Vision Denmark DTU Project: STYLE

“Investigation of the Practical Implementation Possibilities of Motion Style Transfer Algorithms”

*This project has been funded by Vision Denmark. The results presented in this report have been produced in collaboration with DTU and Marionette.

Introduction

The goal of this project was to develop a method for processing monocular motion capture footage and transforming it into high-quality animation. Our objective is to streamline the passage from raw mocap data to usable animation, allowing for faster production of animations that match the desired style with minimal intervention from the end user. Our research into the animation industry revealed that the most important objective is to obtain clean motion capture data that can be easily edited and stylized by animators. This is the main technical challenge in motion capture performed on a monocular video feed, such as the one produced by Marionette.

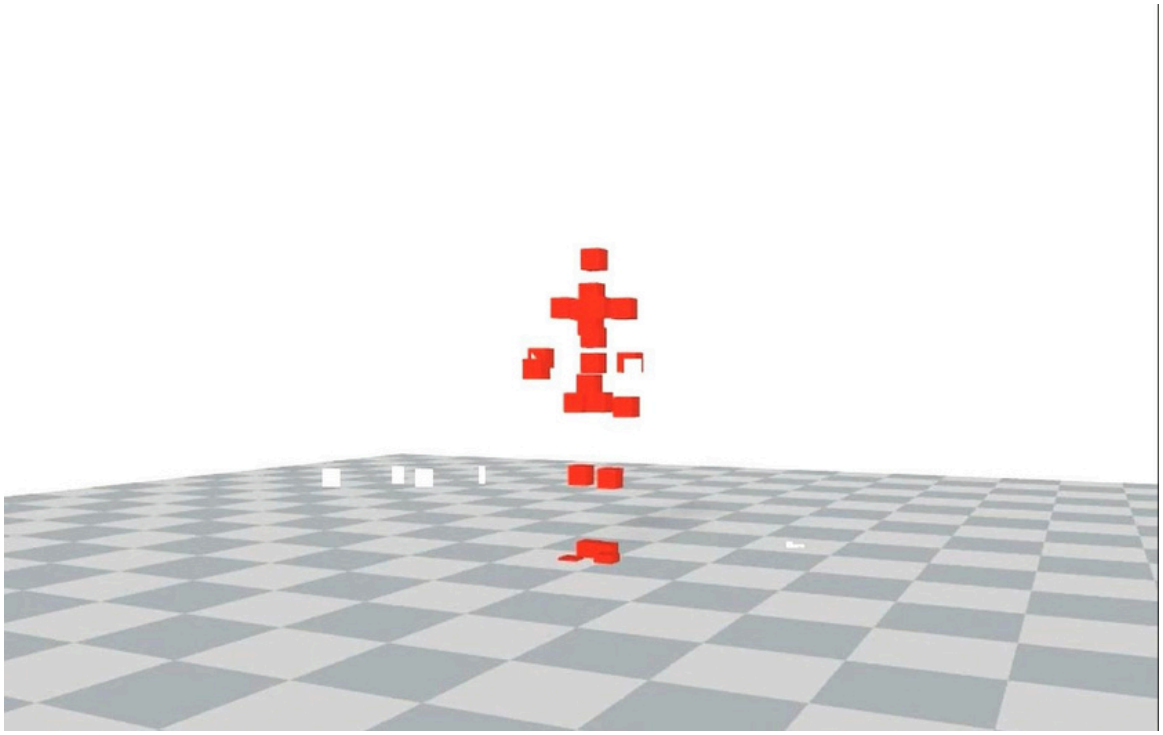

Figure 1: Sample Motion Capture Data

Motion capture data (Figure 1) obtained in such a way presents various imperfections that need to be corrected before it’s possible to utilize it to produce animations.

Firstly, the extracted animation can present incorrect orientation of the limbs during complex motions, such as dance moves or motions where the subject is prone on the ground. This is referred to as “limb flipping” and results from inaccuracies in the extracted 2D keypoints that cannot be corrected by the inverse kinematics (IK) step. Secondly, motion capture data can be prone to noise (also caused by natural inaccuracies during the detection of 2D keypoints) that makes the resulting animation jittery.

Once these two issues are solved, the resulting animation looks natural and can be further processed to apply style features. We chose to focus on tackling these two important challenges in motion capture data post processing.

AGroL

After some initial research, we landed on modifying and re-training the Avatars Grow Legs model from Facebook Research (https://github.com/facebookresearch/AGRoL) to suit our purpose.

The original paper presents a novel conditional diffusion model for the task of reconstructing full-body motion from sparse inputs, such as that of a VR headset only tracking head and hands position. AGroL is able to produce smooth, natural, accurate full-body motion from a reduced set of inputs and can run in real time. The architecture consists of a series of blocks of MLPs (multilayer perceptrons), a simple neural network structure, combined with a conditioned diffusion model for motion synthesis.

We elected to use AGroL as the basis for our method as it is able to solve both of the issues previously presented: being based on a diffusion model, it is resistant to noise in the input data (as well as missing frame data for some joints) and can still output smooth and accurate motions despite potential inaccuracies in the input. Moreover, given its initial purpose of generating precise and believable lower body motion data from a set of sparse inputs, it also acts as a “built-in” IK step to correct joint orientations and prevent limb flipping.

Since our input data is more detailed than the one available from tracking a headset, we modified the original model to work on more input joints (specifically, the first 23 joints of the SMPLX body model). This produces an output that is very close to the original motion, removing any “guesswork” about how the lower body parts move, while still leveraging the benefits of conditional diffusion.

We expanded both the input layer and the inner layer sizes, to account for the larger number of joints processed. After training it to convergence on a version of the AMASS dataset expanded to include dancing motion data, we reached very satisfying results - in our test video (Figure 2), the modified AGroL model was able to efficiently remove noise and correct limb flipping even in a complex “rolling on the ground” motion, with

accuracy metrics comparable or better than the ones in the original paper (as expected, since the task of reconstructing the motion was made easier).

Figure 2: Stylistically Consistent Animation

Conclusion

In conclusion, this project successfully developed a model for processing monocular motion capture footage, tackling critical challenges that typically arise in motion capture data. By addressing issues such as limb flipping and noise in the data, we were able to create a more accurate and natural animation output. The modified AGroL model proved to be an effective solution, as it seamlessly handled both noise reduction and joint orientation correction, ensuring that even complex movements such as rolling on the ground were accurately captured and processed. This model, which builds on the original AGroL framework, was adapted to handle a more detailed input, processing the first 23 joints of the SMPLX body model, thus improving motion fidelity while still leveraging the benefits of conditional diffusion.

The results achieved through the use of AGroL show a significant improvement in motion capture data quality, meeting the objectives of streamlining the clean-up step of the mocap-to-animation pipeline. The successful integration of these techniques sets the stage for the next step in the project: applying style features to the processed data, further streamlining the animation production process. By reducing the manual effort needed for motion correction, animation workflows can be significantly accelerated, allowing studios to generate high-quality, stylistically consistent animations with minimal intervention. Ultimately, the project not only resolves key technical challenges but also provides a foundation for further innovation in motion capture-based animation production.

VISION DENMARK
Med støtte fra Uddannelses- og Forskningsstyrelsen