Optical Human 3D Motion Capture in Professional Football

David Björkstrand, Ind. PhD Student, Tracab and KTH
Dept. of Robotics, Perception and Learning
Supervisors: Assoc. Prof. Josephine Sullivan, Dr. Tiesheng Wang (Tracab) and Dr. Lars Bretzner (Tracab)

Background

Human 3D motion capture has numerous applications in fields such as augmented and virtual reality, animation, robotics, and sports. In football specifically, there are a number of use cases such as Semi-Automatic Offside Technology. Using a multi-view human 3D motion capture system installed in football arenas all over the world, Tracab [1] has the capability to do human 3D motion capture at scale.

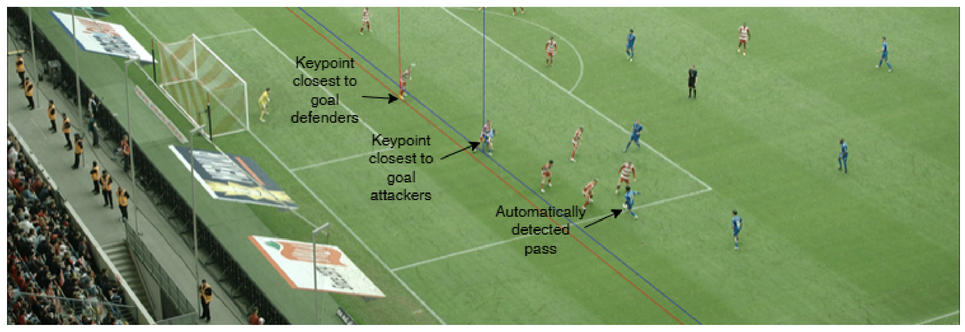

Figure 1: Semi-Automatic Offside Technology.

Challenges

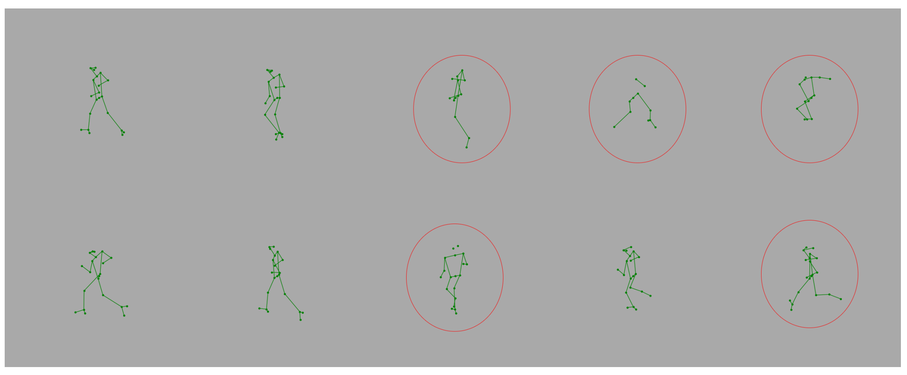

Figure 2: Noisy skeletons from Tracab's multi-view tracking system.

Even though state-of-the-art systems are impressive today, they still stuffer from artifacts. Two common artifacts are missing joints and noisy joints positions as can be seen in Figure 1. There are many potential causes for these, such as camera obstructions (smoke is common), failures of pose estimation models and very challenging scenarios as depicted in Figure 3.

Figure 3: A challenging crowded scenario.

Proposed solution: XMAE

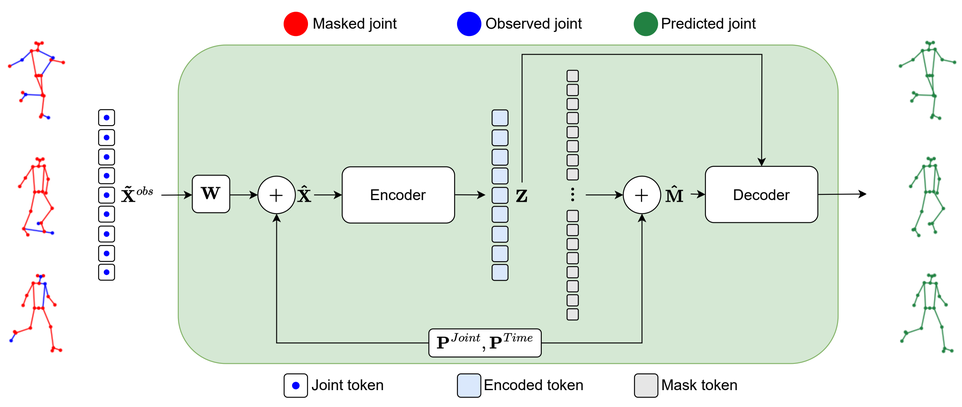

Figure 4: The Cross-attention Masked Auto-Encoder (XMAE)

A common way to try to mitigate artifacts is by using another post-processing method, which tries to predict the motion explained by the noisy observations. In Figure 4 is our contribution, the Cross-attention Masked Auto-Encoder (XMAE) [2], a regression model for short sequences.

Results

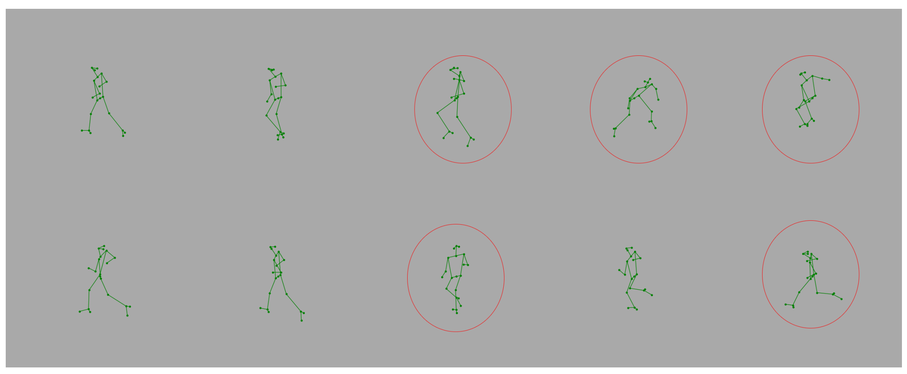

Figure 5: Skeletons from Figure 1, post-processed by XMAE.

Compared to other state-of-the-art methods across three datasets, XMAE was very successful at correcting low to moderate levels of artifacts. It additionally was effective on Tracab's real-world football data (Figure 5) and is currently used in production.

> Motion is best view in motion. Please talk to David Björkstrand if you would like to view more examples in video form.

Current & Future work

Diffusion models for multi-modal modeling Under severe artifacts regression models fail due to the multi-modality of the distribution of motions explained by the observations. In these cases, the best we can do is to generate a plausible motion that is explained by the noisy observations. For this we need a generative model.

Global translation If global translation is not handled properly, new artifacts, such as foot skating, can be introduced into the data. While not an issue for XMAE, tackling longer sequences with severe artifacts makes this a challenge. This could potentially be further exacerbated by the high speeds typical in football. As public datasets exhibit a limited range of global translation and speed, we hope to be able to gain unique insights using Tracab's football data.

Multi-human modeling To solve scenarios as depicted in Figure 3, we need models that take into account other nearby persons.

References

[1] Tracab
https://tracab.com/

[2] Cross-attention Masked Auto-Encoder for Human 3D motion Infilling and Denoising
David Björkstrand, Josephine Sullivan, Lars Bretzner, Gareth Loy, Tiesheng Wang
BMVC 2023

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM