

Chapter 1

Introduction

As computers are becoming more widespread in the society, so is the need for more advanced interfaces. This has led to a new research field known as Human-Computer-Interaction (HCI). One problem with current interfaces is that the communication is mainly done via devices non-intuitive to humans, e.g., mouse and keyboard. The communication is on the terms of the computer rather than natural human terms such as speech and body-language.

Since it is hard for humans to learn the "language" of computers the solution must be to develop computers that are able to communicate on terms of humans. Among other things, this means that computers should be equipped with devices providing the ability to interact in a human-like manner. These devices could be: a camera to enable it to see, a microphone to enable it to hear, and a loudspeaker to enable it to speak. Other devices might be introduced in the future but the three modalities mentioned are the primary ones. The technologies providing the speak- and hear abilities are currently leaving the labs around the world and entering the market. The see-ability, however, is still lacking due to the high complexity involved in the image interpretation task. This has led to a new research field known as "Looking at People". It covers topics such as face detection and recognition, small-scale body language understanding, such as hand-gesture recognition and facial expression recognition, and large-scale body language understanding, such as action recognition. The initial problem of this thesis is focused around the general problem of large-scale body language understanding and formulated as

how can computer vision be applied to understand large-scale human body language?

One way of structuring an investigation into this problem is to divide the problem of understanding large-scale human body language into two separate parts: a low-level process to capture the motion of the human and a high-level process to interpreting the captured motion. A large body of work exists within the field of general pattern recognition, which can be applied to support the interpretation problem. For the motion capture (MoCap) problem, however, basic research is still required.

1.1 Model-Based Approaches

Many different approaches have been suggested for computer vision-based human MoCap. The more successful approaches apply a priori knowledge, usually in the form of a geometrical model, i.e., applying a model-based approach. Different configurations of the model is synthesised and compared with the image data. The configuration most similar to the current image data defines the current state of the model, i.e., its pose. This principle is known as Analysis-by-Synthesis (AbS). Temporal AbS provides a sequence of poses, which define the captured motion.

Each degree of freedom in the geometrical model is represented by one variable. The set of variables in the model spans a coordinate system - *the state-space* - wherein one point corresponds to a particular configuration of the model. The major problem related to model-based approaches is the potentially high dimensionality of the state-space which contains too many configurations to evaluate for each image, hence exhaustive search is not possible. To overcome this problem, only a part of the state-space is investigated, i.e., a local search. The standard approach is to assume that a high sampling rate is present, hence the configuration does not change much between two consecutive images. The current configuration is therefore assumed to be in the proximity of the configuration in the previous image. This is known as temporally-based pruning of the state-space and is carried out by predicting the current configuration and searching (exhaustively or iteratively) in a local region around the predicted configuration until an extremum is found. Unfortunately, an investigation into the state-of-the-art computer vision-based MoCap systems shows that the approach of temporally-based pruning has some inherent problems: 1) no prediction can be provided in the first image, hence an initialisation of the configuration is required, 2) if tracking is lost the prediction will be incorrect, hence continuous tracking is assumed, and 3) the wrong state might be estimated if the solution space is multi-modal, hence a uni-modal solution space is assumed.

1.2 The Focus of the Thesis

The inherent problems related to applying temporally-based pruning calls for alternatives. This directly leads to the final problem formulation for this thesis, namely **how can the dependency on temporally-based pruning be circumvented in model-based computer vision systems in general, and in model-based computer vision-based human MoCap in particular?**

To answer this question this thesis suggests to apply spatially-based pruning as an alternative or complement to temporally-based pruning. To this end, we will investigate two different issues both of which are rooted in how spatial information can be applied to prune the state-space. These are the spatial relation of image features and the intrinsic and extrinsic spatial characteristics of the object to be

motion captured. Both are formulated as hypotheses.

Hypothesis 1 The inclusion of low-level image features allows for a more compact state-space representation, hence a pruning of the original state-space.

Hypothesis 2 The inclusion of intrinsic and extrinsic spatial object characteristics allows for a definition of constraints, which can significantly prune the state-space.

To investigate the two hypotheses a concrete MoCap problem is addressed. As the primary information conveyed in body language is done by the arms, the MoCap of a human arm seems to form a relevant case study. Furthermore, a monocular sensing approach is of special interest in MoCap due to its generality. Altogether, the focus of this thesis can therefore be stated as

Monocular computer vision-based MoCap of the human arm utilising spatially-based pruning of the state-space.

1.3 The Outline of this Thesis

The thesis consists of nine chapters. The first chapter is the current chapter which contains the introduction. The following seven chapters form the main body of the thesis, while chapter nine contains the conclusion. The following seven chapters are divided into three parts:

- I. **Human Motion Capture**
- II. **Spatially-Based Pruning of the State-Space**
- III. **Applying the Spatially Pruned State-Space Representation in a Model-Based Framework**

In the first part a general introduction to Human MoCap is given. In the second part the two hypotheses are investigated, and in the third part the pruned state-space is tested in different systems. The seven chapters are partly texts written directly for this thesis and partly texts previously published. Each chapter is initiated by a synopsis briefly stating the purpose of the chapter and, if already published, explains the context of the publication. Each chapter is followed by its own bibliography. In the following each chapter is summarised.

Part I. Human Motion Capture

Chapter 2. Interacting with a Virtual World through Motion Capture

To give a broader introduction to the topic of the thesis, this chapter describes, defines, and discusses human MoCap in general, and relates it to HCI.

First different MoCap devices are described. The devices are divided into those based on active sensing and those based on passive sensing. Furthermore, the complexity of the devices is related to the required post-processing of the devices' output in order to achieve usable MoCap data. Next the usage of MoCap data in HCI applications is described with respect to whether synchronous or asynchronous interacting is carried out. During the entire chapter a number of examples are given in order to emphasise different aspects.

Chapter 3. A Survey of Computer Vision-Based Human Motion Capture

In this chapter, in-depth knowledge about computer vision-based MoCap approaches is presented through a comprehensive survey. The survey outlines the different problems in the field and hereby motivates the approach taken in this thesis.

More than 200 different papers are included in the survey. To structure all this information, a detailed taxonomy is developed. Its primary categories are: initialisation, tracking, pose estimation, and recognition. Each of the categories is divided into sub-categories in order to identify the general aspects of the individual MoCap papers, and hence, allowing for a comparison between different papers. Three general application areas are identified and related to the taxonomy and the state-of-the-art within the three application areas is presented and discussed. Finally the general problems in this field are identified and summarised.

Part II. Spatially-Based Pruning of the State-Space

Chapter 4. Deriving a Compact State-Space Representation by Applying Low-Level Image Features

In this chapter the first hypothesis is investigated. Concretely it is described how to obtain a more compact state-space representation by including low-level image features.

The chapter first identifies the primary degrees of freedom (DoF) in the arm and shoulder. Four DoF for the arm and two DoF for the shoulder. Then a number of different state-space representations of the arm is presented and evaluated with respect to their size. Concretely, it is described how a camera calibration and the

position of the hand in the image can be combined with the screw axis representation to obtain a very compact state-space representation, denoted the *local screw axis model*. It can model all possible configurations of the arm utilising just two parameters, (α, H_z) . In the light of the local screw axis model, the two DoF in the shoulder are revisited. Through a study of the anatomy and the kinematics of the shoulder, a modelling scheme is proposed that couples the two DoF in the shoulder directly to (α, H_z) . That is, the two parameters (α, H_z) in the local screw axis model are sufficient to model all six DoF in the arm and shoulder.

Chapter 5. Pruning the State-Space Representation using Extrinsic and Intrinsic Object Characteristics

In this chapter the second hypothesis is investigated. Concretely, it is described how to apply kinematic- and physical constraints to further prune the compact state-space representation derived in the previous chapter, i.e., (α, H_z) .

Six constraints are introduced: four to prune H_z and two to prune α . For each of the constraints a minimum, a maximum, and an average pruning effect is calculated. Hereafter the effects of the six constraints are combined and the overall minimum, maximum, and average pruning effects of the state-space (α, H_z) are calculated to 87.9%, 100%, and 97.3%, respectively.

Part III. Applying the Spatially Pruned State-Space Representation in a Model-Based Framework

Chapter 6. Estimating the 3D Shoulder Position using Monocular Vision and a Detailed Shoulder Model

The aspects of the model developed in chapter 4 that deals with the shoulder is tested in this chapter. The test is in the context of estimating the 3D position of the human shoulder given a sequence of 2D hand positions. A user is asked to move his/her outstretched arm in a circle-arch. If the position of the shoulder was fixed, this movement would result in a circle in 3D and an ellipse in the image which could be used to calculate the 3D shoulder position. However, due to the complex nature of the shoulder this is not the case and the circle is rather an ellipse in 3D. The model developed in chapter 4 is therefore applied to correct the image data, resulting in an ellipse in the image plane and, hence, a 3D circle in space.

Chapter 7. 3D Human Pose Estimation using 2D-Data and an Alternative Phase Space Representation

In order to test the concept of spatially-based pruning of the state-space developed in chapter five, a system is build. This chapter describes the design, implementation, and test of this system. The system uses the compact state-space and spatially-based pruning to reduce the number of possible configurations for a given image. All non-pruned configurations are synthesised and compared with the image data, i.e., an exhaustive search. In this system silhouettes are applied for comparison. That is, the silhouette of the human arm is extracted in each image and compared with the silhouette of each synthesised configuration. The similarity measure used when comparing the synthesised and extracted silhouettes is implemented as a combination of an AND-operation and a bounding box-matching.

Chapter 8. Improving Sequential Monte Carlo Tracking by Bootstrapping

In this chapter the spatially-based pruning derived earlier is used to bootstrap the temporally-based pruning. The spatially and temporally-based pruning are not alternatives, but rather cooperating approaches. Besides this, the chapter also shows how to use an *approximated* exhaustive search in the pruned state-space, as opposed to an exhaustive or an iterative. The approximation is achieved by applying a Sequential Monte Carlo (SMC) approach. The theory behind this approach and a concrete implementation is described in this chapter. Applying SMC in the pruned state-space is presented as bootstrapping the SMC approach. The bootstrapping refers to the fact that the compact state-space allows for focusing the search due to the spatially-based pruning. In this implementation the similarity measure between image and model is based on comparing orientations of the arm, found by applying the dynamic Hough transform to temporal edge pixels.

Chapter 9. Conclusion

This chapter concludes the work presented in this thesis. It includes a discussion of whether the two hypotheses have been proven and an identification of the primary contributions of this Ph.D.-work. Finally, different directions for further research are suggested.

1.4 The Publications of this Ph.D.-work

Below are listed the relevant publications produced during this Ph.D.-study. After each publication the amount of work conducted by me is stated in percentage. In all essence the contributions reported in the publications are included in this thesis and only the thesis has thus been submitted as documentation for the Ph.D.-study. However, the highlighted entries number 2, 4, 6, 10, and 13 appear directly as chapters 8, 3, 2, 6, and 7 in this thesis.

Journal papers

1. T.B. Moeslund, C.B. Madsen, and E. Granum. Modelling the 3D Pose of a Human Arm and the Shoulder Complex utilising only Two Parameters. Submitted to *International Journal on Integrated Computer-Aided Engineering* (95%)
2. **T.B. Moeslund and E. Granum. Improving Sequential Monte Carlo Tracking by Bootstrapping. Submitted to *Journal on Applied Signal Processing. Special Issue on Particle Filtering in Signal Processing* (95%)**
3. T.B. Moeslund and E. Granum. Modelling and estimating the pose of a human arm. *Machine Vision and Applications*, 14(4), 2003 (95%)
4. **T.B. Moeslund and E. Granum. A Survey of Computer Vision-Based Human Motion Capture. *Computer Vision and Image Understanding*, 81(3), 2001 (90%)**

Book chapters and papers in Lecture Notes

5. T.B. Moeslund, M. Störring, and E. Granum. A Natural Interface to a Virtual Environment through Computer Vision-estimated Pointing Gestures. In Ipke Wachsmuth and Timo Sowa, editors, *Gesture and Sign Language in Human-Computer Interaction*, number 2298 in LNAI. Springer, 2001 (50%)
6. **T.B. Moeslund. Interacting with a Virtual World through Motion Capture. In Lars Qvortrup, editor, *Interaction in Virtual Inhabited 3D Worlds*, chapter 11. Springer-Verlag, 2000 (100%)**

Peer reviewed conference papers

7. T.B. Moeslund and E. Granum. Sequential Monte Carlo Tracking of Body Parameters in a Sub-Space. In *International Workshop on Analysis and Modeling of Faces and Gestures*, Nice, France, October 2003 (95%)

8. T.B. Moeslund and E. Granum. Bootstrapping Sequential Monte Carlo Tracking. In *Scandinavian Conference on Image Analysis*, Göteborg, Sweden, June 2003 (95%)
9. T.B. Moeslund, C.B. Madsen, and E. Granum. Modelling the 3D Pose of a Human Arm and the Shoulder Complex utilising only Two Parameters. In *Conference on Model-based Imaging, Rendering, image Analysis and Graphical special Effects*, INRIA Rocquencourt, France, 10-11 March 2003 (95%)
10. **T.B. Moeslund, M. Vittrup, K.S. Pedersen, M.K. Laursen, M.K.D. Sørensen, H. Uhrenfeldt, and E. Granum. Estimating the 3D Shoulder Position using Monocular Vision and a Detailed Shoulder Model. In *International Conference on Imaging Science, Systems, and Technology*, Las Vegas, USA, June 24-27 2002 (50%)**
11. T.B. Moeslund and E. Granum. Pose Estimation of a Human Arm using Kinematic Constraints. In *The 12th Scandinavian Conference on Image Analysis*, Bergen, Norway, 2001 (95%)
12. M. Störring, E. Granum, and T.B. Moeslund. A Natural Interface to a Virtual Environment through Computer Vision-estimated Pointing Gestures. In *4th International Workshop on Gesture and Sign Language based Human-Computer Interaction*, City University, London, UK., April 2001 (50%)
13. **T.B. Moeslund and E. Granum. 3D Human Pose Estimation using 2D-Data and an Alternative Phase Space Representation. In *Workshop on Human Modeling, Analysis and Synthesis at CVPR*, Hilton Head Island, South Carolina, June 16 2000 (95%)**
14. T.B. Moeslund and E. Granum. Multiple Cues used in Model-Based Human Motion Capture. In *The fourth International Conference on Automatic Face and Gesture Recognition*, Grenoble, France, March 2000 (75%)

Non-reviewed conference papers

15. T.B. Moeslund and E. Granum. Bootstrapping the Condensation Algorithm. In *The 11th Danish Conference on Pattern Recognition and Image Analysis*, Copenhagen, Denmark, August 2002 (95%)
16. T.B. Moeslund, M. Störring, and E. Granum. Vision-Based User Interface for Interacting with a Virtual Environment. In *The 9th Danish Conference on Pattern Recognition and Image Analysis*, Aalborg, Denmark, 2000 (50%)
17. T.B. Moeslund and E. Granum. Visual Motion Capture as a means of Control in Telepresence. In *The 9th Danish Conference on Pattern Recognition and Image Analysis*, Aalborg, Denmark, 2000 (95%)

18. T.B. Moeslund. The Analysis-by-Synthesis Approach in Human Motion Capture: A Review. In *The 8th Danish Conference on Pattern Recognition and Image Analysis*, Copenhagen, Denmark, 1999 (100%)

Technical reports

19. T.B. Moeslund. Improving Sequential Monte Carlo Tracking by Bootstrapping. Technical Report CVMT 02-02, Laboratory of Computer Vision and Media Technology, Aalborg University, Denmark, 2002 (100%)
20. T.B. Moeslund. Modelling the Human Arm. Technical Report CVMT 02-01, Laboratory of Computer Vision and Media Technology, Aalborg University, Denmark, 2002 (100%)
21. T.B. Moeslund. Pruning the Possible Configurations of a Human Arm using Kinematic Constraints. Technical Report CVMT 01-01, Laboratory of Computer Vision and Media Technology, Aalborg University, Denmark, 2001 (100%)
22. T.B. Moeslund. Estimating the Configuration of a Human Arm using Computer Vision. Technical Report CVMT 00-02, Laboratory of Computer Vision and Media Technology, Aalborg University, Denmark, 2000 (100%)
23. T.B. Moeslund. Computer Vision-Based Human Motion Capture - A Survey. Technical Report LIA 99-02, Laboratory of Image Analysis, Aalborg University, Denmark, 1999 (100%)
24. T.B. Moeslund. Summaries of 107 Computer Vision-Based Human Motion Capture Papers. Technical Report LIA 99-01, Laboratory of Image Analysis, Aalborg University, Denmark, 1999 (100%)

