human locomotion. It is shown that additional unknowns such as alignment, scale, or more problem specific properties such as walking-speed can be incorporated into the framework in a straightforward manner.

The inference framework is also described in [Jaeggli et al., 2005], and sets the theoretical background of the learning-based tracking approach introduced in chapter 4, where a model is learned from joint observations of body pose and body shape. At tracking time only shapes can be observed from images, while body poses are unobserved, but can be estimated given the learned model.

2.2 Related Work: Partial Observations

In many fields, researchers have been confronted with the problem of partial observability. An in-depth review of the topic, in particular for the field of statistical modelling of anatomical shapes, is found in [Zsemlye, 2005].

The general solution to the simplest case considered in this work, is known as Gaussian Conditioning and related to multiple linear regression ([Anderson, 1958; Draper and Smith, 1998], see section 2.3.2 for the equations). The method allows to predict unobserved vector elements given partial observations and a Gaussian model of the joint distribution of all elements.

An application area that is strongly related to the main topic of this thesis is investigated in [Leventon and Freeman, 1998]. From 2D image measurements 3D body motions are estimated. A prior model of likely body poses and motions is learned, the image coordinates can thus be seen as incomplete
observations where the depth would be the missing information. The posterior
is obtained in closed form for Gaussian distributions, and is also a Gaussian
pdf (probability density function). In [Howe et al., 1999], the approach is ex-
tended to Gaussian Mixture Models (GMM) for the prior. Given the output
of a 2D tracker, the most probable 3D reconstruction is obtained using Ex-
pectation Maximisation (EM). In [Grochow et al., 2004] the most likely body
motion that satisfies incomplete 3D measurements of body locations is found.
The method is based on a Gaussian Process Latent Variable Model (GPLVM,
[Lawrence, 2005]) framework and performs an iterative optimisation in a latent
space that generates the original 3D poses.

For statistical shape modelling from partial measurements, several approaches
have been proposed [Hug et al., 2000; Hug, 2001; Rajamani et al., 2004;
Zsemlye, 2005; Hwang et al., 2000; Blanz and Vetter, 2002; Blanz et al., 2004].
These approaches model a prior pdf of expected shapes on a lower dimensional
subspace [Cootes and Taylor, 1992; Blanz and Vetter, 1999], and find the best
parameters of the shape model given measurements for a small number of
points on the shape. Shape modelling is either done in 2D [Hug et al., 2000;
Hug, 2001; Hwang et al., 2000] or 3D [Rajamani et al., 2004; Blanz and Vet-
ter, 2002; Blanz et al., 2004]; in the latter case the measurements are typ-
ically available in terms of 2D image coordinates [Blanz and Vetter, 2002;
Blanz et al., 2004], where in addition to the unknown shape points, the depth
measurements are missing. In [Hug, 2001; Hug et al., 2000; Hwang et al.,
2000] the measurement constraints are required to be satisfied exactly while
maximising the probability of the shape given the training examples, i.e. min-
imising the Mahalanobis distance to the mean shape in closed form. These ap-
proaches are shown to be equivalent to ‘multiple linear regression’ in [Zsemlye,
2005]. A Bayesian approach is followed in [Blanz and Vetter, 2002; Blanz et al.,
2004], that finds a trade-off between the most likely shape, and the best recon-
struction of the measurements, and thus does allow for noisy measurements.
Particularly suited for interactive shape initialisation, in [Hug et al., 2000;
Rajamani et al., 2004] an iterative method is proposed, where the remaining
variability is eliminated progressively by adding image measurements one by
one.

Many approaches perform alignment in a preprocessing step ([Zsemlye, 2005]).
Offset correction, rotation and scale estimation are considered in [Blanz et al.,
2004] by modifying the subspace in which the result must lie. Additionally, the
measurement constraints can be relaxed by specifying straight lines on which a
certain model point must lie. In the approach proposed here, these two types
of alignment transformations are both included in the representation of the
incomplete measurement, without assuming a model-subspace.
The main contribution proposed here is the description of a unifying approach that can serve as a general tool, based on simple and intuitive geometrical reasoning. This allows for an extension to more complex cases where low-dimensional subspaces, mixture models, alignment transformations and noisy measurements are involved. The approach includes many of the mentioned approaches as special cases, while being independent from an actual application. Furthermore, in contrast to previous work, entire posterior pdfs are derived rather than just expected values, and closed form solutions are given.

2.3 Inferring Missing Observations

We consider $D$-dimensional vectors $\mathbf{x} \in \mathbb{R}^D$. A partial observation is denoted $\tilde{\mathbf{x}}$ and is characterised by a set $\mathbf{x}_J$ of observed vector elements $x_j, j \in J$ and a set $\mathbf{x}_{\sim J}$ of unknown elements $x_i, i \notin J$, where $J$ is a subset of the indices $\{1, \ldots, D\}$. $\# J$ is the number of observed elements.

Additionally, we are provided with a model of the probability density function $p(\mathbf{x})$, typically learned from training data in a preprocessing step. In the simplest case considered here, $p(\mathbf{x})$ is a multivariate normal distribution with mean $\mu$ and covariance matrix $\Sigma$.

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mu, \Sigma)$$  \hspace{1cm} (2.1)

Given this information, the unknown elements of $\tilde{\mathbf{x}}$ are inferred. Instead of just seeking the most likely completion of $\tilde{\mathbf{x}}$, we are interested in the entire probability distribution over $\mathbf{x}$, conditioned on the measurements of a subset of its elements $\mathbf{x}_J$.

2.3.1 Partial Observations as Affine Linear Subspaces

Given an actual measurement $\mathbf{a} \in \mathbb{R}^{\# J}$ for the known vector elements, the set of possible completions of $\tilde{\mathbf{x}}$ that fulfil the $\# J$ constraints determined by $\mathbf{a}$ can be written as an affine linear subspace $M$.

$$M = \{ \mathbf{x} | x_j = a_j, j \in J \} \text{ with } J \subset \{1, \ldots, D\}$$  \hspace{1cm} (2.2)

$M$ has dimension equal to $D - \# J$, and can alternatively be written using a parametric representation,

$$M = \{ \mathbf{m}_0 + M_a \mathbf{z} | \mathbf{z} \in \mathbb{R}^{D-\# J} \},$$  \hspace{1cm} (2.3)
where \( \mathbf{z} \) are the local coordinates in the subspace, \( \mathbf{m}_0 \) is its origin and contains
the known entries \( \mathbf{a} \) while the unknown entries are chosen arbitrarily. \( M_u \) is a
matrix containing the canonical basis vectors \( \mathbf{e}_i \) of \( \mathbb{R}^D \) that correspond to the
unknown measurements, \( i.e. \ i \not\in J \).

### 2.3.2 Combining Observation with Gaussian Prior Model

The affine subspace \( M \) describes the hard constraints determined by the actual
observation \( \mathbf{a} \). By including the learned model \( p(\mathbf{x}) \), the conditional pdf \( p(\mathbf{x}|\mathbf{a}) \)
is obtained.

By definition

\[
p(\mathbf{x}|\mathbf{a}) \propto p(\mathbf{x})\theta_M(\mathbf{x}),
\]

where \( \theta_M(\mathbf{x}) = 1 \) iff \( \mathbf{x} \in M \). Eq. (2.4) restricts \( \mathbf{x} \) to lie in the subspace \( M \),
other vectors have probability equal to zero. If \( p(\mathbf{x}) \) is Gaussian, it follows from
the affinity of \( \theta_M(\mathbf{x}) \) that \( p(\mathbf{x}|\mathbf{a}) \) takes a Gaussian form. In local coordinates
of \( M \), its mean \( \mu_M \) and covariance matrix \( \Sigma_M \) are computed by intersecting
\( M \) with the model pdf. See also Fig. 2.1 for an illustration.

\[
\mu_M = \left( M_u^T \Sigma^{-1} M_u \right)^{-1} M_u^T \Sigma^{-1} (\mu - \mathbf{m}_0)
\]

\[
\Sigma_M = \left( M_u^T \Sigma^{-1} M_u \right)^{-1}
\]

(2.5)

The complete pdf \( p(\mathbf{x}|\mathbf{a}) \) in the original observation space \( \mathbb{R}^D \) is given as
\( \mathcal{N}(\mathbf{x}; \mu_x, \Sigma_x) \), with

\[
\mu_x = \mathbf{m}_0 + M_u \mu_M, \quad \Sigma_x = M_u \Sigma_M M_u^T.
\]

(2.6)

The Gaussian section is not a proper pdf, since it does not integrate to 1.
However, \( p(\mathbf{x}|\mathbf{a}) \) is known to be a Gaussian pdf, and its scaling (normalisation)
factor is thus determined by the mean and covariance matrix.

This process corresponds to Gaussian Conditioning, where given a joint nor-
mal distribution of a set of variables, the distribution of any subset of these
variables conditioned on the remaining variables can be computed [Anderson,
1958]. Say that the vector \( \mathbf{x} \) has been partitioned into sub-vectors \( \mathbf{x}_1 \) and \( \mathbf{x}_2 \)
(in our context they correspond to the observed and unknown elements re-
spectively), and similarly the covariance matrix is split into sub-matrices \( \Sigma_{11}, \Sigma_{22}, \Sigma_{12} \) and \( \Sigma_{21} \), then

\[
p(\mathbf{x}_2|\mathbf{x}_1) = \mathcal{N}(\mathbf{x}_2; \mu_{2|1}, \Sigma_{2|1})
\]

\[
\mu_{2|1} = \mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (\mathbf{x}_1 - \mu_1)
\]

\[
\Sigma_{2|1} = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}.
\]

(2.7)
2. Inference from Partial Measurements

If one is only interested in the expected value of the unknown elements $\mu_{2|1}$, the solution is a linear regression, where the $\Sigma_{21}\Sigma_{11}^{-1}$ are the regression coefficients. The equivalence of Gaussian Conditioning and the proposed subspace intersection is proven in Appendix A.

### 2.3.3 Unknown Alignment Transformations

In addition to the fact that only a subset of the dimensions of $x$ are observable, it might be necessary to transform the observation in order to bring it into the coordinate system of the learned model. For example, if $x$ consists of the pixels of a greyscale image that is only partially visible, the pixel intensities might be subject to an unknown scaling because the exposure time of the photo camera changed between two takes, or because the pixel intensities were expressed in different ranges (e.g. in the interval $[0,1]$ instead of $[0,255]$). Such alignment parameters should be included as degrees of freedom into the definition of $M$, next to the missing observations. Affine linear alignment transformations such as unknown scaling or offset correction (translation) operations can be written as affine linear subspaces, in which the transformed observation must lie, similar to the affine subspace of the missing observations. Consequently,
those degrees of freedom can be included by adding appropriate basis vectors to $M_u$ in (2.3), yielding

\[
M = \{m_0 + \begin{bmatrix} M_u & A_u \end{bmatrix} z\} = \{w_0 + W_u z\}. \tag{2.8}
\]

Here, $A_u$ is a matrix consisting of vectors that perform the desired alignment transformations. If scale is considered as a parameter to be estimated, $m_0$ is included in the $A_u$ matrix, and $M$ becomes a linear subspace rather than an affine one, i.e. it contains the origin of $\mathbb{R}^D$. In this case, $w_0$ vanishes, otherwise $w_0$ is equal to $m_0$. Strictly speaking, $M$ also allows for ‘impossible’ configurations such as negative scale; those solutions will however be eliminated by the inclusion of the learned model $p(x)$.

Again, inference is performed according to (2.5), where the local variable $z$ additionally parametrises the unknown alignment operations.

### 2.3.4 Extension to Gaussian Mixture Models

The approach can easily be extended to models of $p(x)$ that are approximated by a mixture of Gaussians (GMM) rather than a single normal distribution.

\[
p(x) = \sum_{k=1}^{K} \alpha_k \ p_k(x) \tag{2.9}
\]

\[
p_k(x) = \mathcal{N}(x; \mu_k, \Sigma_k)
\]

Here, $K$ is the number of Gaussian components, $\alpha_k$, $\mu_k$ and $\Sigma_k$ are the mixture proportions, mean and covariance matrices of the components respectively. The distribution conditioned on the incomplete observation now becomes

\[
p(x|a) = \sum_{k=1}^{K} \beta_k \ p_k(x|a), \tag{2.10}
\]

where $p_k(x|a)$ are the intersections of the Gaussian components with the subspace $M$, according to section 2.3.2. The new mixture proportions are $\beta_k = \alpha_k p_k(M)$, where $p_k(M) = \int_M p_k(x)dx$ denotes the total probability that any measurement in the subspace $M$ can occur according to $p_k(x)$. It can be seen that $p_k(M)$ is thus also the scale of the unnormalised Gaussian section of component $k$. 

2.3.5 Uncertainty in Observations

So far we have assumed that the observation $a$ of the known elements $x_J$ is noise free. Consequently, only completions of $x$ that exactly fulfil the constraints determined by $a$ (or $M$) were considered. As seen in the previous sections, this results in a hard intersection of an affine subspace and a parametric pdf. In practice, however, those observations may be subject to uncertainty as well.

If the distribution of the true values of the observed elements $x_J$ is related to the actual observation $a$ by the addition of zero mean i.i.d. Gaussian noise with variance $\sigma$, we obtain

$$ p(x|a) = \int_{x_J} p(x|x_J)p(x_J|a) $$

$$ = \int_{x_J} p(x|x_J)N(x_J; a, \sigma I), $$

where $p(x|x_J)$ is computed using the hard intersection of (2.5).

To generalise the formulation even further, instead of distinguishing between known and unknown elements of $x$, it may be useful to specify a confidence value for each element, that ranges from ‘very certain’ (low variance) over ‘uncertain’ (larger variance) to ‘unobserved’ (infinite variance) for elements that have not been observed at all. In analogy to the affine subspace $M$ that represents the hard constraints of noise free observations, the ‘soft’ constraints of noisy and thus uncertain observations can be intuitively represented by a Gaussian distribution $p_a(x)$. Similar to (2.4), $p(x|a)$ is now given by the product of two Normal distributions.

$$ p(x|a) \propto p(x)p_a(x) $$

The covariance matrix of $p_a(x)$, denoted $\Sigma_a$, has finite entries that correspond to the noise variance of the known elements $x_J$, and infinitely large variances for the unknown elements. It is thus a degenerate Normal distribution. In practice, only the inverse covariance matrix $\Sigma_a^{-1}$ is needed for the computation of (2.12); it simply has zero 0 entries for the unknown vector elements. The unknown elements thus lead to degeneracies of $p_a(x)$ that are parallel to the axes of the coordinate system, and a diagonal inverse covariance matrix $\Sigma_a^{-1}$ with entries that are equal or larger than 0. When including additional degrees of freedom that are e.g. related to alignment transformations, additional degeneracies of $p_a(x)$ are introduced, and the diagonal structure of $\Sigma_a^{-1}$ may be lost.
In general, expressing the degrees of freedom introduced by missing observations and unknown alignment transformations as a degenerate Gaussian is straightforward. If $W_u$ is the basis of the subspace representation of (2.8), the inverse covariance function of the degenerate Gaussian $p_a(x)$ is given by

$$\Sigma_a^{-1} = U(\sigma I)^{-1}U^T,$$

(2.13)

where $U$ is an orthonormal basis of the nullspace of $W_u$. That is, when $W_u$ represents the degrees of freedom of the subspace $M$, $U$ represents the constraints determined by $M$. The uncertainty of these constraints is denoted by $\sigma$, here chosen to be equal for all of them for notational simplicity. The respective ranks of $U$ and $W_u$ sum to $D$, the total dimensionality of the observation space. The mean of $p_a(x)$ can be chosen to be any point on the subspace $M$, e.g. $w_0$.

Recall the formulas for the product of the two normal distributions in (2.12):

$$p(x|a) \propto \mathcal{N}(x; \mu_x, \Sigma_x)$$

$$\Sigma_x^{-1} = \Sigma^{-1} + \Sigma_a^{-1}$$

$$\Sigma_x^{-1} \mu_x = \Sigma^{-1} \mu + \Sigma_a^{-1} \mu_a$$

(2.14)

As can be seen from these equations, the covariance matrix of the degenerate Gaussian (which does not exist) is not needed, only its inverse that is computed according to (2.13). Again, $\mathcal{N}(x; \mu_x, \Sigma_x)$ is equal to the product (2.12) up to scale, and the normalisation factors are determined by $\mu_x$ and $\Sigma_x$, since $p(x|a)$ is known to be a Gaussian distribution.

Fig. 2.2 shows on a toy example, how the resulting pdf $p(x|a)$ changes for different choices of $\sigma$, the variance of the known observations. As expected, in the two extreme cases where the measurement uncertainty $\sigma$ is either equal to 0 or $\infty$, the solution converges to the result of the hard intersection of equations (2.4) to (2.6), or to $p(x)$ respectively.

### 2.3.6 Working on a Low-Dimensional Subspace

Often, the dimensionality $D$ of $x$ is very high, and it is impractical to carry out the computations in the original observation space. Furthermore, when the number of available training data is low in comparison with $D$, it is not possible to learn a reliable (non-singular) model of $p(x)$ using Normal distributions or GMMs.

Instead, the modelling can be done on a linear subspace of lower dimensionality, e.g. obtained by Principal Component Analysis (PCA). Such a strategy
Figure 2.2: Posterior pdf $p(x|a)$ given an uncertain measurement $a$ of the known elements $x_J$. Four plots are shown, for a toy example with a two dimensional Normal distribution, with different values for the measurement uncertainty $\sigma$ (top: $\sigma = 0.001$, bottom: $\sigma = \{0.01, 0.1, 1\}$). For increasing $\sigma$, the posterior converges towards the prior $p(x)$, i.e. the influence of the measurement diminishes.

has been followed in much of the work on statistical shape modelling [Hug et al., 2000; Blanz and Vetter, 1999; Cootes and Taylor, 1992]. The statistical model $p(y)$ is now learned on a model-space $V$ of dimensionality $d < D$, related to the original space by a linear projection.

$$y = V_u^T(x - v_0), \quad (2.15)$$

where $V_u, v_0$ describe the (PCA-) subspace $V$. In order to express the constraints of the observation $a$ in terms of the subspace $V$ an intermediate step is necessary: By intersecting $M$ and $V$ we obtain a new subspace $M_V$ that contains all $\{x| x \in M \land x \in V\}$, expressed in local coordinates of $V$. $M_V$ is thus the restriction of $V$ to vectors that contain the observed values $a$. Computing this intersection amounts to solving a linear equation system. If the
subspaces do not intersect, i.e. \( M \) is the empty set, the observation \( \mathbf{a} \) cannot be satisfied exactly by the model.

In a similar fashion, if the measurement constraints are expressed by a degenerate Gaussian \( p_{\mathbf{a}}(\mathbf{x}) \) rather than an affine subspace \( M \) (see section 2.3.5), another Gaussian \( p_{\mathbf{a},V}(\mathbf{y}) \) can be computed, that expresses the measurement constraints on the model-space \( V \). This is done by intersecting \( V \) and \( p_{\mathbf{a}}(\mathbf{x}) \); the intersection of a Gaussian pdf and an affine subspace is shown in (2.5). It results in another degenerate Gaussian, if \( M \) and \( V \) intersect, and in a regular Gaussian otherwise.

Now, the remaining steps are performed on the lower dimensional subspace \( V \), by taking into account the learned model, and result in \( p(\mathbf{y}|\mathbf{a}) \), a pdf over model/PCA parameters \( \mathbf{y} \).

2.4 Example Applications

2.4.1 3D Face Reconstruction

Illustrative experiments have been conducted, in order to show how a low number of measurements (observations) can be integrated with a high dimensional model of human faces, using our framework. The final application in mind would be the reconstruction of human faces given a number of observations, e.g. 3D locations of interest points on the skull in the case of post-mortem face reconstruction from skull data [Claes et al., 2006]. We employed a morphable model of human faces that is based on 3D scans of 39 human faces. Each scan consists of approx. 60000 3D points. Due to this high dimensionality, combined with a relatively low number of training data, we utilize the subspace-technique described in section 2.3.6. Computations were performed in a 38-dimensional PCA-subspace. We manually specified values of certain key face points (e.g. tip of the nose, elevation of upper and lower lip, chin etc.), simulating measurements that could come e.g. from a human skull. By sampling from the inferred probability distribution, a number of faces is then obtained, reflecting the remaining uncertainty after integrating the statistical prior model and the specified constraints. Some example results are shown in Fig. 2.3. It can be seen that the reconstruction becomes noisy when specifying constraints that are unlikely according to the learned prior model. This artefact is mainly due to the low number of training data and to misregistrations between the individual training faces.
Figure 2.3: Face reconstructions given measurements for a number of face locations: a) mean face. b-c) modified chin position. d-e) modified lip position. f) noisy reconstruction (an unlikely combination of values for nose, chin, upper and lower lip was specified).
2.4. Example Applications

Figure 2.4: Scale is correctly estimated from the foot trajectory alone. The ellipses visualise the uncertainty of the individual body locations.

2.4.2 Body Pose Estimation

Another application is the problem of 2D motion analysis from images. A model of human locomotion is first learned, then, given a subset of observed body locations the most likely body movement and the remaining uncertainty are computed using the described framework. This example adumbrates a way to deal with occlusions and self-occlusions as they often occur in body motion analysis. A more detailed description can be found in [Jaeggli et al., 2005].

An instance of running or walking is described as a sequence of body configurations at consecutive points in time. The body configurations are represented as absolute positions of certain body locations (cf. section 3.4.1). The horizontal and vertical components of the image coordinates are denoted $u$ and $v$, an observation vector thus takes the form

$$x = [u_{11}v_{11}u_{12}v_{12} \ldots u_{ln}v_{ln}]^T,$$

where $n$ is the number of body part locations that are used to describe a single body configuration and $l$ is the number of discrete timestep that are used to represent an entire walking or running cycle. The dimension of the entire observation space is thus $D = 2ln$. Given a set of training vectors that were obtained by manual annotation, the model pdf $p(x)$ is approximated with a multivariate normal distribution. The training examples are spatially and temporally aligned and of equal scale. They are also normalised for translation
2. Inference from Partial Measurements

![Figure 2.5: Prediction of occluded features. The partial measurement consists of trajectories of foot and shoulder, and of the shoulder only, when the subject is behind the fence. The known features are marked with a dot on top of the white cross.](image)

speed, that is, a constant displacement in $u$-direction is subtracted from the trajectories, such that they form closed curves.

**Subspace Representation.** The uncertainty that is introduced by missing body part observations can be characterised by an affine subspace $W$. Because the training data is spatially aligned and normalised for scale and speed, a number of alignment parameters has to be estimated in addition to the missing elements of a new observation vector. According to (2.8), the basis of the subspace $W$ is thus given by the canonical basis vectors that correspond to the missing entries in the vector $x_{\sim,j}$, as well as additional basis vectors that account for the alignment transformation. The spatial alignment consists of a $u$ and $v$-offset and can be described by the two vectors $t_u = [1010\ldots]^T$ and $t_v = [0101\ldots]^T$. Unknown speed is considered by a vector $p$, where all the $u$-entries contain a value that is proportional to the phase within the walking cycle, and the $v$-entries are zero. Unknown scale is accounted for by adding $x_0$ (the vector containing the actually observed values) to the basis of the subspace.
Experiments. For a set of video sequences, partial observations were manually obtained. In a real tracking framework, these measurements could stem from body part detectors. Inference was done using the basic method from section 2.3.2. Alternatively, the subspace method (section 2.3.6) could be applied to address the high dimensionality of $x$.

Fig. 2.4 shows one frame of a walking cycle that was reconstructed given a partial measurement consisting only of the trajectory of the foot. The predictions of the observed features as well as the alignment parameters are good. In particular, scale was accurately estimated. Generally, scale can be well estimated if multiple body locations are observed in at least one frame, whereas a shoulder or hip trajectory is not strongly correlated with the size of the figure in the images. A more realistic sequence is shown in Fig. 2.5, where the larger part of the subject is occluded by a fence. The pose of all limbs was predicted on the basis of the observed shoulder positions and of the foot before it disappears behind the fence. The posterior distribution can be used to infer the observed activity, where the reconstruction uncertainty is taken into account. We modelled two activity classes, walking and running. For this simple data set, all the reconstructions and example sequences were correctly classified as running or walking respectively.
3

Statistical Approach to Tracking and Pose Estimation

3.1 Problem Statement

A prototypical vision problem. The estimation of human body poses from video images has received a lot of interest from the computer vision community in recent years. This is due to several reasons. For one, such algorithms have many potential applications that justify the research effort. In most areas where motion data is needed, e.g. for animation purposes, for bio-medical analysis of sport performers or pathological body motion, expensive and cumbersome motion capture systems are used today. The availability of markerless tracking and 3D pose estimation systems that work with standard consumer video cameras would make modern analysis methods more widely applicable. One step further, being able to understand how people move their bodies in video sequences is an integral component of more general scene understanding, and could eventually lead to automated systems that help to avoid accidents and dangerous situations, and thus increase the safety in traffic situations or public places, as well as in hospitals or homes for elderly people.

Furthermore, and maybe most interestingly, the pose estimation problem combines many aspects of machine vision, and is thus an excellent research topic with the potential to bring the entire field of computer vision forward. As a matter of fact, it is maybe the most typical example of a very broad class of vision tasks, namely the possibly ambiguous inference of high-dimensional continuous quantities from images. It is also a very representative subtask of vision in the sense that it shares many of the difficulties that occur in image analysis, such as noise and occlusions, as well as the fact that it is very efficiently and seemingly easily solved by biological systems like the visual human perception. The latter can also be seen as an indication that an approach that is based on learning, experience and training is eventually more likely to lead to success than purely geometric reasoning. We believe that the frameworks
discussed here are general enough to be applied to many other sub-problems of computer vision with minor changes.

**What do we want to infer.** Many factors affect the appearance of a moving human person in video sequences. Most obviously, body pose and global body orientation have a drastic influence on how a person looks on photos or videos. Other causes of variance in appearance include clothing, limb dimensions and general physical constitution of the subject. There are many applications that require automatic understanding of such footage using computer vision methods. Depending on the application, some of the aforementioned causes of variation may be of interest, other may not need to be inferred from image data.

In our context, understanding how people move their bodies in video images means that the motion can be expressed using an abstract representation that concentrates on what seems important, and ignores less important aspects. These abstract representations of human body pose typically take the form of stick figures and are thus based on a model of the human body consisting of a number of rigid limbs connected by joints. Such representations concentrate on the angles between neighbouring body joints and subject-specific limb sizes while ignoring other characteristics of the subject appearance such as the colour and texture of the clothes.

Tracking of people through video sequences can occur at different levels of detail. On the crudest level, the pedestrian is just followed as a blob, by estimating the centre and size of a *bounding box* that encloses the subject’s image appearance. This process is referred to as *pedestrian detection* when applied to single frames, with a direct extension to tracking when applied to sequences. One step further, one can estimate the walking direction of the pedestrian, which involves some basic understanding of the scene geometry, for instance in terms of a ground plane and the relative viewpoint of the observing camera. An even more detailed analysis allows for the estimation of the body configuration of the tracked subject in terms of the relative pose of its body’s limbs. Here, the local body pose and the global position of the subject in the world can be jointly expressed in 3D world coordinates, and the relation to the 2D image coordinates is obtained using an image projection model and camera calibration. There are however reasons to decompose the overall task into 2D image location tracking on one hand, and the estimation of 3D articulations on the other hand, a choice that was also made in the proposed approaches. This decoupling allows to investigate the body poses independently from the location in 3D space where they occur, which is particularly important for the statistical modelling and learning approaches presented here.
Why adopt a learning based approach. Previous tracking algorithms often work with hand-crafted geometric body models that are rendered and compared to input images in order to verify body pose hypotheses (e.g. [Deutscher et al., 2000; Sigal et al., 2004; Sidenbladh et al., 2000; Sminchisescu and Triggs, 2003b; Urtasun and Fua, 2004a]). These body models have many parameters such as limb lengths and widths that all have to be known or estimated, typically in an initialisation procedure or even on-the-fly. Furthermore the appearance models are often defined in an ad hoc manner.

As opposed to geometric approaches, a learned model, as proposed in this thesis, allows to predict the appearance while reflecting the variance that is present in the training set. In other words, a distribution over appearances can be computed for a given pose hypothesis, rather than just one precise appearance prediction. Probabilistic Machine Learning methods thus naturally offer the possibility to learn the dependencies of body pose and its appearance while generalising over irrelevant variation of appearance and inter-person variance.

Since the models can be learned from real image data, they potentially outperform hand-crafted geometric models in terms of generalisation and accuracy. Rather than hard-coding ad-hoc prior knowledge about human bodies and their movements, learning provides means to introduce prior knowledge in a principled way; either in form of the training data, or by imposing a certain structure on the learned statistical model, which reflects independence assumptions about the involved random variables.

While most authors have described their poses with angles between rigid body limbs, the approach presented here offers greater flexibility with respect to the pose descriptors, ranging from low-level parametrisations (joint angles etc.) to representations that are more related to high-level gesture understanding. To illustrate this, we use different pose representations, either based on joint locations or on a temporal description of periodic locomotion, i.e. the phase within a walking cycle.

Monocular vs. Multi-Camera. It has to be distinguished between multi-camera tracking scenarios and the monocular case. While in principle the same algorithms can be applied, a few extra difficulties have to be taken into account in monocular tracking. Multi-camera setups often imply a somewhat controlled environment with e.g. known backgrounds, cameras that are carefully placed around the observed subjects etc. This leads to a situation where a large amount of high-quality image information is available and can serve as a solid basis of the pose estimation process. In particular, due to the baseline between the different cameras, an estimate of the three-dimensional geometry of the images and thus of the subject can be derived. In contrast, when
the scene is observed by a single camera, depth information can not easily be computed. Worse, some parts of the subject can not be observed at all, because they are occluded by the body itself. Even the pose of visible body parts can often not be determined non-ambiguously by pure geometrical reasoning. Learned models of prior knowledge can alleviate this problem to a certain extent by compensating for the missing information, and help to resolve ambiguous situations. There are however ambiguities that are intrinsic to the body pose estimation problem and have to be taken into account for both the statistical modelling and the inference algorithms. This means that the relationship between body pose and the image appearance of the same body, both viewed as random variables, can not be expressed as a functional mapping from appearance to pose. The relationship corresponds to a one-to-many or many-to-many mapping, since different body poses may, at least temporarily, be very similar in image appearance. This excludes the very attractive option of directly modelling the discriminative inference of body poses from monocular images as a simple functional mapping, a method that has been successfully applied for multi-view tracking scenarios. From the point of view of the tracking algorithm, the possibility of having multiple valid pose hypotheses, \textit{i.e.} unresolvably ambiguous situations, has to be considered. This suggests the use of probabilistic inference algorithms that can deal with non-Gaussian and multi-modal posterior distributions. They have the advantage of providing means to express and deal with uncertainties that occur during tracking, due to the above-mentioned problems, rather than just returning a solution that may be unreliable or arbitrary.

**Machine Learning Pipeline.** We proceed in a manner that is common for machine learning applications, starting by computing a number of features from the images, followed by a statistical modelling and analysis stage. The feature computation is strongly dependent on the specific task at hand and has the goal to extract the relevant information from images while being invariant to clutter and noise. The second stage often consists of a fairly general machine learning machinery that can be applied to many tasks, provided it is fed with suitable data, \textit{i.e.} image features.

The main focus of this thesis lies on the statistical modelling stage and the question, how machine learning techniques and tools can be combined in order to build models that reflect the specific characteristics of the body tracking task, and how to apply them to new image sequences.

**Image information.** A large amount of the information about the body posture is encoded in the overall shape of the subject, \textit{i.e.} its silhouette. The
use of colour is not very promising, since, with the exception of the hands and some facial areas, the subject as well as the image background can be of virtually any colour. When focusing on image edges, it can be seen that the most interesting and consistent edges occur at the border between figure foreground and background, while edges within the silhouette are often related to texture and clothing and thus less reliable for appearance modelling that generalises to other persons. Therefore, image descriptors that are computed from the human silhouette are an appealing choice for encoding image appearance. However, they leave certain aspects unobservable, and thus add to the aforementioned problem of ambiguous pose estimation from images. This phenomenon is known as Necker Reversal and refers to the (two-fold) ambiguous interpretation of silhouettes or wireframes that are orthographic projections of three-dimensional objects (see [Hartley and Zisserman, 2003]).

**High dimensionality.** A further difficulty when aiming at the vision based estimation of body postures is the high dimensionality of the involved representations. Typical parametrisations for articulated bodies have between approximately 20 and 80 degrees of freedom. The search for one or multiple optima in a continuous space of this dimensionality is challenging for any inference algorithm. When considering a statistical learning based approach, as we do here, the situation gets even worse, since the high dimensionality also makes the estimation of reliable statistical models more difficult, and drastically increases the amount of required training data. Furthermore, next to the pose representation, the images are also considered as (high-dimensional) random variables that are included in the statistical models. However, statistical analysis also proposes solutions to these problems. Hence, manifold learning, dimensionality reduction algorithms and feature extraction steps are integral components of the tracking framework that we present here.

### 3.2 Related Work

A vast amount of research has dealt with the task of articulated tracking and pose estimation in recent years (see [Forsyth et al., 2006; Moeslund et al., 2006] for a comprehensive survey). In this section, part of that work will be reviewed, starting with a general overview over different approaches for different settings and scenarios, and then focusing on the application of statistical learning techniques to this task. Then, extensions to multiple person tracking and prior work related to action recognition and activity classification - relevant for chapters 6 and 7 - are discussed. Finally, an overview of widely used machine learning techniques is given.
3.2.1 Articulated Tracking

One of the main difficulties of the task comes from the fact that images do not provide all the necessary information for 3D articulated tracking due to their 2D nature. Some authors thus use a body pose representation that is defined in the 2-dimensional image plane [Cham and Rehg, 1999; Felzenszwalb and Huttenlocher, 2005; Hua et al., 2005; Ramanan and Forsyth, 2003; Ren et al., 2005b]. While avoiding the problems related to depth estimation, these approaches have to deal with limb foreshortening, have difficulties to explain missing limbs and self-occlusions by occlusion reasoning, and do often not generalise to arbitrary viewpoints. Summarising, modelling 3D effects in 2D basically just relocates the difficulties, which might suggest to use 3D body pose representations.

For the estimation of 3D body poses, many approaches are based on multiple image streams that come from a set of synchronised cameras that all observe the same scene [Ren et al., 2005a; Kehl et al., 2005; Carranza et al., 2003; Sigal et al., 2004]. Multiple cameras and a controlled environment lead to high-quality images with implicit depth information, hence ambiguities (and thus multimodalities) can be limited, and accurate tracking results can be obtained.

This thesis mainly focuses on realistic scenarios with noise and occlusions, where the scene is observed by a single camera. In contrast to the multi-camera scenario, this monocular setting leads to unfavourable observation likelihood functions that may exhibit many local minima. This in turn puts high demands on the inference algorithm that has to be able to maintain multiple interpretations of the images. Multiple-hypotheses approaches and Monte-Carlo sampling techniques have been proposed to tackle this difficulty [Cham and Rehg, 1999; Sidenbladh et al., 2000; Deutscher et al., 2000; Sminchisescu and Triggs, 2001; Sigal et al., 2004].

Many existing articulated tracking inference algorithms can either be described as model-based generative top-down methods or part-based bottom-up approaches. These two categories will be discussed in turn.

**Part-Based Inference.** By representing body postures redundantly as a collection of rigid body parts and pairwise constraints between them, the tree-like kinematic structure of the human body can be exploited to obtain efficient inference algorithms. This approach relies on the factorisation of the posterior into pairwise functions of compatibility between neighbouring body parts, and limb-wise observation likelihoods. Hence it avoids a high dimensional representation of the joint configuration of all body parts. Inference in such
3.2. Related Work

Graphical tree models can be performed using algorithms like dynamic programming, Viterbi [Forney, 1973], belief propagation or variational (e.g. mean field) methods. Except for a few simple cases (such as Gaussian compatibility and likelihood functions) these algorithms operate in discrete state spaces. Using different variants of algorithms from the belief-propagation (BP) family, either per-limb marginals are computed (sum-product), or the most likely overall configuration of all nodes is found (max-product/min-sum, Viterbi). See [Kschischang et al., 2001; Yedidia et al., 2002] for an overview and comparison of BP algorithms. In [Felzenszwalb and Huttenlocher, 2005], the discretisation of the state space is achieved by exhaustively enumerating the possible locations and orientations of body parts represented as 2D rectangles in the images. Straightforward dynamic-programming-like inference has a complexity of $O(m^2n)$, for $n$ body parts with $m$ possible states each. By imposing a particular form on the compatibility functions between neighbouring body limbs, these can be implemented using generalised distance transforms, and their quadratic complexity can be removed, leading to a complexity of almost $O(nm)$.

Other part-based approaches often have a strong bottom-up component, where a first stage identifies body part candidate locations, that are then assembled using kinematic constraints of the human body models. In [Ramanan and Forsyth, 2003; Ramanan et al., 2005], the subject-specific appearance of body parts is learned by clustering part candidates or by matching a template in a specific ‘stylized’ body pose that often occurs. Given the discriminative colour model, articulated 2D bodies are then assembled using configuration and velocity constraints. In [Hua et al., 2005] candidate part locations are generated by sampling from bottom-up importance functions that are based on face detection, skin colour, edges and some heuristics guided by implicit prior knowledge. From these part hypotheses, likely part configurations are then assembled using belief-propagation with pairwise constraints between body parts, leading to a Monte-Carlo Belief Propagation approach.

The combination of Monte-Carlo and belief propagation methods can be driven further, based on the non-parametric belief propagation frameworks from [Isard, 2003; Sudderth et al., 2003]. These can be seen as a generalisation of particle filtering from Markov chains to arbitrary graphical models, as well as a generalisation of BP to continuous state spaces using Monte-Carlo sampling. The multi-camera tracking approach presented in [Sigal et al., 2003; Sigal et al., 2004] relies on such an inference scheme. Messages between graph nodes are represented as sample sets, where a certain percentage of the samples is generated by bottom-up importance sampling. One of the problems of local part-wise likelihood computation is that self-occlusions are not taken into account, hence a set of pixels or image features can be explained multiple
times by different body parts, which violates the geometric principles of image formation. These issues have been addressed by adding distributed occlusion reasoning in [Sudderth et al., 2004; Sigal and Black, 2006].

In part-based assembling strategies, the body pose priors are limited to pairwise constraints between neighbouring body parts in a graphical model representation of the human body, since the factorisation of the posterior is key to reduce the combinatorial complexity (see also [Ren et al., 2005b]). This implies that strong pose priors that model more complex relations between body parts can not be imposed; for this reason, the mentioned approaches either estimate 2D poses from monocular input, or base their 3D estimates on range data or multi-camera input. For monocular 3D articulated tracking, global pose representations allow for capturing the necessary prior knowledge about feasible body poses, often based on learning methods.

**Search in the entire pose space.** When using a global representation of the entire body configuration, the high dimensionality of the search space is challenging for the inference algorithm. Many approaches use geometric models of the human body that allow for an appearance prediction given a pose hypothesis. These models range from simple cylinder models over e.g. [superquadric] ellipsoids ([Sminchisescu and Triggs, 2003a; Urtasun and Fua, 2004a]) to realistic 3D computer graphics human models ([Carranza et al., 2003]). Using such generative appearance models, body poses that best predict observed image features then have to be found in a deterministic or stochastic manner. In controlled multi-camera settings, the resulting cost functions are often smooth and can efficiently be minimised by numerical optimisation [Carranza et al., 2003]. [Kehl et al., 2005] use ‘stochastic meta-descent’ randomised optimisation. Joint limit violations and body part self-intersection can be avoided by defining a simple (ad-hoc) prior.

In the monocular case, as mentioned before, the observation likelihoods typically have multiple local optima - a multimodal inference algorithm is necessary. The particle filter (or sequential Monte-Carlo method [Doucet et al., 2000a], CONDENSATION [Isard and Blake, 1998a]) often serves as a basic framework. However, as shown in [Sidenbladh et al., 2000], a straightforward application to high dimensional articulated tracking using simple dynamic models is not optimal. A huge amount of samples would be necessary to find several possibly peaky likelihood modes. Several extensions of the particle filter have been proposed, with better sampling strategies. While many of these algorithms maintain a proper probabilistic interpretation, they often may be interpreted as multi-hypothesis randomised search that heuristically explores the observation likelihood. [Deutscher et al., 2000] propose a multi-stage filtering approach called ‘annealed’ particle filtering, where the samples
are guided towards peaky likelihood modes by a cascade of weighting with more or less smoothed versions of the likelihood function, resampling and random perturbation. Using this algorithm, they fit a 3D body model to observations from multiple cameras. ‘Partitioned sampling’ is another method to lower the amount of required samples for high dimensional search spaces [MacCormick and Isard, 2000]. Hereby, the state variable is partitioned into two or more parts; sampling is then done sequentially in these sub-spaces, with resampling steps in between, that avoid wasting samples in areas of low likelihood. This requires that marginal versions of the weighting function are available, that only involve a subset of dimensions. In the ‘hybrid Monte-Carlo’ approach [Choo and Fleet, 2001], particle filtering and Markov-chain Monte-Carlo (MCMC) is combined, and applied to 3D people tracking from point-light displays with a dramatic speed-up when compared to plain particle filtering. Here, every sample of the particle filter serves as the starting point of a MCMC chain, each of which independently produces a certain amount of samples from the posterior.

Several authors have proposed to combine particle filtering with local optimisation; this can be seen as an extension of particle filtering to more efficiently identify likelihood peaks, or, as an extension of optimisation-based approaches to multiple hypotheses to avoid getting trapped in local optima. This approach originates from [Cham and Rehg, 1999], and has been formulated in an importance sampling framework in [Bray et al., 2007], for 3D hand tracking from range data. Additionally, in [Sminchisescu and Triggs, 2001], the local shape of the cost function is analysed in order to place the samples more efficiently in areas of large uncertainty. In combination with heavy tailed sampling and explicit handling of ambiguities related to kinematic flipping of limbs [Sminchisescu and Triggs, 2003b], good starting points for local optimisation are obtained, that eventually lead to the modes of the posterior and result in robust tracking through challenging sequences.

3.2.2 Learning Based Tracking

Many recent approaches to articulated tracking are based on learning techniques, that is, they make use of example data in order to improve or enable the inference of human body pose representations. While the term ‘learning’ is sometimes used in a broader sense in the literature, the focus here is on methods that typically consist of a learning stage, where example data (training data) is analysed and brought into a suitable form, and an inference stage, where the learned knowledge is applied to the tracking or pose estimation task.
**Body Pose and Motion Models.** Models of body pose and motion are one of the main areas where learning has been successfully used. This is appealing to compensate for information that cannot be extracted from the images themselves, for instance in the case of occlusions, or when estimating 3D body poses from monocular videos. Or, more generally, such models can constrain the large high-dimensional spaces of body pose parametrisations. In [Leventon and Freeman, 1998] a model of short fragments of 3D body motion is learned using a Gaussian distribution, with an extension to Mixtures of Gaussians in [Howe et al., 1999]. This model is then used to ‘lift’ a 2D body motion into 3D, by viewing the reconstruction as a probabilistic inference problem.

Activity-specific motion and pose models are learned in [Sidenbladh et al., 2000; Urtasun and Fua, 2004a; Jaeggli et al., 2005] for periodic activities. By the means of Principal Component Analysis (PCA) of a set of training vectors that each represent a walking or running cycle, a morphable model is obtained, that can blend between different styles and types of locomotion. This model is then used in the framework of probabilistic sample-based tracking in [Sidenbladh et al., 2000] or deterministically fitted to observed 2.5D reconstructions in [Urtasun and Fua, 2004a].

In the part-based approach of [Sigal et al., 2004] the relative position of neighbouring body parts is learned with Gaussian Mixture Models, to obtain prior constraints on the body pose. However, as mentioned above, in part-based approaches the restriction to pairwise relations between neighbouring body parts precludes the learning of powerful priors of overall body poses.

In [Demirdjian et al., 2003] a nonlinear model of valid body poses is learned by training a Support Vector Machine (SVM) classifier on a collection of motion capture data. During tracking, feasible body poses are enforced by solving a constrained optimisation problem.

Numerous authors have presented methods to characterise human body configurations and their dynamics by determining low dimensional embeddings and manifolds. [Brand, 1999] models human motion with a Hidden Markov Model (HMM) whose states correspond to neighbourhoods where the relationship of position and velocity is roughly linear. This yields a configural and dynamical manifold that summarises the target systems behaviour. Given a state sequence, inferred using Viterbi from image information, a motion trajectory is found that goes through the regions of high probability in pose (target) space, and agrees with the learned dynamical model.

The dimensionality of the body pose representation is reduced in [Li et al., 2006] with a mixture of factor analysers (MFA). Using the model of [Roweis et al., 2001], a globally consistent parametrisation is obtained, with a bi-directional mapping between the original space and the low-dimensional one.
The results demonstrate robust multi-hypothesis tracking. In [Li et al., 2007], a piecewise linear model simultaneously learns a low-dimensional embedding and the dynamics. In [Pavlovic et al., 2001] a set of linear dynamical models are learned and used in combination with an inference algorithm that switches between them. The parameters of this Switching Linear Dynamical System (SLDS) are learned using Expectation Maximisation (EM). A piecewise linear autoregressive dynamical model was learned in [Agarwal and Triggs, 2004c].

Using the Gaussian Process Latent Variable Model (GPLVM, [Lawrence, 2005]), low-dimensional embeddings of MoCap data are learned in [Grochow et al., 2004; Urtasun et al., 2005]. Well defined motions like walking and golfing motions can be expressed using only 2 latent dimensions, by learning from a small training set. Given measurements from 2D body part trackers, or 3D trajectories of a subset of body parts, naturally looking 3D body motions can be inferred. This approach was subsequently extended by adding the dynamics to the model [Wang et al., 2008], leading to a model that simultaneously learns a low-dimensional embedding, the mapping into the original pose space, and the dynamical mapping in the low-dimensional space that allows for temporal pose predictions. Again, this model is applied to tracking given 2D body part trajectories in [Urtasun et al., 2006]. Similarly, in [Sminchisescu and Jepson, 2004] a non-linearly embedded low-dimensional model is proposed, learned using Laplacian Eigenmaps [Belkin and Niyogi, 2003]. The model contains a layered generative prior and a dynamical model, and is designed to support efficient continuous search.

Furthermore, various approaches make use of motion capture database directly, in order to support the tracking task. In these example based methods, rather than learning parametric models from the data, the data themselves serve as the model. In [Sidenbladh et al., 2002], short fragments of motion capture data constitute an implicit model that allows for synthesising body motions from training data in a probabilistic manner. Similarly, for animation purposes, sophisticated tools exist that assemble naturally looking animated sequences from a database of real motion capture data, often in a semi-automatic way (see e.g. [Pullen and Bregler, 2002]).

**Learning Appearance Models.** So far we have revised learning methods for pose models and dynamics. The following approaches include the image appearance of human bodies in the learning procedure.

From a set of annotated ground-truth images, image statistics are learned in [Sidenbladh and Black, 2003] as discrete probability distributions of filter responses. Learned likelihood models are then derived by combining multiple image cues, and applied in a Bayesian tracking framework. [Ren et al., 2005a]
use AdaBoost feature selection to learn image descriptors that drive their motion-graph based multi-view tracking system.

Related to the field of people tracking, pedestrian detectors are typically learned from an annotated set of images that do or do not contain humans. To this end, [Viola et al., 2003; Leibe et al., 2004; Dalal and Triggs, 2005] train systems that extract relevant features from images and perform a categorisation based on that information. They thus use learned appearance models to solve their task, which, however, is easier than the one analysed here, involving only a binary label rather than a continuous high-dimensional representation of the result. In fact, the localisation of persons in the images is a sub-part of the considered task of articulated tracking.

In [Giebel and Gavrila, 2002] a learned deformable shape model (point distribution model [Cootes and Taylor, 1992]) is used to define an appearance model for blob-level tracking, while accounting for dynamic shape deformations of the tracked subject. Tracking is done in a state space that encodes 2D image location and a low-dimensional parametrisation of the deformable shape. Similarly, human faces are tracked in [Liu et al., 2006] by simultaneously inferring location and appearance parameters. The appearance manifold is modelled using the global coordination system from [Roweis et al., 2001], allowing for the appearance parameters to be approximately marginalised out analytically, which leads to a Rao-Blackwellised Particle Filter [Murphy and Russel, 2001].

Low-dimensional manifolds of subject specific appearance are also modelled in [Lim et al., 2005; Lim et al., 2006], using Locally Linear Embedding (LLE) and a robust system dynamic identification technique based on Caratheodory-Fejer (CF) interpolation. Experiments are shown to support the claim that the learned dynamics of appearance changes improve the reliability of particle filter based tracking.

[Elgammal and Lee, 2004a] learn appearance manifolds from encoded silhouettes of walking persons using LLE. Different viewing points are explicitly modelled, and in [Elgammal and Lee, 2004b] subject specific motion styles are taken into account and separately parametrised. The one-dimensional nature of walking manifolds is exploited by fitting cubic splines to the manifold and by performing search along this 1-dimensional parametrisation. Using Radial Basis Function (RBF) interpolation, the mapping from the embedding space to the original silhouette representation is learned. Additionally, the mapping to body pose configurations is learned. This results in a system that can infer body poses from input silhouettes entirely relying on a learned statistical model of the relation between silhouettes and 3D pose.
3.2. Related Work

In [Grauman et al., 2003a] the relation of a multi-view silhouette descriptor and 3D body pose is captured by a mixture Probabilistic PCA [Tipping and Bishop, 1999]. Similarly, in [Grauman et al., 2003b] a related quantity - the visual hull of a person - is inferred from input silhouettes. Similar in spirit, many subsequent approaches have adopted a discriminative strategy to directly infer poses from silhouette descriptors by regression. When modelling the bottom-up mapping from silhouettes to body poses, ambiguities and multimodal posteriors can be avoided by capturing the person with a multi-camera setup as e.g. in [Sun et al., 2006].

Another way to enforce unimodality in a regression based approach is proposed in [Agarwal and Triggs, 2004b], where dynamical information is included in the learned regression model. The knowledge of the previous states helps to select the correct mode. This allows for a single mode (which is determined at initialisation time) to be tracked consistently through a sequence. In practice, this algorithm performs better than [Agarwal and Triggs, 2004a], where a single regressor from silhouettes to poses often underestimates pose angles. These linear or non-linear regression models are based on damped least squares or Relevance Vector Machines (RVM). In [Williams et al., 2006] a sparse Gaussian Process (GP) regressor is learned in a semi-supervised manner to discriminatively infer gaze positions from images in real-time.

Many authors estimate human poses from silhouette descriptors, in particular shape context distributions [Belongie et al., 2002]. An exception is made in [Agarwal and Triggs, 2006a], where a regression based approach uses Histograms of oriented gradients (HOG) to encode the images of the upper body, similar to SIFT [Lowe, 1999]. Feature selection is done using nonnegative matrix factorisation (NMF), and experimental results are shown, where upper body poses are inferred from images with cluttered background using a single-valued regressor.

For monocular tracking, a number of approaches have addressed the problem of learning the non-functional one-to-many mapping from appearance to pose. In the ‘specialized maps’ architecture [Rosales and Sclaroff, 2001], a number of forward mapping functions are learned with artificial neural networks, each function specialised for a certain ‘branch’ of the multi-valued overall mapping. The model is learned in an EM framework to determine the mappings and their domains. An inverse function, i.e. a model-based prediction of appearances, is then used to select among hypotheses. The overall technique thus combines discriminative hypothesis generation and generative top-down verification. A similar approach is used in [Thayananthan et al., 2006], consisting of a mixture of sparse RVM regressors. Hausdorff matching scores are used as image features. [Curio and Giese, 2005] use a set of parallel bottom-up Sup-
port Vector Regressors (SVR) to initialise and support model-based top-down
tracking in a low-dimensional posture space.

[Agarwal and Triggs, 2005] and [Sminchisescu et al., 2005b] both learn the
dependencies between pose and appearance with a mixture of regressors (ex-

erpts). These frameworks support nonlinear mappings, sparsity, and define
gating functions to select among regressor outputs given an input appearance
descriptor (see also [Jacobs et al., 1991; Bishop and Svensén, 2003]). In [Smin-

chisescu et al., 2005b] the temporal dependencies and image-pose dependencies
are learned in a single model. The distributions are propagated analytically. In
[Agarwal and Triggs, 2005], a pdf over possible poses is inferred given an input
silhouette. The analytical inference procedure does not include any temporal
aspects. For tracking, a particle filter in high dimensions is used, where the
inferred pdf is treated as the observation likelihood. [Navaratnam et al., 2006]
propose a semi-supervised strategy to improve a mixture model of the joint
pdf of pose and appearance, by adding easily obtainable marginal training
examples of pose or appearance alone.

While the described discriminative approaches lead to straightforward infer-
ence algorithms and are self-initialising, they have to deal explicitly with am-
biguities of the one-to-many discriminative mapping. The number of compo-
nents in the mixture models is a main parameter of the learning procedures.
Furthermore, these conditional models can not support the localisation of the
subjects, since the observation itself is not modelled. Therefore, the subject’s
2D image location is assumed to be known beforehand, although the localisa-
tion is not a trivial task for challenging scenarios such as the ones considered
here.

The approach presented in chapter 4 of this thesis, and in [Jaeggli et al., 2006;
Van Gool et al., 2006], is related to these last-mentioned discriminative ap-
proaches, in that a mixture model is learned to predict poses from appearances.
However, rather than learning a purely conditional model, the joint distribu-
tion is approximated, providing a base for estimating the 2D image location
of the persons along with their body pose. A hybrid parametric/sample-based
representation and inference scheme is proposed, in a Rao-Blackwellised Par-
ticle Filter, to combine the non-parametric posteriors with efficient discrimi-
native pose estimation. The task is formulated as a filtering problem with a
temporal prior that enforces feasible body poses and temporal smoothness.

Generative tracking approaches with learned appearance models, on the other
hand, suffer from the high dimensionality of the body pose space, which poses
problems for both the learning and the generative tracking algorithms. Their
performance can thus be improved by a suitable dimensionality reduction.
[Lee and Elgammal, 2007] first learn such a low-dimensional pose represen-
3.2. Related Work

3.2.1 Pose and Appearance Estimation

Illustration and then model the mappings into the pose and appearance spaces, as well as the pose dynamics, using kernel regressors. [Navaratnam et al., 2007] propose an integrated formulation that obtains a dimensionality reduction in a Gaussian Process framework by estimating a low-dimensional latent space which simultaneously maps into the pose and appearance spaces, and in [Ek et al., 2008] additionally takes into account the dynamics. These models are based on the idea of a shared latent space, proposed in [Shon et al., 2006]. Roughly simultaneously, a conceptually very similar approach with nonlinear manifold learning, dynamical model and generative appearance prediction has been presented in [Jaeggli et al., 2007a; Jaeggli et al., 2007b; Jaeggli et al., 2008], which is also described in detail in chapter 5 of this thesis. It is based on RVM regression; a Gaussian Process version is presented in [Gammeter et al., 2008] and chapter 6.

[Sminchisescu et al., 2006] propose to combine conditional and generative modelling into a framework that supports consistency feedback (and thus 2D image localisation) while maintaining the efficiency of feed-forward recognition-based pose inference.

Rather than learning global appearance models for the entire figure, appearance modelling can also be applied on a local level. In part based approaches, dedicated body part detectors are learned, whose part hypotheses are subsequently assembled to form tree-like body structures. Head or limb detectors are learned e.g. using AdaBoost or SVM classifiers from training examples in 2D or 3D [Hua et al., 2005; Ronfard et al., 2002; Bhatia et al., 2004]. Considering even more local appearance patches, the voting-based pedestrian detector of [Leibe et al., 2004] can be extended to articulated tracking [Demirdjian and Urtasun, 2007]. [Andriluka et al., 2008] combine per-part appearance models based on local appearance patches and a dynamical articulation model learned with GPLVM. The system is applied to the inference of simple 2D articulated skeletons from videos with multiple persons viewed from the side.

Finally, impressive pose estimation have been obtained with example-based matching approaches ([Stenger et al., 2003; Shakhnarovich et al., 2003]), where the main challenge is the efficient organisation of the example data in trees or using hashing techniques, to enable the storage and search in the huge amounts of data.

3.2.3 Multi-Person Articulated Tracking

While most articulated tracking approaches consider only single persons, several methods have also been proposed for multi-person scenarios. In [Mitchelson and Hilton, 2003], multiple independent articulated trackers are initialised
on different persons. [Ramanan et al., 2005] automates the initialisation stage by detecting ‘stylized’ poses for 2D body pose estimation. Several approaches have demonstrated 3D body pose estimation in static camera surveillance scenarios [Lee and Nevatia, 2006; Zhao and Nevatia, 2004]. Most directly related to our approach of chapter 6 ([Gammeter et al., 2008]), [Zhao and Nevatia, 2004] applies a multi-object tracker to identify individual trajectories and estimate each tracked person’s body poses over time. Their tracking approach relies on background modelling, and the pose estimation process is relatively simple, based on a coarse discretisation of the pose space. The combined tracking and detection approach of [Andriluka et al., 2008] manages to track multiple persons in cluttered traffic scenes, and estimate simple 2D articulation models. In contrast, 3D articulated multi-body tracking from a moving, human-level perspective still remains an open issue. [Gammeter et al., 2008] are among the first to present results for this challenging scenario (see chapter 6).

3.2.4 Activity and Motion Recognition

Next to body articulations, many works have aimed at a high-level interpretation of visual motion in terms of activity classes or motion categories. A classical pipeline for the recognition of human motion first derives a medium-level representation (e.g. joint angles, joint trajectories) from images, that serves as the input for the actual classification [Fanti et al., 2005]. While this approach simplifies the classification task, it involves tracking or pose estimation, which are difficult tasks themselves. In [Urtasun and Fua, 2004a; Urtasun and Fua, 2004b] a parametrised model of human locomotion is learned from training data using PCA. When fitted to an observation sequence, its estimated parameters can also be used to perform identification or recognition of activities in a nearest neighbour manner.

Image-based methods directly infer class labels from appearance or low-level features, without any notion of body posture. In [Mori et al., 2003] optic-flow features are computed from a temporal window around the current frame and subsequently fed into an nearest-neighbour classifier. [Laptev and Lindeberg, 2003] represent video sequences as a sparse set of spatio-temporal interest points, which are found with a 3D version of the Harris corner detector. Classification is done at sequence level, by nearest-neighbour matching, or with a SVM [Schüldt et al., 2004].

Temporal aspects of motion sequences are often modelled using discrete-valued chain models of motion class labels. In a Hidden Markov Model (HMM) the observations are modelled generatively, and class labels are inferred using
Bayes’ rule \(\textit{e.g.} \) [Gong and Xing, 2003]). Alternatively, in a Conditional Random Field the observations are conditioned upon without being modelled [Sminchisescu \textit{et al.}, 2005a], allowing for taking into account temporal context in the observation.

Model switching mechanisms can perform recognition when the different models correspond to discrete states with a semantic meaning, \textit{i.e.} in a supervised setup. [Isard and Blake, 1998b] have proposed a probabilistic state switching mechanism, where different dynamical models are chosen, depending on a discrete state variable. In chapter 7 of this thesis, [Jaeggli \textit{et al.}, 2007b; Jaeggli \textit{et al.}, 2008], body poses and discrete activity labels are jointly inferred using similar model switching, where the different states (activities) involve separate models for pose, dynamics and appearance.

### 3.2.5 Learning Methods

Several machine learning techniques have been used for the realisation of statistical tracking algorithms. Here, the relevant literature is briefly reviewed, while more technical short introductions to those techniques are provided if needed in the respective technical chapters 4, 5 and 6, where the different tracking approaches are presented.

Mixture models and in particular mixtures of Gaussian distributions (GMM) are a widely used method for the approximation of multivariate probability density functions. Their parameters are typically estimated using the Expectation Maximisation (EM) algorithm, a well known iterative procedure to simultaneously estimate model parameters and a set of related hidden variables [Dempster \textit{et al.}, 1977]. Despite their simplicity, finding the modes of a GMM or even the number of modes is challenging, as argued in [Carreira-Perpiñán, 2000; Carreira-Perpiñán and Williams, 2003]. For modelling the multi-valued mapping between two sets of variables, mixture models are employed that consist of a set of regressors or experts [Jacobs \textit{et al.}, 1991; Bishop and Svensén, 2003].

Kernel methods [Shawe-Taylor and Cristianini, 2004] have become very popular for modelling nonlinear regression. Support Vector Regression (SVR) is the regression variant of the Support Vector Machine (SVM, [Cristianini and Shawe-Taylor, 2000]). During training, a subset of the training are selected as ‘support vectors’, this sparsity leads to very efficient computation of predictions for new input values. The Relevance Vector Machine (RVM, [Tipping, 2001]) uses a similar formulation for its prediction formula. It is based on Bayesian theory, has fewer parameters to set, and produces
even sparser results for large training sets. There is a very active field a research related to Gaussian Processes and in particular to their application for nonlinear regression [Rasmussen and Williams, 2006] and dimensionality reduction [Lawrence, 2005]. From a practical perspective, the attractiveness of these Bayesian methods has been raised recently by extensions that allow for sparsity and thus larger datasets [Snelson and Ghahramani, 2006; Lawrence, 2007].

For dimensionality reduction, linear Principal Component Analysis (PCA, [Hotelling, 1933]) is the best known statistical method. A version of PCA of binary data has been proposed in [Schein et al., 2003], based on Bernoulli distributions.

By the use of kernel methods, a nonlinear variant of PCA is obtained. In so-called kernel PCA (kPCA, [Schölkopf et al., 1998]), kernel functions are used to form the mapping from the original space to the low-dimensional embedded space.

In contrast, the Gaussian Process Latent Variable Model (GPLVM, [Lawrence, 2005]) models the mapping from the low-dimensional space back to the original space using nonlinear Gaussian Process regression, while optimising for the low-dimensional latent embedding of the data. In order to ensure that data points that are close in the original space remain close on the embedded manifold, an extension has been proposed that also enforces a functional relationship in the opposite direction, from the original to the embedded space, using kernel regression [Lawrence and Quiñonero-Candela, 2006]. Other extensions take into account temporal relationships between data points [Wang et al., 2008], or estimate a latent space that is shared between two related datasets [Shon et al., 2006].

Further nonlinear dimensionality reduction methods include the Generative Topographic Mapping (GTM, [Bishop et al., 1998]), that also learns the reconstruction mapping along with the embedding. Another class of algorithms is based on the local relationship of neighbouring data points, often encoded as a proximity matrix that takes into account the nearest neighbours of each data point. The data are then expressed in a space of lower dimensionality while approximately retaining the neighbourhood relationship. These methods include Locally Linear Embedding (LLE, [Roweis and Saul, 2000]), Isomap [Tenenbaum et al., 2000], Laplacian eigenmaps [Belkin and Niyogi, 2003] and semidefinite embedding [Linial et al., 1994], and do not learn an explicit transfer function between the spaces, which could be applied to new data points, unknown at manifold learning time.
3.3 Proposed Overall Framework for Tracking and Pose Estimation

The tracking approaches presented in chapters 4, 5 and 6 can be viewed as specific realisations of a general modelling framework. They all consist of the same basic building blocks, differ however in the way the relationship between the components is modelled. An overview of the modelling framework is shown in Fig. 3.1.

The statistical framework aims at capturing prior knowledge about the two random variables $X$ and $Y$ that encode human body pose respectively the image appearance thereof, as well as the statistical relationship between these random variables. The structure of the model is subject to a number of design choices. Within this thesis different variants have been implemented and tested, as reported in the following chapters. Given the model structure, the actual statistics are learned from a set of training data, consisting of pairs of body pose representations and their corresponding appearance in video images. The specific representations may vary due to the availability of training data, as well as theoretical and practical considerations. Here, it should be noted that both pose and appearance are treated as multivariate random variables, i.e. a concrete realisation of these variables simply takes the form of a
vector consisting of a number of real values. The training data is thus available as matrices $X$ and $Y$ containing many observed realisations of these random variables.

**A: Relationship of Pose and Appearance.** The model as illustrated in Fig. 3.1 provides the necessary information to infer body pose hypotheses from new previously unseen appearance descriptors. When the 2D position and size of the person in the video sequence is already known, the sequence of appearance descriptors can directly be computed from the images and the known trajectory of the bounding box locations. The subsequent pose inference step is then basically a discriminative task. Body poses can be inferred, no matter whether the model is built in a discriminative or generative way. Under some assumptions and approximations, the computations can even be performed in a parametric or semi-parametric way.

However, in the scenario considered in this work, the 2D bounding box of the moving person is not known beforehand, nor is it a trivial task to obtain this trajectory in a preprocessing step. To the contrary, both 3D pose estimation and 2D tracking on a bounding box level are difficult tasks, that are strongly interconnected and may benefit from each other. For example, a known body pose can help to find the precise position of the figure in the image, and consequently a learned dynamical model on the body pose level might help to track through difficult subsequences with noise or occlusions. Similarly, the pose estimation is also strongly dependent of a precise localisation of the bounding box. A statistical model that follows a purely discriminative strategy to capture the dependency of $X$ and $Y$ might not be optimal for this scenario, since it does not capture the information that is needed for localising the figure in the image.

Furthermore, while the relationship of pose and appearance descriptors may be modelled in a parametric way, things get more complicated when an unknown bounding box location is taken into account. Due to the discrete nature of digital images, image noise etc., the 2D bounding box estimation problem is unlikely to be solved well using parametric posterior distributions. This is one of the underlying reasons for the success of Monte Carlo methods as compared to Kalman Filters, which can be observed in the vast amount of published research in visual tracking.

For these reasons, this thesis investigates combinations of discriminative and generative methods, as well as combinations of analytical inference and sample based Monte Carlo inference. The general framework discussed in this section does not yet specify whether the relationship of body pose and appearance is modelled discriminatively or generatively, which is reflected by the
bi-directional arrow connecting the body pose and image appearance variables in Fig. 3.1.

**B: Pose and Motion Prior.** Prior knowledge about typical human poses and motion patterns can also be captured by the learning framework. Here, the temporal structure of the tracking problem is accounted for, by learning a temporal prior, *i.e.* a probability distribution over body poses given the history of image observations or previously estimated poses. During tracking, this temporal model helps to enforce smooth motions. In principle we could also model the dynamics in the appearance space, but in all our implementations, the temporal aspects are modelled on the level of body poses, which is where we actually want to do inference, *i.e.* the search space of the problem. The pose manifolds are smoother than appearance manifolds, and suffer less from noise and self-intersections. Also, poses and their dynamics can be analysed independently from the camera viewpoint, whereas an instance of an appearance descriptor is always dependent of the relative pose of the camera and the subject.

**C: Dimensionality Reduction.** The modelling and tracking framework contains dimensionality reduction steps for both pose and appearance representations. Many of the statistical methods applied here suffer from the high dimensionality of the random variables. The models might not be learned well, suffer from overfitting and poor generalisation. Also the high dimensionality can lead to numerical problems and poor computational efficiency both in terms of speed and memory usage. The dimensionality of the search space is in particular also a problem for the Monte-Carlo tracking approaches that we use. Linear as well as nonlinear dimensionality reduction methods will be considered that do or do not provide mappings between the high-dimensional and the low-dimensional spaces. In Fig. 3.1 these mappings are indicated with arrows denoted *dim. red.* and *reconstruction*, in practice we sometimes use approximations or learn the respective mappings in a separate step.

### 3.4 Representations

The structure of the proposed statistical model is quite general and can in principle be applied to other vision based estimation problems. The framework is adapted to the specific task at hand - pose estimation - by choosing suitable representations for the body pose and image appearance. The selected representations thus build the interface between the generic framework and the specific task it is applied to. In this section we introduce these representations.
3.4.1 Body Pose Representations

Previous work in visual 3D tracking has often reported the results using an abstraction of the human body, consisting of a number of rigid parts that are connected by rotational joints. Such representations are intuitive and to some extent supported by bio-mechanical findings and the fact that there are rigid bones in a human’s limbs. However, the limitations of such models, like to accurately represent the possible movements of a torso, are also obvious. They are nevertheless precise enough to reflect the level of detail that can be achieved by visual 3D tracking, and are also employed for state-of-the-art motion capture applications.

We distinguish between angular and positional representations, hierarchical and non-hierarchical ones, and whether single body poses (snapshots) or entire motion sequences are expressed. Furthermore, some representations are linked to specific activities while others are generic.

Angular Parametrisation. For many applications (e.g. animation) an angular parametrisation has been used. The body is represented as a kinematic tree, where one body part (typically the hip) is chosen as the root, and the pose of each remaining body part is expressed with respect to its direct parent in the body tree structure. The pose of a given limb in 3D world coordinates is then obtained by concatenating all transformations along the chain from the root to this limb. The process of computing the position of an end effector such as the hand from an angular representation is also referred to as forward kinematics.

There are 6 degrees of freedom (DOF) for the relative position of a body part in the coordinate system of its parent. The translational parameters (3 DOF that also determine the size of the limb) typically remain constant for a given person, and the rotational parameters vary over a sequence to allow for body motions.

The 3 DOF rotation can be expressed in different ways, including Euler angles, quaternions, or an angle-axis representation (see [Goldstein, 1980]). They each have their disadvantages, such as non-uniqueness of the representation, discontinuities when the angles wrap around from 360° to 0°, or gimbal locks. Furthermore, it is not obvious how to define a useful distance measure in these non-Euclidean angle spaces. The problem of periodic wrap-around can be avoided by replacing each angle by its \( \sin \) and \( \cos \), thus doubling the dimensionality of the representation [Agarwal and Triggs, 2006b].

Such representations allow for an easy introduction of simple prior knowledge about the kinematics of the articulated body. For instance, the knee and elbow
joint might be constrained to 1 rotational DOF, and some hard joint limit can be imposed on that rotation angle, to avoid that the knee is bent in the wrong direction.

Using the kinematic tree parametrisation, it is easy to make sure that the limbs stay connected to each other at their joints, and that the limb sizes remain constant throughout a motion sequence. Alternatively, the transformation of the local coordinate system of each body part can also directly be expressed with respect to the world coordinate system. This avoids some complex dependencies between the position and orientation of the different joints and body parts that might affect a statistical analysis or dimensionality reduction step.

The mentioned representation of the human skeleton as a kinematic tree allows for a separation of the subject-specific characteristics like limb dimensions and thus the overall height of the person on one hand, and the motion on the other hand. In articulated tracking, the limb dimensions are often estimated in an initialisation procedure and then kept static during the tracking (e.g. [Kehl et al., 2005; Carranza et al., 2003]).

**Positional Parametrisation.** In this thesis, we mainly use a body pose parametrisation that is based on the 3D joint locations of the human skeleton. It consists of a list of joints that determine the overall body pose and allow to draw a stick figure of the skeleton in a similar way as with the angular parametrisation. The spatial joint positions are expressed in a local coordinate system rooted e.g. at the hip centre or at its projection onto the ground plane. The local coordinate system is rotationally aligned with the subject’s direction of motion, that is, the X axis points forward, the Y axis points sideways, and the Z axis points upwards. See Fig. 3.4 for an example of a skeleton and the list of its joint centres.

Using this pose parametrisation, problems of angular representations can be avoided, and in contrast to a hierarchical approach each joint location is represented independently. There is however no intrinsic mechanism that enforces constant limb size or at least avoids the dissociation of the body parts. In our framework this is accounted for by learning a statistical model of body poses, which simultaneously enforces body poses that are mechanically possible (e.g. joint limits, connectedness of limbs) and that correspond to motions that are actually performed by humans. These constraints are expressed in a probabilistic way, without any distinction between bio-mechanical feasibility and the likeliness of a certain body pose to occur.

In addition to the local body pose as described above, the global position of the person in 3D world coordinates or in 2D image coordinates is encoded, as
well as the rotation of the subject around its own axis. By choosing a local representation that is invariant to global position and orientation, the amount of training data that is necessary for a statistical analysis is reduced. The choice is also motivated by practical considerations - motion captured training data are often acquired on a treadmill where the global position and orientation are fixed. Also, the probability distribution over global positions and orientations is often assumed to be uniform for good reasons, and independent of the local body pose. This should be taken into account by normalising the training data for global position and orientation, otherwise arbitrary pdfs over the global pose might be estimated.

Temporal Parametrisations. Many typical human motion patterns are essentially one-dimensional, i.e. it is possible to determine a prototypical motion pattern that captures the basic motion while ignoring the subject-specific characteristics of a concrete instance of that motion. For instance, a periodic motion like walking and running can be parametrised by a single periodic variable. Other activities like dancing or many sports consist of a sequence of well-defined motion-segments that are concatenated and that can be explained with a one-dimensional parametrisation each. Some of our experiments (cf. section 4.4.3, 5.4.1) are based on such a temporal parametrisation.

Motion segments. So far we have concentrated on representations that operate at a snapshot-level, where each body pose is regarded as one observation. This means that part of the variation in training data sets stems from the body pose variation during the execution of a specific motion, whereas another part stems from the differences how these motions are executed by different subjects. Alternatively, short motion sequences can be analysed, e.g. a sequence of poses corresponding to one walking cycle can be regarded as one observation. A model of such ‘segments’ can then be learned, for instance using Principal Component Analysis or a Gaussian Mixture Model. This has been done e.g. in [Jaeggli et al., 2005; Urtasun and Fua, 2004a; Sidenbladh et al., 2000] for activity-specific models, and requires a temporal alignment of the training data. Such an analysis allows for a separation of intra-subject variation within a walking cycle and inter-subject variation. See also section 2.4.2 for an example where such a parametrisation was used. Similarly, but without the restriction to a single well-defined activity and temporal alignment, in [Leventon and Freeman, 1998; Howe et al., 1999] a corpus of training data was cut into short motion segments, and a probabilistic model over such short segments was learned. This approach combines motion and pose information, where the variation within a motion segment is parametrised
by a time variable, and the variation between motion-segments is captured by the probabilistic model.

### 3.4.2 Shape Representations

In this section we examine how the image information that is relevant for tracking the pose of human bodies can be extracted and brought into a suitable form for the further statistical processing, *i.e.* a vector of numbers. The image descriptors should be invariant to the image location of the subject and only encode local shape. Image cues that have been used for visual tracking and detection of pedestrians include image colour, edges, or Haar-like features [Viola *et al.*, 2003]. It has been found that for pedestrians (in contrast to *e.g.* faces) most of the discriminative information is located at the border edges of the person. Therefore, as mentioned before, in this work the focus is on descriptors that encode the overall shape of the figure, that is, they are computed from the silhouettes of the persons. The representations either focus on the silhouette itself, *i.e.* the contour that encloses the entire person, or describe the two regions that are separated by the silhouette (segmentation), which is in principle equivalent and complementary, but leads to different representations. In the presence of noise, obtaining a single closed contour can be difficult, while region descriptors may be less sensitive to mis-segmentations.

**Silhouette descriptors.** In [Grauman *et al.*, 2003a], the descriptor is an ordered list that consists of a fixed number of 2D points that are equidistantly sampled on the silhouette, and expressed in coordinates relative to the bounding box. Similar polygon based representations have also been used for statistical shape modelling for other applications [Hug *et al.*, 2000]. They are limited to representing the outermost enclosing contour of the objects, ignoring the fact that the background is not necessarily a single connected component. For the example of a moving person, such a representation leads to discontinuities in the motion trajectories, for instance when the legs visually separate from one frame to the next, as shown in Fig. 3.2. Because of the increased total length of the second silhouette and the equidistant sampling, the point lists of the two silhouettes do not correspond.

These difficulties are avoided by adopting a local encoding of the silhouettes, based on shape context [Belongie *et al.*, 2002] histograms in [Agarwal and Triggs, 2004a; Agarwal and Triggs, 2004b; Agarwal and Triggs, 2005; Agarwal and Triggs, 2006b; Sminchisescu *et al.*, 2005b; Ek *et al.*, 2008]. Two levels of histogramming result in an approximately 100-dimensional powerful descriptor.
Region descriptors. Choices that encode the foreground region rather than the silhouette include lower-order moments [Curio and Giese, 2005; Freeman et al., 1996; Rosales and Sclaroff, 2001] such as Alt [Alt, 1962] or Hu [Hu, 1962] moments. As argued in [Curio and Giese, 2005], this representation is relatively unspecific, and thus better suited for a coarse initialisation of pose estimation than for precise inference of body articulations.

Most simply, the foreground segmented binary image contained in the bounding box is directly used as a descriptor, by forming a long vector containing all the pixels. Obviously this representation is very high dimensional (a few thousand pixels for very low resolution silhouettes) and highly redundant with strong correlations between the pixels. Using dimensionality reduction algorithms, this can be accounted for. In some of our experiments, Principal Component Analysis (PCA, [Hotelling, 1933]) was used to this end. For the considered data sets, about 80% of the variation could be captured using 10 principal components. While that may seem a low percentage, the experimental results suggest that these principal components indeed capture the relevant discriminative variation that permits to infer body poses.

Binary Principal Component Analysis. It can easily be seen that the binary foreground/background data do not satisfy the Gaussian assumption that underlies PCA. In [Schein et al., 2003; Zivkovic and Verbeek, 2006] binary data is modelled with Bernoulli distributions, leading to Binary Principal Component Analysis (BPCA). Despite being based on a linear model, like ordinary PCA, the projection on the lower-dimensional BPCA subspace defined by a number of components is an iterative procedure. A short-cut notation for this operation is introduced.

\[ y_{BPCA} = BPCA(z), \]  

(3.1)
3.4. Representations

Figure 3.3: Distance transformed silhouette of a human figure as shape descriptor. The representation has positive values inside the silhouette, negative values outside, and 0 on the silhouette, scaled here to the interval \([0, 255]\) for visualisation. The corresponding body pose stick figure is shown in Fig. 3.4.

where \(z\) is the original high dimensional descriptor, and \(y_{BPCA}\) is its projection on the lower dimensional BPCA subspace. We also consider the inverse operation that projects the descriptor \(y_{BPCA}\) back into the original high dimensional pixel space and transforms it into binary images or foreground probability maps. BPCA reconstruction is based on the sigmoid function \(\sigma(a) = (1 + e^{-a})^{-1}\) and the vector containing the so called log-odds parameters \(\theta\).

\[
\theta = V^T y_{BPCA} + \mu \\
P(z_k = 1|y_{BPCA}) = \sigma(\theta_k)
\]

(3.2)

Here, \(\mu, V\) are the mean and basis vectors of the Binary PCA, and \(\theta_k\) is the log-odds parameter corresponding to pixel \(k\). \(\sigma(\theta_k)\) is the probability that pixel \(z_k\) lies on the foreground (i.e. takes the value 1), as well as the expected value of that pixel, in the interval \([0, 1]\). See [Zivkovic and Verbeek, 2006] for the mathematical details and derivation.

Signed Distance Functions. Another class of descriptors is based on discrete representations of signed distance functions. While encoding the silhouette itself (more precisely the distance to the silhouette), this representation uses a regular sampling of points within the bounding box, thus avoiding aforementioned correspondence problems that occur when points are equidistantly sampled from the contour itself.

The signed distance function representation is obtained by computing distance transforms \(y_{DT}\) of the silhouettes or foreground segmented bounding box images, similarly to the chamfer image transform [Rosenfeld and Pfaltz, 1966;
To compute $y_{DT}$, signed distances are computed on a grid of equidistantly spaced sample points inside the bounding box of the segmented object. Each sample point is assigned a value that is proportional to the distance to the closest point on the silhouette, and whose sign indicates whether the sample point lies inside or outside the silhouette. For the experiments of section 4 and 5 a hybrid distance measure was used as an approximation to the real Euclidean distance, for efficiency reasons. See [Bailey, 2004] for an overview of algorithms. An example of a distance transformed silhouette is shown in Fig. 3.3. Since the distance value of neighbouring sample points are strongly correlated, the resulting grid of sample points can be represented more compactly by applying PCA dimensionality reduction, without significant loss of information. The following notation is introduced for the computation of distance transforms and projection on a PCA subspace:

$$y_{DT} = V(dt(z) - \mu)$$  \hspace{1cm} (3.3)

where $z$ and $y_{DT}$ are the original image and the distance transformed descriptor respectively, and $\mu$ and $V$ are the mean and basis vectors that determine the PCA subspace. Given the distance transformed representation, the original silhouette or segmented figure can in principle be obtained by looking at the sign of the stored values. Here, we choose a probabilistic interpretation, based on the intuition that the foreground/background probabilities are higher far away from the silhouette, and lower ($\approx 0.5$) very close to the silhouette. The reconstruction is thus based on the sigmoid function $\sigma(.)$.

$$y' = V^T y_{DT} + \mu$$

$$P(z_k = 1|y_{DT}) = \sigma(y'_k)$$  \hspace{1cm} (3.4)

where $y'_k$ is the sample point of the distance transformed image with index $k$ and $z_k$ is the corresponding pixel of the reconstructed image.

A similar representation was used in [Elgammal and Lee, 2004a] to encode human silhouettes. A priori it can be noted that signed distance functions can be sensitive to noise and mis-segmentations when computed bottom-up, since even few wrongly segmented pixels can affect an entire neighbourhood of computed signed distances and thus corrupt the descriptor. In this thesis these problems are dealt with by using these representations in a top-down manner to predict silhouettes, or by first cleaning the segmented images in a preprocessing step. See section 5.3.1 for a short discussion of the empirical experience with the different shape descriptors.
3.5 Training Data Sets

The statistical models for human body pose estimation were trained on data sets consisting of body pose data obtained from a motion capture system and of corresponding images of a human body in the same body pose.

3.5.1 LocoETH Dataset

The dataset concentrates on the typical visual motion patterns for two types of human locomotion, walking and running. Multiple subjects were recorded under laboratory conditions performing those activities at different speeds. The resulting three-dimensional motion data was then further processed, and transferred into representations that are suitable for the planned experiments. This section recapitulate the recording setup and processing steps.

Motion Capture Setup. The sequences were recorded at the Motion Capture (MoCap) laboratory at ETH Zurich. This lab is equipped with an optical MoCap system\(^1\) with 6 cameras that operate in the near-infrared range. In order to reconstruct the 3D body motions, 41 infrared-reflective markers are attached to the skin of the test subject, according to a specific protocol. The trajectories of these markers are then tracked in the individual camera streams and integrated into a 3-dimensional representation. Finally, an abstract body model (bi-ped, 17 rigid limbs) is used to interpret that data and solve for body poses. The system operates at 120Hz, and its spatial accuracy is better than 1cm. Such MoCap systems are often used in the film industry, but also for medical purposes and biomedical analysis or therapy, and thus meet the demands of highly delicate applications. The working volume of the setup is limited (approximately 2 by 2 meters), therefore a treadmill was used to allow for locomotion patterns of a certain duration (\(i.e.\) multiple running or walking cycles) without leaving the working area.

Set of Motions. Six subjects, male, between 20 and 40 years of age, of average physical constitution, and in good health were asked to perform a set of activities on the treadmill. They were allowed to acclimate to moving on the treadmill, which may be a bit cumbersome at first and lead to biased or unnatural motions. The subjects were asked to walk and run at three different speeds each, for about 10 seconds. The speeds ranged from slow walking (2.5 km/h) over average speed to fast walking (4.2 and 6 km/h). Running

\(^1\)www.vicon.com
was performed at 8, 10 and 12 km/h. The result of this stage are motion capture sequences (36, 6 subjects at 6 speeds), represented either as marker trajectories, or as a kinematic tree with 6DOF transformations indicating the relative pose of each limb with respect to its parent limb, or with respect to a global coordinate system. This data can then be converted to any of the parametrisations discussed in section 3.4.1 or other.

**Normalisation of the data.** Since the data was recorded on a treadmill, the orientation was fixed and the global translation was approximately constant. In a normalisation step the subject was fixed at its hip, its global motion was thus completely removed, including the periodic vertical motion that is characteristic for human locomotion. Furthermore, the motion data was normalised for limb lengths, i.e. all the motion sequences are based on a single skeleton template. This allows us to concentrate on the body motion and body configuration rather than on anatomical characteristics of the individual test subjects.
3.5. Training Data Sets

Figure 3.5: Fraction of the data variance that is captured by a certain number of principal components. The 99% level is reached at 14 principal components.

After converting into a location-based pose parametrisation, we end up with a 60-dimensional representation that consists of 3D locations such as the head, shoulders, elbows, wrists, and hips. Fig. 3.4 shows an example of a skeleton from this data set.

**Basic Statistical Analysis.** To gain a better understanding of the data at hand, a simple statistical analysis in terms of principal modes of variation and intrinsic dimensionality was carried out. We investigated the data using linear tools. The principal components of the 60-dimensional pose examples were extracted. As expected, this analysis showed that our 60-dimensional representation is highly redundant. Indeed, even using a simple linear method such as PCA, 99% of the variation of the data can be explained using only 14 principal components (see Fig. 3.5), which corresponds to a reduction of the original dimensionality by more than 75%.

Another interesting question is whether the motion patterns of the different subjects occupy distinct spaces in the PCA-reduced pose space, e.g. if there is
Figure 3.6: a) Walking and running sequences of a single subject at all speeds. We plotted the first principal component against the second. The second component separates walking and running motions. b) The figure shows a plot of the first against the third principal component, for a different subject. In both (a) and (b) the three speeds for each motion pattern can easily be distinguished.
3.5. Training Data Sets

Figure 3.7: The first principal component captures the main forward/backward motion of the legs. The stick figures were produced by varying the first principal component, while fixing the remaining components at their average value.

Figure 3.7: The first principal component captures the main forward/backward motion of the legs. The stick figures were produced by varying the first principal component, while fixing the remaining components at their average value.

a linear separation that allows for distinguishing between subject or activities given new observations of body poses. In the following, we give a brief overview over the first principal components (PC), aiming at interpreting them in a manner that is meaningful and intuitive for human observers.

- PC 1: This PC encodes the main forward and backward motion of the legs during a walking or running cycle. Its amplitude corresponds to the step length and thus correlates with the walking and running speeds. See also Fig. 3.6 and Fig. 3.7.

- PC 2: Variation of this PC results in stick figures that are more or less bent forward and whose knees are bent more or less. This component allows for discrimination between running and walking. Especially for the running examples, it is also strongly dependent on the individual locomotion style of a person, whereas in the walking examples this component mainly varies with the phase of the walking cycle for all subjects (Fig. 3.6 (a) and Fig. 3.8 ).

- PC 3: This component periodically varies throughout the walking or running cycle; its amplitude tells how much the free leg is lifted during the floor contact phase of the other leg. The amplitude is consistently
3. Statistical Approach to Tracking and Pose Estimation

Figure 3.8: Varying the second principal component.

Figure 3.9: Varying the third principal component results in lifting the legs.

higher for running examples than for walking examples; it however depends on the individual running styles and is also positively correlated with the running speed (Fig. 3.9).

To summarise, we can state that the first and the third principal components constitute a prototypical walking or running motion, whereas the remaining components mainly encode subject specific aspects of the locomotion, as well as the transition from walking to running motions.

Synthetic Appearance Data. Matching shape descriptors for the body motion capture data are synthetically generated using a computer graphics program (*Maya*²). A 3D skin mesh was attached to the moving skeleton, in order to generate realistically looking human shapes. Collecting real appearance data would be very laborious and time-consuming: Since we need appearance data from different viewing directions, a large number of cameras would have to be placed at different positions around the moving subject, all synchronised with the MoCap system. Also, the attached markers would compromise the quality of the images.

²www.autodesk.com/maya
3.5. Training Data Sets

![Figure 3.10: Example rendering and silhouettes from the LocoETH data set.](image)

Using synthetically rendered figures, these problems can be avoided, in favour of a flexible approach that allows us to easily modify selected aspects of the figure appearance. This process yields large amounts of high-quality shape silhouettes that are largely noise-free which in turn facilitates statistical modelling and training.

For this data set, 36 virtual cameras were placed around the animated realistic 3D human figure, every $10^\circ$. Due to this choice of training data, the tracking algorithms currently assume approximately horizontal camera viewing directions, \textit{i.e.} no top-view. An orthographic camera projection model was chosen, ignoring any perspective distortions that occur in real (\textit{e.g.} wide angle) images. Example silhouettes that were used for training are shown in Fig. 3.10.

**Final Data Set**

The individual training motion sequences were subsampled at 30 Hz, and truncated to segments of 3 seconds, allowing for multiple walking/running cycles at each speed. The resulting data set consists of 2178 body poses for each of the activities walking and running. Multiplied by the number of discrete viewing directions, this leads to 78408 shape examples for each of the activities.
3.5.2 CMU MocapData

This dataset was created in a manner similar to the LocoETH set. Motion capture samples were obtained from a public database\(^3\). For the pose representation, the 3D locations of a set of 13 body joints were computed by fitting a kinematic model to the 3D marker trajectories. Ten walking sequences from 10 different subjects were then rendered, each from 36 different cameras on a horizontal circle around the human model, using a specialised software (MotionBuilder\(^4\)). The motion data was normalised for limb sizes, *i.e.* all the rendered images are based on a standard skeleton with a skin model. Binary foreground/background masks were then obtained from the output images (see Fig. 3.11). This data set consists of 1013 pose examples and 36468 corresponding shape examples.

\(^3\) mocap.cs.cmu.edu

\(^4\) www.autodesk.com/motionbuilder
Combining Discriminative and Generative Inference in a Rao-Blackwellised Particle Filter

4.1 Introduction

The core of the approach described in this chapter is a model of the statistical dependencies between body poses and their appearance, which is learned from training data. Using Gaussian Mixture Models (GMM), the joint probability distribution between body poses and corresponding shape descriptors is approximated. Such a model captures the necessary information to infer all the quantities that are of interest here, given a new image sequence, i.e. the body pose, the global orientation, and the 2D image coordinates of the bounding box that encloses the tracked person.

Fig. 4.1 illustrates the learned model on toy data. Like for the real data sets we consider, there is no direct functional mapping from the shape space to pose space in this illustrative example. However, the Gaussian components of the GMM can act as linear regressors in a discriminative sense. The entire GMM thus amounts to a mixture of linear regressors that can predict multiple pose hypotheses for a given shape instance. The Gaussian components also provide a way to choose among the set of hypotheses, by defining gating functions. When marginalising the model of the joint distribution with respect to pose or shape (i.e. projecting the GMM on one of the axes in Fig. 4.1), prior distributions over likely body poses and likely shapes are obtained. Both are important for 3D tracking; the prior model over poses allows for inferring poses from a given shape in a Bayesian manner, while the distribution over likely shapes helps to locate and follow the humans in the image on a bounding box level, by discriminating between humans and non-human shapes. In contrast to a model of the joint pdf, a purely discriminative model does not provide this last-mentioned piece of information.
Figure 4.1: Illustration: A toy data set consisting of two one-dimensional variables, ‘pose’ and ‘appearance’. The mapping from appearance to pose is multi-valued and can thus not be written as a function. The joint distribution of pose and appearance is approximated by a Gaussian Mixture Model. By marginalising the joint distribution with respect to one of the variables, prior pdfs over pose resp. appearance are obtained.

If the image locations of the tracked persons are known beforehand, the inference problem can be solved parametrically or semi-parametrically. That is, using closed-form computations, parametrised representations of the posterior distributions (mixtures of Gaussians in our case) can be derived. The task however becomes non-parametric when the image location has to be estimated as well. The described approach accounts for the fact that a sub-problem can be solved analytically, by proposing the use of a Rao-Blackwellised Particle Filter, that combines sample based and analytical inference. The flexibility of the approach also allows for a purely sample-based inference strategy that is reasonable when using a low-dimensional parametrisation of the body pose, or for a combination of analytical computations and a preprocessing step that performs the 2D bounding box tracking.
4.2 Modelling and Learning

We want to learn the dependencies of body pose and its appearance in images. The state space for the body pose is given by the variables $\omega$ and $x$, the global orientation of the body relative to the camera and its local pose, i.e. the configuration of its limbs. Under the assumption that the camera is in an approximatively horizontal position, at face or shoulder level, the global orientation can be described with a single parameter that determines the position on a circle around the object from which the latter is observed. We therefore face the problem of learning the joint pdf $p(\omega, x, y)$, where $y$ is an observation, i.e. a descriptor that is computed on the input image. In order to simplify the learning problem, we rewrite this pdf as a mixture of $C$ view-dependent models $p_c$ that each cover a section of possible view directions/global orientations.

$$p(\omega, x, y) = \frac{1}{C} \sum_{c=1}^{C} p_c(\omega, x, y) \quad (4.1)$$

Within the view-dependent models, there is little variation of view direction, so the view angle can be assumed independent from local pose and observation, which enables us to rewrite equation (4.1) as

$$p(\omega, x, y) = \frac{1}{C} \sum_{c=1}^{C} p_c(\omega)p_c(x, y), \quad (4.2)$$

where $p_c(\omega)$ is a one-dimensional Gaussian $N(\omega; \omega_c, \sigma_\omega)$ and $p_c(x, y)$ is the joint pdf of pose and appearance for a certain view direction; this pdf will be learned from training data. Within a view-dependent model the view angles are distributed normally around the mean $\omega_c$, with means $\omega_c$ uniformly spaced over the interval $[0, 2\pi]$, and variances chosen such that adjacent models overlap, and the whole domain of $\omega$ is uniformly covered. In practice the model consists of 36 view-dependent models (see also Fig. 4.2).

The view-dependent models $p_c(x, y)$ themselves are approximated by a mixture of Gaussians (GMM), estimated using e.g. an EM algorithm. The joint distribution over orientation, pose and appearance is thus a mixture of mixtures of Gaussians.

$$p(\omega, x, y) = \frac{1}{C} \sum_{c=1}^{C} \left[ p_c(\omega) \sum_{s=1}^{S} w_{c,s} N(\mu_{c,s}, \Sigma_{c,s}) \right] \quad (4.3)$$

Here, $S$ is the number of Gaussian components in each $p_c$, and $w_{c,s}$ are their weights estimated by the EM algorithm in the learning phase ($\sum_{s=1}^{S} w_{c,s} = 1$). To simplify the learning problem, the number of components $S$ is chosen.
4. Combining Discriminative and Generative Inference in a Rao-Blackwellised Particle Filter

Figure 4.2: Each view dependent model $p_c(\omega)$ covers a range of the possible view angles. Summing up the Gaussian distributions results in an approximately uniform overall distribution $p(\omega)$ over all view angles.

identical for all view-dependent models. $\mu_{c,s}$ and $\Sigma_{c,s}$ are the parameters of the Gaussian components.

Note that even though the omnidirectional model $p(\omega, x, y)$ consists of a discrete number of almost unidirectional models, we have defined a smooth and continuous overall model that covers the entire state space.

4.3 Bayesian Recursive Filtering

According to Bayes’ rule, the tracking problem can be formulated as Recursive Bayesian Filtering in a (hidden) Markov Chain:

\[ p(\theta_t | y_{1:t}) \propto p(y_t | \theta_t) p(\theta_t | y_{1:t-1}), \tag{4.4} \]

where $\theta_t = [x_t, \omega_t]$ is the state variable we want to infer from aggregated observations $y_{1:t}$, and $t$ is the discrete time index. The image likelihood is directly obtained from the learned model of $p(\theta, y)$ by

\[ p(y_t | \theta_t) = p(\theta_t, y_t) / p(\theta_t). \tag{4.5} \]

The temporal prior $p(\theta_t | y_{1:t-1})$ is based on the transition function $p(\theta_t | \theta_{t-1})$ which is not learned explicitly. It should include static information about likely
Figure 4.3: This figure illustrates the learned prior, as a pdf over the state space $\theta_t$, visualised as a 2-dimensional space here. The overall prior (shaded) is defined as the product of the motion model (dashed, for this illustration a Gaussian pdf around the state of $t-1$) and the static learned prior $p(\theta)$.

body poses as well as a motion model of the temporal evolution of poses, which encourages smoothness of the inferred body motions. The temporal behaviour is thus modelled as a Brownian motion around the expected new position, with an additional bias towards likely body configurations. It is defined as the product of the normally distributed temporal prediction and the time-independent (static) pose prior $p(\theta_t) \propto p(x_t)$, where the prior over view angles $p(\omega)$ is uniform. See Fig. 4.3 for an illustration.

$$p(\theta_t|\theta_{t-1}) := k(\theta_{t-1})p(\theta_t)\mathcal{N}(\theta_t; A\theta_{t-1}, \Sigma_T)$$

(4.6)

Here, $A$ specifies the linear autoregressive transition between subsequent states, and $k(\theta_{t-1}) = 1/\int_{\theta_{t-1}} p(\theta_t)\mathcal{N}(\theta_t; A\theta_{t-1}, \Sigma_T)$ is a normalisation factor. Using this definition, we obtain

$$p(\theta_t|y_{1:t-1}) = \int_{\theta_{t-1}} p(\theta_t|\theta_{t-1}) p(\theta_{t-1}|y_{1:t-1})$$

$$= \int_{\theta_{t-1}} k(\theta_{t-1})p(\theta_t)\mathcal{N}(\theta_t; A\theta_{t-1}, \Sigma_T)p(\theta_{t-1}|y_{1:t-1}).$$

(4.7)

The factor $k(\theta_{t-1})$ depends on $\theta_{t-1}$ which makes analytic integration intractable; it can however be computed explicitly in a sampling based approach.
as shown below. We propose a slightly different definition that is suitable for both analytic and Monte-Carlo integration.

\[ p(\theta_t | y_{1:t-1}) := K p(\theta_t) \int_{\theta_{t-1}} N(\theta_t; A \theta_{t-1}, \Sigma_T) p(\theta_{t-1} | y_{1:t-1}) \]  

(4.8)

This formulation corresponds to first propagating the old posterior according to the motion model, and then eliminating unlikely body poses. Both (4.7) and (4.8) define a pdf over \( \theta_t \) that takes into account temporal as well as static prior information.

### 4.3.1 Monte Carlo integration

With the previous definitions and learned models, the inference problem can be solved using sample-based Monte-Carlo integration. In the sequential Monte-Carlo (or particle filter) algorithm [Doucet et al., 2000a], the posterior distribution is approximated by a set of samples (particles) \( \theta^i_t \) with associated weights \( w_t^i \).

\[ p(\theta_t | y_{1:t}) \approx \sum_{i=1}^{N} w_t^i \delta(\theta^i_t, \theta_t) \]  

(4.9)

Here, \( i \) is the index of the particle, and \( \delta(a, b) = 1 \) if \( a = b \) and 0 otherwise. The weights are normalised in order to sum to 1.

The algorithm starts by drawing samples from a proposal density (importance function) \( q(\theta_t) \). Then the importance weights are computed by multiplying the weights from the previous timestep by the ratio between the posterior weight and the value of the proposal density function. The posterior weight is computed as the product of transition probability and image likelihood given the sample.

\[ w_t^i = w_{t-1}^i \frac{p(y_t | \theta^i_t) p(\theta^i_t | \theta^i_{t-1})}{q(\theta^i_t)} \]  

(4.10)

By substituting the learned likelihood model (4.5) and transition density (4.6) into (4.10), \( p(\theta^i_t) \) cancels out and we obtain

\[ w_t^i = w_{t-1}^i k(\theta_{t-1}) \frac{p(\theta^i_t, y_t) N(\theta^i_t; A \theta^i_{t-1}, \Sigma_T)}{q(\theta^i_t)}. \]  

(4.11)

There is a resampling step at each iteration of the tracking algorithm, i.e. \( N \) samples are drawn from the current sample set, each with a probability proportional to its weight. The new set replaces the old one, and all the
samples obtain an equal weight of $w_t^i = \frac{1}{N}$. The multiplication with the previous weight $w_{t-1}^i$ is thus no longer necessary. In our particular case we sample the new particles by using $\mathcal{N}(\theta_t; A\theta_{t-1}, \Sigma_T)$ as the proposal density. Hence, the computation of the weights reduces to the evaluation of the learned model $p(\theta_t, y_t)$ and the computation of the scaling factor $k(\theta_{t-1}^i)$

\begin{equation}
    w_t^i = k(\theta_{t-1}^i)p(\theta_t^i, y_t) = k(\theta_{t-1}^i) \sum_{c=1}^C p_c(\omega_t^i)p_c(x_t^i, y_t).
\end{equation}

### 4.3.2 Analytical Pose Inference

Analytically, $\theta_t$ can be inferred for any given sequence of observations $y_{1:t}$ by combining likelihood (4.5) and prior (4.8).

\begin{equation}
    p(\theta_t | y_{1:t}) \propto p(y_t | \theta_t) \int_{\theta_{t-1}} \mathcal{N}(\theta_t; A\theta_{t-1}, \Sigma_T)p(\theta_{t-1} | y_{1:t-1})
\end{equation}

\begin{equation}
    \propto p(\theta_t, y_t) \int_{\theta_{t-1}} \mathcal{N}(\theta_t; A\theta_{t-1}, \Sigma_T)p(\theta_{t-1} | y_{1:t-1})
\end{equation}

The term $p(\theta_t, y_t)$ is computed according to section 2.3.5, by treating the variable $\theta_t$ as missing information. In order to account for noisy observations, $p(\theta_t, y_t)$ is computed by multiplying the learned model with a Gaussian pdf $p(y_t | y_{\text{obs}})$ around the actual observation $y_{\text{obs}}$, and then marginalising over $y_t$ to obtain a pdf over $\theta_t$. This is illustrated in Fig. 4.4. Marginalisation of a GMM is straightforward; the marginal mixture has the same number of Gaussian components as the original joint mixture with the same weights. The means and covariances of the marginal mixture are simply the means and covariances of the original mixture with all elements corresponding to the variable $y$ removed.

The integral in (4.13) can be calculated in closed form and will result in a Gaussian mixture, so the result of (4.13) is the product of two mixtures and thus a mixture itself. However, the number of mixture components will grow exponentially over time. Hence, a mixture simplification step reduces the number of Gaussian components at each iteration, by pruning components with very low weights and replacing clusters of components by their ‘average’ Gaussian.

\footnote{Alternatively, by using the second definition of the temporal prior (4.8), the normalisation constant $k(\theta_{t-1}^i)$ is not needed.}
Figure 4.4: Multiplication of the Gaussian distribution around the observation $y_{obs}$ and the model $p(\theta, y)$ yields $p(\theta, y | y_{obs})$. By marginalisation, the pdf $p(\theta | y_{obs})$ over body poses is then obtained. Note the multimodality of $p(\theta | y_{obs})$.

4.3.3 Rao-Blackwellised particle filter

The described analytical and sample-based inference approaches both assume that a sequence of shape descriptors $y_{1:t}$ is available as the input of the algorithm. In practice, these image descriptors first have to be computed from the input images, at the image location where the subject is found.

This requires that the bounding box containing the person is either known beforehand or estimated in some way. For the silhouette based image descriptor, one could imagine an ad-hoc algorithm for this 2D tracking problem that tracks the person on a blob-level as a preprocessing step. In general, however, we want to support multiple hypotheses and handle uncertainty for estimation of the 2D location variables, as well as dependencies of body pose estimation and the 2D tracking. The image bounding box coordinates $l = [u, v, w, h]$ (position, width and height) thus have to be included in the state variable and inferred by the overall tracking algorithm. The particle based approach can easily be extended accordingly by adding these location-variables to the state space. In the case of the analytic approach however, there is no straightforward extension since the posterior over this extended state space is unlikely...
4.3. Bayesian Recursive Filtering

Figure 4.5: Graphical structure obtained by partitioning the search space into two parts. This is the setting in which the Rao-Blackwellised particle filter operates. Two temporal slices are displayed, arrows represent conditional dependencies. In the specific case analysed here, $x$ does not directly depend on the node $(\omega, l)$ (dashed arrow).

to have a parametric form. We therefore propose to partition our state space into a part that is solved using a particle filter and a part that is solved analytically using our learned models. In such a way the non-parametric nature of the inference problem can be accounted for, while keeping the advantages of analytic computations where possible. By the chain rule of probability, the posterior over $x, \omega, \text{and the location variable } l$ can be written as

$$p(x, \omega, l | y) = p(x | \omega, l, y)p(\omega, l | y), \quad (4.14)$$

where the temporal aspects of the problem are omitted for notational simplicity. Given our learned model, $p(x | \omega, l, y)$ can be inferred analytically and described parametrically, whereas $p(\omega, l | y)$ lacks an analytic solution. However, due to its low dimensionality, it can be handled by a particle filter. The Rao-Blackwellised Particle Filter (RBPF, [Murphy and Russel, 2001]) offers a framework for inference, when a part of the state space can be marginalised analytically. Fig. 4.5 shows the graphical structure of this setting as a Bayesian network.
The marginal posterior distribution over $\omega_t$ and $I_t$ will be approximated by a set of particles:

$$p(\omega_t, I_t | y_{1:t}) \approx \sum_i w^i_t \delta(\omega^i_t, I^i_t; \omega_t, I_t), \quad (4.15)$$

where $w^i_t$ is the weight of the $i$th particle, and $\delta(A, B; a, b) = 1$ if $a = A$ and $b = B$ and 0 otherwise. The marginal on $x_t$ is approximated as

$$p(x_t | y_{1:t}) \approx \sum_i w^i_t p(x_t | \omega^i_{1:t}, I^i_{1:t}, y_{1:t}). \quad (4.16)$$

In practice, each particle will consist of a sample for $I_t$ and $\omega_t$, a parametric pdf $p(x_t | y_{1:t}, I^i_{1:t}, \omega^i_{1:t})$ and a weight $w^i_t$. There are thus three tasks: proposing samples $[I^i_t, \omega^i_t]$, computing the weight for each sample, and perform the exact step to update the parametric pdf $p(x_t | y_{1:t}, I^i_{1:t}, \omega^i_{1:t})$. The latter follows the derivation for analytic density propagation (4.13), except that we only infer the variable $x_t$, and that the expression is additionally conditioned on $\omega^i_{1:t}$ and $I^i_{1:t}$. We will denote as $y^i_t$ the image descriptor computed from the bounding box coordinates encoded by the sampled location $I^i_t$, and obtain

$$p(x_t | y^i_{1:t}, \omega^i_{1:t}) = \frac{1}{L^i} p(y^i_t | x_t, \omega^i_t) p(x_t | y^i_{1:t-1}, \omega^i_{1:t-1})$$

$$\propto \frac{K^i}{L^i} p(x_t, \omega^i_t, y^i_t) \int_{x_{t-1}} p(x_{t-1} | A x_{t-1}, \Sigma_T) p(x_{t-1} | y^i_{1:t-1}, \omega^i_{1:t-1}). \quad (4.17)$$

Here, we used the independence of $x_t$ and $\omega^i_t$ and the uniformity of $p(\omega^i_t)$. $K^i$ is the scaling factor from the prior (4.8), and the normalisation factor $L^i$ is equal to the likelihood of the observation given the $i$th sample.

$$L^i = \int_{x_t} p(y^i_t | x_t, \omega^i_t) p(x_t | y^i_{1:t-1}, \omega^i_{1:t-1})$$

$$= p(y^i_t | y^i_{1:t-1}, \omega^i_{1:t}) = p(y_t | y^i_{1:t-1}, I^i_{1:t}, \omega^i_{1:t}) \quad (4.18)$$

Hence, as in standard particle filtering, if we choose the transition function $p(I_t, \omega_t | I_{t-1}, \omega_{t-1})$ as a proposal function, the weights $w^i_t$ are given by the observation likelihood, i.e. the normalisation factor $L^i$ [Murphy and Russel, 2001]. Marginalising over $x_t$ to compute $L^i$ is the key element of the RBPF; it is simply done by explicitly computing the scale of the unnormalised pdf $p(x_t | y^i_{1:t}, \omega^i_{1:t})$ in (4.17).
4.4 Implementation and Experimentation

The previous sections derived the theory of body pose estimation and tracking, independently of any specific choices for the used image descriptors, body pose parametrisation, and training data. In this section, an implementation and experimental validation is presented, which serves as a proof of concept and aims at illustrating the potential and flexibility of the overall approach.

4.4.1 Representations and Dimensionality Reduction

Shape descriptors. The chosen image descriptors are based on the silhouette of the tracked person. Using a stationary camera, the segmentation in the input images is obtained via background subtraction. To encode these segmented images using a descriptor with desirable properties, we use signed distance functions, that assign to every point on a grid a signed value indicating the distance to the closest point on the silhouette. See section 3.4.2 for a description of this shape representation. Equation (3.3) shows the computation of this descriptor, including a PCA dimensionality reduction step. For the reported experiments, the 15 dominant principal components were retained.

Pose representation. To describe the body articulations as a vector of real numbers, a positional parametrisation is used (see section 3.4.1). It is based on the 3D locations of 13 body joints such as ankles, knees, hips, shoulders, elbows and wrists, expressed in a local coordinate system rooted at the hip of the subject, and oriented in according to the body orientation. The resulting descriptor is fairly high-dimensional (39 dimensions) which makes it difficult to model with Normal distributions. Furthermore we believe that the intrinsic dimensionality of the training data is much lower, and in particular for the considered locomotion patterns there are strong correlations between the positions of the individual joints and limbs. Therefore, the dimensionality can be reduced to 15 dimensions using linear PCA, without significant loss of information.

In the case of gait or human locomotion we have the strong intuition, that this class of motion is essentially a one dimensional manifold, a closed curve that is run through once every period. This structure could be identified automatically by a nonlinear dimensionality reduction step, or, since we are aware of it, explicitly imposed by using an appropriate high-level parametrisation of the body pose. In one of the experiments, such an action specific pose representation was used, which essentially consists of the phase $\gamma$ within a walking cycle, and is thus a number between 0 and $2\pi$. However, the discontinuity of
4. Combining Discriminative and Generative Inference in a Rao-Blackwellised Particle Filter

this parametrisation, that occurs when the values ‘wrap around’ from $2\pi$ back to 0 is a problem for the learning procedure. Therefore we use a redundant encoding using two dimensional coordinates on the unit-circle,

$$\Gamma = [\cos(\gamma), \sin(\gamma)]. \quad (4.19)$$

4.4.2 Dataset and Training

Human locomotion is chosen as a test case. The statistical models are learned by using the CMU MocapData dataset from section 3.5.2 as training examples. 36 view dependent models, spaced every $10^\circ$ on a circle around the subject, were learned. For each of them, the joint distribution of appearance and pose descriptors was approximated by a GMM using an EM algorithm. Empirically, GMMs with 11 components were found to yield good results.

4.4.3 Tracking Experiments

**Rao-Blackwellised particle filter.** Using a plain particle filter in combination with this pose representation would require a large number of particles. The following results were obtained using the algorithm from section 4.3.3, where a part of the inference problem is solved analytically. Samples for $l$ and $\omega$ are generated from a temporal prior that assumes constant velocity respectively Brownian motion. The temporal model for pose $x$ assumes Brownian motion in body pose space, with a covariance matrix learned from the training data. For the initialisation of the 2D location variables $l$, an ad hoc proposal function at the centre of gravity of the segmented image was used to sample from, $x$ and $\omega$ were initialised using the analytical inference equation (4.13), by assuming a uniform temporal prior.

The posterior distributions are stored as a set of samples, where each sample has its own pdf over body articulations. This representation is very powerful, allowing for multimodality and for expressing posterior uncertainties. However, it is also cumbersome to deal with, and in particular suboptimal when a single estimate is required as the output of the algorithm. In the figures of this section, the sample with the highest weight was chosen for the visualisation of the orientation and the bounding box. For the body pose, the mean of the Gaussian component with the highest weight was chosen, from the pdf that corresponds to the sample with the highest weight. While this visualisation does generally not correspond to the maximum a posteriori estimate, it is still useful to judge how well the tracking works. To improve the visualisation, in section 5.3.2 a consistent motion trajectory is extracted by optimising over the entire sequence.
Figure 4.6: Pose and bounding box estimates for the diagonal synthetic sequence. The view angle is approximately 45°.
Figure 4.7: Reconstruction vs. ground truth for best (a) resp. worst (b) reconstruction of the fronto-parallel synthetic sequence.

Figure 4.8: Deviation from ground truth for two synthetic sequences. Euclidean distance (centimetres) between reconstructed joints and the ground truth, averaged over the sequences. The error is largest at the extremities (ankle, wrist).
4.4. Implementation and Experimentation

First, the algorithm was applied to synthetically generated image sequences, for which ground truth body pose trajectories are available. The synthetic images were obtained using the same computer graphics pipeline that was used to generate the training data. They are based on MoCap locomotion data of two subjects that are not contained in the training database. There are two sequences, one with diagonal walking and a fronto-parallel one. A few frames of the diagonal input sequence with estimated poses are shown in Fig. 4.6. For these two sequences, Fig. 4.8 compares the estimation errors, in terms of distances between the estimated joint position and their ground truth, averaged over the sequence. For the diagonal walking sequence the errors are larger than for the fronto-parallel sequence. Two reconstructed poses of the fronto-parallel sequences, the best and the worst estimations of that sequence, are shown in Fig. 4.7, next to their groundtruth. In Fig. 4.9 the estimated foot trajectories of the diagonal sequence are plotted against the ground truth.

The estimated view angle trajectory $\omega$ is shown in Fig. 4.10, for the fronto-parallel sequence, as well as for a third synthetic sequence with varying view angle. The estimation follows the overall rotation. Largest deviation from this ground truth is about $15^\circ$, the average absolute error is $5^\circ$. However there seems to be no systematic misestimation since the mean difference from

Figure 4.9: Trajectory of the feet (ankle) through a sequence of diagonal walking. The figure shows the value of the x-coordinate (that points in walking direction) of the estimated ankle positions, and the ground truth.
Figure 4.10: Estimated view directions \( \omega \) (radians) for two synthetic sequences with ground truth (dashed curve). The figure shows the angle encoded by the sample with the highest weight.
ground truth is only 1°, the negative and positive deviations basically sum to zero. These results are rather convincing, when considering that it is very difficult, even for humans, to perceive the orientation of a body from the silhouette alone.

Fig. 4.11 shows tracking through a real office sequence. The images were recorded with a DV camera at a frame rate of 25 fps and segmentation was obtained using background subtraction. For each of the selected frames, the left column of the figure shows the segmented input images, or more precisely the content of the bounding box that corresponds to the sample with the highest weight. The other columns show the estimated pose from side view resp. 45°. Note that between frame i) and j) the posterior mode that corresponds to stepping forward with the left leg suddenly becomes more likely. This results in a sudden switch between the two main modes of the posterior, which can only be seen from the 45° view, and is essentially an artefact of choosing the sample with the highest weight for visualisation. It also shows that the intrinsic multimodality of the tracking problem is correctly represented by the posterior distributions. The reconstructed poses and motion look natural. Occasionally (e.g. frame c), the reconstruction of the arms is imprecise, especially when they are occluded by the torso, i.e. not visible in the silhouette images. In such cases the pose prior alone is responsible for the estimation of the arm pose.

Tracking results for another real sequence are displayed in Fig. 4.12. The person walks diagonally to the camera’s view direction, and there are perspective distortions due to the wide angle lens used. While strictly spoken this is a violation of the assumed orthographic projection (the training images use orthographic projection), it does not seem to affect the performance substantially.

**Low dimensional pose parametrisation - particle filter.** For another experiment, a high-level encoding of the body pose of a walking subject was used. This experiment illustrates how the proposed framework can be used with different kinds of pose descriptors. The training database was annotated with a phase label corresponding to each rendered body pose, by manually identifying walking cycles in the motion data. A two-dimensional pose descriptor was then computed according to (4.19). For each view dependent model, the joint distribution of appearance and pose descriptors was approximated by a GMM with 10 components using EM.

Due to the relatively low dimension of the search space (2D location, width and height of the bounding box, angle ω, and pose representation γ), a simple particle filter is suitable for inference. Note that for sampling, the one-dimensional
Figure 4.11: Tracking through a real sequence. See text for explanations.
Figure 4.12: Diagonal walking sequence. The estimated bounding boxes and 3D poses are shown. For visualisation, the samples with the highest weight are selected for each timestep.
Figure 4.13: Estimated phase (low dimensional body parametrisation). For each frame, the figure shows the proposed samples (small dots), the distribution after resampling (larger dots), and the sample with the highest weight (solid trajectory). At frame 35 the estimate wraps around from phase 1 to 0. Despite some problems around frame 25, the tracker can recover.

parametrisation $\gamma$ is used, and then converted to the two-dimensional representation $\Gamma = [\cos(\gamma), \sin(\gamma)]$ for likelihood evaluation. As a temporal prior, standard Brownian motion models were used for 2D location $l$ and $\omega$. For the temporal pose parametrisation, a meaningful temporal prior can be defined very easily, by stating that the phase label moves forward in time on average. The weights of the samples were computed according to (4.12). The particle set consisted of 100 particles that were initialised manually. This variant of the tracking algorithm was tested on the real office sequence of Fig. 4.11. Fig. 4.13 visualises the propagation of the particles by the particle filtering algorithm. See also Fig. 4.14 for a few estimated body poses. By definition, this method can only recover the basic motion, and ignores all other characteristics of movements, such as individual walking styles etc.

4.5 Discussion

The presented approach, based on the Rao-Blackwellised Particle filter, combines sample-based Monte-Carlo inference with analytical computations. While
4.5. Discussion

Figure 4.14: Reconstructed poses using the low dimensional pose parametrisation. The frames shown here correspond to frames b,d,f,h,j,k of Fig. 4.11.

The former operates in a generative top-down manner, the analytical inference is essentially a discriminative bottom-up process. Both sub-problems are guided by the (generative) model of the joint pdf of body pose and shape.

This combination allows for inference of multimodal and nonparametric posteriors despite the high-dimensional state space parametrisation. Nevertheless, the mixed particle-based/parametric representation of the posteriors can be cumbersome in practice, e.g. for visualisation of the results or for quantitative evaluation. These problems can be partly avoided by the approach described in chapter 5, where purely sample-based inference is used in combination with a stronger dimensionality reduction. Furthermore, that approach includes a global optimisation which yields a single consistent result for an input sequence.

In general, it is hard to evaluate the results quantitatively, because they are available in the form of probability distributions. Furthermore, they are expected to reflect uncertainties and multimodalities. Thus, just comparing the most likely explanation to ground truth is insufficient, since it is known that
there are remaining uncertainties that the algorithm can not (and thus must not) resolve given silhouettes as observations.

In the sense of a qualitative evaluation, it is easy to identify cases where the performance of the tracker is not satisfying. Reliably estimating the orientation of the subject from silhouettes turns out to be rather difficult, for different reasons. First, the shape of observed real humans may differ from the ones used for training. In particular the shape of corpulent or skinny subjects may be explained by the algorithm by inferring a wrong orientation. Second, the temporal prior on orientation may not reflect the characteristics of human motion optimally, since it discourages from abrupt turns. Even the Markov assumption may not be appropriate here; in order to decide whether a motion fragment looks natural, one must often consider what happens before or after that segment. An online tracking algorithm like particle filtering, which only takes into account observations up to the current timestep, may easily be distracted. Some kind of look-ahead mechanism could be of help here. Learning the relationships between the body configuration and variations of the viewpoint angle seems promising, but requires much larger and more complex training databases. Another practicable way is the decoupling of the pose estimation, and the tracking of location and view-direction, as done in chapter 6. In such a way, more complex global reasoning can be issued for the view-direction, which would not be feasible for the entire pose space, due to complexity reasons.

Modelling the joint distribution with Gaussian Mixture Models is attractive due to their simplicity, and because they capture all the necessary prior information in a single model. Nevertheless, being a collection of linear components, they are somewhat limited in modelling the complex nonlinear relation between appearance and pose. Hence, when visualising the tracking results, the transition between mixture components sometimes causes noticeable jitter in the joint trajectory reconstructions. Again, extracting a consistent and smooth optimal trajectory as the final result helps to get smoother motions (see chapter 5). Furthermore, introducing a learned model of the dynamics of human locomotion further improves the results, as shown in the following chapters of this thesis.
Generative Regression-Based Model for Pose Estimation

5.1 Introduction

Much of the work on human body pose has been based on generative models and inference algorithms. This approach is often used in conjunction with geometrical models of the body. On one hand these generative methods have been successful to some extent. On the other hand, learning based statistical approaches do have advantages, as argued in section 3, since the necessary information can be automatically learned from realistic examples.

Here, the generative methodology is combined with a learning based statistical approach. The mapping from pose to appearance is single-valued and can thus be seen as a nonlinear regression problem. In contrast, the discriminative mapping from appearance to pose, that would allow for more direct inference, is multivalued and ambiguous, as illustrated in Fig. 5.1. We approximate the generative mapping with a Relevance Vector Machine (RVM) kernel regressor [Tipping, 2000] that is efficient due to its sparsity. Although single-valued, the appearance prediction will be subject to uncertainty, because other factors than just the body configuration (pose) may affect appearance (clothing, physical constitution, lighting conditions etc). This is taken into account by learning a prediction variance matrix of the mapping.

The observations are available in the form of roughly segmented monocular image sequences that are obtained by a pre-processing step such as motion segmentation, background subtraction or other. A main focus of the proposed approach lies on the ambiguities and uncertainties that are inherent in body tracking from such input. Recursive Bayesian Sampling [Isard and Blake, 1998a; Doucet et al., 2000a] offers a framework for dealing with non-Gaussian and multimodal body pose posteriors and allows us to integrate the nonlinear learned dynamical model. However, sampling-based algorithms are generally
Figure 5.1: Illustration of a discriminative one-to-many mapping with a mixture of linear regressors (a), and of a generative mapping from pose space to appearance space with a single nonlinear regressor (b).
not applicable for inference in high-dimensional state spaces like the space of body poses. We therefore use Locally Linear Embedding (LLE, [Roweis and Saul, 2000]) to find a low-dimensional embedding of our 60-dimensional pose parametrisation. With 4 LLE dimensions, the considered motions can be captured reasonably well.

To summarise, when comparing to the general framework for statistical tracking of Fig. 3.1 and the system of the previous chapter, the regression-based approach discussed here has the following specific characteristics:

- The generative top-down mapping from body articulations to body shape is learned.
- Learning a low-dimensional pose manifold is an integral component of the algorithm, necessary to render the generative learning and inference feasible.
- The dynamics of body poses are learned.
- A postprocessing step extracts a globally optimal trajectory through an entire video sequence.

This approach differs from most previous work on statistical tracking in that it simultaneously tracks in a state space that includes body pose and 2D bounding box location. It can be extended to additionally infer a discrete activity label, i.e. perform simultaneous activity recognition (see chapter 7). Furthermore, the pipeline is entirely based on learned models, with generative rather than discriminative modelling of the appearance. The system is built in a modular manner. Some choices of precise statistical methods that are applied for the individual modules are mainly based on practical considerations (e.g. efficiency, sparsity). They could be substituted by equivalent methods, like e.g. Isomap [Tenenbaum et al., 2000] instead of LLE, and regularised kernel regressors or Gaussian processes instead of RVMs.

## 5.2 Modelling and Learning

Fig. 5.2 shows an overview of the tracking framework. Central element is the low-dimensional body pose parametrisation, with learned mappings back to the original pose space and into the appearance space. In this section all elements of the framework will be described in detail. The models were trained on motion capture data sets of different subjects, running and walking at different speeds.
5. Generative Regression-Based Model for Pose Estimation

Figure 5.2: An overview of the tracking framework. Solid arrows represent signal flow during inference, the dashed arrow stands for the nonlinear dimensionality reduction during training. The figure refers to equations in Sections 5.2 and 3.4.2.

5.2.1 Relevance Vector Machines

The relations between the different variables of the framework are learned by modelling the regression from one variable to the other. The machine learning technique of choice here is the Relevance Vector Machine (RVM, [Tipping, 2000]). The RVM is based on Bayesian theory, and its prediction formula has a functional form identical to the Support Vector Machine (SVM, [Cristianini and Shawe-Taylor, 2000]) that, while best known for classification, also has its regression variant (Support Vector Regression, SVR). RVM learning produces even sparser solutions than SVR, and has fewer parameters, with a clear probabilistic interpretation. In this section a short introduction to the use of RVMs is given, for their derivation and theoretical background the reader is referred to [Tipping, 2000; Tipping, 2001; Quiñonero-Candela and Rasmussen, 2005; Rasmussen and Williams, 2006].
Using a RVM, predictions of the function value $f(\mathbf{x})$ are made by a linear combination of the responses of a set of basis functions.

$$f(\mathbf{x}) = \mathbf{W} \Phi(\mathbf{x}),$$

(5.1)

where $\mathbf{W} \in \mathbb{R}^{d \times m}$ are the regression weights, $d$ the number of output dimensions and $m$ the number of basis functions. $\Phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \ldots, \phi_m(\mathbf{x})]^T$ is the vector with the responses of all basis functions for the input $\mathbf{x}$. The basis functions are centred at the training inputs. Learning the RVM is based on optimising the RVM marginal likelihood that relies on the likelihood of the dataset given the weights, assuming i.i.d noise, and a Gaussian prior over the weights. Learning amounts to optimising the marginal likelihood with respect to the variances of the individual weights in the prior. This process prunes away many of the basis functions, which leads to a sparse $\mathbf{W}$. The surviving $m$ basis functions are centred at the corresponding training inputs, called Relevance Vectors. In the multivariate case we are interested in here, the sparsity for the different output dimensions has to be linked, making sure that entire columns of $\mathbf{W}$ will consist of zeros.

Widely used basis functions (or kernels) include linear, polynomial and Gaussian functions. In this thesis Gaussian radial basis functions are used,

$$\phi_j(\mathbf{x}) = k(\mathbf{x}_j, \mathbf{x}) = \exp\left(\frac{-||\mathbf{x} - \mathbf{x}_j||^2}{\sigma^2}\right).$$

(5.2)

### 5.2.2 Low-dimensional Pose Manifold

Representations for the full body pose configuration are high dimensional by nature; our current representation is based on 3D joint locations of 20 body locations such as hips, knees and ankles (see Fig. 3.4), but any other representation (e.g. based on relative orientations between neighbouring limbs) can easily be plugged into the framework. To alleviate the difficulties of high dimensionality in both the learning and inference stages, a dimensionality reduction step identifies a low dimensional embedding of the body pose representations. We use Locally Linear Embedding (LLE) [Roweis and Saul, 2000], which approximately maintains the local neighbourhood relationships of each data point and allows for global deformations (e.g. unrolling) of the dataset/manifold.

LLE dimensionality reduction is performed on all poses in the data set and expresses each data point in a space of desired low dimensionality (see also Fig. 5.3). However, LLE does not provide explicit mappings between the two spaces, that would allow to project new data points (that were not contained in the original data set) between them. Therefore, we model the reconstruction
Figure 5.3: Low-dimensional manifold of walking data obtained by Locally Linear Embedding. Three dimensions of the four-dimensional representation are visualised here, from different views. The different colours indicate different walking speeds (red: 2.5 km/h, green: 4.2 km/h, blue: 6 km/h).
5.2. Modelling and Learning

Figure 5.4: Reconstruction errors of the mapping from LLE coordinates to the original pose representation, as a function of the number of LLE dimensions. The errors are computed in terms of Euclidean distances of the reconstructed joints from their ground truth, averaged over all joints and poses.

The training examples form a periodic twisted ‘ring’ in LLE space, with a curvature that varies with the phase within the periodic movement. A linear dynamical model, as often used in tracking applications, is not well suited to predict future poses on this curved manifold. We view the nonlinear dynamics
as a regression problem, and model it using another RVM regressor, yielding the following *dynamic* prior,

\[
p_d(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_{t-1} + f_d(\mathbf{x}_{t-1}) \Delta_T, \Sigma_d),
\]

where \( f_d(\mathbf{x}_{t-1}) = \mathbf{W}_d \Phi_d(\mathbf{x}_{t-1}) \) is the nonlinear mapping from poses to local velocities in LLE pose space, \( \Delta_T \) is the time interval between the subsequent discrete timesteps \( t - 1 \) and \( t \), and \( \Sigma_d \) is the variance of the prediction errors of the mapping, computed once for the entire data set, from a hold-out data set that was not used for the estimation of the mapping itself.

Not all body poses that can be expressed using the LLE pose parametrisation do correspond to valid body configurations that can be reached with a human body. The motion model described so far only includes information about the temporal evolution of the pose, but no information about how likely a certain body pose is to occur in general. In other words, it does not yet provide any means to restrict our tracking to feasible body poses. This is an issue, because the learned regressors can produce erroneous outputs when they are applied to infeasible input poses, due to the limited extrapolation capabilities of kernel regressors to regions without any training data. Therefore, the additional prior knowledge about feasible body poses, or likely poses for the modelled activity, is introduced as a *static* prior that is modelled with a Gaussian Mixture Model (GMM),

\[
p_s(\mathbf{x}) = \sum_c p_c \mathcal{N}(\mathbf{x}; \mu_c, \Sigma_c),
\]

with \( C \) the number of mixture components and \( p_c, \mu_c \) and \( \Sigma_c \) the mixture proportions and parameters of the Gaussian components. The influence of this pose prior can be kept low, avoiding a distortion of the tracking results towards typical average motion. We introduce a weighting factor \( \lambda > 1 \) and obtain the following formulation for the temporal prior by combination with the *dynamic* prior \( p_d(\mathbf{x}_t | \mathbf{x}_{t-1}) \),

\[
p(\mathbf{x}_t | \mathbf{x}_{t-1}) \propto p_d(\mathbf{x}_t | \mathbf{x}_{t-1}) \, p_s(\mathbf{x}_t)^{\frac{1}{\lambda}}.
\]

The temporal prior thus amounts to a learned motion model with a slight bias towards likely body motions.

### 5.2.4 Appearance Model

The representation of the subject’s image appearance is based on a rough figure-ground segmentation. Under realistic imaging conditions, it is not possible to get a clean silhouette, therefore the image descriptor has to be robust
to noisy segmentations to a certain degree. We consider two types of image descriptors, *distance transforms* \( dt(Y) \) of segmented figures with a subsequent linear PCA dimensionality reduction step, and a representation obtained by applying *Binary PCA (BPCA)* to binary foreground images. These appearance encodings are introduced in section 3.4.2. Both image descriptors are computed from the content of a bounding box around the centroid of the figure, and 10 to 20 PCA resp. BPCA components have been found to yield good reconstructions. The descriptors are denoted \( y_{BPCA} \) and \( y_{DT} \).

As we will see later, it is useful in some situations to consider the reverse operation that projects the image descriptors \( y_{DT} \) and \( y_{BPCA} \) back into high dimensional pixel space and transforms it into binary images or foreground probability maps. From the descriptors we compute probability maps via the sigmoid function \( \sigma(\cdot) \), as shown in equation (3.2) for the BPCA descriptor and in (3.4) for signed distance functions.

Given the operations to compute the image descriptors and convert them back into segmented images, we will look how the image appearance is linked to the LLE body pose representation \( x \). The *generative* mapping from pose \( x \) to image descriptors \( y \) is modelled, which allows to predict image appearances given pose hypotheses and fits well into generative inference algorithms such as recursive Bayesian sampling. In addition to the local body pose \( x \), the appearance depends on the global body orientation \( \omega \) (rotation around vertical axis).

\[
p(y|x, \omega) = \mathcal{N}(y; f_a(x, \omega), \Sigma_a)
\]

\[
f_a(x, \omega) = W_a \Phi_a(x, \omega)
\]

(5.7)

Here, the functional mapping \( f_a(x, \omega) \) is approximated by a sparse kernel regressor (RVM) with weight matrix \( W_a \) and kernel function responses \( \Phi_a(x, \omega) \). \( \Sigma_a \) is the prediction variance matrix, it indicates which dimensions of the descriptor \( y \) can be well predicted and which cannot, and thus accounts for the fact that the prediction of \( y \) will always be subject to uncertainty. \( \Sigma_a \) is estimated from a hold-out set of the original training data and is restricted to a diagonal matrix for simplicity.

### 5.3 Tracking and Pose Inference

#### 5.3.1 On-Line Sample-Based Tracking

In this section it is shown how the 2D image position, body orientation, and body pose of the subject are simultaneously estimated given a video sequence,
by using the learned models from the previous section within the framework of recursive Bayesian sampling. Both pose estimation and image localisation can benefit from the coupling of pose and image location. For example, the known current pose and motion pattern can help to track through occlusions and distinguish subjects from each other. We therefore believe that tracking should happen jointly in the entire state space \( \Theta \),

\[
\Theta_t = [\omega_t, u_t, v_t, w_t, h_t, x_t],
\]

(5.8)

consisting of the orientation \( \omega \), the 2D bounding box parameters (position, width and height) \( u, v, w, h \), and the body pose \( x \).

Despite the reduced number of pose dimensions, we face an inference problem in 9-dimensional space. Having a good sample proposal mechanism like our dynamical model is crucial for the Bayesian recursive sampling to run efficiently with a moderate number of samples. For the monocular sequences we consider, the pose posteriors can be highly multimodal. For instance a typical walking sequence, observed from a viewing direction orthogonal to the walking direction, has two obvious posterior modes, shifted 180 degrees in the phase of the walking cycle, corresponding to the left respectively the right leg swinging forward. When considering other viewing directions, the situation gets even more complex, with an additional orientation ambiguity. Our experiments have shown that a strong dynamical model is necessary to avoid confusion between these posterior modes and to reduce ambiguities. Some posterior multimodalities do however remain, since they correspond to a small number of different interpretations of the images, which are all valid and feasible motion patterns. In section 5.4.1 the ambiguities of the learned appearance model are studied experimentally on a concrete example.

In detail, the inference algorithm is very similar to classical \textit{CONDENSATION} [Isard and Blake, 1998a], with normalisation of the weights and resampling at each time step. The prior and likelihood for our inference problem are obtained by extending (5.6) and (5.7) to the full state space \( \Theta \). In our implementation, the dynamical prior \( p_d(\Theta_t | \Theta_{t-1}) \) serves as the sample proposal function. It consists of the learned dynamical pose prior from eq. (5.4), and a simple linear motion model for the remaining state variables \( \theta = [\omega_t, u_t, v_t, w_t, h_t] \).

\[
p_d(\Theta_t | \Theta_{t-1}) = p_d(x_t | x_{t-1})N(\theta_t; \theta_{t-1}, \Sigma_\theta)
\]

(5.9)

In practice, one may want to use a standard autoregressive model for propagating \( \theta \), omitted here for notational simplicity. We assume statistical independence between the body pose \( x \) and the state variables \( \theta \) in (5.9), since modelling these dependencies would imply restricted or known camera motions (e.g. static camera). The static prior over likely body poses (5.5) and
5.3. Tracking and Pose Inference

the likelihood (5.7) are then used for assigning weights $w_i$ to the samples. The static prior over orientations and 2D image locations, $p_s(\theta)$, is assumed to be uniform.

$$w_i \propto p(y_i^T \Theta_i)p_s(\Theta_i)^{\frac{1}{\lambda}} = p(y_i^T | x_i^T, \omega_i)p_s(x_i^T)^{\frac{1}{\lambda}}$$  \hspace{1cm} (5.10)

Here, $i$ is the sample index, and $y_i^T$ is the image descriptor computed from the current image $I_t$ at the sampled bounding box coordinates $[u_i^T, v_i^T, w_i^T, h_i^T]$. Note that our choice for sample proposal and weighting functions differs from CONDENSATION in that we only use one component ($p_d$) of the prior (5.6) as a proposal function, whereas the other component ($p_s$) is incorporated in the weighting function.

Likelihood computation in image space or on a PCA subspace.

Our framework has a generative flavour, since we model the pdf of the appearance given the body pose in a top-down manner. The computation of the image descriptor and projection on the subspace and back can be issued in both directions, as seen in equations (3.1) to (3.4). The image likelihood can thus be evaluated either in the low-dimensional space of the image descriptor, or in the original image space. The first possibility is to compute the image descriptors in a bottom-up manner and project them onto the PCA or BPCA subspace (3.1, 3.3), where the likelihood is then directly obtained using (5.7).

Alternatively, in a purely generative top-down manner, we can predict whether we expect a certain pixel to be foreground or background given a pose hypothesis. This is done by concatenating the mapping $f_a(x, \omega)$ from (5.7) and the projection of the appearance descriptor into full appearance space (image space) (3.2, 3.4). This yields a discrete 2D probability distribution of foreground probabilities $Seg$ over the pixels $p$ in the bounding box. From this pdf, a likelihood measure can then be derived by comparing it to the actually observed segmented image $Obs$, also viewed as a discrete pdf, using the Bhattacharyya similarity measure $BC$ [Bhattacharyya, 1943] which measures the affinity between distributions.

$$Seg_i^t(p) = P(p = 1 | f_a(x_i^T, \omega_i^T))$$

$$Obs_i^t(p) = P(p = 1 | I_t, u_i^T, v_i^T, w_i^T, h_i^T)$$

$$BC_i^t = \sum_p \sqrt{Seg_i^t(p) Obs_i^t(p)}$$  \hspace{1cm} (5.11)

$I_t$ is the raw unsegmented input image at time $t$.

Both alternative ways of likelihood computation have advantages and drawbacks. The bottom-up variant requires binary images to compute the image
descriptors, whereas the top-down variant can handle continuous foreground probabilities $\text{Obs}_t^i(p)$. Often the foreground segmentation is available in the form of probability maps, and thresholding it may cause an unnecessary loss of information and yield unsatisfying results. On the other hand, the evaluation of likelihood on the (B)PCA subspace can benefit from the learned variance matrix $\Sigma_a$. Also, the bottom-up computation of descriptors can be disturbed by noisy segmentations. This holds particularly for the distance transformed image descriptor $y_{DT}$, where a few missegmented pixels can spread due to the distance transform operation and thus corrupt the descriptor. In the case of the descriptor based on BPCA, the projection on the subspace is iterative and therefore slow, which in this case reduces the attractiveness of the bottom-up variant from a practical perspective. Experimentally, the combination of distance transformed descriptors and bottom-up descriptor computation fails when the input image segmentation is very noisy. The other three combinations, i.e. the top-down variants and the BPCA bottom-up variant, perform similarly well.

5.3.2 Off-Line Global Optimisation

The described sample-based tracking algorithm provides a set of $N$ samples with corresponding weights for each frame of the sequence. This representation of the posterior is not suitable for many purposes, even visualisation is difficult. Furthermore, the posteriors are computed on a per-frame level, i.e. at time step $t$ we compute $p(\Theta_t|I_{1:t})$ given the history of observed images $I_{1:t}$. Often we are interested in a consistent trajectory through the entire image sequence, i.e. in the maximum of the posterior $p(\Theta_{1:T}|I_{1:T})$ over the states of all time steps, given all observations. In other words, we are interested in the value for $\Theta_{1:T}$ with maximal probability rather than marginals for each $\Theta_t$.

In our framework this is achieved by a postprocessing algorithm that finds optimal paths through the set of samples. As shown in [Doucet et al., 2000b] the MAP estimate of the state sequence is obtained by a Monte-Carlo (forward) filtering stage, followed by a Viterbi algorithm [Forney, 1973] that operates on the samples of the particle filter. In the approach proposed here, the Viterbi algorithm is replaced by the max-product algorithm, which is a generalisation of the Viterbi to soft outputs instead of hard decisions [Wiberg, 1996; Kschischang et al., 2001]. The max-product algorithm is a variant of the standard belief propagation algorithm (or sum-product algorithm). See [Kschischang et al., 2001] or [Yedidia et al., 2002] for belief propagation algorithms. These algorithms are discrete by nature, i.e. each node of the Markov chain (each time step, see also fig. 5.5) has a number of discrete states that in our case is equal to the number of samples $N$ of the particle filter tracking
5.3. Tracking and Pose Inference

The algorithm will thus choose one sample per node to form a trajectory through time and state space that best satisfies both observation likelihood and temporal prior. In practice we use the numerically more stable counterpart of max-product, the min-sum algorithm that performs the computations in negative log space instead of probabilities. Instead of finding the optimal trajectory for the entire sequence, the algorithm can also be applied to sub-sequences, in a sliding-window fashion.

In [Isard, 2003; Sudderth et al., 2003] unifying frameworks have been presented, that generalise belief propagation to continuous state spaces using Monte-Carlo sampling. They perform filtering and smoothing, forward and backward propagation in a single formulation. These methods are however not applicable here, since they are based on the sum-product algorithm and therefore compute per-node marginals. Similarly, in [Isard and Blake, 1998c] a backwards sample re-weighting step is introduced to compute marginal ‘smoothing’ distributions for each time step. In the two-stage method proposed here, the particle filtering stage provides the discretisation of the state space that is required by the second stage. What might come a bit counterintuitive at first is the fact that this discretisation is non-uniform, different for each node, and in fact reflects the sample proposal distributions of the filtering stage. Having a look at the algorithm, it is however clear that the max operation (in contrast to the marginalisation in the sum-product algorithm) is insensitive to the varying density of the sampling, as long as there are sufficient samples in the area of interest.

More formally, the goal is to find a sequence of state variables \( \Theta_{1:T} \) that maximises the global function \( p(\Theta_{1:T}) \), which is factorised into the product

\[
\Theta_{t-1} \quad \psi_t \quad \Theta_{t} \quad \psi_t \quad \Theta_{t+1}
\]

**Figure 5.5:** Graphical model of the Markov chain in which the global optimisation is performed.
Figure 5.6: Final trajectory through the LLE pose space obtained by the global optimisation step. A subsequence of 32 frames, roughly one walking cycle, is shown here. The resulting curve is significantly smoother than the one connecting the particles with the highest weight of each timestep of the online tracking algorithm. The dot clouds indicate the sample distribution at frame 20 and 30 of this subsequence.

of observation functions $v$ that take into account the image information, and compatibility functions $\psi$ of temporally adjacent nodes,

$$p(\Theta_1:T) = \frac{1}{Z} \prod_{t=2}^{T} \psi(\Theta_t, \Theta_{t-1}) \prod_{t=1}^{T} v(\Theta_t), \quad (5.12)$$

where $Z$ is a normalisation constant. The equations from the recursive tracking can be reused, as the global function uses the same terms. The observation functions $v(\Theta_t)$ are computed according to (5.10). In fact we can directly reuse the sample weights computed during tracking. The compatibility between neighbouring nodes is given by (5.9). The max-product respectively min-sum algorithm performs inference in this chain graph by propagating local messages between the neighbouring nodes of the graphical model shown in Fig. 5.5. In Fig. 5.6 a globally optimised trajectory through LLE pose space is shown. It is much smoother than the trajectory that connects the samples with highest weight of each timestep.
5.4 Experiments

5.4.1 Empirical Ambiguity Analysis

One of the difficulties in monocular pose estimation, in particular from silhouettes, are ambiguities. That is, for a given input image sequence, there are additional alternative interpretations, next to the correct solution. These alternative solutions may provide an explanation that explains the image data nearly equally well as the expected solution does, or even better.

In this section, two concrete examples illustrate how tracking may go wrong, and why. These experiments are based on a one-dimensional pose parametrisation of walking motions that consists of the phase within a walking cycle. In the experiments, only the generative appearance modelling is examined, while LLE manifold learning and the dynamics model are factored out. The learning is based on the CMU MocapData set, with semi-automatic phase annotations as in the experiment of Fig. 4.14 in the previous chapter. This dataset is well suited for the analysis, because the search space consists of only two dimensions, phase $\gamma$ and view angle $\omega$, and thus allows for easy visualisation.

Multimodalities of the image likelihood. The tracking is done via particle filtering, i.e. in a Bayesian framework, by combining a temporal prior and an image likelihood. In Fig. 5.7 the image likelihood for the entire state space is visualised, for a sequence of fronto-parallel walking. The expected pose trajectory would thus have constant view angle $\omega = 0$, and a phase $\gamma$ that moves forwards in time (upwards in the figure) with constant speed. In

**Figure 5.7:** Likelihood maps for a sequence of fronto-parallel walking, and a two-dimensional state space. The intersection of the dotted lines marks the expected values (ground truth).
5. Generative Regression-Based Model for Pose Estimation

Figure 5.8: Erroneous tracking: The swinging of the leg is interpreted as a rotation of the whole body, which can be clearly seen at the shoulders. Nevertheless, for most of the individual frames the poses seem reasonable.

the first likelihood map shown, there are two relatively clear modes, one corresponding to the true pose, and another one shifted by half a period, which corresponds to the obvious labelling ambiguity of the two legs. In the two other likelihood maps however, there are circle-like ridges of almost equal likelihood level, suggesting that there are more complex ambiguities that also involve the rotation $\omega$, without any clear modes. When visualising these likelihood maps as an animation, one can observe the expected upwards-movement, as well as a second movement in the opposite direction, i.e. backwards in time. This means that common walking can be explained fairly well by backwards walking as well. This is not surprising, since the main shape deformation of the silhouette involves a scissor-like opening and closing of the legs, and is thus temporally symmetric.

Tracking failure. The learned generative model with the one-dimensional pose parametrisation was also applied to a synthetic sequence, shown in Fig. 5.8. As a temporal prior, Brownian motion was used for the viewing angle and the phase, imposing smooth movements. The chosen parametrisation also guarantees that only ‘valid’ body poses can be expressed. Nevertheless, the estimated poses for this particular experiment do not correspond to a valid walking movement. Instead, the stick figure first moves correctly, then suddenly starts rotating without moving the legs, and finally locks on the second mode caused by the Necker reversal. However, surprisingly, when inspecting
the estimated poses on a frame-by-frame level, all individual poses except for a short subsequence of a few frames, seem valid interpretations of the silhouette. As a human observer, the inadequacy of this result is immediately noticed when playing back the poses as an animation. The erroneous pose estimate thus violates some characteristics of human motion, that were not encoded in the temporal prior. For this simplified example, the problem can be fixed relatively easily, by imposing a temporal prior that encourages the phase $\gamma$ to move forward in time. Hence, the shown failure certainly motivates the use of strong dynamical models, as provided by the described framework. In general however, formulating meaningful priors, in particular for the view angle $\omega$, turns out to be rather difficult, and is to some extent an unsolved issue.

5.4.2 Tracking Results

Training. For the main tracking experiments, the described models were trained on the LocoETH dataset, consisting of Motion Capture pose data and corresponding synthetically rendered silhouettes (see section 3.5). Two separate models were learned, on either walking or running examples. Due to the choice of training data, our system currently assumes that the camera is in an approximately horizontal position, and ignores perspective projection effects of the camera. All the kernel regressors were trained using a multivariate extension of the Relevance Vector Machine algorithm [Tipping, 2000], with Gaussian Kernels. Different kernel widths were tested and compared using a crossvalidation set consisting of 50% of the training data, in order to avoid overfitting. The sparseness priors of the RVM led to approximately 150-250 basis functions (relevant vectors in analogy to support vectors) for the training set of 1089 items.

Tracking The tracking algorithm was evaluated on a number of different sequences. The main goals were to show its ability to generalise to real human shapes, and to deal with noisy sequences with poor foreground segmentation, image sequences of very low resolution, and varying viewpoints through the sequence. The figures in this section show the body poses of the optimal trajectory that was computed according to section 5.3.2, based on the samples from the recursive Bayesian sampling algorithm.

The particle filtering was performed using a set of 500 samples, leading to a computation time of approximately 2-3 seconds per image frame in unoptimised Matlab code. The sample set is initialised in the first frame as follows. Hypotheses for the 2D bounding box locations are either derived from the output of a pedestrian detector that is run on the first image, or from a simple
Figure 5.9: Circular walking sequence. The figure shows full frames (top), and cutouts with estimated bounding boxes in original or segmented input images, as well as stick figures of the estimated body poses. For the visualisation of the 3D stick figures, body limbs that are closer in depth appear darker in the plot.
Figure 5.10: An extract from a soccer game. The figure shows original and segmented images, with estimated bounding boxes, and estimated 3D poses.
Figure 5.11: Circular walking sequence, original resp. segmented input images with estimated bounding boxes, and estimated poses.
Figure 5.12: Diagonal walking sequence. Estimated bounding boxes and poses. The intensity of the stick figure limbs encodes depth; lighter limbs are further away.
Figure 5.13: Estimated view directions, for the circular walking sequence of Fig. 5.9 (top), and the diagonal walking sequence of Fig. 5.12 (bottom).
procedure to find the centre of gravity of the largest connected component in the segmented image. Pose hypotheses $x_0$ are difficult to initialise, even manually, since the LLE parametrisation is not intuitive. Therefore, we randomly sample from the entire space of feasible poses in the reduced LLE space to generate the initial hypotheses. Thanks to the low-dimensional representation, this works well, and the sample set converges to a low number of clusters (typically two) after a few time steps, as desired.

The first experiment (Fig. 5.9) shows tracking on a standard test sequence\(^1\) from [Sidenbladh et al., 2000], where a person walks in a circle. We segmented the images using simple background subtraction, yielding noisy foreground probability maps. The main challenge here is the varying viewing angle that is difficult to estimate from the noisy silhouettes. Also, during the phase where the person is seen from behind, the image information gets less discriminative, and the tracker mainly survives thanks to the dynamical model. Fig. 5.11 shows another publicly available sequence\(^2\). Here we used only one camera, while this sequence has been mainly used for multi-camera tracking (e.g. [Sigal et al., 2004; Sun et al., 2006]). Fig. 5.10 shows an extract from a real soccer game with a running player. The sequence was obtained from www.youtube.com, therefore the resolution is low and the quality suffers from compression artefacts. We obtained a rough foreground segmentation by masking the colour of the grass.

Tracking through another publicly available sequence from the HumanID corpus is shown in Fig. 5.12. The subject walks in an angle of approximately 35° to the camera plane. In addition it is viewed from a slight top-view and shows limb foreshortening due to the perspective projection. These are violations of the assumptions that are inherent to our choice of training data, where we used horizontal views and an orthographic projection. Nevertheless the tracker performs well.

The sequences of Figures 5.14 and 5.15 were recorded in a real traffic environment with a webcam. The image resolution is 320 × 200 pixels, with subjects as small as 40-50 pixels in height. Furthermore, the image quality is unfavourable due to severe MPEG compression artefacts and noisy foreground segmentation that was carried out by subtracting one of the frames at the beginning of the sequence. In Fig. 5.14 the person carries an umbrella that could be misinterpreted as a leg, and a bag that distorts the overall shape of the pedestrian. The subject also turns away from the camera over the duration of the sequence. Our experiments showed that such a challenging sequence, which combines different kinds of difficulties, can only be tracked thanks to the

\(^1\)http://www.nada.kth.se/~hedvig/data.html
\(^2\)http://www.cs.brown.edu/~ls/
Figure 5.14: Real traffic scene with low resolution input images, noisy segmentations, disturbing objects (umbrella, bag), and varying view angle. Original frames (top) and cutouts.

dynamical model, since the information from individual images is unreliable and therefore has to be accumulated over time. The tracker could otherwise be distracted by the noisy segmentations and the multimodal per-frame likelihoods.

5.5 Discussion

The experiments have shown that the proposed approach can deal with challenging input sequences that exhibit noisy foreground segmentations and low resolution. Furthermore, the method generalises from training shapes to real observed human shapes, that may be affected by carried objects, coats etc. The learned strong prior, and in particular the dynamical model, are crucial elements that enable pose estimation from such sequences. Even for humans it is hard to identify people and their poses in the noisy foreground segmented images. Details such as the arm position are often not visible and are thus ‘made up’ by the algorithm, by choosing the most likely arm pose given the legs and the learned prior. It is thus doubtful whether the algorithms (or any cur-
5.5. Discussion

Figure 5.15: Traffic scene with low resolution images and noisy segmentations.

rent algorithms) can be applied to sequences that are even more challenging, in terms of resolution, low contrast and image noise.

The main reasons for failures of the tracking algorithm are excessive noise in the segmented images, especially if the false segmentations are due to occlusions or background objects and thus not randomly distributed. Furthermore, it is very difficult to estimate body poses if the walking direction and view direction of the camera coincide. In such front-views, the image variation that is caused by the body motion is very small, typically much smaller than the image noise, and does thus not allow for successful tracking.

The learning and inference stages are designed in particular with regard to ambiguities and multimodalities that are inherent in the monocular 3D tracking from silhouettes. It has been shown that a learned prior is necessary to infer meaningful results despite the strongly ambiguous appearance model. In practice, the inference algorithm – particle filtering – often does not quite manage to maintain multiple modes of the sample distribution over long sequences. It is a known phenomenon [Doucet et al., 2000a] that all the samples of the set tend to converge to a single cluster, which can lead to the extinction of posterior modes. This selection of one single posterior mode out of the true
posterior is often rather arbitrary, incidental and guided by noise. Here, it would be desirable to add a mechanism that encourages multimodality, e.g. through constant re-initialisation by bottom-up sample proposal functions. The extraction of the globally optimal sequence could then be extended in order to identify and extract multiple alternative trajectories that all satisfy the prior and explain the input sequence.

The framework was implemented using a combination of various machine learning techniques, many of which could be substituted by an equivalent method. The resulting algorithm is thus a rather arbitrary combination of techniques such as LLE, RVM, GMM, PCA that all have their own origins and may or may not be probabilistically motivated. Furthermore, each sub-model is learned separately, even though they may mutually depend on each other. This shortcoming is addressed in chapter 6, where a model is proposed that is conceptually very similar, but uses a unified formulation of the entire model based on Gaussian Process regression.
Articulated Multi-Body Tracking Based on Gaussian Processes

6.1 Introduction

One of the goals of research in articulated tracking is the automatic interpretation of complex multi-person scenarios such as street scenes. This task is however extremely challenging, and many factors contribute to this difficulty. Next to the general problems of articulated tracking, such a system for instance has to deal with mutual occlusions of pedestrians, non-static cluttered backgrounds, low resolution and unexpected human shapes due to bags, coats etc. Towards solving this challenging task, in this chapter a system is presented which combines several state-of-the art methods into an integrated pipeline.

Articulated tracking algorithms that support multiple hypotheses can in principle deal with multiple persons. However, most approaches do not explicitly distinguish between competing pose hypotheses for a single person in the image and different persons that are simultaneously visible. In e.g. a particle framework, clusters of particles belonging to different subjects would compete against each other, with the risk that one of the modes (one of the subjects) eventually dies out. Also, relations between different subjects, such as temporary occlusion, cannot be modelled that way. A straightforward extension of a probabilistic inference algorithm to multiple subjects with occlusion reasoning requires a joint representation for the state space of multiple subjects [Hue et al., 2002; MacCormick and Blake, 2000], leading to an exponential increase in computational complexity for Monte-Carlo methods. Explicitly modelling multiple targets and their interactions on the level of articulated tracking is thus currently intractable.

Instead, we propose an approach to overcome those difficulties in a system’s context. The idea is to carry out the global occlusion and multi-object reasoning on a coarser level and to only perform a more detailed articulated analysis
Figure 6.1: An example for the challenging articulated multi-person tracking scenarios considered in this chapter. The proposed system addresses the difficulties of this task by first applying a robust multi-body tracker to handle the data association problem and identify individual tracks. An articulated tracker is then applied to each single-person track independently to infer precise body poses, which are in turn fed back to improve the observation model. As can be seen from the results, this procedure allows robust performance despite the presence of multiple people, temporary occlusions, scale changes, and camera motion.

on the output trajectories of the higher-level multi-body tracker (see Fig. 6.1). This allows us to also impart the articulated tracker with important information from trajectory analysis, such as a person’s 3D walking direction, speed, and the knowledge when a trajectory is occluded. However, even a sophisticated multi-body tracker cannot solve the entire problem. Data association remains a challenging task: especially when multiple persons are walking close to each other, their limbs are often hard to distinguish. We address this issue by providing the articulated trackers with a guided segmentation that incorporates top-down knowledge from human detection. Together with a dynamic shape prediction from tracking, this observation model provides sufficiently precise measurements to support articulated multi-body tracking in very challenging street scenes.

Regarding the articulated tracking stage, we conceptually follow the generative method of the previous chapter, with the following modifications. The approach presented in chapter 5 is based on a combination of machine learning techniques that were used to model a subset of the model components each. In this chapter, while sticking to the same graphical structure
of the learned statistical model, the entire learning process is carried out within the framework of Gaussian Process regression (GP). A unified formulation of the whole learning pipeline is presented, that includes models of the relation of pose and appearance as well as of the temporal evolution, and a low-dimensional representation that is derived within the proposed model. Due to recent extensions that allow for sparse and thus efficient evaluation and training of the regressors [Snelson and Ghahramani, 2006; Lawrence, 2007], Gaussian Process regressors have become an alternative to previous regression models such as ridge regression, RVM and Support Vector Regression (SVR), which is more than equivalent, with practical advantages next to the theoretical ones.

6.2 Modelling with Gaussian Processes

The statistical modelling of body motion and appearance is similar to the generative framework described in chapter 5. The main difference here is the use of Gaussian Process regression [Rasmussen and Williams, 2006] instead of RVM and GMM. The tracking algorithm operates in a low-dimensional representation $x$ of the body poses that is learned from training data. We then model the reconstruction of the original representation of the articulations, the prediction of the corresponding human shape in image space, and the temporal evolution (dynamics) of the body poses over time using Gaussian Process regression. This model is illustrated in Fig. 6.2.

6.2.1 Gaussian Process Regression

Gaussian Processes (GPs) define probability distributions over functions. When conditioned on example data they can be used to model the regression between two variables, say an independent variable $p$ and a dependent variable $q$ (see [Rasmussen and Williams, 2006]). GPs are Bayesian models and thus model the probabilistic uncertainty of regression analysis from noisy data. In our case the reconstruction from the low-dimensional pose space $x$ to the original articulation representation $s$, and the prediction of appearances $y$ from body configurations $x$ will be modelled.

Given a covariance function $k(p_i, p_j)$ and a set of training pairs, a posterior pdf over expected regressor outputs $q^*$ can be computed for any input point $p^*$. Training a GP regression model consists in finding good parameters $\beta$ of the covariance function (model selection). An elegant way to do so is to maximise the marginal likelihood of matrix $Q$, containing the dependent training examples, with respect to the covariance parameters $\beta$ (explained shortly).
6.2.2 Pose, Appearance and Dynamics

Pose. According to the generative modelling framework of chapter 5 (see Fig. 5.2), we model the mappings from the low-dimensional pose representation $x$
6.2. Modelling with Gaussian Processes

Figure 6.3: The prediction of the shape $y$ depends on the low-dimensional body pose variable $x$ and the orientation $\omega$, while the body articulation $s$ is only modelled as a function of $x$.

to the pose and appearance spaces $s$ and $y$. The pose reconstruction mapping is learned by substituting the datasets $X$ and $S$ in (6.1), and then optimising the marginal likelihood. The mapping is based on the covariance matrix $K_{rec}$ with covariance parameters $\beta_{rec}$.

Dynamics. In addition, the model is able to temporally predict future body poses according to a transition model $p(x_{t+1}|x_t)$. Similarly to (6.1), the marginal likelihood $P(X|\beta^{dyn})$ is derived for the regression from $x_t$ to $x_{t+1}$ (see [Wang et al., 2006]), and optimised with respect to the parameters of the dynamics covariance function $\beta^{dyn}$.

$$P(X|\beta^{dyn}) = p(x_1) (2\pi)^{-\frac{d_x(N-1)}{2}} |K_*|^{-\frac{d_x}{2}} exp\left(-\frac{1}{2} tr\left(K_*^{-1}X_*X_*^T\right)\right). \quad (6.3)$$

Here, $K_* \in \mathcal{R}^{N-1 \times N-1}$ is the kernel matrix constructed from the inputs to the dynamical mapping, that is $\{x_1, \ldots, x_{N-1}\}$ if the training data consists of a single continuous sequence of body poses. Similarly, $X_* = [x_2, \ldots, x_N]^T$ are the outputs of the mapping. The prior $p(x_1)$ is assumed to be isotropic Gaussian.

Appearance. In contrast to the pose reconstruction, the shape prediction additionally depends on the orientation $\omega$ of the subject with respect to the observing camera (see Fig. 6.3). In our training data, every body pose has a number of corresponding silhouettes, each viewed from a different angle. This results in $NM$ training examples, where $N$ is the number of poses and $M$ the number of viewing directions in our training database. For the regression model, we thus have to optimise the marginal likelihood $P(Y|\Omega, X, \beta^{app})$, where $\Omega$ contains the viewing angles of the training shapes. Using a straightforward implementation, the complexity of the GP training algorithm scales with $(NM)^3$, since it involves the inversion of the covariance matrix $K_{app} \in \mathcal{R}^{NM \times NM}$. This is impractical for the large datasets we use. We thus propose
a covariance function that allows the covariance matrix to be written as a Kro-
necker tensor product, reducing the complexity to $O(N^3 + M^3)$ instead of the
original $O((NM)^3)$. This can be done by defining the appearance covariance
function as a product of a pose covariance function $k_{\text{pose}}(x_i, x_j)$ (e.g. squared
exponential) and an orientation covariance function $k_{\text{ori}}(\omega_i, \omega_j)$,

$$k_{\text{app}}(x_i, \omega_i; x_j, \omega_j) = k_{\text{pose}}(x_i, x_j)k_{\text{ori}}(\omega_i, \omega_j).$$ (6.4)

For every pose $x \in X = \{x_1, \ldots, x_N\}$ there are silhouettes for all possible
viewing directions $\omega \in \Omega = \{\omega_1, \ldots, \omega_M\}$, the appearance covariance matrix
can be written as:

$$K_{\text{app}} = K_{\text{pose}} \otimes K_{\text{ori}}.$$ (6.5)

Complexity can be further reduced by replacing the orientation covariance
function with a delta function $k_{\text{ori}}(\omega_i, \omega_j) = \delta_{\omega_i, \omega_j}$. This makes sense dur-
ing training of the GP regression, because the training samples only involve a
number of discrete viewing directions $\omega \in \Omega$. Once the regression parameters
have been learned with this additional approximation, the orientation covari-
ance function can then be replaced by one with a larger support (e.g. a Von
Mises distribution [Evans et al., 1993]), in order to allow for interpolations
between the discrete viewing directions $\omega \in \Omega$.

### 6.2.3 Learning the Embedding

The marginal likelihood of the entire learned model can now be written as the
product of the marginal likelihoods

$$P(S, Y, X|\Omega, \beta) = P(S|X, \beta^{\text{rec}})P(Y|\Omega, X, \beta^{\text{app}})P(X|\beta^{\text{dyn}}),$$ (6.6)

where $\beta$ is a vector of all covariance parameters. Rather than just optimising
the regressors (that is, the parameters $\beta$), (6.6) can be optimised with respect
to the latent positions $X$ as well. In such a way, the algorithm learns a shared
latent space of body pose and appearance, that simultaneously maps into pose
and appearance space, and additionally enforces smooth manifolds by requiring
$X$ to be well modelled by the dynamics mapping. A similar approach has
been followed in [Shon et al., 2006; Navaratnam et al., 2007; Ek et al., 2008].
This model amounts to a multi-set extension of the Gaussian Process Latent
Variable Model [Lawrence, 2005] with separate covariance functions for each of
the mappings. In such a way, all the components of the model can be learned
simultaneously rather than sequentially, and the globally optimal combination
of embedding and regressors can be found. On the downside, when additionally
optimising the marginal likelihood with respect to the latent variables $X$, the
number of unknowns in the optimisation problem raises from less than ten to
6.3 Integrated Multi-Person Tracking System

The focus of this thesis is on modelling body articulations. Here, the integration of these models into a multi-person tracking system is shown. Except for the articulated tracking (which is similar to the tracking algorithm of chapter 5), the description of the system modules is brief and kept on a high level of abstraction, concentrating on the interfaces between the components in a system context.

6.3.1 System Overview

Fig. 6.4 shows the schematical layout of the multi-body articulated tracking system. A small-baseline (40cm) calibrated stereo rig mounted on a mobile platform captures two image streams and passes them on to a human detection module. Based on the obtained bounding boxes and rough stereo depth information, a multi-body tracker (section 6.3.2) finds consistent object trajectories in 3D. Each trajectory is then passed to a single-person articulated tracker (section 6.3.3), which estimates the person’s 3D articulation based on a

Figure 6.4: Overview of the multibody articulated tracking system.

several thousand. The optimisation thus becomes harder, and due to local minima a good initialisation is important.

In the performed experiments, LLE body pose coordinates served as an initialisation for the latent variable $X$. However, using the optimised latent variables in tracking experiments did not noticeably improve the results, when compared to directly using the LLE coordinates as latent variables. The reported tracking results thus do not include an optimisation with respect to the latent variables in the learning stage.
learned statistical representation. The estimation is made robust by a guided segmentation stage (section 6.3.4) that combines the pedestrian detector’s top-down segmentation with bottom-up image cues and a shape prediction inferred from the current state of the articulated tracker. This results, for every frame of the sequence, in one body pose estimate per tracked person, located in 3D world coordinates.

While in the proposed system, stereo-based depth computation contributes to finding the silhouettes of the subject (see section 6.3.4), the accuracy of the depth information is limited by the small baseline between the cameras and does not allow for further disambiguation of the pose estimates (as would be possible in a true multi-camera setup [Carranza et al., 2003; Sigal et al., 2004; Ren et al., 2005a]). The articulated pose estimation algorithm thus relies on image descriptors that are computed from the subject’s silhouette. We do however take into account both image streams of the binocular sequences, which helps to alleviate problems that are caused by image noise; i.e. when one camera stream is temporarily corrupted by noise or occlusion, the algorithm can base its pose estimation on the second camera.

6.3.2 Multi-Body Tracker

In order to reliably handle the complex interactions between multiple objects, we first address the task of tracking multiple pedestrians without taking into account the articulations, effectively factorising the state space into independent “tracklets” for each visible pedestrian. We adopt a tracking-by-detection approach similar to [Leibe et al., 2007], but extended to incorporate stereo depth in order to make it robust enough for mobile applications. As input, the multi-person tracker takes two video streams recorded with our small-baseline stereo rig. A global world coordinate frame and groundplane are recovered using structure-from-motion and stereo depth. Pedestrians are detected at each time-step with an ISM detector [Leibe et al., 2005] and are placed in this global frame. Based on the space-time detections, an overcomplete set of trajectories is tracked with independent Kalman filters. The best subset of this pool of hypotheses is selected through a global optimisation procedure, which enforces physical exclusion constraints, resulting in a consistent scene interpretation. The tracker is able to automatically initialise new tracks (usually, after about 5 detections) and to recover temporarily lost trajectories, thus enabling the system to track through occlusions.

The output of this tracking module is a trajectory for each pedestrian in 3D world coordinates (including the person’s 3D orientation, velocity, and bounding box), as well as the information when the person was occluded.
As the articulated tracker is currently only trained on walking people, objects below and above a certain speed threshold are discarded. The unoccluded parts of each remaining trajectory (the “tracklets”) can be processed independently by the subsequent articulated tracking module, which would otherwise become intractable. We want to point out, however, that data association between those tracklets still remains a challenging problem, as the limbs of adjacent persons may easily get confused. Section 6.3.4 therefore introduces a guided segmentation, which combines top-down information from the human detector with bottom-up image cues and which considerably improves the observation process.

6.3.3 Articulated Tracking

The articulated tracking algorithm operates on the output trajectories of the multi-body pedestrian tracker of section 6.3.2, which delivers 2D image locations, scales and orientations of the tracked persons. Its observations are automatically estimated pedestrian silhouettes, obtained through the guided segmentation procedure of section 6.3.4. Many ambiguities and temporary occlusions are already resolved and accounted for by the previous stage. A particle filter serves as an overall framework, where at time $t$ the body pose hypotheses $x^i_t$ are propagated in the low-dimensional pose space according to the learned dynamical model. For each particle $x^i_t$, a shape $y^i_t$ can be predicted by taking into account the 3D track orientation $\omega_t$, estimated by the multi-body tracker. The particles are then weighted with their image likelihoods, obtained by comparing the predicted shape to the actually observed shape $y^{obs}_t$:

$$w^i \propto p(y^{obs}_t | \omega_t, x^i_t) = \mathcal{N}(y^{obs}_t; \mu^i_t, \Sigma^i_t),$$

(6.7)

where $\mu^i_t$ and $\Sigma^i_t$ are the mean and covariance matrix of the predicted shape.

Finally, once the particle filter has been run on all images of a tracklet, a Viterbi algorithm extracts a smooth and consistent trajectory through the particle set (this can in practice be approximated with a fixed temporal look-ahead). Again, the transition costs between neighbouring states are based on the learned dynamical model. In order to account for variations in the framerate of the sequence and the walking speed of the subjects, this step additionally chooses between different scaling factors of the predicted velocities, i.e. accelerated and slowed-down variants of the dynamical model.
6.3.4 Guided Adaptive Segmentation

As an interface between the multi-body tracker and the articulated tracker, we are using a set of automatically estimated figure-ground segmentations for each tracked person. In the majority of previous works [Elgammal and Lee, 2004a; Navaratnam et al., 2007; Zhao and Nevatia, 2004], silhouettes are assumed to be available, and are in practice often obtained using background modelling. Since we are dealing with a moving camera setup, we cannot use this option. Instead, we propose to obtain the segmentations by fusing top-down cues (from the detector and the articulated tracker) with bottom-up image information (from colour and stereo depth). Keeping in line with previous work by several authors [Boykov and Lea, 2006; Cremers et al., 2007], the segmentation is formulated as an energy function that is minimised with respect to the foreground/background labelling $C = \{c_0, \ldots\}$ of all pixels.

$$E(C) = \sum_i R(c_i) + \lambda \sum_{i,j \in \mathcal{N}} B(c_i, c_j) \tag{6.8}$$

In the above equation, $R(c_i)$ denotes the region term for a pixel with index $i$, which has a label $c_i$ (figure/ground). $R(c_i)$ is based on the top-down segmentation map $f$ of the detector and the shape prediction map $\pi = \sum_j w^j_t \mu^j_t$ of the articulated tracker, where $w^j_t$ is the weight of sample $j$ and $\mu^j_t$ is its predicted shape from (6.7).

$$R(c_i) = -\log(P_{\pi}(c_i) \ P_f(c_i)) \tag{6.9}$$

Here, $P_{\pi}$ and $P_f$ are the probabilities of a certain label given the segmentation maps $\pi$ and $f$ from the articulated tracker and from the detector respectively:

$$P_{\pi}(c_i) = \begin{cases} \pi_i & \text{if } c_i = 1 \\ 1 - \pi_i & \text{if } c_i = 0 \end{cases} \tag{6.10}$$

The boundary term $B(c_i, c_j)$ encodes the belief that region boundaries typically coincide with intensity and depth discontinuities. It is defined on the 4-neighbourhood $\mathcal{N}$ and penalises neighbouring pixels with different labels but similar colours $\vec{I}_i$ and depths $\mathcal{D}_i$. Due to the delta function $\delta_{c_i \neq c_j}$ this term vanishes when the neighbouring pixels are assigned the same label $c_i = c_j$.

$$B(c_i, c_j) = e^{-\frac{|\vec{I}_i - \vec{I}_j|^2}{2\sigma^2_i}} e^{-\frac{|\mathcal{D}_i - \mathcal{D}_j|^2}{2\sigma^2_d}} \delta_{c_i \neq c_j} \tag{6.11}$$

The resulting cost function can be minimised efficiently using standard graph-cut methods [Boykov and Lea, 2006], yielding the binary foreground mask $y^obs_i$. Together with the bounding box position and motion direction from
6.4 Experiments

Training.

For training the articulated tracker, walking data from the LocoETH dataset was used. A three-dimensional LLE of the body pose data
Figure 6.6: Articulated multi-person tracking results for test sequence #1. The last row shows a 3D visualisation of the estimated world state in the three images of the second row.
Figure 6.7: Articulated tracking results for test sequence #2. Note the robust articulation estimates of tracked pedestrians under scale changes and egomotion.

serves as the low-dimensional pose space for the GP regression model. The marginal likelihood of the regressors was optimised with scaled conjugate gradients using the ‘FITC’ sparse approximation with 200 inducing variables [Snelson and Ghahramani, 2006; Lawrence, 2007].

Tracking. We demonstrate our approach on 3 challenging video sequences showing real-world inner-city scenes. These videos were captured at about 13–14 fps from a mobile recording platform. Such a low framerate complicates the articulation reasoning considerably. Table 6.1 gives an overview over the sequences shown in this chapter.

The first sequence (Fig. 6.6) was taken on a busy sidewalk. Even though the camera itself is standing still, traditional background subtraction would be difficult due to small camera shake, as well as passing trams and cars. While most people move sideways, they still occur at different depths and often have a slightly tilted trajectory, which we can account for by tracking
directly in 3D. In the sequence’s 454 frames, our system tracks 20 out of 23 people successfully with the multi-body tracker. One of the missed pedestrians runs too fast, and another one is at all times occluded by other persons. In addition to the 20 correct tracks, the system yields two additional tracks that contain errors due to wrongly estimated orientation or scale. Counting each person individually, this amounts to a total track length of 932 frames, where a detection is available in 86% of the cases. We visually inspected all the resulting segmentations and found that 55% of these are well-defined (meaning the entire person is covered) in at least one camera. For the individual cameras, only 41% were well-defined. This underlines the usefulness of a stereo system in such real-world scenarios with frequent occlusions. While these numbers might seem low, we would like to note that the articulated tracker can also operate if only parts of the body are segmented correctly (most importantly, the legs). Based on these segmentations, the system tracks 74 walking cycles, 54 out of which were entirely correct. The remaining 20 cases mainly occur at the end of longer trajectories and are mostly due to multiple, consecutive bad segmentations or occlusions. Note, however, that the silhouettes generally did not contain enough information to unambiguously recover the arm positions, which additionally differed from the training examples due to the fact that many people were carrying shopping bags or similar accessories. Example pose estimates of our system are shown in Fig. 6.6.

For the remaining sequences, qualitative results are shown in Fig. 6.7 and 6.8. As the multi-body tracker takes care of the mapping between the world coordinate frame and the local articulated trackers, the system can be applied to scenes captured under significant egomotion. Fig. 6.7 shows an example of such a case, where people enter the scene from several directions and undergo large scale changes. As the multi-body tracker provides orientation and scale estimates, we can still obtain acceptable results on such data. A challenging case is shown in Fig. 6.8. Here, the system has to cope with more extreme scale changes and people moving in many different directions, while following one person through a busy pedestrian zone. In particular movement parallel to the camera’s viewing direction is highly ambiguous, but the correct walking cycle is still identified in most cases.

6.5 Discussion

The proposed system achieves good results in challenging real-world scenarios by factorising the problem into separate tasks of multi-body tracking under occlusion and articulated body pose estimation for individual trajectories. As was shown, this formulation allows the articulated tracker to benefit
from trajectory-level information about the tracked person’s speed and walking direction, which considerably simplifies inference and renders the problem tractable. Further, a way was shown to integrate an articulated tracker based on Gaussian Processes into a context-aware system that gathers and accumulates different kinds of scene information. This tracking framework can be applied under egomotion with the help of a guided top-down/bottom-up segmentation module. Experimental results confirm the viability of the proposed approach.

Currently, due to the choice of training data, the system is limited to well defined human motions and cannot yet recover arbitrary body poses. Next to increasing the amount of training motions, the results of our estimation could be used to learn specialised colour models for different body parts, which then support more general pose recovery [Ramanan et al., 2005]. As a possible extension of the system, additional feedback mechanisms would allow for a further integration of the modules. In such a way, feedback from the articulated tracker and its dynamical model could help detection and data-association.

<table>
<thead>
<tr>
<th>Seq.</th>
<th># Frames</th>
<th>Pedestrians</th>
<th>Found by MBT</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>454</td>
<td>23</td>
<td>20</td>
</tr>
<tr>
<td>#2</td>
<td>173</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>#3</td>
<td>242</td>
<td>21</td>
<td>17</td>
</tr>
</tbody>
</table>

Table 6.1: Sequences used for evaluating the proposed system. We report the number of (walking) persons, and the ones actually found by the multi-body tracker (MBT).
Figure 6.8: Articulated tracking results for test sequence #3. This sequence shows a very challenging scenario with considerable egomotion and many pedestrians entering the visible scene at various distances and from different directions.
7

Inferring Activity Categories

The tasks of body pose estimation and activity recognition are strongly related. On one hand, a sequence of inferred body poses might be used for activity recognition, which is the classical pipeline for gait recognition. On the other hand, a known activity class can also help the pose estimation, e.g. by selecting an appropriate context specific prior. The method proposed here estimates the 3D body pose and action categories simultaneously. It is integrated into the modelling and tracking framework of chapter 5, but could equally well be applied in conjunction with the other tracking approaches of this thesis. We learn strong dimensionality-reduced models of feasible body poses that belong to a certain activity or motion pattern, as well as the temporal evolution of the body poses over time. Then, the transition functions between different activities are learned from training data too. Typical human motion patterns such as walking and running are investigated experimentally.

7.1 Modelling and Learning

Modelling and inferring activity categories is introduced as an extension of the generative tracking approach of chapter 5. Rather than learning a unified model that contains all motions in the training data set, separate activity specific models are learned that allow for an explicit recognition of the performed activity along with the pose estimation, using a switching mechanism of the inference algorithm. The approach will be validated using walking and running as example activities. It can however be extended to more and different categories. For notational convenience, this section will refer to the considered action classes, without loss of generality.

According to section 5.2.2, separate low-dimensional embeddings are learned for the different activity classes, using Locally Linear Embedding (LLE). The resulting LLE spaces do not have any direct relationship, an can in principle even be of different dimensionality. For each activity-specific representation,
the reconstruction of the original pose representation is then approximated using nonlinear regression. For the activities considered in the following experiments, walking and running, the reconstruction mappings are denoted $f_p^w(x^w)$ and $f_p^r(x^r)$ respectively. The superscript refers to the activity category and will be omitted if the same formulation holds for all categories.

In these low-dimensional LLE spaces, the dynamics are learned separately for the different action categories, leading to the temporal priors $p^w(x_t|x_{t-1})$ and $p^r(x_t|x_{t-1})$. Similarly, the generative mappings from pose space to shape space ($f_w^a$ and $f_r^a$) are learned.

We also want to model the transition between the considered action categories, that each have their own low dimensional pose parametrisation expressed in distinct LLE spaces. Informally, we want to find walking poses that are very similar to a given running pose and vice versa, since we know that the transition is performed smoothly, without any sudden or jerky ‘jump’ of the body configuration. Given our distinct training sets of walking and running poses, two sets of training pairs are generated by looking for the most similar running pose for every walking pose and vice versa, where the similarity measure is the Euclidean distance in the original pose representation. The nonlinear mapping between these pairs is modelled using two sparse kernel regressors $f_{sw}^r\rightarrow w(x^r)$ and $f_{sw}^w\rightarrow r(x^w)$. This can be generalised to more action categories, however, the number of transitions grows quadratically with the number of categories, which should therefore be kept low. The following motion model is obtained, where the state space is augmented by a discrete state variable $a_t$.

$$p(x_t, a_t|x_{t-1}, a_{t-1}) \propto \begin{cases} p_{\sim sw} p^{a_t}(x_t|x_{t-1}) & \text{if } a_t = a_{t-1} \\ p_{sw} p^{a_t\rightarrow a_{t-1}}(x_t^{a_t}|x_{t-1}^{a_{t-1}}) & \text{else} \end{cases}$$  (7.1)

Here, the motion model for the case of activity switching $p^{a_{t-1}\rightarrow a_t}(x_t^{a_t}|x_{t-1}^{a_{t-1}})$ is modelled as a normal distribution where the mean is the pose predicted by the regressor $f_{sw}^{a_{t-1}\rightarrow a_t}$. The probabilities that an activity transition does or does not occur are denoted $p_{sw}$ and $p_{\sim sw}$. In the case of more than two activity categories, these transition probabilities can be represented as a transition matrix with the $p_a^{a_{sw}}$ of the different categories $a$ on the diagonal.

### 7.2 Tracking with Activity Recognition

The sample-based tracking from section 5.3.1 is extended in order to take into account multiple activity classes. To this end, the state parametrisation is extended by adding the discrete activity label $a_t$.

$$\Theta_t = [a_t, \omega_t, u_t, v_t, w_t, h_t, x_t]$$  (7.2)
The sample proposal function is adapted according to eq. (7.1). A sample \( i \) undergoes an activity switch with probability \( p_{sw} \). In our experiments, the scheme is demonstrated for two activity categories, therefore we set \( p_{sw}^{w\rightarrow r} = p_{sw}^{r\rightarrow w} = 1 - p_{\sim sw} \). In case of an activity switch, the sample \( i \) is initialised with a value in LLE pose space of the new activity \( a_t \) by sampling from the activity transition function \( p^{a_{t-1} \rightarrow a_t}(x_t^{a_t} | x_{t-1}^{a_{t-1}}) \). In such a manner, at each time step a number of samples are generated that allow for a smooth transition into the other activity. If these hypotheses are supported by the image information, they will be selected in the subsequent resampling step and take the upper hand. The percentage of samples of a certain activity category is a measure for the algorithm’s belief about the currently observed action. The image support for the hypotheses is given by the observation likelihood, which is always based on the action specific appearance model (i.e. using the mappings \( f_a^w \) or \( f_a^r \) resp.). Unlike previous particle filters with model switching like [Isard and Blake, 1998b; Heap and Hogg, 1998], the algorithm thus switches between dynamical models, appearance models, and even between state parametrisations.

### 7.3 Experiments

Two activity-specific models were trained on the LocoETH dataset, which was split into two parts containing only walking or only running poses. The main difference to the training stage of section 5.4.2 were the additional cross-activity transition models, that link the 4-dimensional LLE spaces of the walking and running models. Then, the particle filter tracking algorithm with the additional switching mechanism was applied to video sequences of humans that switch between walking and running motions.

The particle set was initialised with 50% walking and 50% running samples. The sample transition probability \( p_{sw} \) was empirically chosen. At time \( t \) the fraction of samples that encode a specific activity are considered as the current belief of the algorithm that the subjects performs that activity. Since at each timestep, it is likely that a number of samples undergo an activity switch, the algorithm’s confidence for a certain activity rarely reaches 100%, by design. Finally, a discrete activity label is inferred by the global optimisation algorithm (see section 5.3.2).

Fig. 7.1 shows an extract from a treadmill sequence that was 1660 frames long in total. In this sequence, the subject initially walks and switches to running and back to walking several times. The figure shows a few frames from the transition from running to walking; the first two frames clearly contain running poses, then the arms are lowered and the last 3 frames show walking. The plot
Figure 7.1: Transition from running to walking. The original sequence is 1660 frames long, this figure shows selected frames from the transition phase between frame 921 and frame 936. See also Fig. 7.2 for a plot of the estimated activity categories.

in Fig. 7.2 shows the estimated running probabilities throughout the sequence. Even for humans, it is not obvious to identify the exact moment of activity change, there is typically a transition phase of about 0.5 seconds. In this experiment, the activity switch was always detected within this transition phase, as desired. Note that we do not take into account the typical periodic motion in vertical direction that distinguishes running from walking. The activity is correctly estimated from the local shape and its deformation over time alone.

A challenging realistic traffic scene is shown in Fig. 7.3, with low resolution and noisy input. The pedestrian suddenly starts to run when crossing a street.
7.4 Discussion

The presented approach jointly tracks in a space that encodes the activity class as well as body pose and 2D image location. Coupling pose estimation and activity recognition is an attractive approach with advantages for both subtasks. If one is only interested in activity recognition, a purely appearance-based approach might be more suitable, since the pose estimation is a difficult problem on its own, and coupling the two tasks also means that the failure of one task can imply that the other task fails as well. Experimentally, for the activity classes of walking and running, we have observed a recognition rate that is close to perfect, provided that the tracking itself works well. The failure modes are thus the same as for the single-activity tracking.

Additional experiments with more subtle activity classes are subject of current research. One of the expected challenges is the limited capability of particle
7. Inferring Activity Categories

Figure 7.3: Real traffic scene with a transition from walking to running. Full frames (top) and cutouts with estimated poses. Fig. 7.4 shows the inferred activity categories of this sequence.

filters to maintain multiple posterior modes over many frames, i.e. entire subsequences. This might pose problems for input sequences that cannot be unambiguously assigned to a single activity class. In such cases one of the modes might take the upper hand, whereas the other(s) die out.

While the activity transitions are in general accurately detected, the applied transition model is currently very simple. As there are no activity transitions in the training corpus, the transition itself is not learned. Instead, the transition behaviour is modelled by incorporating the obvious assumption of smooth motion across the activity change, as shown in section 7.1. The results show that the algorithm is able to reliably detect an activity switch and to temporally locate it precisely. Furthermore, the tracked body motion shows a smooth
transition from one activity into the other and looks natural. As a possible extension of the system, the actual transition phase could be modelled more accurately by learning from training data as well, including additional body postures that are neither walking nor running poses but occur only during the transition phase.

![Activity Category Transition](image)

**Figure 7.4:** Estimated activity category for the traffic sequence of Fig. 7.3
Conclusion

After the general theoretical focus of chapter 2, this thesis concentrates on the visual analysis of human motion - tracking, pose estimation and action recognition. In this chapter, the different proposed approaches to body pose estimation are summarised and compared. Finally, we give perspectives for pursuing the presented work.

8.1 Summary and Comparison of Frameworks

Table 8.1 gives an overview of some key properties of the different approaches. We will denote the three implementations as the RBPF-, RVM- or GP-method respectively, in the order they appear in chapters 4 to 6, for additional comparative comments.

The RBPF-method is based on linear tools. Linear PCA reduces the dimensionality of both pose and shape descriptors. Non-linearities and multimodalities of the relation of pose and shape are accounted for by learning the joint distribution with a mixture of Gaussian distributions. In mixture models, setting the number of mixture components is a critical parameter, trading off the ability to model complex relationships against computational efficiency and robustness to overfitting. The RVM- and GP-methods are based on nonlinear models of regression and dimensionality reduction.

While modelling the joint pdf is a generative approach, pose estimation is done discriminatively in the RBPF-method by conditioning on the observed shape in closed form. In contrast to a purely discriminative learned model, our approach also models the observation, which is necessary for tracking the 2D location of the subject. When simultaneously estimating 2D location and body pose, in order to be able to represent the non-parametric posteriors and to maintain the efficiency of closed form pose estimation, a hybrid inference scheme is applied, using a Rao-Blackwellised particle filter.
<table>
<thead>
<tr>
<th></th>
<th><strong>RBPF</strong></th>
<th><strong>Generative (RVM)</strong></th>
<th><strong>Generative (GP)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>statistical modelling</strong></td>
<td>joint pdf (mixture model)</td>
<td>generative appearance model (regression)</td>
<td>generative appearance model (regression)</td>
</tr>
<tr>
<td><strong>dim. reduction / manifold</strong></td>
<td>linear (PCA)</td>
<td>nonlinear pose manifold (LLE)</td>
<td>manifold, LLE initialisation shared latent space</td>
</tr>
<tr>
<td><strong>dynamical model</strong></td>
<td>Brownian motion</td>
<td>learned velocity field</td>
<td>velocity field GPDM</td>
</tr>
<tr>
<td><strong>pose inference</strong></td>
<td>analytical</td>
<td>sample-based</td>
<td>sample based</td>
</tr>
<tr>
<td><strong>2D location estimation</strong></td>
<td>sample based</td>
<td>sample-based</td>
<td>sample based</td>
</tr>
<tr>
<td><strong>pose initialisation</strong></td>
<td>automatic bottom-up initialisation</td>
<td>random initialisation</td>
<td>random initialisation</td>
</tr>
<tr>
<td><strong>computation of shape descriptor</strong></td>
<td>bottom-up</td>
<td>bottom-up/top-down</td>
<td>bottom-up/top-down</td>
</tr>
<tr>
<td><strong>multimodal posteriors</strong></td>
<td>hybrid, parametric/sample based</td>
<td>sample based</td>
<td>sample based</td>
</tr>
<tr>
<td><strong>global optimisation</strong></td>
<td>-</td>
<td>max-product / Viterbi</td>
<td>Viterbi</td>
</tr>
</tbody>
</table>

**Table 8.1:** Overview over the different approaches to statistical pose estimation presented in this thesis.
The proposed generative *RVM*- and *GP*-approaches model the mapping from pose to shape, i.e. the conditional probability distribution over shape descriptors given pose hypotheses. Pose tracking relies on sample-based inference. The predictive appearance model directly fits into the particle filtering framework as the observation likelihood, while the learned predictive dynamics serve as the sample proposal function. Efficient particle filtering requires a pose representation of even lower dimensionality than the one obtained by linear PCA in the *RBPF*-method. In the *RVM*-pipeline, Locally Linear Embedding is applied to obtain a low-dimensional representation of body pose. In the *GP*-formulation the low-dimensional representation acts as an abstract latent space that is shared between the related spaces of body pose and appearance.

The learned dynamical model of the *RVM*- and *GP*-approaches adds a lot to the performance of the tracker, as the experiments have shown. While the particle filter framework can deal with the nonlinear dynamical model, the analytical integration of the *RBPF*-method in eq. (4.13) relies on a linear dynamical model, which limits the strength of the prior.

The initialisation of the particle set is always an issue in sampling based approaches. A pedestrian detector can be used to create hypotheses of human occurrences in images. For the body pose, thanks to the relatively low dimensionality of the representation, random initialisation works well for the *RVM*- and *GP*-approaches. For the *RBPF*-method, pose initialisation is not an issue, since the pose estimation process is essentially a discriminative one, and the entire pose space is thus explored at each timestep. Hence, the distributions over poses are constantly reinitialised, which also implicates that the intrinsic multimodality of pose estimation is well represented in the posteriors. In the particle filtering of the generative approaches, in contrast, it can often be observed that only one of the posterior modes survives, while other valid interpretations vanish and are thus missed.

While a potentially multimodal representation of the posterior is desirable to reflect the nature of the problem at hand, it is often necessary to eventually make a hard decision, i.e. determine a single body pose per frame as a final result. Given the purely sample-based posterior representation of the *RVM*- and *GP*-approaches, the extraction is easily achieved in a postprocessing global optimisation. This leads to smooth motion trajectories that are consistent throughout a tracked sequence. Jitter and shaky results can be efficiently reduced in that way, when compared to visualising e.g. the sample with the highest weight at each time step, as done in chapter 4 for the *RBPF*-method. As it turns out, the hybrid sample-based/parametric representation is not very convenient from a practical perspective, for visualisation and reporting results. However, it can be valuable for further processing steps, such as any
higher-level recognition that builds on top of the articulated tracking, where the multimodality or remaining uncertainty may influence the outcome.

Experimentally, the generative methods \((RVM)\) and \((GP)\) perform similarly well, and outperform the \(RBPF\)-method, when applied to challenging sequences, for several reasons. First, the lack of a strong dynamical model poses problems for sequences with unreliable segmentations and/or a low framerate. Second, an offline global optimisation step, ensuring smoothness and consistency, is not present in the \(RBPF\)-framework. Third, Gaussian Mixture Models in high-dimensional spaces are cumbersome to work with, because the probability mass drops very quickly with increasing distance from the mean of its components. Fourth, the bottom-up computation of the signed-distance shape descriptors is more sensitive to noise than top-down shape prediction and likelihood computation in image space.

8.2 Perspectives

As discussed in the previous chapters, discriminative and generative statistical approaches to 3D articulated tracking both have advantages and disadvantages. A promising line of possible future research addresses possible combinations of these two methodologies, while at the same time being theoretically well motivated and practical. Possible approaches include importance sampling methods, and inspirations for combined top-down and bottom-up processing can be found in [Rosales and Sclaroff, 2001; Curio and Giese, 2005; Thayananthan et al., 2006; Sminchisescu et al., 2006].

The intrinsic ambiguities and multimodalities of the pose estimation task were mentioned several times in this thesis. The design of the modelling and inference stages accounts for this fact, inter alia by allowing for multimodal posteriors. In practice however, sample-based recursive Bayesian filtering often fails to consistently maintain multiple posterior modes with equal importance. Concerning this matter, the performance of the sample based approaches \((RVM)\) and \((GP)\) could potentially be improved by making adequate modifications to the inference algorithm, e.g. by including a bottom up estimation mechanism through periodical reinitialisation or importance sampling. In addition, the offline global optimisation of the \(RVM\)- and \(GP\)-approaches currently extracts a single consistent trajectory. Here, an extension could explicitly identify multiple modes (e.g. as clusters of samples), and accordingly extract multiple trajectories that correspond to different possible interpretations of the image sequence. The \(RBPF\)-method would equally benefit from the estimation of one or multiple globally consistent trajectories, which in this case is complicated by the hybrid sample-based/parametric posterior representation.
For learned appearance models, the design of shape or appearance descriptors is crucial. In the presented learning algorithms and in most related work by other authors, these representations are ‘hand-crafted’ or learned in simple unsupervised procedures using PCA or clustering. Including feature extraction and cue selection in the overall learning algorithm, which is usually formulated as a supervised learning problem, seems a promising strategy. In such a way, those image features that provide maximum information about the body pose can automatically be identified and extracted from a large pool of image cues, yielding a compact, efficient and powerful representation for the appearance of humans in images.

Other directions of interesting future work, and partially ongoing research, include more extensive qualitative and quantitative experimental evaluation, the application of the proposed methods to larger training data sets with more general body motions, and a larger number of more subtle activity categories than the ones used for the human locomotion experiments.
In this chapter the equivalence of the Gaussian section method of section 2.3.2 and Gaussian Conditioning is shown. We restate the equation (2.5) for the intersection of a Gaussian model with an affine subspace $M$:

$$
\mu_M = \left( M_u^T \Sigma^{-1} M_u \right)^{-1} M_u^T \Sigma^{-1} (\mu - m_0) \\
\Sigma_M = \left( M_u^T \Sigma^{-1} M_u \right)^{-1}
$$ (A.1)

Without lack of generality, we choose a partitioning of the vector $x$ into a first part $x_1$ that is observed, and an unknown second part $x_2$. $M_u$ is a matrix containing the canonical basis vectors corresponding to the unknown vector elements, and $m_0 = [x_1^T \ 0]^T$ contains zeros for the unknown elements, and the measured values for the known vector entries. Note for generality that any other basis and parametrisation for the subspace $M$ could be chosen, the described parametrisation is however the most convenient here, since it directly yields a posterior pdf over the unknown elements. The same notation is used for the Gaussian Conditioning equations (2.7), repeated here:

$$
\mu_{2|1} = \mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (x_1 - \mu_1) \\
\Sigma_{2|1} = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}
$$ (A.2)

It has to be shown that $\mu_{2|1} = \mu_M$ and $\Sigma_{2|1} = \Sigma_M$. This can be done using the matrix inversion lemma (e.g. [Petersen and Pedersen, 2008]):

$$
\Sigma^{-1} = \left[ \begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{array} \right]^{-1} = \left[ \begin{array}{cc} C_{11} & C_{12} \\ C_{21} & C_{22} \end{array} \right] \\
= \left[ \begin{array}{cc} (\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21})^{-1} & -\Sigma_{11}^{-1} \Sigma_{12} (\Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12})^{-1} \\ -(\Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}) \Sigma_{21} \Sigma_{11}^{-1} & (\Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12})^{-1} \end{array} \right]
$$ (A.3)

1 For an arbitrary parametrisation, the projection back to the original observation space (2.6) followed by a marginalisation over the known elements is necessary.
Due to the choice of the basis $M_u$, the expression $(M_u^T \Sigma^{-1} M_u)$ in (A.1) selects those columns and rows from the inverse covariance matrix that correspond to the unknown vector entries. It thus directly follows from (A.3) that

$$\Sigma_M = C_{22}^{-1} = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} = \Sigma_{2|1} \tag{A.4}$$

Similarly, for the mean $\mu_M$, $M_u^T \Sigma^{-1}$ selects the lower part of the inverse covariance matrix.

$$\mu_M = \Sigma_M M_u^T \Sigma^{-1} (\mu - m_0)$$

$$= C_{22}^{-1} \begin{bmatrix} C_{21} & C_{22} \end{bmatrix} \begin{bmatrix} \mu_1 - x_1 \\ \mu_2 - \mu \end{bmatrix}$$

$$= \begin{bmatrix} C_{22}^{-1} C_{21} & I \end{bmatrix} \begin{bmatrix} \mu_1 - x_1 \\ \mu_2 \end{bmatrix}$$

$$= \mu_2 + C_{22}^{-1} C_{21} (\mu_1 - x_1)$$

$$= \mu_2 - (\Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12})^{-1}(\Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}) \Sigma_{21} \Sigma_{11}^{-1} (\mu_1 - x_1)$$

$$= \mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (x_1 - \mu_1)$$

$$= \mu_{2|1} \tag{A.5}$$
Bibliography


[Barrow et al., 1977] H. Barrow, J. Tenenbaum, R. Boles, and H. Wolf. Parametric correspondence and chamfer matching: Two new techniques for im-


Proceedings of the IEEE Workshop on Motion and Video Computing (WACV/MOTION), 2005. 6, 17, 30, 44


the Ninth International Workshop on Artificial Intelligence and Statistics, 2003. 38, 46


# Curriculum vitae

**Name:** Tobias Jaeggli  
**Date of Birth:** July 26, 1976  
**Place of birth:** Aarau (AG), Switzerland  
**Citizenship:** Swiss

## Education

<table>
<thead>
<tr>
<th>Year</th>
<th>Institution</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004-2008</td>
<td>ETH Zurich (CH), Computer Vision Lab, PhD Student.</td>
<td></td>
</tr>
<tr>
<td>1998-2002</td>
<td>Université de Fribourg (CH), Studies of Computer Science with a minor in Media and Communication Sciences. Graduation with the degree <em>Master of Science</em>.</td>
<td></td>
</tr>
<tr>
<td>1997-1998</td>
<td>SAE (School of Audio Engineering), Zurich (CH), Graduation with the degree <em>Audio Engineering Diploma</em>.</td>
<td></td>
</tr>
</tbody>
</table>

## Experience

<table>
<thead>
<tr>
<th>Year</th>
<th>Institution</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004-2008</td>
<td>ETH Zurich (CH), Computer Vision Lab, Research and teaching assistant.</td>
<td></td>
</tr>
<tr>
<td>2003-2004</td>
<td>Katholieke Universiteit Leuven (B), Computer Vision Lab, Research assistant.</td>
<td></td>
</tr>
<tr>
<td>2000-2002</td>
<td>Université de Fribourg (CH), Teaching assistant.</td>
<td></td>
</tr>
</tbody>
</table>