Popular Summary

Human motion has fascinated scholars and artists since ancient times. The advent of photography and film, allowing to capture movement accurately, paved the way for modern scientific study of motion. Modern computer systems can now identify where a person's body parts are located in images, creating simple stick-figure representations that track how people move, enabling automatic analysis. But a single camera viewpoint provides limited information: it is impossible to tell how far away objects are, and objects or other people may block the view. With several cameras capturing a scene from different angles, we can estimate true 3D information, much like how our two eyes work together to give us depth perception. This thesis proposes methods for building practical, robust systems that can perceive and understand human activities and scenes in 3D using multiple cameras.

Before multiple cameras can work together, their exact positions and orientations must be known. This can be done by *calibrating* the cameras, a process often involving special equipment, like checkerboard patterns, waved through the scene following specific procedures, which often requires a trained operator to achieve good accuracy. This can be long and expensive, making it impractical for scenarios requiring fast installation. In this thesis, we present a new calibration method that uses the people moving in the scene as calibration patterns, enabling faster installation without equipment: all it takes is for someone to walk around the scene for a while something that might happen naturally in a new installation.

In controlled laboratory conditions, perfectly calibrated cameras would allow flawless reconstruction of human poses in 3D. But real-world environments present significant challenges. Furniture, other people, or even the environment itself may hide parts of a person, lighting conditions may vary, and the initial position estimates can be noisy or inaccurate. Simply combining these imperfect observations can result in distorted or absurd body poses. To reconstruct accurate 3D human poses even under challenging conditions, we trained a neural network to pay attention to the geometric relationships between views and to the movements of 2D poses over time. It learns to create coherent and accurate 3D human posture and motion even when some body parts are not visible to any of the cameras, filling in the gaps with realistic predictions based on how people naturally move.

Once we have accurate 3D movement sequences, we can teach computers to recognize different activities, like "walking" or "doing push-ups". Training com-

puters for *activity recognition* usually requires large amounts of *labeled data*, where humans manually mark and name each activity in thousands of motion sequences. This is labour-intensive and very expensive. What's more, recognizing a new activity requires collecting and labeling new data from scratch. On the other hand, just collecting *unlabeled* movement data is easy and inexpensive. We developed a *self-supervised learning* method that allows computers to learn meaningful patterns from large amounts of unlabeled movement data. For example, it learns that an action remains fundamentally the same even when viewed from different angles, or when performed at different speeds. After this initial learning phase, called *pre-training*, the system is very adaptable: it can quickly learn to recognize new activities using just a few labeled examples, significantly reducing the time and cost of training.

Truly understanding human behaviour requires more than just tracking body movement. We interact constantly with our environment, and our movements take on specific meaning depending on the context: standing in front of a refrigerator suggests that you will open it; moving towards a chair indicates that you are about to sit down. Without seeing the surrounding objects, these movements could mean almost anything. We developed a method for localizing objects in 3D using just few camera views and everyday language descriptions, like "a blue bowl" or "an office chair". This enables the system to find objects it has never been trained to recognize, providing rich environmental context for human activity.

Together, these contributions help move multi-camera 3D perception systems from the lab to the real world, making them more accessible, efficient, and effective for diverse applications such as sports analysis, safety monitoring, entertainment, and beyond.