Synth2Track Editor for Efficient Match-Animation

Jakob Buhmann
Disney Research Studios
Switzerland

Douglas L. Moore
Industrial Light & Magic
United States of America

Dominik Borer
Disney Research Studios
Switzerland

Martin Guay
Disney Research Studios
Switzerland

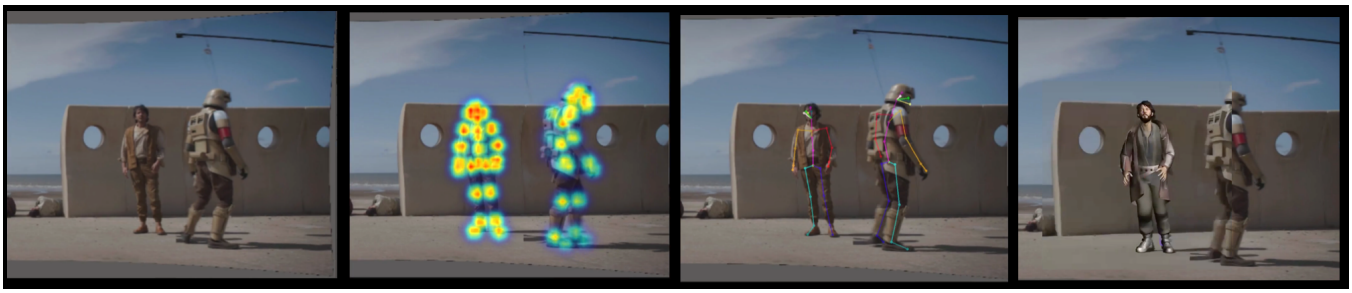

Figure 1: Match-Animation is the semi-manual process of capturing actors in videos for augmentation with 3D costumes, assets, and visual effects. We introduce a novel interactive workflow based on neural network predictions with user-specified cues.

ABSTRACT

A critical step in VFX production is capturing the movement of actors to integrate 3D digital assets into live-action footage. In recent years, advances in regression-based computer vision models such as human detection and motion models have enabled new workflows to emerge where parts of the Match-Animation process are automated. However, challenging shots that contain ambiguous visual cues, strong occlusions, or challenging appearances can cause automated systems to fail and users must revert to manual specification or to the previous generation of semi-automatic tools based on local feature-based tracking [Bregler et al. 2009; Sullivan et al. 2006].

Our key insight is that regression models can be used not only at the beginning of the process, but throughout by using manually specified cues. For example, given a partially detected actor, the user can specify a few landmarks manually, which once re-injected into a model, will yield new detections for the rest of the body. Based on this insight, we developed new tools that significantly reduces the time required for complex shots, combining automation with human expertise to overcome the limitations of current markerless motion capture systems.

ACM SIGGRAPH 2025, August 10–14, 2025, Vancouver, BC

ACM ISBN 978-1-4503-14811-12/2025/08
https://doi.org/3721239.3734082

ACM Reference Format:
Jakob Buhmann, Douglas L. Moore, Dominik Borer, and Martin Guay. 2025. Synth2Track Editor for Efficient Match-Animation. In *Proceedings of Proceedings (ACM SIGGRAPH 2025).* ACM, New York, NY, USA, 2 pages. https://doi.org/3721239.3734082

1 INTRODUCTION

Synth2Track is a markerless motion capture toolkit that performs 2D landmark detection and 3D motion inference, developed at DisneyResearch|Studios in collaboration with Industrial Light & Magic. It can be used as a fully automatic motion capture system with a single or multiple cameras. While it can capture actor movements fully automatically in many conditions, challenging situations that contain extreme occlusions or multiple actors interacting with each other can result in failures to detect 2D landmarks, and of the downstream 3D inference.

Instead of manually fixing the final 3D motion, our key insight is that a few of the landmarks can be specified by the user, and then re-inserted into our detection and completion models, resulting in new and more accurate predictions. This results in a simple and efficient user-guided 2D editing workflow for cleaning Mocap and effectuate Match-Animation. We made this new iterative workflow available in a software called Synth2Track *Editor* and used it for several recent VFX productions.

2 SYNTH2TRACK MOCAP

Synth2Track is a fully automatic markerless Mocap system that supports single and multi-camera setups. The system, as outlined in Figure 2, starts with 2D landmark detection, followed by 2D pose completion, and then predicts a 3D pose that is jointly optimized with the camera parameters and ground plane to ensure conformity to the 2D landmarks in each camera.

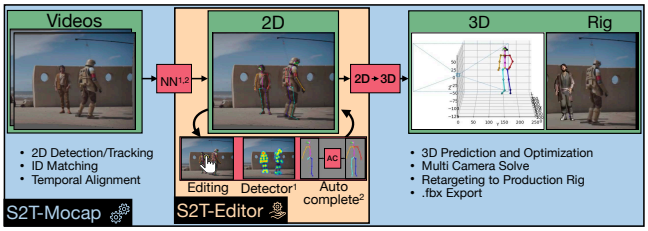

Figure 2: Overview of Synth2Track, an automatic markerless Mocap pipeline (blue), and the interactive Synth2Track Editor (orange) allowing for iterative user guidance.

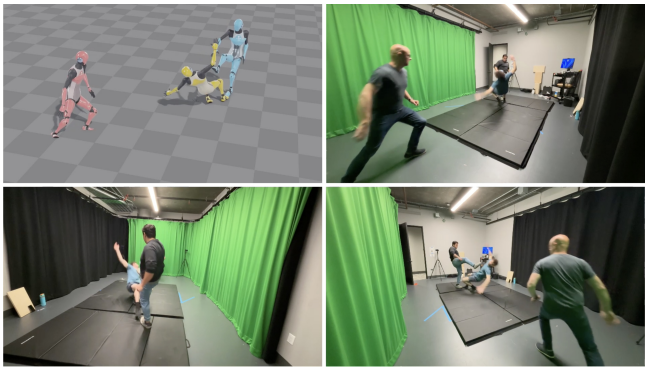

Figure 3: Synth2Track can be used with multiple self-calibrating cameras, which is particularly useful to capture multiple actors performing fast and dynamic movements.

3 SYNTH2TRACK EDITOR

Synth2Track works well in most cases. However, similar to other vision-based systems[Bregler et al. 2009; MoveAI 2025; Sullivan et al. 2006], on-set conditions can challenge high quality capture and tight match-animation. In these cases, the Synth2Track *Editor* enables user-guidance for both our 2D landmark detector and our 2D pose completion model (see orange box in Figure 2). Lastly, the former can also be permanently improved for a given appearance by finetuning. For better visual demonstration of the editing modes, we refer to the accompanying video.

Pose Detection. First, our 2D landmark detector is a neural network consisting of a sequence of 2D predictors. Inserting user-specified landmarks in early stages can result in new landmark discoveries—thanks to learned correlations in later stages of the model, as illustrated in Figure 4. Thus a user can either use the already predicted or manually added landmarks to iteratively update the predictions which progressively helps to find more landmarks. While false positives may occur, we observed that the added predictive capability leads to faster 2D pose detection.

Pose Completion. However, ambiguous footage containing occlusions or misleading color patterns can cause false landmark predictions. In such cases, a temporally-aware 2D pose model is particularly effective in maintaining tracking consistency despite challenging or missing visual cues. Users can correct or remove incorrect 2D landmarks, as depicted in Figure 5, after which our pose completion model predicts a new temporally coherent pose.

Finetuning. Due to costumes, an actor's appearance may be outside the range of appearances that were trained for. To improve robustness, our editor allows finetuning the 2D landmark detector with new annotations coming from the manual specifications, or from self-labeling. This is done automatically on a subset of frames, and results in better detection across the entire shot and is particularly useful for multiple shots with the same costume.

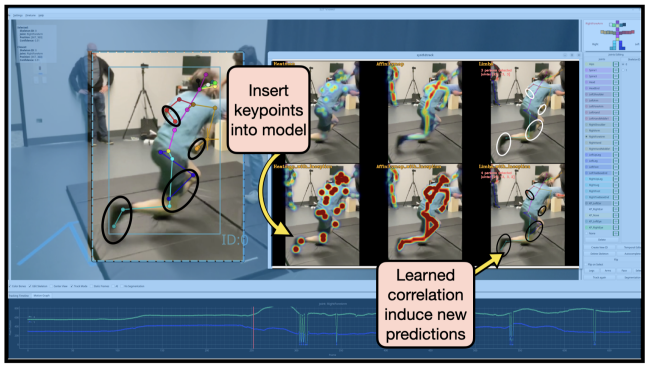

Figure 4: In difficult shots where detection fails, the user would have to manually add landmarks for all the body parts. In contrast, we allow the user to add only a few landmarks, that we feed into our pose detection model along the previous partial detections. This helps the model predict additional landmarks due to the learned correlations in the model between visual features and landmark activations.

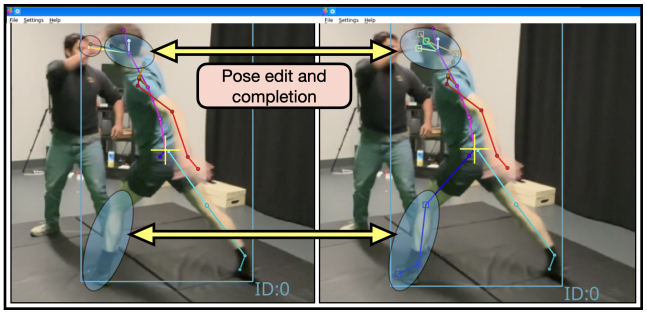

Figure 5: Sometimes landmark detection fails due to occlusions, resulting in missing landmarks—such as the left leg here, or falsely predicted points such as the right arm. In these cases, the user can create new landmarks, delete or move landmarks, and re-launch pose completion in order to get statistically and temporally meaningful predictions.

REFERENCES

C. Bregler, K. Bhat, J. Saltzman, and B. Allen. 2009. ILM's Multitrack: A new visual tracking framework for high-end VFX production. *SIGGRAPH Talk* (2009).
MoveAI. 2025. *https://move-ai-v2.webflow.io/* (2025).
S. Sullivan, C. Davidson, M. Sanders, and K. Wooley. 2006. Three-dimensional motion capture. *USPatent US7848564B2* (2006).