

Chapter 9

Conclusion

The initial problem inspiring the research of this thesis concerned the question of how computer vision can be applied to understand large-scale human body language. To respond to this question we divided the notion of understanding large-scale human body language into two separate problems, namely that of capturing the human motion (MoCap) and that of interpreting the captured motion. The former was chosen to be the focus on this thesis. We therefore conducted an investigation (chapter two) into general MoCap techniques. From here a need for a marker-free computer vision-based human MoCap solution became clear. In chapter three a survey was conducted to obtain more knowledge of the current state of the art within computer vision-based human MoCap techniques. From the survey a number of important problems were identified. Among these were the problems related to the use of temporally-based pruning of the state-space in model-based approaches. These problems inspired the formulation of the specific problem to be addressed in 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 we suggested to investigate the use of spatially-based pruning. Concretely we formulated the following two hypotheses to structure our efforts:

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 definition of constraints, which can significantly prune the state-space.

In chapter four the first hypothesis was investigated and a compact state-space model based on the measured position of the hand in each individual image, and a screw

axis representation was suggested. This compact state-space was denoted the *local screw axis model*. In chapter five the second hypothesis was investigated by applying kinematic and physical constraints to prune the local screw axis model. In chapter six those aspects of the local screw axis model related to the shoulder modelling were evaluated. In chapter seven the spatially-based pruning developed in chapters four and five was tested. In chapter eight the spatially-based pruning was combined with the standard temporally-based pruning. Furthermore, an approximated exhaustive search was tested in the context of the pruned local screw axis model.

9.1 Contributions

The primary contributions of this thesis are considered to be the following (in order of appearance)

The survey on computer vision-based human MoCap. The most comprehensive so far both in terms of the number of included papers and in terms of the time period it covers. Furthermore, the survey contains a number of taxonomies and tables which allow other researchers to quickly get an overview of different papers. The survey is complemented by a technical report [4] containing summaries of more than 100 of the papers in the survey.

The inherent problems related to temporally-based pruning. Identifying the fact that temporally-based pruning may have fatal limitations and that an alternative/supplementary approach is possible.

Spatially-based pruning based on image features. Showing how low-level image features allow for a more compact state-space representation of the human arm. Introduction of the local screw axis model.

Model of the shoulder complex. Modelling the DoF of the shoulder complex by a ball-and-socket joint together with two prismatic joints.

Relating the shoulder DoF with the DoF in the arm. Linking the displacements of the GH-joint to the parameters of the arm. That is, no additional parameters are required to model the shoulder complex.

Spatially-based pruning based on intrinsic object characteristics. By investigating the structure of an articulated object the already compact state-space can be pruned even more.

Bootstrapped tracking. The bootstrapping approach combines the spatial and temporal pruning by using spatially-based pruning to bootstrap the temporally-based pruning.

Local maximum a posteriori (MAP) estimate. Suggesting a local method for estimating the MAP (or more correctly the most likely a posteriori (MOLAP)) which is superior to global methods.

9.2 Discussion

9.2.1 The hypotheses

In chapter 4 we showed that a more compact state-space can be defined by including measurements from the individual images. This corresponds to making a new instantiation of the model for each image. Our local screw axis model reduces both the dimensionality and the number of possible configurations compared to the standard state-space representation consisting of four Euler's angles. Hence the first hypothesis is considered to be validated in chapter 4.

The second hypothesis states that the intrinsic spatial object characteristics can significantly prune the state-space. From the literature it is known that kinematic constraints can be used to prune the state-space independently on the representation. Therefore the question is not whether or not we have been able to prune the state-space, but rather if the pruning achieved is significant or not.

We found that the state-space of the local screw axis model can be pruned 97.3% in average which we consider significant. However, evaluating the significance with respect to the standard Euler's angles representation might provide better insight.

If we apply the static constraints from table 5.1 to the Euler's angles we can calculate the total number of non-pruned¹ configurations as $\Delta\theta_1 \cdot \Delta\theta_2 \cdot \Delta\theta_3 \cdot \Delta\theta_4 = 180 \cdot 235 \cdot 145 \cdot 135 = 8.3 \cdot 10^8$. In the case of the pruned screw axis model we showed in chapter 5 that 1253 different non-pruned configurations exist on average. In result, the Euler's angles representation contains $6.6 \cdot 10^5$ times more non-pruned configurations than our model does. Clearly a significant pruning.

Even though most human MoCap systems merely apply the static kinematic constraints to prune their solution space, we also compare our pruning results to that obtained by applying our temporal angular constraints, derived in section 5.5.1, to the Euler's angles representation. That is, we found the average angular displacements of each Euler angle by inserting each static interval from table 5.1 into equation 5.35, and multiplying. This yields, $\Delta\theta_1^{avg} \cdot \Delta\theta_2^{avg} \cdot \Delta\theta_3^{avg} \cdot \Delta\theta_4^{avg} = 28.5 \cdot 29.6 \cdot 27.3 \cdot 26.8 = 6.17 \cdot 10^5$. In result, the Euler's angles representation still contains 492 times more non-pruned configurations than our model does. This is also considered to be a significant pruning.

¹All pruning examples are calculated for an angular resolution of 1° and a spatial resolution of 1 cm .

9.2.2 Generality of the Approach

In this work we have validated our hypotheses on a somewhat simple object, the human arm. An interesting question is of course if our findings scale to more complex objects. We believe very much that the first hypothesis can be applied to other, both rigid and articulated, objects. We think of our approach as a way of combining the two normally distinct low-level and high-level approaches. Many systems exist that seek to recover the structure of objects from low-level features such as edges or corners. Many high-level model-based approaches exist which use the AbS-approach. In many applications a bootstrapping approach can be applied to combine the two approaches, hence a generalisation of our approach. In the case of the entire human body low-level image features such as the hands, head, feet, shoulders, armpits, and crotch, may be detected prior to applying the AbS-approach. For example, if the human body is modelled by six DoF for the torso, four DoF for each arm and leg, and zero DoF for both the hands, feet, and head, then 22 DoF is present. If the head, feet, and hands can be detected, then around twelve DoF would be sufficient to describe a compact state-space, i.e. a large reduction of the dimensionality.

The second hypothesis cannot be generalised for non-articulated objects. But in the case of articulated objects the six constraints can more or less be applied directly, as many articulated objects can be divided into an open-looped kinematic chain modelled by screw axes. In cases of closed-looped kinematic chains, care must be taken, but still we believe that many of the constraints can be reused.

9.2.3 Uncertainties

We assume the image features are deterministic, where in fact they are not. This results in uncertainties both with respect to the compact state-space model and with respect to the pruning. The latter has already been discussed in chapter 5. Only the former will therefore be discussed here.

When calculating the 3D location of the hand (and shoulder) an amount of uncertainty will be present. The uncertainty transforms the α -circle into a torus where each point inside the torus is an elbow candidate. An investigation into this showed that not much can be gained using this, more correct, pose candidate selection. Also, the uncertainty due to the variation of the clothes is larger than the extra precision gained by using a torus instead of a circle. Therefore it was decided to use the less complex model of the elbow candidates.

9.2.4 Probabilistic Pruning

To reduce the size of the state-space further, other types of pruning could be applied. We considered including pruning mechanisms such as specific motion types and the probabilities of different arm configurations. For example, if we know that the

user is currently involved in explaining something, he is likely to have his arms in front of his body, while a human running is likely to move his arms in certain patterns in plans perpendicular to the torso. Furthermore, the overall likelihood of the arm being in different configurations could be estimated and included to weight the likelihood of the different configurations. We have not included these types of pruning for two reasons. Firstly, a huge amount of experiments would have to be conducted to obtain sufficient statistics and in the end we would probably end up with a system that only allows certain types of movements, hence a lack of generality. Secondly, due to the statistical nature of training data we would prune with respect to plausible configurations as opposed to possible configuration. This would to some extent bring us back to the problems associated with temporally-based pruning, as the pruning would be depending on, e.g., the recognition and prediction of certain motion patterns and the quality of the trained motion patterns. In more dedicated MoCap systems, however, probabilistic pruning should always be considered, as it can often increase the pruning effect.

9.3 Further Research Topics

A number of topics in this thesis could deserve more attention. The primary ones are briefly described below.

9.3.1 Static Torso

In chapters seven and eight we introduced the assumption that the pose of the torso is known initially and fixed during the entire test sequences. This allowed us to evaluate our approach. For usage of our arm model in the context of pose estimation of the entire (upper) body, this assumption is acceptable as the torso will be estimated by other means. Using the arm model independently, however, requires either the inclusion of additional parameters in the model for the torso, or estimating its pose from images (or a combination). Both of which could be based on the detection of the position of the head and/or shoulder in the image. Perhaps also the size of the head could be used.

9.3.2 The Shoulder Model

Even though our results suggest that modelling the shoulder complex via five DoF is sufficient, more tests are required to draw a final conclusion. This could be done by comparing our model with clinical data.

When modelling the shoulder, the displacement in the z-direction (towards the camera) is not included due to its relatively small influence. This means that our model can be slightly improved by including this displacement as well.

9.3.3 Dependency between the Joint Angles

In table 5.1 the different intervals for each of the four joint angles are listed. In all aspects of pruning we have assumed that these intervals were fixed. However, this is not true and in fact the interval of one joint angle depends on the current values of the other joint angles. If we could incorporate this information into our constraints we could increase the pruning effect.

Describing the dependencies between the four joint angle is not trivial and incorporating the dependencies into the constraints is even more difficult, if at all possible. Comparing this to the expected benefit resulted in the assumption of independent joint angle limits.

Nevertheless, it could be interesting to see if it is possible to incorporate the dependencies and if so, how big the effect will be. Inspiration to such a study can perhaps be found in the literature dealing with the problem of modelling the dependencies between the four joint angles, see e.g., [1, 2, 3, 5].

Furthermore, the dependencies between the limits on the velocity and acceleration might also depend on the current values of the joint angles. Further work in this direction might also produce interesting results. The main problem, besides those mentioned above, is how to quantify the dependencies. When reviewing papers in the area of kinematics we have not encountered a single reference describing this topic.

9.3.4 Pruning H_z using Temporal Constraints

It turned out to be very difficult to analytically describe the pruning of the position of the hand utilising a temporal constraint. Assumptions were therefore introduced. These were defined in such a way that the true analytical pruning always will be better, hence if an analytical solution can be found it will provide better pruning results. The main problem with the current approach is that we ignore the fact that H_z is located on a camera ray that passes through the hand.

9.3.5 Stochastic Features in Bootstrapping

When the bootstrapping approach is applied in situations with many DoF, many image features will be required to obtain a compact state-space. In this case it is not realistic to assume that each feature is detected with the same accuracy. The bootstrapping approach therefore needs to be expanded in order to associate each feature with an uncertainty. That is, the bootstrapping should allow for stochastic features rather than deterministic features.

References

- [1] A.E. Engin and S.T. Tümer. Three-Dimensional Kinematic Modelling of the Human Shoulder Complex - Part 1: Physical Model and Determination of Joint Sinus Cones. *Journal of Biomechanical Engineering*, 111, 1989.
- [2] L. Herda, R. Urtasun, and P. Fua. An Automatic Method for Determining Quaternion Field Boundaries for Ball-and-Socket Joint Limits. In *International Conference on Automatic Face and Gesture Recognition*, Washington D.C., USA, May 20-21 2002.
- [3] W. Maurel. *3D Modeling of the Human Upper Limb including the Biomechanics of Joints, Muscles and Soft Tissues*. PhD thesis, Laboratoire d'Infographie - Ecole Polytechnique Federale de Lausanne, 1998.
- [4] 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.
- [5] D. Tolani, N. Badler, and J. Gallier. A Kinematic Model of the Human Arm Using Triangular Bezier Spline Surfaces. Submitted to *Journal on Graphical Models*.

